AN EVALUATION OF THE EXTENSIBILITY AND APPLICATION OF OCCUPANCY PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES

By

Hollis J. Wright

A DISSERTATION

Presented To the Department of Medical Informatics and Clinical Epidemiology and the Oregon Health & Science University School of Medicine in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

September 2010

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the PhD Dissertation of

Hollis J. Wright

"An Evaluation of the Extensibility and Application of Occupancy Prediction of Transcription Factor Binding Sites"

Has been approved

Dissertation Advisor – Shannon McWeeney

Committee Member – Aaron Cohen

Committee Member – Kemal Sonmez

Committee Member – Gregory Yochum

Committee Member – Armand Bankhead

TABLE OF CONTENTS

Acknowledgements iii

Abstract iv

Chapter 1: Background and Introduction 1
1.1 Background 1
1.1.1 Difficulties of biochemical and in silico TFBS discovery 1
1.1.2 Matrix-based TFBS prediction 2
1.1.3 Ab initio TFBS prediction 3
1.1.4 Limitations of in silico discovery 4
1.1.5 Integrative methods 4
1.1.6 Unanswered questions about integrative methods 6
1.1.7 Possibilities of integrative methods 8
1.2 Specific Aims and Methods 8
1.2.1 Selected algorithms 9
1.2.2 Selected TFs 10
1.2.3 Selected predictors 11
1.2.4 Selected protein interaction database 12
Chapter 2: Occupancy Classification of PWM-Predicted Transcription Factor Binding Sites 13
2.1 Introduction 13
2.2 Methods 13
2.2.1 Evaluation 15
2.3 Results 15
2.3.1 Comparison of Algorithms 16
2.3.2 Contribution of Feature-Feature Distances to Classification 16
2.3.3 Individual Chromosome Classification 17
2.3.4 Common Predictors Across TFs 17
2.3.5 Cross-Classification Performance 18
2.3.6 Cross-Classification of SRF 19
2.4 Discussion 20
2.4.1 Biological Factors 20
2.4.2 Technical Factors 20
2.4.3 Summary 21
Chapter 3: Generalizable Occupancy Classifiers 22
3.1 Background 22
3.2 Methods 22
3.2.1 Generalizable Classification Schemes 22
3.3 Results 24
3.4 Discussion 24
Chapter 4: Similarity of protein interaction networks derived from occupancy classification to those derived from biochemical data 26
4.1 Background 26
4.2 Methods 26
4.2.1 Network Similarity Metric for Connected Target Gene Nodes 27
4.2.2 Network Similarity Metric for Hub Preservation 29
4.3 Results 29

4.3.1 Edge Preservation Analysis 29
4.3.2 Hub Preservation Analysis 30
4.4 Discussion 30
Chapter 5: Use-case of protein interaction data to prioritize selection of c-Myc/TCF4 shared target candidates and confirmation of TCF4 and c-Myc binding in human colorectal cancer cells 32
5.1 Background 32
5.2 Methods 33
5.2.1 Analysis of External Supporting Evidence for Candidate Genes 34
5.2.2 Prioritization of Candidate Genes 34
5.2.3 Confirmation of TCF4 and c-Myc binding by immunoprecipitation in HCT116 human colorectal cells 35
5.3 Results 36
5.3.1 Results of External Support Analysis and Candidate Selection 36
5.3.2 Chromatin Immunoprecipitation of predicted TCF4 and c-Myc target genes 36
5.3.3 Improvement of correlation of classifier agreement pair score with support 37
5.3.4 Overall performance of protocol 37
5.4 Discussion 39
5.5 Discussion of potential biological significance of confirmed target genes 39
Chapter 6: Discussion of results, relevance, and future directions 41
6.1 Significance of initial classification and cross-classification experiments 41
6.1.1 Limitations of analysis 41
6.1.2 Comparison to existing work 42
6.1.3 Summary of significance 44
6.2 Significance of generalizable occupancy classification experiments 44
6.2.1 Comparison to existing work 45
6.2.2 Summary of significance 45
6.3 Significance of network analysis and shared c-Myc and TCF4 target prediction 45
6.3.1 Significance of network similarity analysis 46
6.3.2 Significance of common c-Myc and TCF4 target selection and prioritization 47
6.4 Future Directions and Potential Extensions/Applications of Occupancy Classification 48
6.4.1 Extension to additional TFs, cell types/species, and genomic contexts 48
6.4.2 Algorithmic and methodological questions of interest 49
6.4.3 Model perturbation 49
6.4.4 Possibilities for network-based and other multi-method analyses 50
Bibliography 51
Tables 62
Figures 75

ACKNOWLEDGEMENTS

I would like to thank my committee, Drs. Aaron Cohen, Kemal Sonmez, Gregory Yochum, Armand Bankhead, and in particular my committee chair Dr. Shannon McWeeney (who has both intelligence and patience with graduate students in ample supply, much to my benefit). Every member of my committee was invaluable in the development of this dissertation, offering both scientific insight and practical advice without fail, despite the obstacles of geography, travel scheduling, and the occasional volcano. I literally could not have done this without their support; my most heartfelt thanks to all of you. I'd also like to give special mention to Dr. Yochum's assistant, Sydney Kyler, for all of her immunoprecipitation work which appears in Chapter 5.

I would also like to thank my colleagues both in the NLM Fellow's program and in Dr. McWeeney's weekly BioDev group, who were testbeds for presentations, sounding boards for questions, and generally excellent people to work with.

I must of course thank the various OHSU, OGI, and PSU faculty I have had the pleasure of learning from and working with over the years. It's too long of a list to enumerate everyone, but hopefully you know who you are. I must thank the DMICE faculty in particular; Their instruction and help over the years helped me build the foundation upon which this work rests. In particular, I'd like to cite Dr. Beth Wilmot, as she was invaluable in helping put together this dissertation in the proposal stages. Also, I'd like to mention the DMICE staff, for being incredibly helpful in many ways over the years. Nothing gets done without them.

I would like to specifically thank some of my mentors and collaborators during my time out at ONPRC: Drs. Sergio Ojeda, Betsy Ferguson, and Christoper Dubay. Besides providing me with a chance to work on some fascinating projects, some of the work I did at ONPRC was in large part the inspiration for this work.

This work was supported primarily by the National Library of Medicine Fellowship Training Grant (no. T15 LM007088); my gratitude to the NLM for giving young scientists in our field such an excellent opportunity to receive top-notch education and to have the freedom to pursue research they find stimulating.

Finally, I'd like to thank my friends and family, largely for putting up with me for the several years this has taken. Annabel, I don't know how you actually live with me, but I love you for it. Also, chu! My parents, Michael and Marsha Wright, have been as loving and as supportive as I think is humanly possible; I love you both. My mother is no longer with us, unfortunately; it is my hope that this work may be some small help in the battle against cancer, and if so I hope it will be an honor to her memory.

Hollis Wright
August 2010

ABSTRACT

Identification of biologically relevant high-occupancy transcription factor binding sites (TFBS) in silico has historically been a difficult problem with a high error rate. Methods which utilize information in addition to the sequence of binding sites (e.g. chromatin information) have been shown to improve performance over strictly sequence-based methods; however, a number of questions about such methods remain unanswered: whether such models are suitable for multiple transcription factors, whether a general model or generalizable approach to the problem is possible, and what the effect of such prediction on biological inference is. In this work, we construct and evaluate a number of classifiers of position weight matrix-predicted TFBS ("occupancy classifiers") based on four distinct transcription factors and demonstrate that such classifiers identify biochemically confirmed high-occupancy sites at a high rate. I contrast and compare the algorithms and predictors used by these classifiers and demonstrate that efficient cross-classification of one factor by a classifier trained on another factor is possible. We then construct generalizable occupancy classifiers, combining data from several transcription factors and intended to operate on potentially arbitrary transcription factors; I evaluate several versions of these classifiers and demonstrate that they can perform comparably to classifiers trained on factor-specific information even when not provided with that information. We then demonstrate that occupancy classification is capable of recapitulating a statistically significant portion of protein interaction networks derived from biochemical studies of transcription factor binding, suggesting that biological inference from occupancy classification may be comparable to that expected from biochemical identification of TFBS. Finally, we use a generalizable classification method to identify novel binding sites for the TCF4 and c-Myc transcription factors in the human genome, prioritize these predictions using classification metrics and protein interaction information, and confirm the predictions via immunoprecipitation. These results demonstrate that occupancy classification can be used for multiple TFs, can be used in cross-classification or to generally classify

TFBS occupancy even in the absence of factor-specific training information, and that occupancy classification reproduces significant portions of a protein interaction network expected from biochemical TFBS identification. Occupancy classification is further validated by identifying novel binding sites forTCF4 and c-Myc which may be of import in various biological processes, including colorectal cancers.

Chapter 1: Background and Introduction

1.1 Background: Transcription factors (TFs) are proteins which bind DNA and affect the rate of transcription of genes and other transcribed elements of the genome. Typically, transcription factors are thought to have an essentially local mechanism of action (affecting the transcription of genes within a few thousand base pairs of the site at which a given factor binds). The precise mechanism by which a transcription factor modulates transcription varies depending on the factor, and one factor may act in multiple ways, depending on the context in which it operates. Factors may participate directly in the assembly of transcription complexes and/or help recruit elements to the site where the factor binds, or they may block access of such proteins and hence repress transcription. Other TFs may be involved in the remodeling of local chromatin structure (such as methyltransferases which modify the methylation state of certain histone residues) and affect the accessibility of DNA indirectly via condensing DNA into inaccessible heterochromatin or freeing it into accessible euchromatin. (For a review of these concepts, see Fickett and Wasserman[1]). For purposes of defining a transcription factor for this dissertation/proposal, any of these mechanisms could define a "transcription factor"; the only requirements for a transcription factor to be of potential interest in this work are that the TF must affect the rate of transcription of some gene product and it must directly bind DNA to be of interest. Typically, eukaryotic gene regulation requires the participation of several TFs. Multiple factors bind at multiple sites upstream from the transcription start site. Because of this complexity, it can be difficult to discern which factors are responsible for regulating which genes; additional complexity is also introduced by the presence of tissue-specific and temporally specific regulation; e.g. the tremendous diversity of roles attributed to the TGF-beta family of factors (for a review, see Massauge[2]).

1.1.1 Difficulties of biochemical and in silico TFBS discovery: Though direct molecular studies are obviously the most accurate method for determining precisely which TFs bind at and may involved in

regulating a specific gene, such studies can be difficult and expensive to perform, and until recently were impossible to perform for entire genomes. As an example, consider Cawley et al.[3], who were restricted to mapping TF binding sites only in chromosomes 21 and 22 due to the limitations of their tiling array process. Even with advances leading to true genome-wide TF binding site elucidation such as the Serial Analysis of Chromatin Occupancy[4] and ChIP-seq[5,6,7], only a handful of TFs have been mapped genome-wide in a small number of model systems and conditions. Historically, computational methods for predicting which TFs may bind and influence the transcription of genes of interest have therefore been important; it is reasonable to believe that they will continue to be of importance for the foreseeable future. Various in silico methodologies have been developed to predict TF binding sites in a given sequence. These can be classified broadly into matrix-based methods, which use predefined profiles of the composition of expected binding sites to scan input sequences looking for matches,and ab initio methods that attempt to derive potential motifs from a statistical overrepresentation of the putative motif's pattern in a set of input sequences.

1.1.2 Matrix-based TFBS prediction: Matrix based methods are perhaps best exemplified by the TRANSFAC database[8] and its related programs such as MATCH[9]. These methods use a database of position weight matrices[10,11] (PWMs) and compare input sequences to each matrix in the database. PWMs are able to contain information regarding the degeneracy (the permissibility of variant bases at a base of the sequence) of the binding motifs; a particular subsequence which scores past a set threshold as a match to a particular matrix is tagged as a match. This approach can be effective at identifying binding sites in silico[7]; however, it does carry with it several disadvantages. First, one can only query for sequences for which a PWM exists; if a matrix does not exist for a particular factor of interest, it is simply not possible for a PWM-based method to discover a binding site for it. The degeneracy of binding motifs also essentially limits the PWM approach, causing high false positive rates possibly

missing true binding motifs[12]. Even attempting to limit the number of results to a manageable level with stringent cutoffs for quality, a query of a sequence can frequently return an overwhelming number of discovered motifs.

1.1.3 Ab initio TFBS prediction: Various ab initio methods for discovering motifs in sequences have been developed; perhaps the classic exemplar in the field is MEME[13], based on expectation-maximization. Other methods include Weeder[14], which relies on an exhaustive classification and analysis of overrepresentation of all n-mers in a set of input sequences, and various methodologies based on Gibbs sampling such as MotifSampler[15]. These methods make no assumptions other than certain initial parameters when exploring sequences for motifs, and hence are suitable for finding similarities in a set of sequences which have not previously been characterized. However, these methods have their own set of limitations; they are dependent on both the choice of input set and choice of parameters, which can limit the ability of the algorithm to discover motifs if either set of inputs is inadequate or incorrect. In particular, most methods assume that the input set comes from a set of transcriptional control elements derived from coregulated genes; however, if some of the genes are not in fact coregulated or if if the mechanisms for coregulation are heterogenous, this assumption and the results of the analysis may be invalid. There is also no guarantee that these methods will find biologically meaningful patterns within the input sequences; the motifs discovered may in fact bind nothing in vivo, and so any discovered motifs usually need to be confirmed experimentally. While this is true of all computational motif discovery methods, in matrix-based approaches at least some evidence of binding is already observed for a given TF at a motif. Alternatively, the motifs discovered can be compared to known motifs to attempt to determine their validity, but this saddles the ab initio methodologies with most of the drawbacks of motif-based methods as well, most crucially such methods' limited search space.

1.1.4 Limitations of in silico discovery: Most current methods of motif finding utilizing any approach allow for considerable room for improvement, due largely to the limitations of the available data. In the study by Tompa et al.[16], the most accurate tool in the group of 13 assessed ab inito motif discovery programs (Pavesi et al.'s Weeder[14]) achieved an average per site specificity of only 0.17. While the Tompa et al. study is not an exhaustive enumeration of available methods (notably, it does not evaluate methods using phylogenetic support or other additional data beyond the input sequences), it does illustrate the considerable room for improvement of motif detection with regard to ab initio tools. The limitations of PWM-based methods have been generally recognized for some time, with such methods exhibiting high sensitivity but low specificity (see again Roulet et al.[12]).

1.1.5 Integrative methods: With the limitations of motif discovery in mind, methods have been developed which attempt to exploit certain properties of regulatory regions to determine the likely functionality of discovered sites; in effect these methods try to integrate additional information to determine the likelihood of a discovered motif being "real," sidestepping the issue of improving potential motif identification as such. Attributes used in such "integrative" methods have included the tendency for TFBS sites of the same type to cluster near one another in cis-regulatory modules[17], used by the ModuleFinder algorithm[18], phylogenetic conservation[19], or coregulation as implied by gene expression assays[20]. These methods can show drastic improvement relative to less informed methods, either matrix based or ab initio; for example, a very stringent phylogenetic approach using mouse, rat, and human sequences in concert with TRANSFAC demonstrated perfect sensitivity and specificity in a 4000 bp region for experimentally confirmed GATA-1 sites[22]. Ultimately, however, the initial identification of potential TF binding sites is either based on matrices or on an ab initio statistical process; depending on the specific method used, these advanced methods themselves suffer the

limitations of their parent methodology.. Furthermore, it is very possible that models of TF binding based upon sequence alone cannot achieve comparable performance to methods taking into account additional genomic and chromatin features due to the dependence of in vivo TF binding on such features. As such, these integrative methods seem to represent the viable future of computational TF binding site discovery. A particularly promising variant of the integrative method is presented by Chen et al.[22], who consider the occupancy of c-Myc binding sites in the human genome as predicted by PWM methods using a Bayesian network classifier[23]; this classifier utilizes multiple lines of information about the genomic context of a predicted site (its proximity to transcription start sites, CpG islands, and hypomethylation regions), the modification states of nearby histones (the distance to the nearest H3K9 acetylation island) and the conservation of the site across species in terms of its PhastCons score[24] to predict whether a computationally derived site is high or low occupancy, with the prediction of high occupancy taken as a proxy to biological functionality. We term this type of TFBS prediction "occupancy prediction." Trained on a set of confirmed binding sites for c-Myc[25] and validated with additional external biochemically derived binding data sets[3,26] as well as corroborating microarray experiments, the classifier achieves reasonably high predictive ability (~71% of biochemically confirmed high-occupancy sites are predicted correctly). Furthermore, the authors believe that the procedure is generalizable, using the classifier to predict the occupancy of CREB binding sites as well; unfortunately, these results are somewhat less conclusive, due to the high false positive rate of CREB site classification. While the classifier successfully identified a large proportion of high-occupancy sites according to the Zhang et al study[27] (1713/2195 correct predictions), it predicted an additional 3621 false positives. Chen et al. argue that many of these false positives may represent unknown true positive CREB binding sites due to the limitations of the Zhang et al. study, but absent further biochemical confirmation it is not possible to evaluate the validity of that claim.

1.1.6 Unanswered questions about integrative methods: Despite the promise of this type of approach, a number of issues specific to this implementation remain unanswered at this time, many of which may impact the method's generalizability to other TFs and to sites located in regions of the genome for which the classifier is not yet evaluated. The use of a single TF as the training example of the classifier (CREB analysis notwithstanding) leaves unanswered questions about the extensibility of the classifier to additional TFs. as does the use of a single genomic context (within 3kb of a transcription start site) for that TF leave questions about the applicability of the classifier for site occupying regions heterogenous to the training examples (such as TF binding sites positioned in intronic regions). It is not clear that the predictors used by Chen et al. would be optimal for occupancy classification for other TFs; a single feature (predicted hypomethylation island distance) provides the bulk of predictive ability. If the prediction proves inaccurate or binding sites for other TFs and/or in other genomic contexts that are less sensitive to hypomethylation, there may be consequences for predictive accuracy if an unmodified version of the Chen et al. Bayesian classifier is used to predict the occupancy of binding sites. Indeed, it is not entirely clear that the Bayesian framework itself is optimal or extensible for occupancy prediction. While Chen et al. do use differences in the distribution of features between high and low-occupancy sites in selecting features for training their classifier, they do not indicate any exploration of alternative machine learning algorithms such as support vector machines (SVM)[28] as alternatives to the Bayesian network they ultimately used. It is possible that different predictor sets or even different classification methods may prove to be optimal for specific TFs or for TF binding sites occupying different genomic contexts. Finally, from the standpoint of a biologist or translational researcher using TF binding site prediction data in their work, it is unknown to what extent introduction of occupancy information as predicted by such a highly-integrative occupancy prediction framework into the construction of a protein interaction network model would actually change or confirm the structure or predictions made using such a network.

1.1.7 Possibilities of integrative methods: Nonetheless, the possibilities of an occupancy classifier integrating multiple sources of information to predict the occupancy of a predicted TF binding site of the kind proposed by Chen et al. are tremendous. A significant improvement in confirming the biological relevance of the results of a computational motif finding process could be of immense practical value to biology, considering the time and effort required to biochemically characterize TF binding motifs and the tremendous number of TFs whose binding sites have not been characterized in many (if any) genomes. A method of better determining the likelihood of binding at a given predicted TFBS and prioritizing predicted binding sites for further investigation and/or biochemical confirmation from the large number of binding sites resulting from matrix-based discovery methods methods, for example, would potentially allow a researcher interested in exploring a particular TF to focus his efforts on the sites most likely to be of interest to his analysis and/or to be biochemically confirmed. More accurate tools would hopefully sallow for biologists to construct more accurate models of transcriptional regulation by allowing more confident integration of binding predictions into such models, as well. As an example, consider the work of Roth et al.[29], who use TRANSFAC-based prediction of TF binding sites as a source of information in constructing a hypothetical transcriptional control network for the initiation of puberty in mice. Though their hypothetical network is guided not only by prediction of TF binding but also considerable experimental evidence, the application of an effective strategy for predicting occupancy of specific sites could allow these or similar researchers constructing such a hypothetical network to have better confidence in the addition or elimination of a specific binding site for a specific TF as a possible effector of transcriptional control of a gene in the network; that said, the practical effect of occupancy prediction itself on the structure of such networks is an open question. It is clear that an accurate and generalizable highly-integrative occupancy classification method could have wide-ranging impact, particularly for fields such as cancer and

development wherein transcriptional control (or its aberrations) are of paramount importance. This suggests that a larger scale evaluation of the best methodologies for, extensibility of and practical applications involving highly-integrative classification methods is a worthwhile endeavor.

1.2 Specific Aims and Methods: With these issues and open questions in mind, in this work we perform a number of experiments to further advance our understanding of the performance of occupancy classification methods, defined as such methods which utilize combined information from multiple data sources about the sequence features and chromatin structure surrounding an instance of a computationally predicted TF binding site to predict the most likely occupancy class of that site (high or low occupancy). The Chen et al. Bayesian classifier and its set of associated predictors is taken as a starting point for these experiments; however, the experiments will investigate both additional, alternative methods of classification and alternative predictors of occupancy. We have set out to accomplish the following specific aims:

Aim 1) *Investigate the best methodologies for and generalizability of the highly-integrative occupancy classification approach* by constructing best classification schemes for classifying occupancy of predicted TF binding sites for the TFs c-Myc, STAT1, GABP, SRF1, and TCF4 for sites in proximity to transcription start sites (defined as within 3kb of a TSS, "in-window) in the human genome. Best classification schemes will be those specific classifiers which achieve best Area Under Curve (AUC) score of classification for a given TF in a given genomic context using one of the following machine learning algorithms:

1. Bayesian network
2. SVM
3. Combination of multiple classifiers of the above types through e.g. stacking[30] to construct a

generalizable occupancy classifier.

and some combination of features from those enumerated below:

1. Features relatedto nearest hypomethylation region
2. Features related to nearest CpG island
3. Features related e to nearest histone acetylation island
4. Features related to nearest histone methylation islands
5. Features related to the nearest TSS

Based on these results, investigate the performance of TF-specific and generalizable occupancy classifiers to determine the viability and overall performance of occupancy classification in cross-classification and generalized occupancy classification scenarios.

Aim 2: *Investigate the effects of occupancy classification on the topology of protein interaction networks* derived from occupancy classification information as the basis of their construction in comparison to those based on biochemical TF binding information, using the best classifiers constructed for the TFs c-Myc and TCF4 in aim 1. In addition, we investigate the effects on topology of a combined network of c-Myc and TCF4 targets incorporating occupancy information and use a combination of prediction quality measures and network-based metrics to make predictions of shared c-Myc and TCF4 predicted targets likely to be involved in colorectal cancer processes of interest. For this analysis, the PathwayCommons[31] protein interaction and pathway database for humans will be used.

1.2.1 Selected algorithms: We have selected the algorithms to be used in this experiment due to their demonstrated performance in difficult classification problems. The selected classification

methodologies are thought to provide a representative spectrum of current classification and machine learning algorithms. Bayesian network and SVM methods are standard algorithms used in machine learning. Bayesian networks, which predict the probability of membership in a particular class for a given test data point based on the probability of the states of associated predictors being in the given state for that class, have been shown to have direct effectiveness in the question of occupancy prediction by Chen et al., and are generally considered to be powerful classification options even in the reduced naïve Bayes form (which ignores interdependencies between predictors). SVM classification algorithms seek to establish a separating hyperplane between instances of training data in the classes which maximizes the separation of the closest instances of each class ("the margin"); they are considered to be among the most powerful machine learning techniques, and are able to adapt to nonlinearities and relationships among the data by projecting data into higher dimensional spaces (the "kernel trick"), as well as make allowances for classes which cannot be linearly separated ("soft margins"). Finally, the combination classification methods used in the construction of generalizable occupancy classifiers involve combination of outputs from multiple classifiers to arrive at a final answer; for purposes of these experiments such classifiers exploit classifier stacking, a method of classifier combination which uses cross validation to construct a classifier (the "level-1" classifier) using the input of several other classifiers ("level-0" classifiers). The specific implementations of each classifier used in the Weka[32] classification package are used throughout this work, with some modifications to adapt to the specific requirements of the experiments.

1.2.2 Selected TFs: The TFs selected for classification in aim 1 have been selected for a combination of practical and theoretical reasons. TCF4 is a primary actor in the Wnt signaling pathway, an important pathway in a number of developmental and carcinogenic processes[33], notably in defining colorectal cancer stem cell populations[34]. Similarly, c-Myc is both the TF for the Chen et al. classifier and a

known target gene for TCF4[35], and both the factors have long been implicated in the causative process of colon cancer[34,36]. STAT1 is a member of the Jak-STAT signaling pathway, a primary pathway for modulating cellular response to cytokine signaling and implicated in a variety of disease processes[37]. Finally, GABP has the somewhat unique property of being a preferential activator of bidirectional transcription start sites[38], and is involved in a number of important biological processes, notably cell cycle control at the neuromuscular junction[39]. All of the chosen data sets have been constructed using the well-annotated human genome for which all potential occupancy predictors we have selected are readily available or derivable; restricting to human-derived data also avoids complications that may be due to species-specific effects, though it unfortunately excludes potentially interesting data sets such as the CREB SACO data set[4]. In addition, all selected TFs are thought to be transcriptional activators (or, in the case of c-Myc, to operate as one in the presence of beta-catenin) and to be generally active in many tissues as opposed to being tissue specific transcription factors. These are important attributes to avoid potential confounding effects by either the inclusion of transcriptional repressors, which may not respond to the same chromatin cues, or by tissue specific effects, since many of the data sets we intend to use were derived from different cell lines. Unfortunately, since the data sets were constructed using different technologies (ChIP-PET for c-Myc, ChIP-chip for TCF4, and ChIP-seq for STAT1, GABP, and SRF) and different cell types, undesirable technical variation may be introduced into our experiments. However, at the time of this study's design, relatively few high quality binding data sets in human beings were available. It was our judgment that these factors represented sufficiently high quality data in conjunction with our desired traits (activating activity, tissue nonspecific action) and potential for relevant biological discovery from our analyses. Critically, all the factors also had position-weight matrices available for initial discovery to be possible.

1.2.3 Selected predictors: The predictors used in this work have been selected due to demonstrated or

anticipated association with regulation of transcriptional activity or TF occupancy. Distances to hypomethylation region, CpG island, histone acetylation island and TSS have been demonstrated to be directly significant at least in the case of c-Myc by Chen et al. Unfortunately, the methods used to predicted CpG island distance and hypomethylation region distance used by Chen et al. have not been made publicly available; however, a number of CpG island prediction schemes exist, including the UCSC Genome Browser's[40] CpG Island prediction track, and a map of genome-wide DNA hypomethylation in human leukocytes has recently become available[41]. For this project the Shann et al.methylation map and the UCSC Genome Browser's CpG Island prediction track will be used. Distances to histone methylation islands have been shown to indicate regions of both reduced and increased transcriptional activity[42], and at least HK9A modification has been already shown as a useful predictor of occupancy for c-Myc[22]; it is reasonable to believe that other modifications are likely to have predictive value for occupancy classification as well. A map of human histone methylation in positions H3K4 and H3K27 in T cells is included in the Roh et al. data set[42]; an expanded number of histone methylation type mappings in the same cell type are available[43].

1.2.4 Selected protein interaction database: Finally, we have chosen the PathwayCommons database for our basis for constructing protein interaction networks. This is due to the integration of multiple data sources in the PathwayCommons database, such as the Human Protein Reference Database (HPRD)[44] and the Reactome[45] database, among others. This integration allows us to cast a wide net in terms of the possible interactions we may incorporate in our networks. Additionally, in future work it would be possible to segregate information from PathwayCommons based on sources (either in terms of database or in terms of methodology).

Chapter 2: Occupancy Classification of PWM-predicted transcription factor binding sites.

2.1 Introduction and Background: In this chapter, we implement several versions of the occupancy classification approach to the problem of identifying true TFBS, in which machine learning techniques integrate information from multiple data sources to predict the occupancy or functionality of a given predicted binding site. Similar techniques that have incorporated additional genomic landscape information have shown improvement in performance over purely sequence-based techniques; however, an evaluation of the applicability of such techniques using multiple machine learning methods and multiple transcription factors has been lacking. We analyze the performance of these occupancy classifier versions, which use multiple machine learning methods as classifiers, several types of chromatin features and DNA sequence information features as predictors, and multiple publicly available chromatin immunoprecipitation (ChIP)-based TF binding data sets as training and test data set, and contrast and compare the results across transcription factors. We also compare the predictors most commonly selected among factors and use the best classifier built for each TF for cross-classifications of each TF, including a held-out TF (SRF). Our results demonstrate the viability of occupancy classification for many TFs and for use of classifiers trained on one TF to cross-classify sites for a different TF.

2.2 Methods: We identified regions of the human genome on chromosomes 1-22 found to bind a TF according to chromatin immunopreciptation for the factors c-Myc[26], TCF4[46], STAT1[6], GABP[7] and SRF[7]. These regions were mapped to the UCSC Genome Database hg18 build of the human genome[40,47], using a custom MySQL database (MySQL AB). We chose these TFs because each had high-quality whole-genome datasets available and all are thought to function primarily as transcriptional activators, either individually or in the beta-catenin/TCF4 complex in the case of TCF4. SRF was held out for later analysis as a blinded TF; no predictive models were constructed using SRF

data. TRANSFAC[8] PWMs for the factors were used to predict TFBS in the genome, using the TFBS Perl package[48] and a 95% similarity threshold; the scope of the analysis was limited to regions within +/- 3kb of an annotated transcription start site (TSS) for the 4 TFs in the analysis. Predicted TFBS within 1kb of the center of an empirically identified TF binding region were considered to be "high-occupancy" sites, while any other predicted site was considered low-occupancy. For each TF, we constructed ten sample data sets, each with 200 high-occupancy and an approximately equal number of low-occupancy sites per sample; there was some variation in the number of low-occupancy sites due to randomization. Individual predicted binding sites may appear in multiple data sets, but are only represented once in a given sample data set. We then trained Bayesian Network and SVM classifiers, using the Weka machine learning environment and the LibSVM SVM library[49], to discriminate between the sites using a variety of features. The features used were distances to nearest histone modification islands[42,43], nearest hypomethylation island as identified in leukocytes[41], and nearest CpG islands and nearest TSS as identified in the UCSC Genome Browser (hereafter referred to as TFBS-feature distances). We additionally incorporated distances between these nearest chromatin features to the nearest chromatin feature of a given type (hereafter referred to as feature-feature distances). These features are visually summarized in Figure 2.1. Feature-feature distances were capped at a maximum of 10kb between features to speed construction of the data sets. The specific classification algorithms used were:

(1)A BN using the K2 network-building algorithm[50], MDL-based discretization for binning[51], and the CFS-subset algorithm[52] for attribute selection.

(2)A linear-kernel SVM using default parameters, with distance features normalized to a 0-1 scale.

(3)A linear-kernel SVM with attributes preprocessed into bins using the same MDL-based discretization technique.

2.2.1 Evaluation: We evaluated the classifiers's ability to discriminate high and low occupancy sites using 10-fold cross-validation of each sample and the area under the curve (AUC) metric (resulting in a 10x10 cross-validation), and compared the algorithm and feature sets used for the best-performing classifier for each TF. For each TF, we also constructed classifiers on a per chromosome basis as described above, extracting training data from the other chromosomes. We also evaluated the performance of each best classifier on the other TFs in the study, using each sample as a training set and classifying each other sample from the other three TFs. Additionally, we evaluated the difference in performance between the classifiers when the feature-feature distances were excluded from the feature set, using cross-validation and the AUC as described previously. Finally, we performed an analysis of the agreement between TFs on relevant features based on number of times of a feature's inclusion in the cross-validation classifier using Cohen's kappa measure[53] as implemented in the e1071 package for R 2.7[54], and performed a cross-classification of the held-out SRF data set with each TF to further examine cross-classification performance.

2.3 Results: For the four TFs we analyzed (c-Myc, TCF4, STAT, and GABP) and features we utilized(distances to histone modifications, DNA hypomethylation, CpG islands and TSS) we found that the BN-based classifiers consistently outperformed SVM-based classifiers for all TFs, and achieved average AUC scores ranging from .71 (TCF4) to .94 (GABP) (See Table 2.1 and Figure 2.2). AUC Scores achieved in per-chromosome classification were comparable to those achieved in cross-validation (Table 2.2). The resultant classifiers have naïve Bayesian network structures. Both TFBS-

feature and feature-feature distances are predictive, but the feature-feature distances appeared to be the dominant predictors for all TFs. No features of either type appear to be universally predictive across all TFs. Classification of other TFs by a classifier trained on a different TF was universally inferior to performance of a classifier trained on the TF of interest, excepting the case of SRF. All TFs were capable of accurate cross-classification of the held-out SRF data set.

2.3.1 Comparison of Algorithms: In all cases, the BN classifier outperformed either of the SVM-based classifiers (see Table 2.1). The resulting network structures of the classifiers were equivalent to naïve Bayesian networks after discretization and attribute selection, with no edges between the predictors despite the use of the K2 algorithm. We attempted to use the Tree-Augmented Naïve Bayes algorithm[23] to induce additional structure in the networks, but this resulted in a marginally worse classification performance. Similarly, use of more sophisticated polynomial and radial basis function kernels for the SVM-based classifiers did not improve performance over the linear kernel, nor did manual adjustment of the slackness parameter of the optimization function. Typical size of the networks was on the order of 50 relevant predictors.

2.3.2 Contribution of Feature-Feature Distances to Classification: A surprising result of the classification experiments was the dominance of feature-feature distances over TFBS-feature distances in the best-performing classifiers. We therefore decided to rerun classification as above using TFBS-feature distances only. For this and all subsequent analyses, we chose to only construct BN classifiers as they had outperformed SVMs previously. In all cases, average AUC of classification improved with the inclusion of feature-feature distances relative to TFBS-feature distances only (see Table 2.3 and Figure 2.1). As with overall classification performance, it is difficult to determine how much of the variance in improvement is attributable to biological differences in the TFs versus technical differences

in the generation of the data sets. In general, however, the gain in performance tends to drop for ChIP-Seq data vs. data generated from other techniques. From a biological standpoint, the gain in performance could be interpreted from the point of view of the "histone code" hypothesis[57]; the predominance of feature-feature distances may be indicative of combinations of histone modifications in the proper proximity to one another to encourage the recruitment of TFs to a predicted site.

2.3.3 Individual Chromosome Classification: We additionally constructed BN classifiers using feature-feature distances for each individual chromosome and TF, generating training data from the other chromosomes. Performance was quite comparable on average to the performance achieved in the randomly sampled data sets, with the exception of STAT(Table 2.2). However, c-Myc and TCF4 showed considerable variance in performance between chromosomes relative to STAT and especially GABP, which were more consistent per chromosome; this effect was likely due to the lesser coverage of the binding data sets for c-Myc and TCF4. It is most likely that this coverage effect does not reflect a true lack of binding sites e.g. the low number of high-occupancy sites identified on chromosomes 21 and 22 for c-Myc does not reflect a genuine dearth of binding sites on those chromosomes (as evidenced by Cawley et al.[3], who discovered considerably more sites on these chromosomes with a high-quality tiling array approach. (See Table 2.4).

2.3.4 Common Predictors Across TFs: To identify whether or not common predictors were shared across the TFs, we performed a frequency count of all predictors which appeared in at least one classifier instance for all four TFs, over all cross-validations.; see Table 2.5 for top ten such predictors. The degree of concordance between the TFs once again appears to have a relationship with the method of generation of the data sets, with GABP and STAT showing relatively higher concordance with one another, using Cohen's kappa calculated over all attributes (average kappa = .47). c-Myc and TCF4

had less concordance with each other or with STAT or GABP (highest average kappa between any other combination of TFs was STAT-TCF4, with kappa = .43, see Table 2.6. Additionally, while the other 3 TFs have at least one predictor that is selected in 80+ builds of the classifier, TCF4's most commonly selected feature (distance to TSS) is selected only in 60 classifiers, and was only rarely selected by the other TFs. No specific predictor appears to be universally applicable to all of the TFs in this study. It is notable that 9 of the top 10 most commonly selected predictors involve distances to H3K4me3 modification islands (in particular, the top 3 feature pairs H3K4me2-H3K4me3, H3K27me1-H3K27me3, and TSS-H3K4me3, as well as H4K20me1-H3K4me3), as presence of H3K4me3 (as well as H3K4me2, H3K27me1 and H4K20me1) histone modification islands have been shown to be correlated with higher gene expression levels[43], which is biologically consistent with the generally accepted functions of the TFs in this study. While individual per-predictor probabilities varied depending on training set and subsequent binning, in general assigned probabilities behaved in a way consistent with biological evidence (e.g., closer distances with respect to the H3K4me histone mark usually assigned higher probabilities for high-occupancy to the TFBS site in question). Average class-conditional probabilities for high-occupancy sites in both the cross-classification and per-chromosome classification for the smallest distance bin in the top 10 most frequently occurring predictors are summarized in Tables 2.7a (cross-classification, reflecting averages only for two-bin cases and not including closest distance bins for three or more bins) and 2.7b (per chromosome, all cases).

2.3.5 Cross-Classification Performance: We additionally explored the classification performance achieved using one TF's classifier on the data for the other three TFs. Cross-classification performance may both be an indicator of the suitability of one TF for predicting occupancy of another (perhaps for exploratory modeling purposes), as well as a general indicator of the commonality of features

influencing occupancy of a TFBS. For this experiment, each of the ten samples from the previous experiments was used to train a BN classifier using the entire TF sample as training data. This classifier was then tested using each sample from each other TF, resulting in 100 AUC values for each TF-TF pair. We compared the average AUC of these values to the average AUC values achieved in cross-validation (Table 2.8 and Figure 2.3). Both TFBS-feature and feature-feature distances were included. In general, the classification performance achieved appears to correspond well with that achieved in the cross-validation experiments. Both STAT and GABP achieve cross-classification performance on one another comparable to that achieved in cross-validation; this result is sensible in light of the number of predictors that the two TFs were found to have in common during the predictor frequency analysis.

2.3.6 Cross-Classification of SRF: SRF represents a unique data set relative to the other TFs in that its dataset appears to contain multiple strong motifs that differ from the “canonical” SRF binding sequence; the TRANSFAC SRF PWM accounts for only about 33% of the sites reported by Valouev et al. After restricting to the 3kb window about the TSS, only 46 high-occupancy and 421 low-occupancy sites were identified. Because of this low sample size, we chose to exclude SRF from the general analyses described above. However, the data set does represent a tractable “use case” scenario for “blinded” occupancy classification; we hence decided to investigate the cross-classification performance of the classifiers trained on the TFs previously analyzed on the SRF dataset. SRF appears to be highly amenable to cross-classification (Table 2.9). While the SRF dataset is quite small, these results both indicate that occupancy classification can operate well on datasets with smaller sample size and an imbalance of high and low-occupancy TFBS sites and that the method used to generate a binding data set likely plays an important role in the accuracy of the evaluation of our classifiers, as the SRF data set is a ChIP-Seq data set and classification performance is comparable to that achieved on GABP and STAT. This is likely due to coverage differences; despite its smaller size, the SRF data set

is likely more complete in terms of covering most SRF TFBS with 95% similarity to the canonical PWM.

2.4 Discussion: While we were able to achieve good performance with occupancy classification overall, considerable variation in performance and in predictors was observed between the TFs. These differences may be attributed to both biological and technical factors affecting the TFs and the predictor data sets.

2.4.1 Biological factors: Biological differences between the TFs are likely at play; for example, GABP is thought to bind at the majority of human bidirectional promoters[36], and this may imply a stronger or more specific dependence on chromatin environment cues relative to the other factors. Conversely, c-Myc has been shown to have considerable binding activity outside of the window of analysis used in these experiments[3], This result may suggest that c-Myc TFBS occupancy in general may not be as sensitive to the chromatin environment local to the 5' region of genes.

2.4.2 Technical Factors: In addition to biological differences between TFs, differences between the methods with which the TF binding data and histone modification data were obtained may have contributed to differences in classification performance. For example, GABP and c-Myc also represent temporal and technical extremes, with c-Myc data generated in 2004 via paired-end ditag techniques vs. GABP data generated in 2008 via ChIP-seq; this illustrates the difficulty of separating technical from biological variation in performance in these results. We suspect that differences between the techniques used to generate the binding data we utilized, particularly with regard to site coverage, explain many of the discrepancies in performance between TFs. This is indicated by both the consistent trend towards better performance on TFs where ChIP-seq-based data sets were available, as well as the

per-chromosome analysis. Another difficulty arises from potential differences in the chromatin environment in the cell lines from which binding data was generated versus the cell lines from which histone modification and hypomethylation data was generated; ideally, all of the data for a given TF and predictors would be derived from the same cell lines. However, such data is not to our knowledge publicly available, and in its absence determining the exact contribution of potential cell line factors to the difference in classification performance for each TF is not possible. There is evidence that variation in histone modification proximal to core promoters or a TSS is less pronounced across cell types as compared to variation at enhancer regions[57], however, and this suggests that cell line variations are may be less likely to have a severe detrimental effect on our classification performance.

2.4.3 Summary: Even with the potential technical and biological limitations of the study, these results demonstrate that occupancy classification can perform quite well at classifying occupancy of TFBS predicted using PWMs in the promoter region of genes. Additionally, we demonstrate that the predictors we use are viable for application in occupancy prediction and that those predictors are selected in a manner consistent with the known biology of TF binding. The importance of feature-feature distances for prediction is also demonstrated. Our comparison of algorithms suggests that Bayesian network methods may be more effective candidates for implementing occupancy classifiers than methods using SVM algorithms. Finally, we demonstrate that accurate cross-classification of TFs by a classifier trained on a different TF is possible, using both our four training TFs on one another and on the SRF factor, which was held out for the entire training phase of all classifiers. This is an important result, as it demonstrates that occupancy classification can potentially operate with good performance on TFs for which training data is unavailable, suggesting that a generalizable occupancy classifier capable of handling many different TFs for which biochemical training data may not be available is a possibility worth further exploring.

Chapter 3. Generalizable Occupancy Classifiers

3.1 Background: Both per-TF occupancy classification and accurate cross classification are possible using the occupancy classification scheme developed in chapter 3. However, the question of generalizability remains unanswered; is it possible to construct a classifier capable of accurate occupancy classification for an arbitrary TF? While in the most general sense this question is likely not possible to exhaustively answer, we endeavor in this chapter to address the general plausibility of the idea of a generalizable occupancy classifier. We do this by utilizing the data sets described in Chapter 3.to construct classifiers which either combine data from multiple TFs into a single training set or combine multiple classifiers into a single classifier via stacking[30]. We compare the performance of these classifiers to classifiers trained on single TFs, and demonstrate that the performance of these classifiers is comparable to that of single-TF classifiers. We believe that this indicates that a classifier capable of achieving good occupancy classification performance on an arbitrary TF is a reasonable proposition, and that the methods we present in this chapter represent viable methods of achieving such a classifier.

3.2 Methods: Data for the c-Myc, TCF4, GABP and STAT TFs was used as in chapter 3.to construct generalizable occupancy classification schemes. SRF was once again held out as a blinded test set. Ten training/test splits were derived randomly from the combination of TF binding data. A training set consisted of 50 low and 50 high-occupancy sites per TF, with a test set consisting of 50 high and 500 low occupancy sites. This skewing of the ratio of high to low occupancy sites was thought to better reflect a biologically relevant ratio of high to low sites as observed in the analysis in Chapter 3, particularly with respect to SRF.

3.2.1 Generalizable Classification Schemes:The generalizable classification schemes we constructed

can be divided into three types:

1) Single-TF classifiers, in which a classifier was trained on only the training data of a single TF in a given training/test split.

2) Combined classifiers, in which a classifier was trained on the combined training data of a given training/test split as a single training set without discrimination between the TFs.

3) Stacking classifiers, in which the stacking method of Wolpert[30] was used to combine disparate classifiers (the "level-0" classifiers) with a classifier trained to discriminate high and low occupancy sites based on the output of the level-0 classifiers (the "level-1" classifier). For this purpose we utilized the Perceptron[53] algorithm as the level-1 classifier; initial experimentation with alternative level-1 classifiers (e.g. Bayesian network, simple voting) did not produce comparable results to those achieved with the Perceptron.

Stacking classifiers are further subdivided into two types; classifiers whose level-0 classifiers consist of classifiers trained on data from individual TFs in a given training-test split ("Stacking 1" classifiers), and those combining a Bayesian network and an SVM as their level-0 classifiers that had been trained on the entirety of the training set for a given train-test split without discriminating between TFs ("Stacking 2" classifiers). The general structure of the classifiers is visualized in Figure 3.1 and Figure 3.2 respectively. Additional variation was applied to examine the impact of specific TFs on classification performance in the form of "holdout" classifiers, in which a specific TF was removed from the training set; otherwise, data from all four TFs, including the one to be tested on, were included in the training set. Both Bayesian Networks and SVMs were employed as combined or level-0 classifiers, while Bayesian networks were used exclusively for the purpose of single TF classification. For purposes of evaluation, the AUC score and TP/FP and TN/FN rates were recorded for each classifier and combined into an average statistic for each method.

3.3 Results: The overall results of the analysis are presented in Table 3.2 and Figure 3.3. In general, the generalizable occupancy classification schemes performed comparably to single TF classification schemes for a given TF, in 3 of 4 cases outperforming the single-TF classifier for a given TF. Generalizable classifiers generally also outperform single-TF classifiers used for cross-classification. Some loss of performance was observed for the single TF classifiers relative to that achieved in chapter 2; this loss of performance may be due to the imbalance of high to low occupancy sites in the test data, or to differences in experimental methodology (cross-validation vs. single classification run). Combined classifiers gave the best overall performance, though the Stacking 1 methodology was capable of achieving superior TN rates. The Stacking 2 methodology was generally inferior to both combined and Stacking 1 methods. The trend across all generalizable classification methods was an increase in TN rate at the cost of some TP performance in comparison to the single-TF classifier, with the notable exception of GABP. Holdout classifiers did suffer some degradation in performance if the TF of interest in analysis was withheld; typically, this resulted in a loss of between .01-.05 AUC relative to a classifier of the same type including the TF of interest in its training set. Performance on SRF with the Combined and Stacking 1 classifiers was comparable for the former and slightly worse for the latter than that observed with single-TF classifiers in Chapter 2 (see Table 3.3)

3.4 Discussion: Overall performance of the generalizable classifiers was highly satisfactory in comparison to single-TF classifiers, suggesting that a generalizable occupancy classification scheme is a reasonable project and that a classifier utilizing either combined TF training sets or a combination of individual TF classifiers is a good candidate for such a scheme. While the Combined type classifiers had best overall performance, depending on the expected ratio of true binding sites to false positives, the more specific Stacking 1 classifiers may be preferable (e.g., in the case of genome wide

classification, potentially). The performance of holdout classifiers in which the TF to be classified was not part of the training set did show some degradation relative to single TF classifiers or when the TF to be classified was included in the training set. This result is not unexpected, however, and the degradation in performance is fairly modest. This is an important result for the applicability of generalizable occupancy classification in a practical setting, as one of the primary use cases for such a classifier would likely be to attempt to classify predicted TFBS for which biochemical data was not necessarily available. To summarize, these results indicate that generalizable occupancy classifiers can perform well in classification of predicted TFBS in the promoter region about a gene, even if lacking direct biochemical examples of high occupancy sites for the TF in question. If TF-specific training dat is available, they may be able to outperform a classifier based on that training data alone, as there appears to be added value in the addition of training data from other TFs in training the classifier. As such, the generalizable occupancy classifiers presented here are excellent candidates for either the extension of incomplete biochemical TF binding data sets or for de novo prediction of TFBS when biochemical TF binding data for a TF is unavailable.

Chapter 4: Similarity of protein interaction networks derived from occupancy prediction to those derived from biochemical TF binding data

4.1 Background: Prediction of TFBS is a difficult task in silico, and despite advances in biochemical methods is still incomplete for many TFs in many organisms and cell types. The occupancy classification paradigm presents an opportunity for researchers to make accurate inference of potential TFBS in silico for purposes of targeting further biological experiments, for hypothesis generation, or for assembly and refinement of biological models. However, genome-wide occupancy classification can produce a large number of potential high-occupancy sites and corresponding target genes, including inevitable false positives. In these circumstances, a method of prioritizing targets for further investigation through a combination of prediction quality metrics and analysis of potential impact of a target gene's biological function on a biological question of interest will be useful to researchers as an additional guide to researchers for the TFBS/target gene predictions in downstream analysis; one such metric of potential impact may be a gene's position in a transcriptional control or protein-protein interaction network. Furthermore, it is important as a confirmation of the validity of occupancy prediction that a reasonable degree of the network structure of e.g. protein-protein interaction networks implied by biochemical analysis of TF binding is recovered from the predicted version of the network as well, even if a large proportion of the initial PWM-based predictions are eliminated by a stringent occupancy classification procedure. We show that the gene targets returned by our occupancy prediction method returns a statistically significant portion of the protein-protein interaction network generated from biochemical binding data as compared to randomly generated networks even at strict thresholds for predicted high-occupancy of nearby TFBS.

4.2 Methods: TFBS targets were identified on chromosomes 1-22 from biochemical binding data for the TFs c-Myc and TCF4 . Predicted TFBS were then identified using the Stacking 1 method described

in 3.2. We chose the Stacking 1 method in preference to the other methods we presented in Chapter 3 for two reasons. First, we believe that in the genome-wide application the superior TN rate of the approach would be beneficial; secondly, the Stacking classifiers produce a more tractable range of conditional probabilities for analysis than the Combined BN classifiers, whose conditional probabilities showed little variance among positive or negative predictions. Ten versions of the generalizable classifier, each with different training sets, were constructed; these were identical to those described in chapter 3 and did include data about the TF of interest to the network analysis. Target genes were identified in the UCSC genome assembly (hg18)[40,47]. Biochemical target genes were identified as any gene with the midpoint of a biochemically identified binding region within 3kb of its TSS. Predicted target genes were identified with two protocols: either *the most equivocal site* (probability of high-occupancy closest to .5) or the site with *the highest probability of high-occupancy* within 3kb of the TSS of the gene were selected as representative TFBS for a given gene, and as prediction thresholds were varied through the experiments genes were identified as target genes as their representative site for the specified protocol met the threshold. Target genes were mapped to ENSEMBL[56] identifiers and then to gene names using the Bioconductor 2.5 R package[54,59]. Interaction networks were then created using the PathwayCommons[31] interaction network database and all interaction types; we chose to use all data sources and interaction types so as to consider as wide a number of sources of data as possible. Networks consisted of the target genes and their first neighbors. PathwayCommons is highly interconnected and use of other graph searches generally resulted in the recovery of a very large portion of the network which showed little variation between predicted and randomly generated networks or even between TFs.

4.2.1 Network Similarity Metric for Connected Target Gene Nodes: To assess the congruency of the predicted networks with those generated with occupancy prediction, we used two methods. Method 1 is

a variation on Balasubramanian. et al. (See Figure 4.1)[60]. To summarize, the method determines the statistical significance of the overlap between two networks by comparing the number of common edges between the two networks of interest versus the number of common edges between one of the networks of interest and a number of randomly generated networks in the following manner:

1) Protein interaction networks are constructed based on two data sources; in this case, the biochemical binding data and the predicted TFBS binding, as described above.
2) The number of edges in common between the two networks is counted
3) A number of randomly generated networks are created, and the number of edges in common with one of the networks of interest is counted; in this case, the biochemically derived network serves as the reference network.

A p-value is then assigned as the (number of random networks with equal or greater number of connected nodes present in original graphs/total number of random networks). In the original method, one of the networks of interest was permuted via either label or edge randomization to generate the random networks; however, as the question of interest is whether occupancy prediction is generating networks which preserve more of the biochemically-derived network structure to those generated by random selection of genes, we generated our random networks by randomly selecting a number of genes equal to the number of target genes selected by occupancy classification from the genes mapped to PathwayCommons and generating the resulting network from this random selection 1000 times. We repeated this process for selection thresholds of conditional probability of each site from 1 to 0 in increments of .05. We refer to this analysis as edge preservation, as it evaluates whether a statistically significant proportion of edges indicated by the biochemical binding data-based network are included in ("preserved") in the occupancy classification-based network. . We also repeated this analysis for c-Myc using the c-Myc Target Database[61]

4.2.2 Network Similarity Metric for Hub Preservation: For method 2, we examined the preservation of hubs between the biochemically derived and predicted networks. For this purpose, we defined a hub gene as a gene with a connectivity in the upper 5% of the connectivity distribution for the biochemical network. The number of genes which were identified as hubs according to this criteria in both the biochemical and the predicted network was counted. We then generated 1000 random networks, as in method 1, and compared this number of hub genes preserved in the occupancy predicted network versus the number of hub genes preserved by the networks in this random ensemble, and varied the selection threshold as described for method 1. Rather than using an exact p-value as in Method 1, significance was determined versus a normal approximation of the distribution of random network hub preservation. We refer to this analysis as hub preservation hereafter.

4.3 Results: For both c-Myc and TCF4, occupancy predicted networks preserved a statistically significant number of connected nodes and hubs versus a randomly selected ensemble of networks at an alpha of .05. This result held at most of the thresholds for sites with positive conditional probability greater than .5 for most variations of the classifier for both the equivocal and highest-probability site selection criteria.

4.3.1 Edge Preservation Analysis: Results of the edge preservation analysis are visually summarized in Figures 4.2-4.7. Lines in each figure represent the average p-value for all ten classifiers at the given threshold. For most of the thresholds, a significant number of edges were preserved on average by the ten classifier variations versus the random ensemble. The statistical significance of this preservation tended to be higher at a given threshold for the highest-probability site selection criteria than the most equivocal, and the threshold at which the preservation was significant tended to be higher for the former selection criteria; these results are intuitively sensible, as the most equivocal site selection

criteria is more restrictive for most of the genes in the analysis than the highest probability site selection criteria. In the regions where significance was not achieved in the extremes of the threshold spectrum, either not enough sites were selected to exceed the edge preservation of the random networks (in the higher end of the threshold spectrum) or that so many sites were selected that the number of preserved edges were essentially identical (in the lower end of the threshold).

4.3.2 Hub Preservation Analysis: Results similar to those observed in the edge analysis were observed in the hub preservation analysis, as can be seen in Figures 4.8-4.11 Once again, these graphs represent the average p-value for all ten classifiers at the given threshold. Thresholds which do not have a corresponding average p-value indicate that either too few or too many sites were selected, and no distribution of random networks different from the predicted network was possible to generate. For the c-Myc biochemically derived network, hubs were defined as nodes with a connectivity (k) of 13 or greater; for TCF4, the required k for a hub was 27. At reasonable predictive thresholds for classification of .4-.8 a significant number of hubs were preserved by occupancy classification versus random networks using both selection criteria for both TFs; this result is somewhat less clear for c-Myc hub preservation using the most equivocal site selection criteria.

4.4 Discussion: In this chapter, we demonstrate that a version of the generalizable occupancy classification scheme developed in Chapter 3 is capable of identifying high-occupancy sites for two disparate TFs such that the target genes identified by such sites reconstruct a statistically significant portion of a protein interaction network arrived at by biochemical identification of binding sites and target genes versus a random network background. This is demonstrated for two network metrics, the number of target genes connected in the biochemically derived network which retain edges in the occupancy predicted network, and the number of hub genes preserved between the two networks.

While the networks generated by low thresholds of occupancy classification also generate statistically significant network preservation in some cases, this is not unexpected since unfiltered PWM results represent the upper bound of potential recall for occupancy classification. In combination, the results from our two preservation analyses suggest that protein interaction networks constructed from occupancy prediction are likely to contain many of the same features as a network constructed from biological data.. These results are significant as they suggest that researchers can be comfortable using stringent thresholds for occupancy prediction and will likely recover a significant portion of a protein interaction network based on a biochemical TF binding experiment..Additionally, it indicates that methods utilizing protein interaction networks which base their inputs on the results of occupancy prediction are likely to capture many of the salient features of a network based on biochemical evidence for a given TF. Since such methods provide important and powerful tools for hypothesis generation or the construction of systems biology models, the ability to use occupancy classification in lieu of or as a supplement to biochemical TF binding data for such methods improves the flexibility of the methods and extends the range of biological problems they may be used to address into problems where TF binding data may be inadequate or even nonexistent.

Chapter 5: Use-case of protein interaction data to prioritize selection of c-Myc/TCF4 shared target candidates and confirmation of TCF4 and c-Myc binding in human colorectal cancer cells

5.1 Background: The number of potential high-occupancy TFBS and subsequent candidate target genes returned from a genome-scale occupancy classification analysis can be quite large, and may contain a significant number of false positive sites and target genes. Without additional guidance, it is possible that a researcher utilizing occupancy classification may have some difficulty in prioritizting these candidate sites/genes for subsequent followup and confirmation. While it is likely that many researchers using occupancy classification may have a driving biological question, such as interest in members of a particular pathway or family of genes, that may inherently limit the number of interesting candidates to a manageable number, such a driving question is not assumed to exist by the occupancy classification method; indeed, some biological questions by their very nature may require an analysis of a large or genome-scale data set, such as the extension of an incomplete TF binding data set. For such questions, it is plausible to assume a researcher may desire some algorithm or heuristic for narrowing down the number of potential candidate genes for carrying forward into additional analysis. We present in this chapter a method utilizing a combination of quality scores, inter-classifier agreement, and metrics based on protein interaction networks to address a specific biological question: the search for potentially important shared c-Myc and TCF4 targets. It has been well established that the b-catenin/TCF4 complex and c-Myc are important regulators of a number of biological processes and are in particular involved in oncogenesis. They are of particular import in colorectal cancers. Previous work has demonstrated that approximately 60% of genes involved in the Wnt signaling pathway in *Apc/Myc* double mutant mouse intestinal cells that were normally upregulated when APC is inactivated are not upregulated in the absence of functional c-Myc; many such genes are thought to be regulated by b-catenin/TCF4. The loss of upregulation without c-Myc, even with an excess of b-catenin due to the loss of APC, suggests that there may be a large number of genes which are targets for both the TCF4

and c-Myc transcription factors[62].It is reasonable to assume that the binding data used to train our occupancy classifiers in this work does not represent a complete picture of the targets of either TF due to both technical limitations of the parent studies and the temporal and contextual nature of TF binding, and neither would the mere intersection of those data sets represent a catalog of the shared targets of the two TFs. With occupancy classification, we have an opportunity to extend the list of targets for both c-Myc and TCF4 and to thereby identify potential shared targets between the two. We show how occupancy classification combined with the selection method described below allows for a principled selection of potential target genes for further analysis, confirm several predicted targets via chromatin immunoprecipitation, and demonstrate how the method serves as a model for similar analyses in the future.

5.2 Methods: Potential candidate genes and protein interaction networks were constructed as described in 4.2 for both c-Myc and TCF4, using ten variations of the Stacking 1 classifier. Each variation was trained on a unique combination of training data from all four TFs utilized in this study, as described in Chapter 3, and used to predict both TCF4 and c-Myc occupancy for chromosomes 1-22. After constructing the resulting protein interaction networks as described in Chapter 4, we decided to focus our attention on genes fitting the definition of "hub" described in 4.2 (e.g. connectivity k > 95% of the connectivity distribution of a biochemically-based network), since such hub genes have the potential to effect a number of biological processes through their multiple protein interactions. Our selection criteria resulted in a k = 27 requirement in the TCF4 network, while a minimum of k = 13 was used to define a hub for c-Myc, as in 4.2. Shared hubs were identified from the intersection of the hubs of the networks resulting from each pair of classifier variation predictions (e.g., Variation 1 c-Myc and Variation 1 TCF4, Variation 1 c-Myc and Variation 2 TCF4, and so on). Because of the non-uniform distribution of these agreement scores, non-parametric methods (e.g. Spearman's rho) were used to

evaluate the classifier pair agreements scores. Some shared hubs were what we term “secondary hubs”: i.e., while heavily connected to predicted target genes, we did not predict direct binding of one or both of the TFs at those genes themselves. The number of pairs of classifiers which identified a shared hub in both networks was noted, assigning one “point” per pair of classifiers, and resulting in a total score n out of 100 possible pairs of classifier variations and transcription factors; we refer to this as the classifier agreement pair score for that predicted hub. We then ranked the predicted hub genes by this score. We repeated this process for high-occupancy conditional probabilities of .5, .6, .7, and .8 using both the most-equivocal and highest-probability site selection criteria. A workflow of the overall protocol used in this and subsequent portions of the analysis can be found in Figure 5.1.

5.2.1 Analysis of External Supporting Evidence for Candidate Genes: To examine the face validity of the predictions and to identify potential targets as good candidates for a further analysis, we examined the support a given prediction had from outside sources. In the case of c-Myc, since we believe the binding data used for classifier construction to be particularly incomplete, we used the c-Myc Target Database[61] as a source of support for a given target. In the case of TCF4, we felt the binding data we chose was sufficiently large that it was a reasonable, albeit incomplete, snapshot of TCF4 binding. Predicted hubs were identified as being supported by none, one, or both of the data sources. From this data, we selected a list of potential targets which we felt were good candidates for further examination.

5.2.2: Prioritization of Candidate Genes: Candidate genes were prioritized according to the following criteria:

1) Biological interest with respect to colorectal cancer: We chose to prioritize predicted targets which had potential function in colon cancer processes; we additionally chose to focus on

targets with transcription factor and signaling activity, as well as targets with potential involvement in cytoskeletal reorganization.

2) Existing support from our data sources: Using the TCF4 binding data and the c-Myc Target Database as references, we chose targets which were not supported for binding of TCF4 and c-Myc in the reference data sources. Some potential target genes (e.g., CDC2/CDK1, SLK) had been shown to bind one or the other of the TFs of interest, but not both; in such cases we felt it reasonable to retain such genes for confirmation both to assess concordance with our supporting information as well as to confirm binding of the factor that did not have previous evidence indicating such binding
3) Classifier pair agreement score: We used this score as a guide to selecting potential target genes for confirmation, preferring genes with a higher score. Interestingly, some highly scoring genes (e.g. CHUK) were secondary hubs with regard to c-Myc; however, the high score of the targets indicates that it is possible for these target genes to be heavily involved with other c-Myc targets without being direct targets themselves, a phenomenon of potential interest; in combination with the biological function of these targets it was thought to be worthwhile to carry some of these genes forward for confirmation. Conversely, some genes of interest (e.g. SP1) had relatively low scores, and were selected primarily on the basis of biological function.

5.2.3 Confirmation of TCF4 and c-Myc binding by immunoprecipitation in HCT116 human colorectal cells: The method of confirmation of TF binding was as described in Bottomly et al[60]: the primary alteration to the protocol was the inclusion of c-Myc immunoprecipitation (Myc antibody Millipore 05-419) alongside TCF4. Otherwise, cells, conditions and preparations were as described. The same control region was used for both TCF4 and c-Myc.

5.3 Results: We decided to identify ~20 potential targets for further analysis, as it was felt this was a reasonable number for ChIP confirmation. During initial review of the data, we found that using the highest-probability site selection criteria led to a very large number of potential candidates (~700, for a threshold of .5). We found that a threshold of .6 in conjunction with most equivocal site selection led to a reasonable number of candidates for analysis (118); this list is reproduced in Table 5.1. We proceeded with support analysis and candidate selection from this basis.

5.3.1 Results of External Support Analysis and Candidate Selection: Initially, a total of 45 (38%) of our predicted targets had preexisting support from either or both of the c-Myc Target Database and the TCF4 binding data set used for training. It is important to remember that the genes identified as target genes in these data sources are likely to be incomplete; indeed, the very lack of support for the predicted targets is a motivation for developing this method. A higher classifier agreement pair score did not appear to significantly coincide with the presence of existing support for binding (Wilcoxon rank sum test in R, p=.12), nor did significant correlation exist with that score (Spearman correlation test in R, p=.12) suggesting that the number of agreeing pairs is best viewed as a heuristic. However, in light of the supporting data sources being almost certainly incomplete, it is possible that true targets might actually have higher classifier agreement pair scores if supporting data set with more complete coverage of (e.g., a ChIP-seq based data set) were used. With this in mind, we proceeded to select our recommended potential targets for further investigation that are annotated as such in Table 5.2; the biological rationale for each selection is also explained in Table 5.2.

5.3.2: Chromatin Immunoprecipitation of predicted TCF4 and c-Myc target genes: We selected a subset of 18 of the predicted target genes in Table 5.2 for confirmation via ChIP, as viable primers could not be designed for all the genes (see Table 5.3 for primer sequences). Initially, we confirmed

TCF4 binding. Of the 18 selected targets, 13 showed binding of at least 1.5 times the control in the experiment; we consider these confirmed predictions of TCF4 binding at the promoters of these genes, giving a confirmation rate of 72%. Of the 13 confirmed genes, 11 had no support in the TCF4 binding data set; both of the target genes with support from the TCF4 . binding data were confirmed (see Figure 5.1). We did not attempt to confirm c-Myc binding at all 18 candidate genes, since 6 of the hub genes identified are secondary hubs for c-Myc. Additionally, due to technical limitations, an additional eight target genes could not be tested for c-Myc binding, as the predicted binding sites were too far apart for the same primers to be used to detect both TCF4 and c-Myc binding. Of the 18 TCF4 targets assayed, 6 were further tested for c-Myc binding. Of those 6 genes, all 6 showed binding of 1.5 times the control or better for c-Myc for a confirmation rate of 100% (see Figure 5.2). In total, these results indicate four direct shared targets of TCF4 and c-Myc of 12 possible targets, with a fifth (CDC2) which has evidence of c-Myc binding in the c-Myc Target Database but which could not be confirmed in this study. Additionally, we predict and demonstrate TCF4 binding in the region near the TSS of the GTF2F2 gene. GTF2F2 is indicated as a c-Myc target gene by the c-Myc Target Database, but we do not predict c-Myc binding at GTF2F2 with our occupancy classification and prioritization protocol. The overall results of the ChIP experiments are summarized in Table 5.4

5.3.3 Improvement of correlation of classifier agreement pair score with support: As noted above, initially, number of classifier pairs in agreement about a gene did not significantly correlate with support from binding data and/or literature. However, if the TCF4 and c-Myc binding data subsequently derived from our predictions is included, the association between classifier agreement pair score and support becomes significant (Wilcoxon rank sum test p-value > .001), with a correlation coefficient of .3 (Spearman's rho, p-value = .001). This suggests that the agreement between multiple occupancy classifiers is a useful metric for prioritization of target selection in future analyses using

similar protocols to this analysis.

5.3.4 Overall performance of protocol: The protocol described here was capable of successfully prioritizing a small number of predictions for validation from a large number of initially predicted high-occupancy TFBS binding sites and associated genes. Depending on the specific classifier used, approximately 3000-4000 genes possessed predicted high occupancy sites and could be mapped to the PathwayCommons database for network analysis. Restriction by requiring a minimum of a .6 conditional probability of high-occupancy at all predicted TFBS and imposing hub criteria reduced this number to 118 candidate genes. Of these candidate genes, 13 of 18 tested sites were confirmed to show TCF4 binding, while 6 of 6 genes assayed for c-Myc showed binding. Of the TCF4 sites, only 2 of the 13 were previously supported by the TCF4 . binding data we utilized for classifier construction, while only 2 of the 6 genes binding c-Myc were previously supported in the c-Myc Target Database. Of 20 total predictions of binding at a gene which were tested, 15 (75%) were confirmed. None of the 4 shared TCF4 and c-Myc targets that demonstrated direct binding could have been identified from the intersection of the c-Myc Target Database and the biochemical TCF4 binding . data we used. Of the 5 TCF4 direct targets which showed up as secondary hubs for c-Myc in our analysis, only one was supported in the TCF4 binding data as a direct target. Interestingly, all five of these TCF4 direct targets show up as secondary hubs for a c-Myc network built using the c-Myc Target Database as the basis and k = 13 as the minimum hub connectivity criteria, suggesting that their interaction with multiple c-Myc targets has some degree of existing support; furthermore, only one of these genes (SP1) is identified as a secondary hub if the biochemical c-Myc binding data is used as the basis for the protein interaction network (i.e., they do not appear as hubs in the biochemically derived network for c-Myc). These overall results, including the existing support for the predicted genes, is summarized in Table 5.5.

5.4 Discussion: We provide in this chapter an example of the application of occupancy classification for purposes of addressing a specific biological question. We demonstrate a method of combining the results of multiple occupancy classifiers, prediction quality data, and protein interaction network data to prioritize the results of a genome wide prediction of target genes for two TFs (c-Myc and TCF4), and confirm a sizable majority of the predictions biochemically for both factors. Furthermore, there is no reason to believe that mutual targeting of genes between two TFs is a phenomenon exclusive to c-Myc and TCF4. This method provides an explicit way to perform occupancy classification experiments to add weight to or explore the possibility of such mutual targeting phenomena. Finally, this method presents occupancy classification as part of an integrated process of analysis of a biological question, incorporating quality metrics internal to the process of occupancy classification along with biological knowledge and graph theoretic analysis, demonstrating occupancy classification as a valuable addition to the arsenal of computational techniques available for researchers interested in computational prediction of transcription factor binding sites and target genes.

5.4.1 Discussion of potential biological significance of confirmed target genes: From the perspective of colorectal cancer, many of the genes identified could have considerable significance. We restrict discussion here to the targets with biochemical evidence from our experiments of both TCF4 and c-Myc binding (TRAF2, SLK, IQGAP1, THRAP3). Recent evidence[63] indicates a role for TRAF2 in prevention of apoptosis of colorectal cancer cells; amplification of its expression by c-Myc and/or an activated TCF4/ß-catenin complex could confer a selective advantage to a tumor cell population. TRAF2 was a known c-Myc target but was not identified as a TCF4 target previously. The SLK kinase has been shown to be important in cell motility in breast cancer[64]; while the Hatzis et al. screen did identify TCF4 binding at this gene, we believe the discovery of c-Myc binding is novel. IQGAP1 is implicated in cell motility and cytoskeletal reorganization, and recently has been shown to be a

prognostic marker for the severity and likely invasiveness of colorectal cancers[65]. While the literature does not have extensive information about the role of THRAP3 in colorectal cancers, a recent study has indicated that THRAP3 complexes with the SNIP1 protein and other proteins, and that this complex may be key in regulating the stability of Cyclin D1 RNA, overexpression of which has been associated with malignancy[66]. Overall, many of the TCF4 and c-Myc targets we identify here are members of the NF-kappa-ß signaling pathway (TRAF2, CHUK) involved in suppression of the pathway, or are otherwise involved in cell cycle control or transcriptional activation (CDC2, CRKRS, THRAP3, SP1). A considerable number of the targets have roles in cell motility and cytoskeletal reorganization (ANP32A, IQGAP1, PAK4, SLK), which has been shown to influence tumor invasiveness. Additionally, we believe many of these sites may be biologically "normal" sites which could be active in non-cancerous tissues, despite identifying these binding sites in a colon cancer cell line. Both the chromatin feature data used to predict binding and the original biochemical binding data used to train our classifiers were derived from multiple cell lines, and recent evidence[56] indicates that at least histone modifications show reduced variability across cell lines at promoter regions. While it is impossible to definitively determine whether or not the identified binding sites are active in normal tissue from these results, our ability to identify these sites with many heterogenous data sources in concert with the conservation of epigenetic modification in promoter regions suggest that it is possible these sites are active in normal tissues as well.

ACKNOWLEDGEMENTS: The author would like to thank Dr. Gregory Yochum and Sydney Kyler for performing the chromatin immunoprecipitation experiments described in this chapter.

Chapter 6: Discussion of results, relevance, and future directions

6.1 Significance of initial classification and cross-classification experiments

Results of the initial classification experiments presented in Chapter 2 indicate that occupancy classification is in fact a viable method of discriminating true high occupancy TFBS sites from low occupancy sites from a set of sites predicted by position-weight matrix and that occupancy classifiers trained on one TF can successfully classify predicted TFBS sites belonging to another TF. While the results of this analysis are necessarily limited in scope, this does not invalidate those results as a validation of occupancy prediction. Furthermore, these results are in accord with previous results and recent developments in the field of TFBS binding prediction and serve to advance the field of research and add to the body of knowledge regarding TFBS binding site prediction in silico.

6.1.1 Limitations of Analysis

Regarding the limitations of the analysis, an obvious limitation is that only five total TFs were used in the study. Unfortunately, at the time of the conception and execution of the study, high-quality TF binding data sets in humans were relatively rare. The decision to limit the analysis to activating transcription factors only further limited the potential selection of TFs. However, it is unrealistic to assume that any analysis of this type could exhaustively cover all possible TFs, and indeed biochemical binding data for all possible TFs would largely eliminate the need for such an analysis altogether. Additionally, the lack of high-quality data sets for binding motivated the use of older data sets whose coverage and resolution are not to the standards of more modern techniques, which in turn required the use of a relatively large window about the center of biochemically defined binding regions to define high-occupancy sites for the analysis. It is possible that this large window resulted in misidentification of sites that may not truly be high-occupancy. The impact of this effect may be mitigated for the higher quality techniques somewhat, as the likelihood of a high-occupancy site not possessing a corresponding biochemical hit is lessened. This means that the odds of a true low-

occupancy site being misidentified in the region surrounding a true high-occupancy site are most likely lower for these techniques; this may be reflected in the generally superior results of the ChIP-seq derived GABP and STAT binding sites, though some of this performance improvement is almost certainly due to the improved coverage of these data sets relative to c-Myc and TCF4 (c.f., the considerable increase in per-chromosome performance variability observed in Chapter 2). Nonetheless, it is unlikely that the large window used to define high occupancy sites constitutes a serious weakness of the analysis; given higher quality data, however, it would be preferable to define as narrow a window for defining a high occupancy site as possible.

6.1.2 Comparison to existing work: In comparison with other work in the field, the analysis presented addresses several unanswered questions about occupancy classification. The most directly comparable work is that of Chen et al., who constructed a c-Myc classifier using a Bayesian network and distances to various DNA and chromatin features as well as sequence conservation. However, while Chen et al. do begin to address the issue of cross-classification by attempting to cross-classify CREB binding sites using their classifier, they do not address the issue of algorithm comparison in any capacity. The analysis presented in this work both addresses the issue of cross-classification of TFs in greater depth than Chen et al. as well as addressing two distinct algorithms for classification. Two additional novel features separate this analysis from that of Chen et al.; the construction of our data sets from raw binding data and the use of feature-feature distances in the classifiers. Chen et al. use a data set which quantitatively identifies the level of several c-Myc binding sites. In contrast, this analysis uses only binding data which is not quantitated beyond presence/absence (i.e, quantitative levels of protein binding are not used to segregate high and low occupancy sties, as in the Chen et al. study), and yet achieves reasonable performance, demonstrating that quantitation is not a prerequisite for training accurate occupancy classifiers. The use of feature-feature distances is to our knowledge unique for

purposes of identifying high occupancy TFBS, and is not present in Chen et al. A more recent work is that of Won, Ren, and Wang[67], which uses a Hidden Markov Model-based approach to accurately identify binding sites for 13 distinct TFs in mouse. The approach of Won, Ren, and Wang has certain advantages over the analysis presented here, notably that it was able to address many more TFs with an overall higher data quality. The HMM-based approach additionally is able to address enhancer regions distinct from promoter regions, which this analysis does not address. With that said, the Won, Ren, and Wang approach by no mean supplants the analysis presented here. Our analyses are performed in distinct species (human vs. mouse) and utilize distinct TFs, though the Won, Ren and Wang analysis does use c-Myc and a STAT family member. It also does not seem as if the HMM-based approach necessarily outperforms the approach presented in this analysis; the AUC values presented in the Won, Ren and Wang analysis do not greatly exceed those achieved by Bayesian networks on the high-quality data sets in this analysis, and the degree to which this improvement is due to differences in the definition of high-occupancy sites, differences in quality of TF binding data, cell line differences in histone modification data, and of course differences between species is not possible to determine. The HMM-based classifier also does not address the issue of cross-classification directly in terms of the effect of one TF being used to train a classifier distinct from the TF the classifier is used to perform occupancy classification on, as it appears that an individual model is trained per TF in HMM-based method. Won, Ren and Wang do not appear to address the use of feature-feature distances in any fashion, though they claim to implicitly capture a periodicity of histone modification signals with their model. Finally, the method presented in this work is agnostic to the method used to identify potential TFBS, whereas the identification of TFBS is intimately tied into the HMM model used by Won, Ren and Wang; the ability to overlay the methods of this work transparently onto any given binding site discovery algorithm may be useful in generalizing occupancy classification or to tailoring it to specific needs. However, the analyses do share important common features, notably the reliance on histone

modifications as primary inputs to the classifier. While the two approaches present competing methods for achieving occupancy classification, they both make important contributions to the literature regarding binding site prediction in silico.

6.1.3 Summary of significance: The work presented in chapter 3 provides considerable additional support to the body of literature demonstrating that occupancy classification is a viable approach to improving in silico TFBS prediction, by confirming that an occupancy classification approach is viable for several different TFs, by comparing multiple algorithms with regard to their suitability for the task of occupancy classification, by showing that quantitated binding data is not necessary to train accurate classification models, through the introduction of feature-feature distances to the repertoire of features for occupancy classification, and by demonstrating that accurate cross-classification is possible for several different cominbations of TFs. This last point in particular is important, as it has not been well demonstrated in the literature to date and opens the door to the possibility of constructing a generalizable occupancy classification scheme.

6.2 Significance of the generalizable occupancy classification experiments: The generalizable occupancy classification experiments presented in chapter 3 represent an additional contribution to the literature of occupancy classification above and beyond the cross-validation and cross-classification experiments of chapter 2. though they share many similar features. Arguably, most of the weaknesses of the analysis of chapter 3 are retained since the generalizable occupancy classifiers derive directly from those developed in chapter 2. however, as discussed in 6.1, most of these weaknesses are unavoidable byproducts of the available data sets and are incidental to the results reported. Many of the strengths of the cross-validation and cross-classification analysis are carried over to this work as well: in particular, the focus on cross-classification of TFs for which the classifiers used are not trained and

the contest of multiple algorithms for the task. These strengths are integral to the unique contributions the work makes to the literature regarding occupancy classification.

6.2.1: Comparison to existing work in the field: Once again a comparison may be drawn between this work and that of Won, Ren, and Wang. As discussed in 6.1.2, the HMM-based approach used by these authors is inherently designed as a generalizable approach, with the advantage of additionally handling enhancer regions in addition to promoter regions. The work here nonetheless possesses unique attributes which set it apart from the work of Won, Ren, and Wang. Notably, the aforementioned paper does not engage in comparison of potential algorithms, as this work does in some detail. Also, the use of cross-classification and "holdout" variations of the classifier to assess the impact of leaving a TF out of the training set and to determine the performance of the method on TFs which were not included in the training data set are not replicated in the work of Won, Ren and Wang, due to the necessity to train individual models per TF in their method. Most of the other salient similarities between the two approaches are summarized in 6.1 and need not be repeated here.

6.2.2 Summary of significance: Ultimately, the primary contribution of the analysis of generalizable occupancy classification presented in this work is the demonstration that a generalizable occupancy classifier is possible. Beyond the cross-classification procedure used in chapter 2, this work demonstrates that combinations of TF training set data are capable of producing generalizable classifiers equalling or exceeding the performance of a single-TF classifier, and of performing with little loss of accuracy in comparison to a single TF classifier if that TF has not been used to train the generalizable classifier in question. The work compares and contrasts a number of approaches to combination of the training data and to algorithms used for classification, illustrating strengths and weaknesses and suggesting the applicability of specific variations to specific needs (e.g., the use of a

Stacking type classifier in genome-wide prediction for purposes of leveraging its high TN rate). This analysis provides strong evidence for the viability of generalizable occupancy classifiers for TFBS as well as demonstrating and comparing several methods for the construction of such classifiers.

6.3 Significance of network analysis and shared c-Myc and TCF4 target prediction: The analysis of the similarity of protein interaction networks derived from biochemical TF binding data vs. those derived from occupancy prediction and the subsequent use of network metrics to predict shared c-Myc and TCF4 targets is to our knowledge unique in the field.

6.3.1 Significance of network similarity analysis: The analysis presented in chapter 4 provides additional support to the assertions of accuracy of generalizable occupancy classification; it demonstrates that occupancy classification reconstructs a statistically significant portion of the network structure as biochemical data in terms of adjacent nodes and in terms of hub identity as opposed to that achieved by random selection of an equal number of nodes to construct a network. In the most permissive case utilizing the highest probability site selection criteria, our best classifier variation identifies 920 hub genes for TCF4 at the .8 high-occupancy threshold, of which 115 (12.5%) are included in the TCF4 binding data. By contrast, the unfiltered TCF4 PWM identifies 1412 hub genes, of which 175 (12.3%) are included in the TCF4 network. Over a third of the genes indicated by PWM hits may be discarded while still retaining two-thirds of the PWM-identified hubs; while this may seem like a modest improvement, the incomplete coverage of the TCF4 binding data must be considered. We believe it is likely that an even higher percentage of the hubs identified by occupancy classification would be confirmed given binding data with better coverage. To some extent, this supposition is borne out given our success in identifying novel TCF4 binding sites in Chapter 5. In contrast to the highest-probability criteria, the more stringent most-equivocal criteria identifies only 47

hub genes at the .8 threshold, but 13 (27%) are included in the TCF4 binding data; the resultant selection is also significant in terms of edge and hub preservation versus the random network background. This result indicates that even if a large number of PWM-identified TFBS are discarded by a stringent occupancy classification threshold, a significant amount of structure in the biochemically derived network is retained. These results are themselves useful contributions to the field, as they suggest network models may be constructed using occupancy classification with some confidence that their results are likely to replicate the results that would be expected from use of biochemical binding data as a basis for the analysis.

6.3.2: Significance of common c-Myc and TCF4 target selection and prioritization: By demonstrating the use of occupancy classification in conjunction with network metrics derived from a protein interaction network to address the biological question of shared c-Myc and TCF4 targets in chapter 5, this work demonstrates a unique way in which occupancy classification can be used to address a specific biological question, as well as providing a framework by which questions of the same type may be addresses in future. The protocol is admittedly arbitrary in its selection of quality and network metric cutoffs for selection and the selection of genes is biased in favor of those most likely to be of import to colorectal cancer processes; as well, the use of the most equivocal selection criteria may bias selection to genes near fewer predicted TFBS. Nonetheless, the potential target list is arrived at in a principled fashion and the cutoffs for selection are obviously adjustable to suit the needs of a particular biological problem. The approach is conceivably directly extensible as is to any pair of TFs that are thought to share common target genes, making it an excellent candidate for both hypothesis generating research and for building additional confidence in or extending existing models of TF cooperativity or competition in silico. Finally, this method identified a large proportion of targets which were shown to be true targets of a TF of interest in the analysis via chromatin immunoprecipitation as well as four

targets which were confirmed to be shared between the two TFs of interest, validating both occupancy classification and the target prioritization algorithm as useful tools for extending the known targets of a given TF, as well as identifying new targets of the TCF4 and c-Myc transcription factors whose potential regulation by these TFs may have important roles in the pathology of colon cancer or in other biological processes.

6.4 Future Directions and Potential Extensions/Applications of Occupancy Classification: At the completion of this work, many questions and potential extensions and applications of the Occupancy Classification paradigm remain unexamined.

6.4.1 Extension to additional TFs, cell types/species, and genomic contexts: Most obviously, there remain a considerable number of TFs which are not addressed in this work, and it is possible that the framework presented here may not be applicable to occupancy classification for all of them. Notably, the study design deliberately rejected TFs with primarily deactivating or expression-suppressing activity. Whether or not the generalizable classification scheme presented here would operate accurately on such factors is unknown. Similarly, the design focused exclusively on a relatively small promoter region about the TSS of genes; whether or not the methods presented here are good fits for predicting occupancy of TFBS in enhancer regions or for TFBS located intergenically or in the 3' regions of genes is an open question. Many of the histone modifications used to make predictions of high occupancy are correlated with higher gene expression[41], which, while biologically sensible for activating TFs, may not be applicable to deactivating factors or to factors operating outside of the promoter regions of genes. Certainly, the question of cell line variation affecting classification accuracy or the viability of interspecies occupancy classification remain unexamined, though such work may require biochemical data which is unavailable at this time (e.g histone modification data derived from

multiple cell lines or in multiple species). Future work examining these issues would be of tremendous value in further developing occupancy classification as a method.

6.4.2: Algorithmic and methodological questions of interest: A number of algorithmic issues with occupancy classification remain unaddressed. Of course, there remain a myriad of machine learning techniques that were not used in the comparison of algorithms presented in this work, and it is entirely possible that there may be approaches which outperform the Bayesian network approach that proved superior here. More interestingly, since the occupancy classification method in this work is agnostic to the source of the TFBS predictions it classifies, the effect of the source of those predictions on the accuracy of occupancy classification is unknown. How occupancy classification operates on predictions from ab initio TFBS discovery methods, for example, is unknown, nor have we addressed the effects of using PWMs derived from alternate methods of PWM construction such as protein-DNA interaction models[68]. The ability of ab initio discovery procedures to identify common potential TFBS between genes also suggests the possibility of using such methods to identify most likely candidates for TFBS for a gene set and then to further use occupancy classification to add weight to or cast doubt upon those predictions. For purposes of exploring novel sets of potentially coregulated genes such an approach could be quite powerful in discovering common TFs underlying coregulation.

6.4.3 Model perturbation: An unexplored facet of occupancy classification in this work involves model perturbation. As many of the occupancy classifiers in this work are essentially probabilistic modesl of the effects of particular chromatin modifications and sequence factors on the occupancy of a given TFBS, it is possible to determine the most likely effect on any given TFBS of a change in the state of those factors using such occupancy classifiers. As epigenetic remodeling increasingly is shown to be a vital factor in developmental and disease processes, the ability to theoretically permute the state of

epigenetic modifications affecting TFBS occupancy and to observe the likely effect of those changes on a given TFBS could be quite valuable in understanding why particular modifications lead to particular perturbations in biological behavior. An accurate occupancy classification model as presented in this work allows researchers to perform exactly this type of experiment and to thereby generate hypotheses or construct models of the effects of epigenetic changes on TFBS occupancy and its effect on biological processes.

6.4.4 Possibilities for network-based and other multi-method analyses: Finally, the possibilities of network-based analyses are by no means completely explored in Chapters 4 and 5 of this work; only the outlines of their possibilities are sketched. Indeed, the question of integrating occupancy classification with other types of information such as expression, genome-wide association and protein interaction data is an extremely fertile field of study. The form and effect of such integration is highly dependent on the question being addressed in the study; while the work presented in Chapter 5 provides a strong framework for addressing the specific question of TFs targeting mutual genes and shows a method for extracting potential targets from genome-wide occupancy prediction for further study in an informed manner, such a framework may not be appropriate for other circumstances or data types. It is the hope of this author that this work, in conjunction with others addressing similar questions, will sufficiently illustrate the possibilities of occupancy classification to not only spur development of occupancy classification methods, but also to find ways to fully exploit occupancy classification's possibilities as part of an integrated approach to modeling gene regulatory behaviors and discovering and elucidating their roles in biological and disease processes of interest.

BIBLIOGRAPHY

1. Fickett, J. W. and Wasserman, W. W. (2000). Discovery and modeling of transcriptional regulatory regions. Curr Opin Biotechnol , 11(1), 19–24.

2. Massague, J. (1998). TGF-beta signal transduction. Annu Rev Biochem , 67, 753–91.

3. Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T. R. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. Cel l , 116(4), 499–509.

4. Impey, S., McCorkle, S. R., Cha-Molstad, H., Dwyer, J. M., Yochum, G. S., Boss, J. M., McWeeney, S., Dunn, J. J., Mandel, G., and Goodman, R. H. (2004). Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. Cell , 119(7), 1041–54.

5. Johnson, D.S., Mortazavi, A., Myers, R. M. and Wold, B (2007) Genome wide mapping of in vivo protein-DNA interactions. Science 316(5830), 1497-502

6. Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-seq data. Nat Methods , 5(9), 829–34.

7. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods , 4(8), 651–7.

8. Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006). TRANSFAC and its module TRANScompel: transcriptional

gene regulation in eukaryotes. Nucleic Acids Res , 34(Database issue), D108–10.

9. Kel, A. E., Gossling, E. Reuter, I, Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res, 31(13), 3576-9.

10. Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982b). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in e. coli. Nucleic Acids Res , 10(9), 2997–3011.

11. Staden, R. (1984). Computer methods to aid the determination and analysis of dna sequences. Biochem Soc Trans , 12(6), 1005–8.

12. Roulet, E., Fisch, I., Junier, T., Bucher, P., and Mermod, N. (1998). Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. In Silico Biol , 1(1), 21–8.

13. Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intel l Syst Mol Biol , 2, 28–36.

14. Pavesi, G., Mereghetti, P., Mauri, G., and Pesole, G. (2004). Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res , 32(Web Server issue), W199–203.

15. Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouz´e, P., and Moreau, Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. Bioinformatics , 17(12), 1113–22.

16. Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., R´egnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of

transcription factor binding sites. Nat Biotechnol , 23(1), 137–44.

17. Lifanov, A. P., Makeev, V. J., Nazina, A. G., and Papatsenko, D. A. (2003). Homotypic regulatory clusters in drosophila. Genome Res , 13(4), 579–88.

18. Philippakis, A. A., He, F. S., and Bulyk, M. L. (2005). Modulefinder: a tool for computational discovery of cis regulatory modules. Pac Symp Biocomput , pages 519–30.

19. Ferretti, V., Poitras, C., Bergeron, D., Coulombe, B., Robert, F., and Blanchette, M. (2007). Premod: a database of genome-wide mammalian cis-regulatory module predictions. Nucleic Acids Res, 35(Database issue), D122–6.

20. Defrance, M. and Touzet, H. (2006). Predicting transcription factor binding sites using local over- representation and comparative genomics. BMC Bioinformatics , 7, 396

21. Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., Scott, G., Steffen, D., Worley, K. C., Burch, P. E., Okwuonu, G., Hines, S., Lewis, L., DeRamo, C., Delgado, O., Dugan-Rocha, S., Miner, G., Morgan, M., Hawes, A., Gill, R., Celera, Holt, R. A., Adams, M. D., Amanatides, P. G., Baden-Tillson, H., Barnstead, M., Chin, S., Evans, C. A., Ferriera, S., Fosler, C., Glodek, A., Gu, Z., Jennings, D., Kraft, C. L., Nguyen, T., Pfannkoch, C. M., Sitter, C., Sutton, G. G., Venter, J. C., Woodage, T., Smith, D., Lee, H.-M., Gustafson, E., Cahill, P., Kana, A., Doucette-Stamm, L., Weinstock, K., Fechtel, K., Weiss, R. B., Dunn, D. M., Green, E. D., Blakesley, R. W., Bouffard, G. G., De Jong, P. J., Osoegawa, K., Zhu, B., Marra, M., Schein, J., Bosdet, I., Fjell, C., Jones, S., Krzywinski, M., Mathewson, C., Siddiqui, A., Wye, N., McPherson, J., Zhao, S., Fraser, C. M., Shetty, J., Shatsman, S., Geer, K., Chen, Y., Abramzon, S., Nierman, W. C., Havlak, P. H., Chen, R., Durbin, K. J., Egan, A., Ren, Y., Song, X.-Z., Li, B., Liu, Y., Qin, X., Cawley, S., Worley, K. C., Cooney, A. J., D'Souza, L. M., Martin, K., Wu, J. Q., Gonzalez-Garay, M. L., Jackson, A. R., Kalafus, K. J., McLeod, M. P., Milosavljevic, A., Virk, D., Volkov, A., Wheeler, D. A., Zhang, Z., Bailey, J. A., Eichler, E. E., Tuzun, E., Birney, E., Mongin, E., Ureta-Vidal, A., Woodwark, C.,

Zdobnov, E., Bork, P., Suyama, M., Torrents, D., Alexandersson, M., Trask, B. J., Young, J. M., Huang, H., Wang, H., Xing, H., Daniels, S., Gietzen, D., Schmidt, J., Stevens, K., Vitt, U., Wingrove, J., Camara, F., Mar Alb`a, M., Abril, J. F., Guigo, R., Smit, A., Dubchak, I., Rubin, E. M., Couronne, O., Poliakov, A., Hubner, N., Ganten, D., Goesele, C., Hummel, O., Kreitler, T., Lee, Y.-A., Monti, J., Schulz, H., Zimdahl, H., Himmelbauer, H., Lehrach, H., Jacob, H. J., Bromberg, S., Gullings-Handley, J., Jensen-Seaman, M. I., Kwitek, A. E., Lazar, J., Pasko, D., Tonellato, P. J., Twigger, S., Ponting, C. P., Duarte, J. M., Rice, S., Goodstadt, L., Beatson, S. A., Emes, R. D., Winter, E. E., Webber, C., Brandt, P., Nyakatura, G., Adetobi, M., Chiaromonte, F., Elnitski, L., Eswara, P., Hardison, R. C., Hou, M., Kolbe, D., Makova, K., Miller, W., Nekrutenko, A., Riemer, C., Schwartz, S., Taylor, J., Yang, S., Zhang, Y., Lindpaintner, K., Andrews, T. D., Caccamo, M., Clamp, M., Clarke, L., Curwen, V., Durbin, R., Eyras, E., Searle, S. M., Cooper, G. M., Batzoglou, S., Brudno, M., Sidow, A., Stone, E. A., Venter, J. C., Payseur, B. A., Bourque, G., L´opez-Ot´ın, C., Puente, X. S., Chakrabarti, K., Chatterji, S., Dewey, C., Pachter, L., Bray, N., Yap, V. B., Caspi, A., Tesler, G., Pevzner, P. A., Haussler, D., Roskin, K. M., Baertsch, R., Clawson, H., Furey, T. S., Hinrichs, A. S., Karolchik, D., Kent, W. J., Rosenbloom, K. R., Trumbower, H., Weirauch, M., Cooper, D. N., Stenson, P. D., Ma, B., Brent, M., Arumugam, M., Shteynberg, D., Copley, R. R., Taylor, M. S., Riethman, H., Mudunuri, U., Peterson, J., Guyer, M., Felsenfeld, A., Old, S., Mockrin, S., Collins, F., and Rat Genome Sequencing Project Consortium (2004). Genome sequence of the brown norway rat yields insights into mammalian evolution. Nature , 428(6982), 493–521.

22. Chen, Y., Blackwell, T. W., Chen, J., Gao, J., Lee, A. W., and States, D. J. (2007). Integration of genome and chromatin structure with gene expression profiles to predict c-myc recognition site binding and function. PLoS Comput Biol, 3(4), e63.

23. Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. Machine Learning , 29(2-3), 131–163.

24. Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res , 15(8), 1034–50.

25. Fernandez, P. C., Frank, S. R., Wang, L., Schroeder, M., Liu, S., Greene, J., Cocito, A., and Amati, B. (2003). Genomic targets of the human c-myc protein. Genes Dev , 17(9), 1115–29.

26. Zeller, K. I., Zhao, X., Lee, C. W. H., Chiu, K. P., Yao, F., Yustein, J. T., Ooi, H. S., Orlov, Y. L., Shahab, A., Yong, H. C., Fu, Y., Weng, Z., Kuznetsov, V. A., Sung, W.-K., Ruan, Y., Dang, C. V., and Wei, C.-L. (2006). Global mapping of c-myc binding sites and target gene networks in human B cells. Proc Natl Acad Sci U S A, 103(47), 17834–9.

27. Zhang, X., Odom, D. T., Koo, S.-H., Conkright, M. D., Canettieri, G., Best, J., Chen, H., Jenner, R., Herbolsheimer, E., Jacobsen, E., Kadam, S., Ecker, J. R., Emerson, B., Hogenesch, J. B., Unterman, T., Young, R. A., and Montminy, M. (2005). Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues. Proc Natl Acad Sci U S A, 102(12), 4459–64.

28. Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer.

29. Roth, C. L., Mastronardi, C., Lomniczi, A., Wright, H., Cabrera, R., Mungenast, A. E., Heger, S., Jung, H., Dubay, C., and Ojeda, S. R. (2007). Expression of a tumor-related gene network increases in the mammalian hypothalamus at the time of female puberty. Endocrinology , 148(11), 5147–61.

30. Wolpert, D. H. (1992). Stacked generalization. Neural Networks , 5, 241–259.

31. Pathway Commons http://www.pathwaycommons.org/pc/ Accessed 6-28-10

32. Witten, I. and Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques . Morgan Kaufmann, 2nd edition.

33. Taipale, J. and Beachy, P. A. (2001). The hedgehog and wnt signalling pathways in cancer.

Nature , 411(6835), 349–54.

34. Vermeulen, L., De Sousa E Melo, F., van der Heijden, M., Cameron, K., de Jong, J. H., Borovski, T., Tuynman, J. B., Todaro, M., Merz, C., Rodermond, H., Sprick, M. R., Kemper, K., Richel, D. J., Stassi, G., and Medema, J. P. (2010). Wnt activity defines colon cancer stem cells and is regulated by the microenvironment. Nat Cel l Biol , 12(5), 468–76.

35. Yochum, G. S., Cleland, R., and Goodman, R. H. (2008). A genome-wide screen for beta-catenin binding sites identifies a downstream enhancer element that controls c-myc gene expression. Mol Cell Biol , 28(24), 7368–79.

36. Field, J. K. and Spandidos, D. A. (1990). The role of ras and myc oncogenes in human solid tumours and their relevance in diagnosis and prognosis (review). Anticancer Res, 10(1), 1–22.

37. Leonard, W. J. and O'Shea, J. J. (1998). Jaks and stats: biological implications. Annu Rev Immunol , 16, 293–322.

38. Collins, P. J., Kobayashi, Y., Nguyen, L., Trinklein, N. D., and My- ers, R. M. (2007). The ets-related transcription factor GABP directs bidirectional transcription. PLoS Genet , 3(11), e208.

39. M´ejat, A., Ravel-Chapuis, A., Vandromme, M., and Schaeffer, L. (2003). Synapse-specific gene expression at the neuromuscular junction. Ann N Y Acad Sci , 998, 53–65.

40. Karolchik, D., Kuhn, R. M., Baertsch, R., Barber, G. P., Clawson, H., Diekhans, M., Giardine, B., Harte, R. A., Hinrichs, A. S., Hsu, F., Kober, K. M., Miller, W., Pedersen, J. S., Pohl, A., Raney, B. J., Rhead, B., Rosenbloom, K. R., Smith, K. E., Stanke, M., Thakkapallayil, A., Trumbower, H., Wang, T., Zweig, A. S., Haussler, D., and Kent, W. J. (2008). The UCSC genome browser database: 2008 update. Nucleic Acids Res , 36(Database issue), D773–9.

41. Shann, Y.-J., Cheng, C., Chiao, C.-H., Chen, D.-T., Li, P.-H., and Hsu, M.-T. (2008). Genome-wide mapping and characterization of hypomethylated sites in human tissues and breast cancer cell

lines. Genome Res , 18(5), 791–801.

42. Roh, T.-Y., Cuddapah, S., Cui, K., and Zhao, K. (2006). The genomic landscape of histone modifi- cations in human t cells. Proc Natl Acad Sci U S A, 103(43), 15782–7.

43. Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cel l , 129(4), 823–37.

44. Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Ra ju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009a). Human protein reference database–2009 update. Nucleic Acids Res , 37(Database issue), D767–72.

45. Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E., and Stein, L. (2007). Reactome: a knowledge base of biologic pathways and processes. Genome Biol , 8(3), R39.

46. Hatzis, P., van der Flier, L. G., van Driel, M. A., Guryev, V., Nielsen, F., Denissov, S., Nijman, I. J., Koster, J., Santo, E. E., Welboren, W., Versteeg, R., Cuppen, E., van de Wetering, M., Clevers, H., and Stunnenberg, H. G. (2008). Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. Mol Cel l Biol , 28(8), 2732–44.

47. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Sto janovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J.,

Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T.,Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Bl¨ocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet,

D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium (2001a). Initial sequencing and analysis of the human genome. Nature , 409(6822), 860–921.

48. Lenhard, B. and Wasserman, W. W. (2002). Tfbs: Computational framework for transcription factor binding site analysis. Bioinformatics , 18(8), 1135–6.

49. Chang CC and Lin CJ,(2001) LIBSVM : a library for support vector machines http://www.csie.ntu.edu.tw/~cjlin/libsvm

50. Cooper, G. F. and Herskovits, E. H. (1992). The induction of probabilistic networks from data. Machine Learning , 9(4), 309–347.

51. Fayyad, U. M. and Irani, K. B. (1993). Multi-interval discretization of continuous valued attributes for classification learning. Thirteenth International Joint Conference on Artificial Intelligence.

52. Hall, M. A. (1998). Correlation-based Feature Subset Selection for Machine Learning . Master’s thesis, University of Waikato, Hamilton, New Zealand.

53. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement , 20(1), 37–46.

54. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

55. Schreiber, S. L. and Bernstein, B. E. (2002). Signaling network model of chromatin. Cell, 111(6), 771–8.

56. Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet , 39(3), 311–8.

57. Rosenblatt, F (1958), The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review, 65(6), 386-408

58. Flicek, P., Aken, B. L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Gr¨af, S., Haider, S., Hammond, M., Howe, K., Jenkinson, A., Johnson, N., K¨ah¨ari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Koscielny, G., Kulesha, E., Lawson, D., Longden, I., Massingham, T., McLaren, W., Megy, K., Overduin, B., Pritchard, B., Rios, D., Ruffier, M., Schuster, M., Slater, G., Smedley, D., Spudich, G., Tang, Y. A., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S. P., Zadissa, A., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fern´andez-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Smith, J., and Searle, S. M. J. (2010). Ensembl's 10th year. Nucleic Acids Res , 38(Database issue), D557–62.

59. Bates D., Carey V., Dettling M., Dudoit S., Ellis B., Gautier L., Gentleman R., Gentry J., Hornik K., Hothorn T., Huber W., Iacus S., Irizarry R., Leisch F., Maechler M., Rossini A., Sawitzki G., Tierney L., Yang J.Y.H., Zhang J. (2002-) Bioconductor (Web Site),http://www.bioconductor.org/. Accessed 6-28-2010

60. Balasubramanian, R., LaFramboise, T., Scholtens, D., and Gentleman, R. (2004). A graph-theoretic approach to testing associations between disparate sources of functional genomics data. Bioinformatics , 20(18), 3353–62.

61. Zeller, K. I., Jegga, A. G., Aronow, B. J., O'Donnell, K. A., and Dang, C. V. (2003). An integrated database of genes responsive to the myc oncogenic transcription factor: identification of

direct genomic targets. Genome Biol, 4(10), R69.

62. Sansom, O. J., Meniel, V. S., Muncan, V., Phesse, T. J., Wilkins, J. A., Reed, K. R., Vass, J. K., Athineos, D., Clevers, H., and Clarke, A. R. (2007). Myc deletion rescues Apc deficiency in the small intestine. Nature, 446(7136), 676–9.

63. Dai, S., Jiang, L., Wang, G., Zhou, X., Wei, X., Cheng, H., Wu, Z., and Wei, D. (2010). Hsp70 interacts with traf2 and differentially regulates TNFalpha signaling in human colon cancer cells. J Cel l Mol Med , 14(3), 710–25.

64. Roovers, K., Wagner, S., Storbeck, C. J., O'Reilly, P., Lo, V., Northey, J. J., Chmielecki, J., Muller, W. J., Siegel, P. M., and Sabourin, L. A. (2009b). The STE20-like kinase SLK is required for ERBB2-driven breast cancer cell motility. Oncogene , 28(31), 2839–2848.

65. Hayashi, H., Nabeshima, K., Aoki, M., Hamasaki, M., Enatsu, S., Yamauchi, Y., Yamashita, Y., and Iwasaki, H. (2010a). Overexpression of IQGAP1 in advanced colorectal cancer correlates with poor prognosis-critical role in tumor invasion. Int J Cancer , 126(11), 2563–74.

66. Bracken, C. P., Wall, S. J., Barre, B., Panov, K. I., Ajuh, P. M., and Perkins, N. D. (2008a). Regulation of cyclin D1 RNA stability by SNIP1. Cancer Res , 68(18), 7621–8.

67. Won, K.-J., Ren, B., and Wang, W. (2010). Genome-wide prediction of transcription factor binding sites using an integrated model. Genome Biol , 11(1), R7.

68. Alamanova, D., Stegmaier, P., and Kel, A. (2010). Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. BMC Bioinformatics , 11, 225.

TABLES

Table 2.1 - Average AUC score for each classifier and TF (10 data sets, 10-fold cross-validation)

TF	BN	SVM	SVM (Discretized)
c-Myc	0.74	0.71	0.69
TCF4	0.71	0.67	0.66
STAT	0.83	0.78	0.75
GABP	0.94	0.91	0.9

Table 2.2 - Average AUC for per chromosome classification experiments

TF	Avg. ROC
c-Myc	0.75
GABP	0.94
STAT	0.83
TCF4	0.83

Table 2.3 - Comparison of average AUC between BN classifiers trained on all available features vs. only TFBS-feature distances

Factor	All	TFBS Only
c-Myc	0.74	0.72
TCF4	0.71	0.69
STAT	0.83	0.82
GABP	0.94	0.92

Table 2.4a-d - Per-chromosome AUC, true positive rate, true negative rate, and High and Low Occupancy Site Count for all TFs

Table 2.4a: c-Myc

Chromosome	AUC	TP	TN	High	Low
1	0.86	0.7	0.96	47	5115
2	0.83	0.77	0.83	13	3399
3	0.84	0.52	0.89	21	2679
4	0.73	0.5	0.84	10	1771
5	0.81	0.33	0.88	9	2360
6	0.86	0.82	0.84	11	2270
7	0.64	0.33	0.92	18	2911
8	0.82	0.43	0.84	14	1997
9	0.79	0.72	0.73	39	2504
10	0.96	0.8	1	10	2471

11	0.88	0.82	0.78	33	3174
12	0.79	0.74	0.84	65	2651
13	0.17	0	0.93	1	1043
14	0.72	0.84	0.36	45	5166
15	0.52	0	1	9	1747
16	0.86	0.77	0.8	64	3220
17	0.57	0.33	0.8	15	3996
18	0.89	0.7	0.86	10	702
19	0.83	0.64	0.84	137	4260
20	0.6	0.27	0.89	11	1581
21	0.63	0.18	0.85	17	929
22	0.89	0.72	0.9	46	2124

Table 2.4b: GABP

Chromosome	AUC	TP	TN	High	Low
1	0.95	0.95	0.84	442	2052
2	0.93	0.96	0.79	328	1247
3	0.94	0.97	0.81	275	1115
4	0.92	0.95	0.78	128	616
5	0.94	0.94	0.83	226	735
6	0.93	0.95	0.78	240	887
7	0.94	0.96	0.81	263	1145
8	0.93	0.93	0.79	213	725
9	0.94	0.94	0.81	239	910
10	0.94	0.93	0.87	175	800
11	0.92	0.95	0.78	365	1314
12	0.95	0.96	0.85	267	1107
13	0.89	0.91	0.79	81	276
14	0.94	0.95	0.84	129	1071
15	0.95	0.93	0.83	147	722
16	0.95	0.95	0.87	287	1332
17	0.93	0.92	0.84	359	1573
18	0.96	0.98	0.87	65	236
19	0.9	0.93	0.76	549	1771
20	0.93	0.95	0.81	173	624
21	0.95	0.94	0.88	47	291
22	0.95	0.98	0.78	155	788

Table 2.4c: STAT

Chromosome	AUC	TP	TN	High	Low
1	0.83	0.63	0.87	1740	15180
2	0.81	0.57	0.92	833	11092

3	0.81	0.64	0.86	953	8100
4	0.81	0.64	0.88	287	5835
5	0.83	0.62	0.92	636	6498
6	0.85	0.69	0.89	849	7589
7	0.82	0.6	0.9	605	7251
8	0.83	0.61	0.9	543	4813
9	0.85	0.66	0.92	656	6467
10	0.76	0.49	0.92	621	6560
11	0.83	0.69	0.85	817	8167
12	0.8	0.59	0.88	968	7488
13	0.87	0.72	0.91	179	2831
14	0.91	0.69	0.95	531	10516
15	0.85	0.62	0.92	481	5551
16	0.86	0.66	0.91	674	5904
17	0.84	0.65	0.88	1034	7325
18	0.82	0.62	0.93	144	2504
19	0.85	0.79	0.77	962	7637
20	0.81	0.59	0.85	445	3583
21	0.83	0.58	0.91	120	1813
22	0.76	0.63	0.8	239	3757

Table 2.4d: TCF4

Chromosome	AUC	TP	TN	High	Low
1	0.71	0.71	0.57	66	10840
2	0.85	0.53	0.9	53	8584
3	0.78	0.58	0.85	113	6539
4	0.91	0.81	0.81	36	4643
5	0.87	0.7	0.87	135	5137
6	0.86	0.78	0.78	106	5599
7	0.51	0.06	0.92	16	5251
8	0.78	0.65	0.82	51	3729
9	0.83	0.63	0.88	41	4832
10	0.76	0.73	0.69	83	4825
11	0.87	0.85	0.76	92	5546
12	0.89	0.83	0.81	80	5492
13	0.91	0.73	0.92	26	2257
14	0.9	0.63	0.9	35	6595
15	0.87	0.82	0.83	55	4224
16	0.76	0.42	0.83	26	3784
17	0.85	0.75	0.76	76	4678
18	0.89	0.86	0.67	28	1890
19	0.86	0.84	0.76	32	4208
20	0.84	0.79	0.81	28	2310

21	0.9	0.71	0.84	7	1298
22	0.87	0.5	0.93	6	2095

Table 2.5 - Top 10 most frequently occurring predictors in the occupancy classifiers (per-TF and cumulative)

Predictor	GABP	c-Myc	STAT	TCF4	Total
H3K4me2-H3K4me3	80	8	83	57	228
H3K27me1-H3K4me3	84	11	87	43	225
TSS-H3K4me3	85	47	50	22	204
H3K79me1-H3K4me3	69	26	79	25	199
H3K79me2-H3K4me3	97	12	55	34	198
H4R3me2-H3K4me3	61	40	74	23	198
H3K79me2-TSS	81	44	51	20	196
H4K20me1-H3K4me3	68	10	82	30	190
H3K79me3-H3K4me3	30	36	72	49	187
H3K9me1-H3K4me3	78	10	77	13	178

Table 2.6 -Pairwise Agreement on Inclusion of Features into Classifiers (Average Kappa, 553 features, n=100 per feature)

TF 1	TF2	Kappa
GABP	STAT	0.47
STAT	TCF4	0.43
GABP	TCF4	0.41
STAT	c-Myc	0.39
c-Myc	TCF4	0.36
GABP	c-Myc	0.34

Table 2.7: Class-conditional probabilities for top predictors

Table 2.7a: Average class-conditional probability of high-occupancy status for smallest distance bin for top 10 most frequently occurring features in cross-validation (two-bin cases only)

Feature	c-Myc	GABP	STAT	TCF4
H3K4me2-H3K4me3	0.65	0.91	0.59	0.54
H3K27me1-H3K4me3	0.38	0.91	0.52	0.42
TSS-H3K4me3	0.82	0.97	0.78	0
H3K79me1-H3K4me3	0.63	0.87	0.58	0.46
H3K79me2-H3K4me3	0.49	0.88	0.55	0.42
H4R3me2-H3K4me3	0.3	0	0.54	0.43
H3K79me2-TSS	0.6	0.93	0.7	0.6
H4K20me1-H3K4me3	0.28	0.91	0.49	0.38
H3K79me3-H3K4me3	0.86	0.82	0.57	0.59
H3K9me1-H3K4me3	0.38	0.92	0.54	0

Table 2.7b: Average class-conditional probability of high-occupancy status for smallest distance bin for top 10 most frequently occurring features per-chromosome(all cases)

Feature	c-Myc	GABP	STAT	TCF4
H4K20me1-H3K4me3	0.79	0.91	0.58	0.67
H3K4me2-H3K4me3	0.6	0.93	0.57	0.7
H3K79me3-H3K4me3	0.81	0.92	0.53	0.72
TSS-H3K4me3	0.94	0.98	0.67	0
H3K79me2-H3K4me3	0.43	0.91	0.56	0.57
H3K27me1-H3K4me3	0.66	0.91	0.57	0.5
H3K79me1-H3K4me3	0.93	0.91	0.55	0.72
H3K79me2-TSS	0.92	0.87	0.68	0.6
H4R3me2-H3K4me3	0	0.82	0.52	0.74
H3K9me1-H3K4me3	0.68	0.92	0.47	0.56

Table 2.8 - Average AUC in cross-classification experiments

Training/Test TF	c-Myc	TCF4	STAT	GABP
c-Myc	x	0.64	0.79	0.86
TCF4	0.65	x	0.78	0.91
STAT	0.69	0.69	x	0.92
GABP	0.67	0.69	0.83	x

Table 2.9 - Average AUC for SRF cross-classification experiments

Classifier	AUC
SRF (cross-val)	0.88
GABP	0.88
c-Myc	0.86
STAT	0.89
TCF4	0.86

Table 3.1 - Results of Generalizable Classification Experiments

Test Set	Train	Average AUC	Average TP	Average TN
c-Myc	Combine (BN) (TCF4 Holdout)	0.7	0.52	0.76
	Combine (BN) (GABP Holdout)	0.69	0.62	0.67
	Combine (BN) (STAT Holdout)	0.69	0.51	0.77
	Combine (BN)	0.69	0.51	0.76
	Stacking 1 (NN) (TCF4 Holdout)	0.68	0.47	0.79
	STAT	0.68	0.6	0.66
	Combine (SVM Variant)	0.67	0.72	0.55
	GABP	0.67	0.34	0.88

	c-Myc	0.67	0.65	0.61
	Stacking 1 (NN) (GABP Holdout)	0.67	0.6	0.66
	Combine (BN) (c-Myc Holdout)	0.67	0.47	0.78
	Stacking 1 (SVM Variant)	0.67	0.43	0.82
	Stacking 2	0.67	0.56	0.72
	Stacking 1 (NN)	0.67	0.47	0.79
	Stacking 1 (NN) (c-Myc Holdout)	0.65	0.42	0.82
	Stacking 1 (NN) (STAT Holdout)	0.65	0.4	0.82
	TCF4	0.62	0.58	0.62
GABP	GABP	0.93	0.89	0.84
	Combine (BN) (c-Myc Holdout)	0.92	0.97	0.72
	Combine (BN)	0.92	0.97	0.7
	Combine (BN) (STAT Holdout)	0.92	0.97	0.71
	Combine (BN) (TCF4 Holdout)	0.92	0.97	0.7
	Stacking 1 (NN) (c-Myc Holdout)	0.92	0.93	0.77
	Stacking 1 (SVM Variant)	0.91	0.94	0.77
	Stacking 1 (NN) (TCF4 Holdout)	0.91	0.94	0.73
	Stacking 1 (NN)	0.9	0.96	0.74
	Stacking 1 (NN) (STAT Holdout)	0.9	0.92	0.78
	Combine (BN) (GABP Holdout)	0.9	0.98	0.6
	STAT	0.89	0.97	0.59
	Stacking 2	0.88	0.96	0.65
	Stacking 1 (NN) (GABP Holdout)	0.87	0.98	0.59
	TCF4	0.84	0.93	0.57
	Combine (SVM Variant)	0.81	0.94	0.48
	c-Myc	0.74	0.82	0.55
STAT	Combine (BN)	0.82	0.63	0.87
	Combine (BN) (c-Myc Holdout)	0.81	0.61	0.89
	Combine (BN) (TCF4 Holdout)	0.81	0.62	0.87
	Combine (BN) (GABP Holdout)	0.81	0.7	0.79
	GABP	0.81	0.45	0.96
	Combine (BN) (STAT Holdout)	0.81	0.61	0.88
	STAT	0.8	0.69	0.77
	Stacking 1 (NN) (TCF4 Holdout)	0.79	0.57	0.89
	Stacking 1 (SVM Variant)	0.79	0.54	0.91
	Stacking 1 (NN) (c-Myc Holdout)	0.79	0.53	0.91
	Stacking 1 (NN) (GABP Holdout)	0.79	0.69	0.78
	Stacking 1 (NN)	0.78	0.59	0.9
	Stacking 2	0.76	0.63	0.83
	Stacking 1 (NN) (STAT Holdout)	0.76	0.5	0.92
	Combine (SVM Variant)	0.75	0.71	0.68
	TCF4	0.74	0.67	0.71

	c-Myc	0.71	0.62	0.72
TCF4	Combine (BN) (c-Myc Holdout)	0.7	0.43	0.86
	Combine (BN)	0.69	0.45	0.85
	Combine (BN) (STAT Holdout)	0.69	0.43	0.85
	GABP	0.68	0.28	0.94
	Combine (BN) (GABP Holdout)	0.68	0.51	0.77
	Stacking 1 (SVM Variant)	0.68	0.37	0.89
	Stacking 1 (NN) (c-Myc Holdout)	0.68	0.36	0.89
	STAT	0.67	0.52	0.75
	Combine (BN) (TCF4 Holdout)	0.67	0.43	0.84
	Stacking 1 (NN) (GABP Holdout)	0.67	0.53	0.76
	Stacking 1 (NN) (TCF4 Holdout)	0.67	0.39	0.87
	Stacking 1 (NN)	0.66	0.4	0.87
	Combine (SVM Variant)	0.66	0.57	0.67
	Stacking 2	0.66	0.46	0.81
	Stacking 1 (NN) (STAT Holdout)	0.65	0.35	0.89
	TCF4	0.65	0.54	0.7
	c-Myc	0.6	0.46	0.72

Table 3.2 - Results of SRF classification with Generalizable Classifiers

SRF Results	Average AUC	Average TP	Average TN
Combined	0.88	0.83	0.83
Stacking 1	0.86	0.71	0.87

Table 5.1- Predicted shared hub genes of c-Myc and TCF4 with all sites exceeding .6 probability of high-occupancy

Gene	No. Agreements	Support
SFRS1	90	3
MAGOH	90	3
PCBP1	90	2
CDC2	90	1
RPL30	90	1
IQGAP1	90	0
CHUK	90	0
EIF3D	90	0
ACTR2	90	0
GTF2F1	90	0
HIST1H2BH	81	1
RUVBL2	81	1

PLEC1	81	0
RPL35	72	1
SRRM2	72	1
RPS11	72	0
EIF4A2	72	0
RPL13A	63	1
TRAF6	63	0
PAK4	63	0
BCLAF1	63	0
PRKDC	54	1
HSP90AB1	54	0
CDC40	54	0
THRAP3	54	0
HNRNPH1	45	2
SLK	45	2
EEF1A1	45	1
PTPN11	45	0
CRKRS	45	0
GRB2	45	0
SHC1	45	0
U2AF2	45	0
HSPD1	45	0
TUBB2C	45	0
ANP32A	40	2
IGF1R	40	0
IKBKB	40	0
RPL37	36	3
TUBB	36	2
ANXA6	36	1
MAPK1	36	0
DHX9	36	0
RPS29	27	3
NCL	27	1
GTF2F2	27	1
HSPA8	27	1
ATP5A1	27	0
ACVR2B	24	0
IKBKE	18	2
TRAF2	18	1
RPS7	18	1
ACTB	18	0
RPL23A	18	0
CPSF2	18	0
HNRNPA2B1	18	0

CCND1	16	3
CDKN2A	16	1
GRSF1	12	1
MAP3K1	9	3
RPL15	9	3
TUBA4A	9	2
EEF1G	9	2
MATR3	9	2
MAPK3	9	1
RPS9	9	1
RPS16	9	1
RPL7	9	1
RPS13	9	1
RPL3	9	1
SNRPD2	9	1
SLC25A4	9	1
U2AF1	9	1
SP1	9	0
EGFR	9	0
HSP90AA1	9	0
CAMK2D	9	0
RUVBL1	9	0
HSPA1L	9	0
TGFBR1	9	0
RPS3A	9	0
LYN	9	0
HNRNPF	9	0
DYNLL1	9	0
PABPC1	9	0
AURKB	9	0
RPS14	9	0
TNFRSF14	9	0
CBL	9	0
RCC2	9	0
HIST1H4A	9	0
CREBBP	9	0
CAPZB	9	0
LUC7L2	9	0
NEK9	9	0
JUN	9	0
UGDH	9	0
SCIN	9	0
LUZP1	9	0
POLR2J	9	0

EIF3K	9	0
TNPO1	9	0
NUDT21	9	0
HNRNPA0	9	0
PABPN1	9	0
SFRS9	9	0
AP2A1	9	0
AP2B1	9	0
AP2M1	9	0
HLA-B	9	0
HSPA5	9	0
BMPR1A	8	0
BTRC	8	0
CSTF3	7	3
HIPK1	7	0
POLR1E	7	0
RB1	6	1
TRIP6	3	0

Support: 1 = c-Myc Target Database[59], 2 = Hatzis et al.[44] binding data, 3 = Both

Table 5.2 - Prioritized shared c-Myc/TCF4 predicted targets for additional analysis

Gene	Biological Import
ACTR2	Component of ARP2/3 complex, involved in cell shape and motility
ACVR2B	TGF-beta superfamily receptor kinase
ANP32A	Implicated in multiple processes including apoptosis, tumor suppression
BCLAF1	Transcription factor, apoptosis inducing protein
CDC2	Cell cycle control, Wnt pathway member
CDKN2A	Kinase, tumor suppressor
CHUK	NF-kappa-B inhibitor
CRKRS	RNA splicing factor, cell cycle related
EIF4A2	RNA helicase
GTF2F1	General transcription factor
GTF2F2	General transcription factor
HSP90AB1	Chaperone, may stabilize mutant oncogenic proteins
IGF1R	Insulin-like growth factor receptor
IKBKB	NF-kappa-B signaling factor
IKBKE	Noncanonical NF-kappa-B factor, implicated in breast cancers
IQGAP1	Cell adhesion and motility
MAPK1	MAP kinase family member
PAK4	Cytoskeletal reorganization
PLEC1	Cytoskeletal element
PTPN11	Transcription regulation, cell migration
RUVBL2	Helicase, DNA repair, oncogenesis

SLK	Kinase, apoptotic regulation
SP1	Transcription factor, cell growth, apoptosis
SRRM2	RNA preprocessing
THRAP3	Transcriptional Coactivator
TRAF2	Apoptosis, MAPK and N-kappa-B signalling
TRAF6	NF-kappa-B pathway member, TNF receptor family
TUBB2C	Tubulin, cytoskeletal involvement

Table 5.3 – Primers used for TCF4 and c-Myc ChIP experiments ((L)eft and (R)ight)

Primer	Sequence
THRAP3 L	TCAATACCCCAGTAGCACCCATTT
THRAP3 R	GGAAAGCCTCAAGCACCCTGAAAG
IKBKE L	GCGTCTGCCACTCATAGCATCTG
IKBKE R	TCCGTCAATCTCTTTCCCAGCATA
ACTR2 L	TGGGCTGACATTGGAGTATGGAAC
ACTR2 R	CAGGGCTTGGTGTGTTATTGCTTC
ACVR2B L	TAAATGACCACTCCCCGCCCTA
ACVR2B R	GCAGAAAGAGGCTGACTTCCCTGA
CDK1 L	ACTGTGCCAATGCTGGGAGAAAAA
CDK1 R	GAAAGAAAGAGGAAAGGGCGGCTA
CHUK L	CATTCACAGAGACACACACGCACT
CHUK R	GTGGGACCTTGGGCAGTATTTGG
SLK L	CCCCTGGTCCTTATCCTGTCCTTC
SLK R	TGTTCCACCGTAAACCCGACTTCT
IKBKB L	TCCTCACTGCCTCCACTTTCTCTG
IKBKB R	TCCCCCTATTCACTGTCCCAAGAT
TRAF2 L	TCTGAATGCTTGGAGGAGACTTACC
TRAF2 R	GCCTTTGGTGAAATGGAGACCTT
TRAF6 L	CCAGCCTTTGTGTATCCCTCCCTA
TRAF6 R	GATTCTCTTGCTCTTCCTTTTCTCCAG
SP1 L	GGCTCCACCAAAACACGGATAAAG
SP1 R	TGAGGCTAAAGTGCGGATAAGTCA
HSP90AB1 L	GCCGACAAGAATGATAAGGCAGTT
HSP90AB1 R	GATAGATGCGGTTGGAGTGGGTCT
PTPN11 L	TGTCTTCTTTTCCTCCTACCCCTCA
PTPN11 R	CGGCTCCCTTCCTTTCCATCTC
GTF2F2 L	AGGCATTTCTCTCTCCAGCAGCAT
GTF2F2 R	GGCACCATCTCAAGTCACACCATTT
ANP32A L	CTGCCCAAACTCCCAACTCCAT
ANP32A R	ATTTCTGCCTCCCTCTCGCTTTCA
IQGAP1 L	AATCTGGTGACTGCTGCCGAATCT
IQGAP1 R	ATGAAAGCCCTCCAACCCCACTCT
CRKRS L	TGAAAGCGAAGCACGAAACATC

CRKRS R	TCCCTCACACAGACCCAGTCACAC
PAK4 L	CCCTAGCGGAGCAGATGAATGAGT
PAK4 R	GCAATACGCCCTCCTTGGGTTTTA

Table 5.4 – Summary of Results of TCF4 and c-Myc ChIP Experiments to Confirm Prioritized Targets

Table 5.4a - Results of TCF4 ChIP

Gene	TCF4 Binding?
HSP90AB	no
TRAF6	no
ANP32A	yes
PAK4	yes
IQGAP1	yes
ACTR2	yes
CHUK	yes
CRKRS	yes
CDC2	yes
PTPN11	no
TRAF2	yes
ACVR2B	yes
SLK	yes
IBKBB	no
IBKBE	no
THRAP3	yes
SP1	yes
GTF2F2	yes

Table 5.3b - Summary of c-Myc binding

Gene	c-Myc
IQGAP1	yes
PTPN11	yes
TRAF2	yes
SLK	yes
IBKBE	yes
THRAP3	yes

Table 5.5 - Summary of existing support for TCF4 and c-Myc binding at prioritized target genes

Gene	Known TCF4	Known c-Myc
HSP90AB	No binding	Not assayed
TRAF6	No binding	Not assayed
ANP32A	Yes	Not predicted
PAK4	No	Not assayed
IQGAP1	No	No
ACTR2	No	Not predicted
CHUK	No	Not predicted
CRKRS	No	Not predicted
CDC2	No	Yes, not assayed
PTPN11	No binding	No
TRAF2	No	Yes
ACVR2B	No	Not assayed
SLK	Yes	No
IBKBB	No binding	Not assayed
IBKBE	No binding	Yes
THRAP3	No	No
SP1	No	Not assayed
GTF2F2	No	Yes, not predicted

“Yes” indicates literature evidence existed previous to our ChIP experiments

“No” indicates that we were unable to find evidence of TF binding previous to our ChIP experiments

FIGURES

Figure 2.1 – Visual representation of TFBS-Feature and Feature-Feature distances

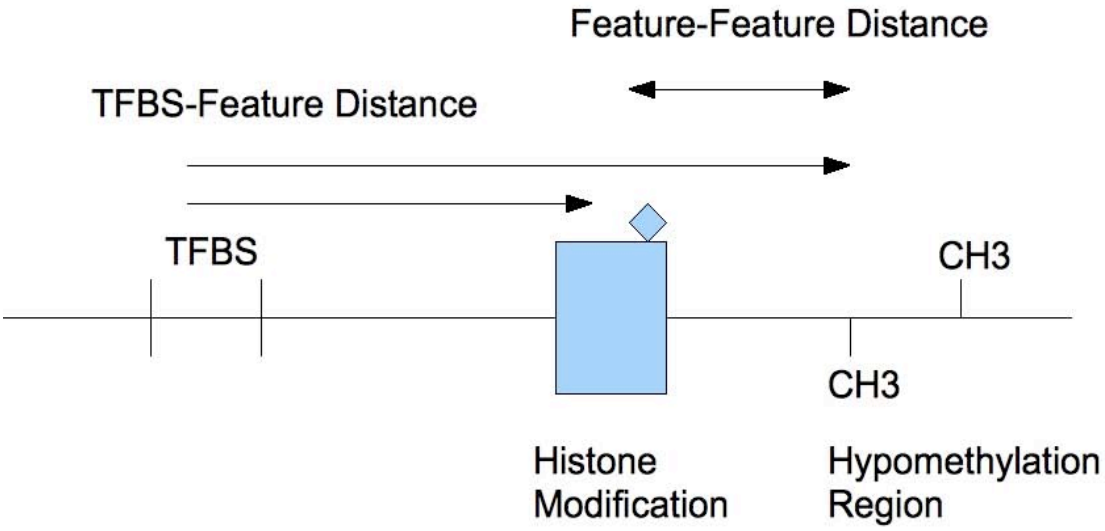


Figure 2.2 – Average AUC values for all classifier variants in single TF classification

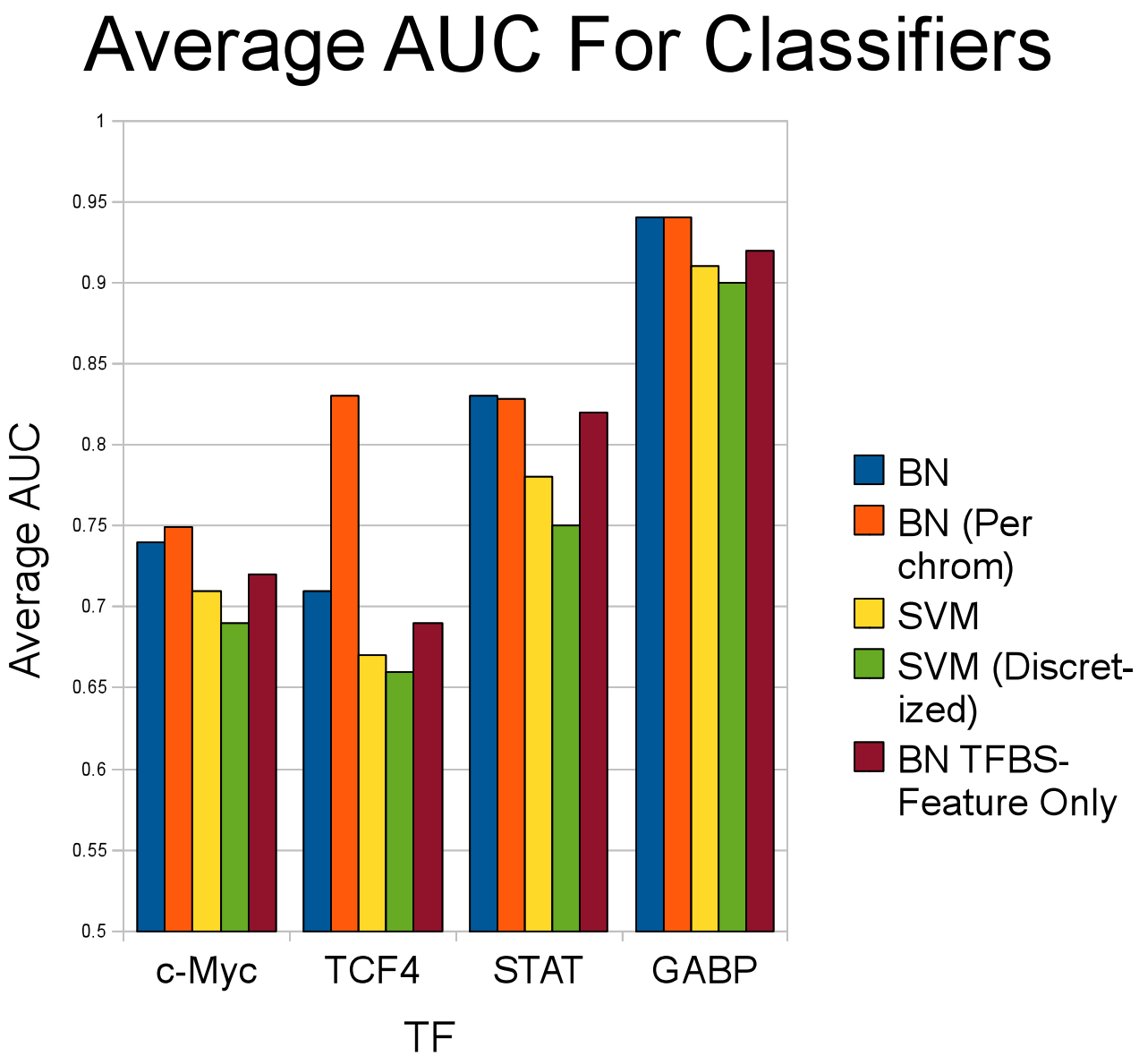


Figure 2.3 – Average AUC values for cross-classification experiments

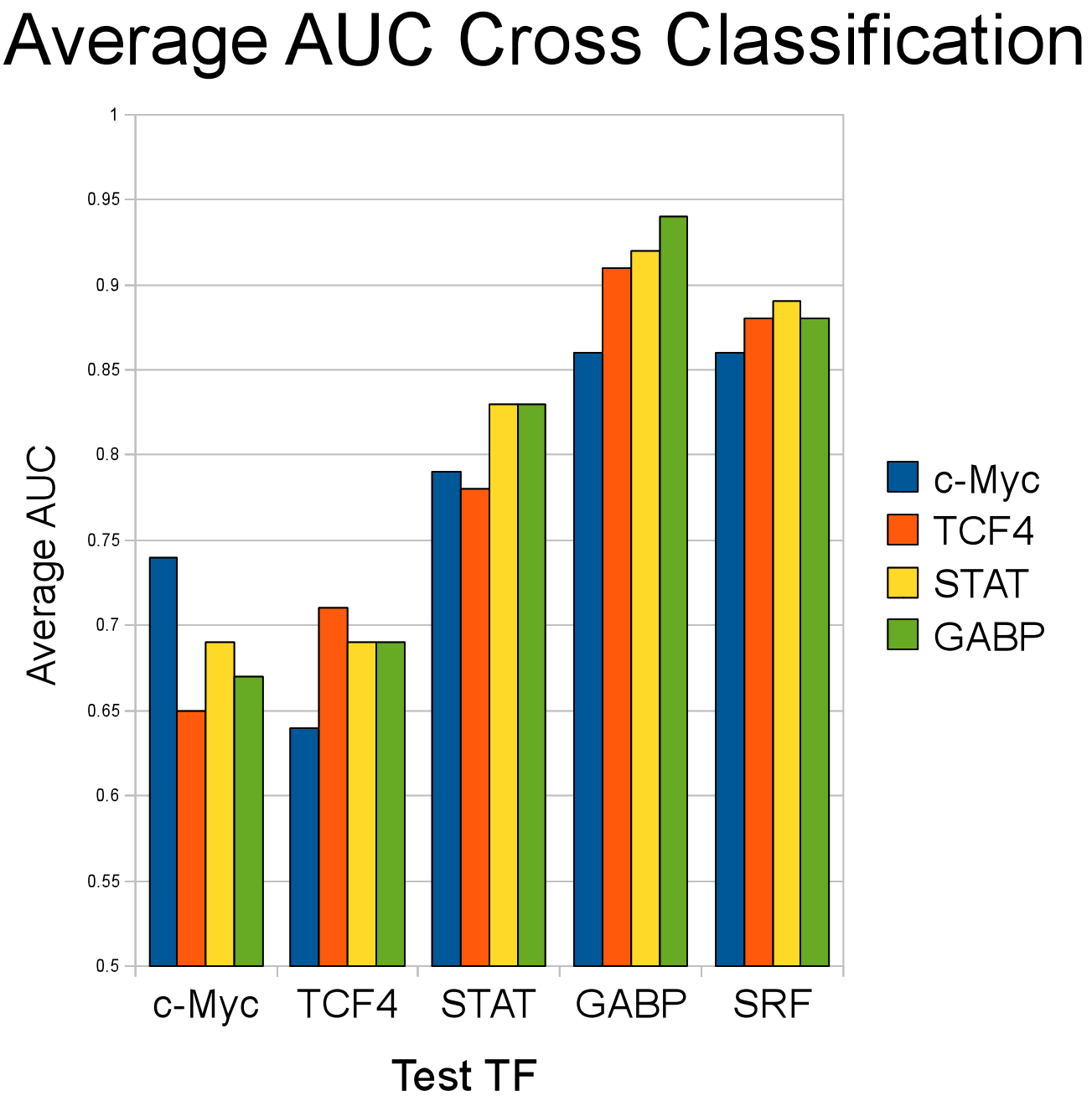


Figure 3.1 – Schematic of Stacking 1 classifier

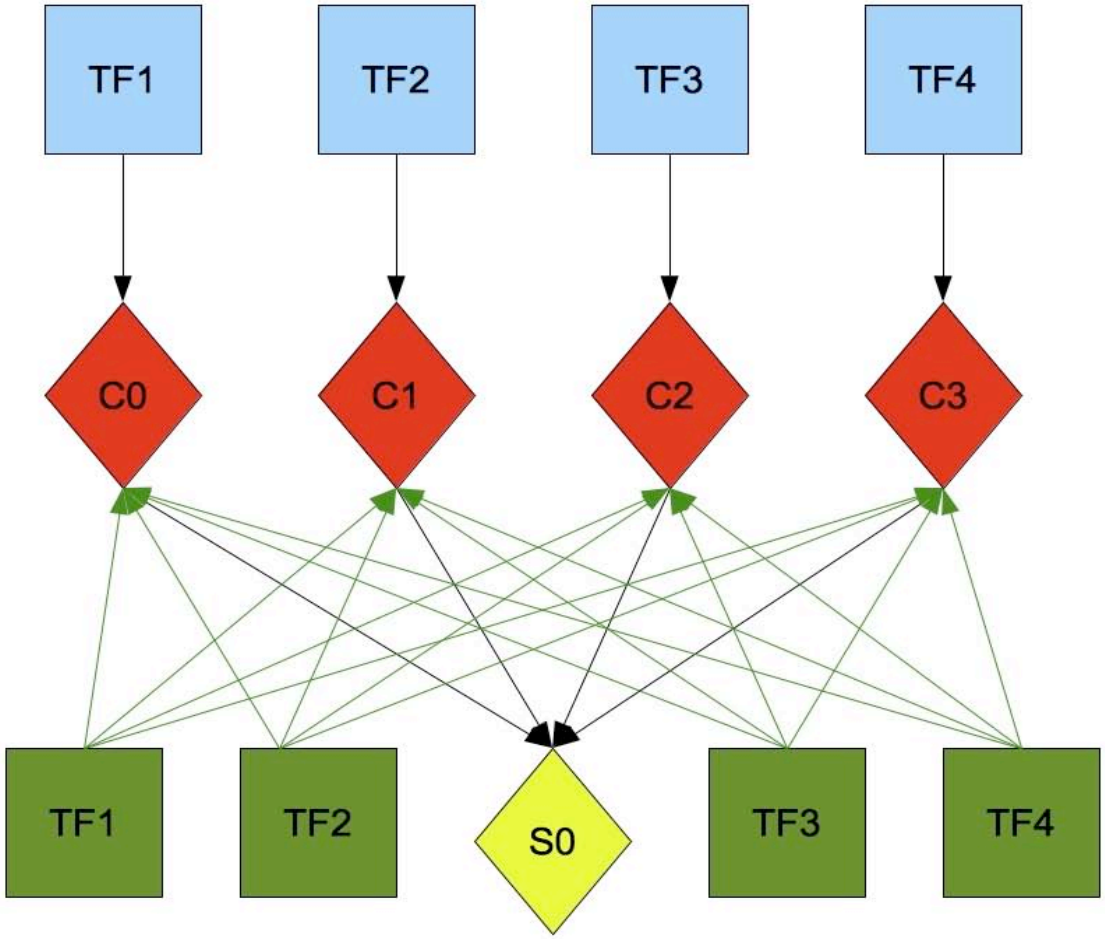


Figure 3.2 – Stacking 2 classifier diagram

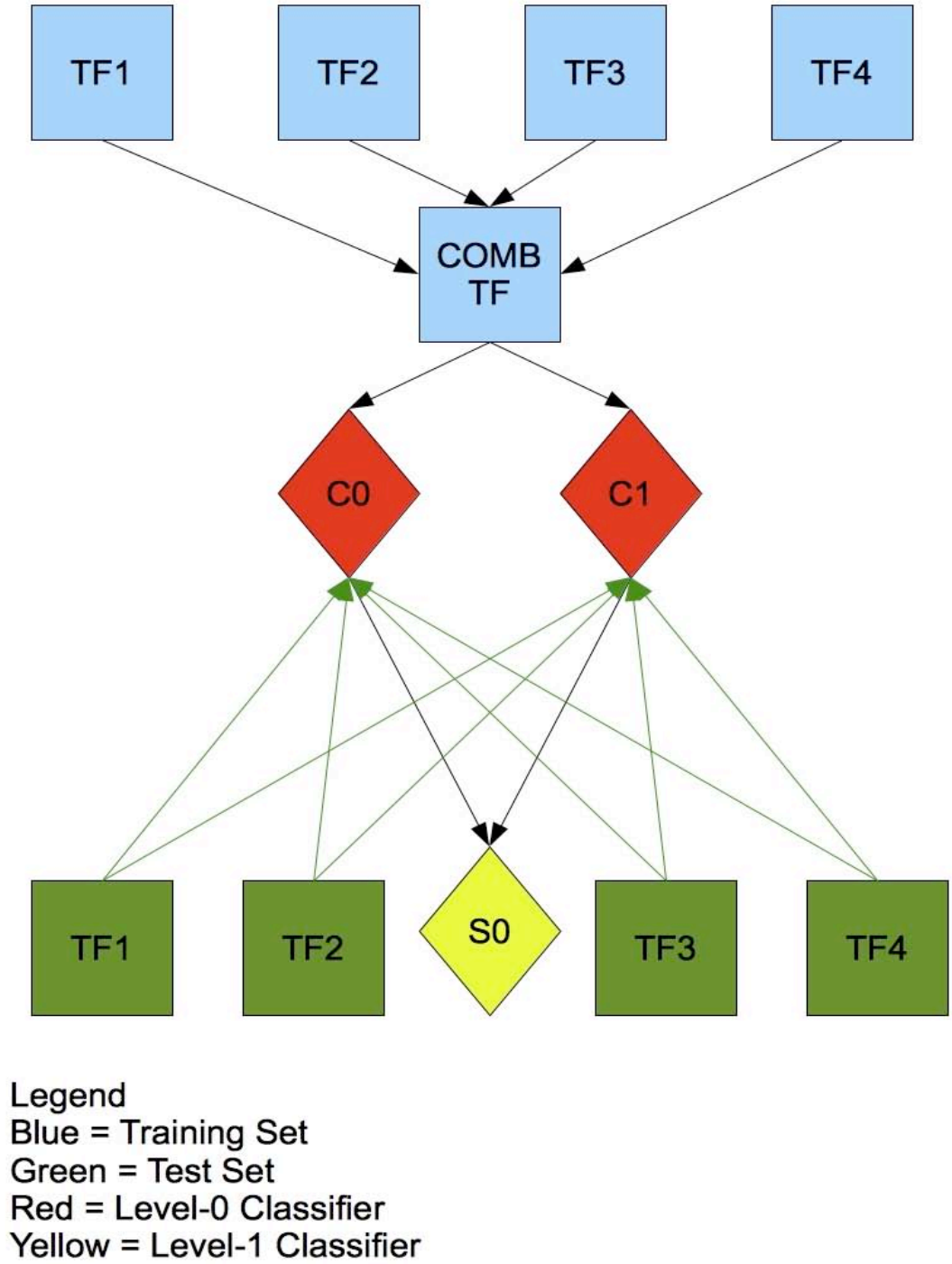


Figure 3.3 – Average AUC for Generalizable Classification Experiments

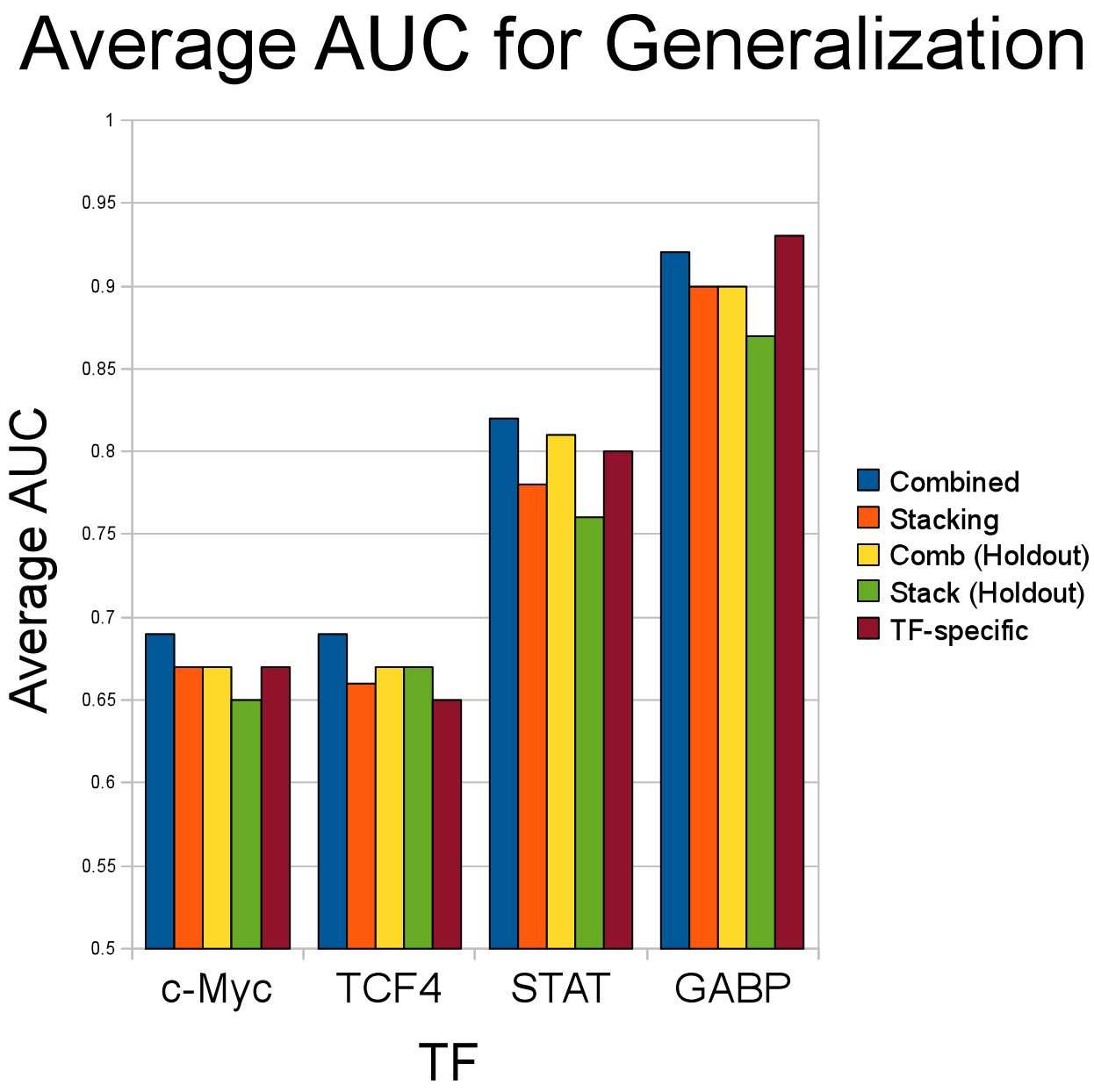


Figure 4.1 – Edge Preservation Analysis Diagram (Adapted from Balasubramanian et al. 2004)[57]

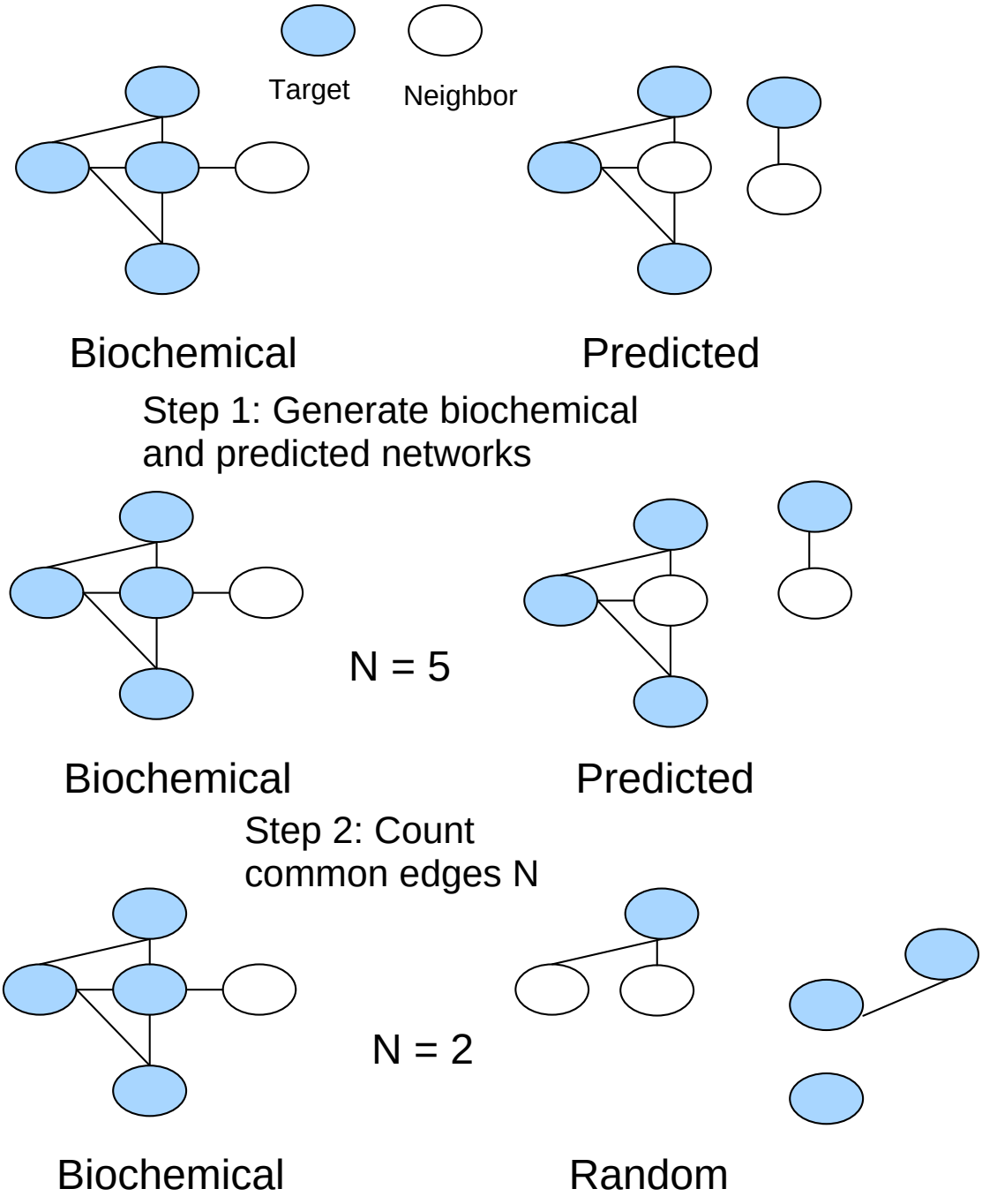
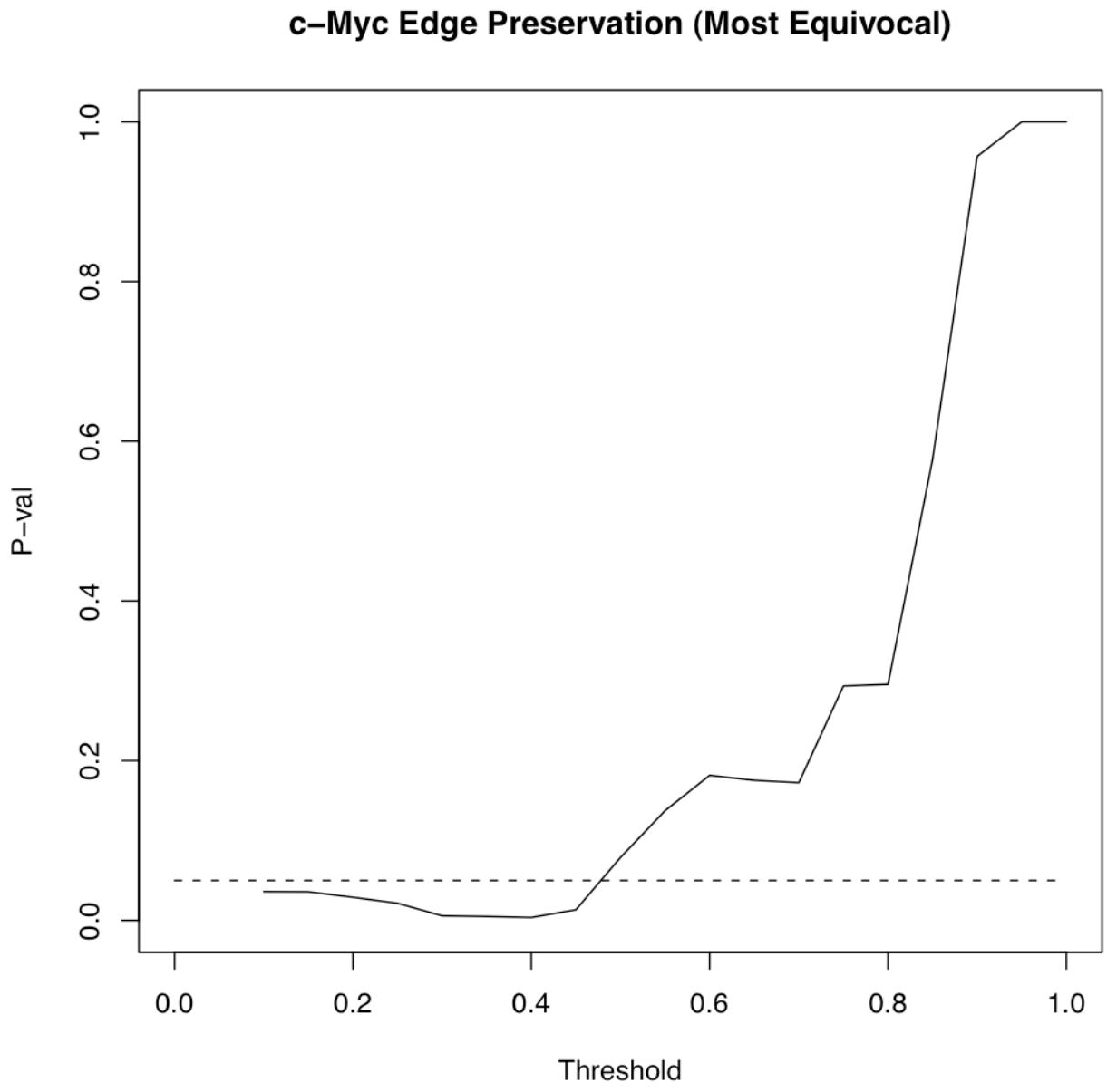
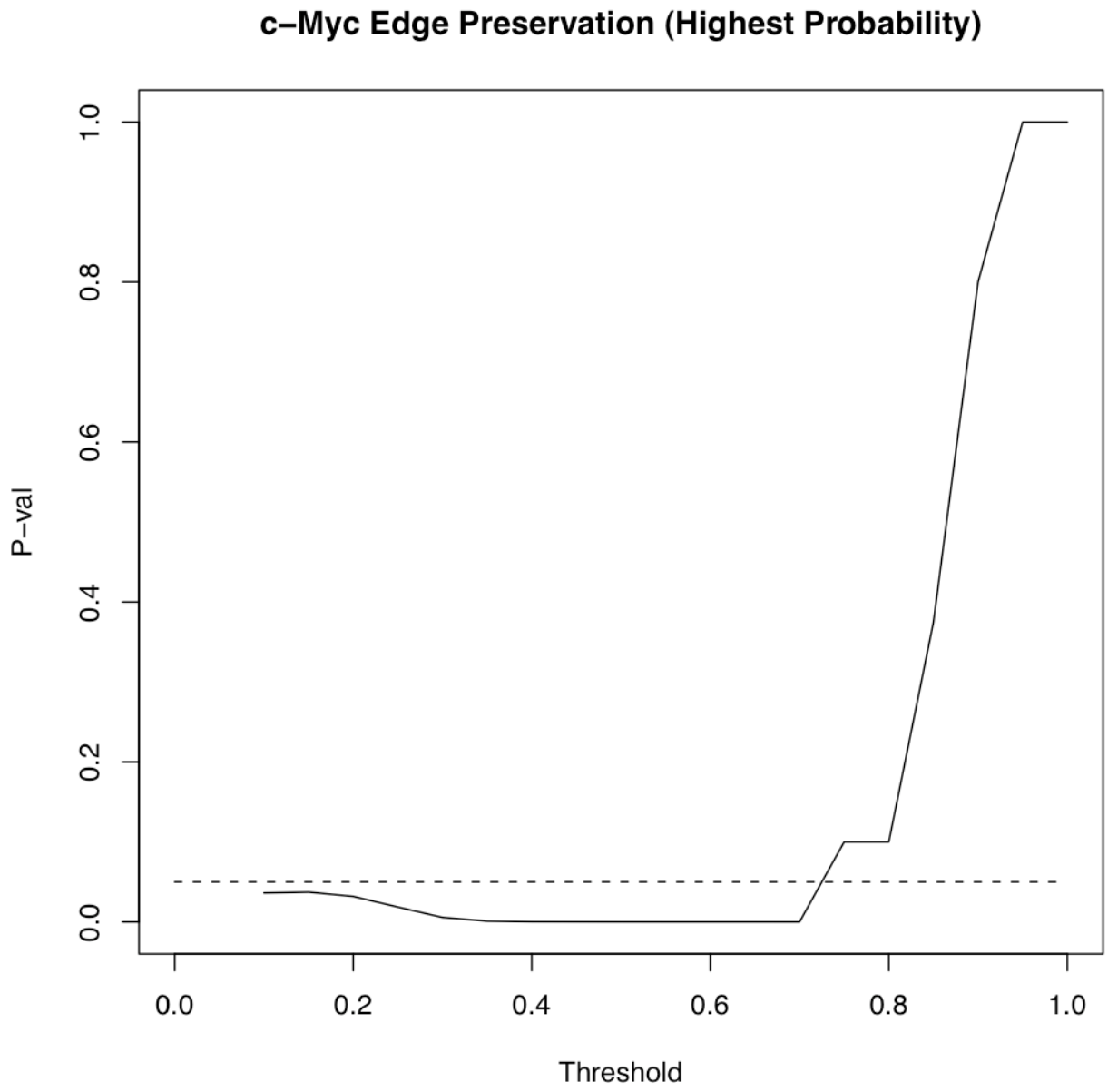


Figure 4.2 - P-values of connected edge preservation analysis, most equivocal site selection criteria (c-Myc)



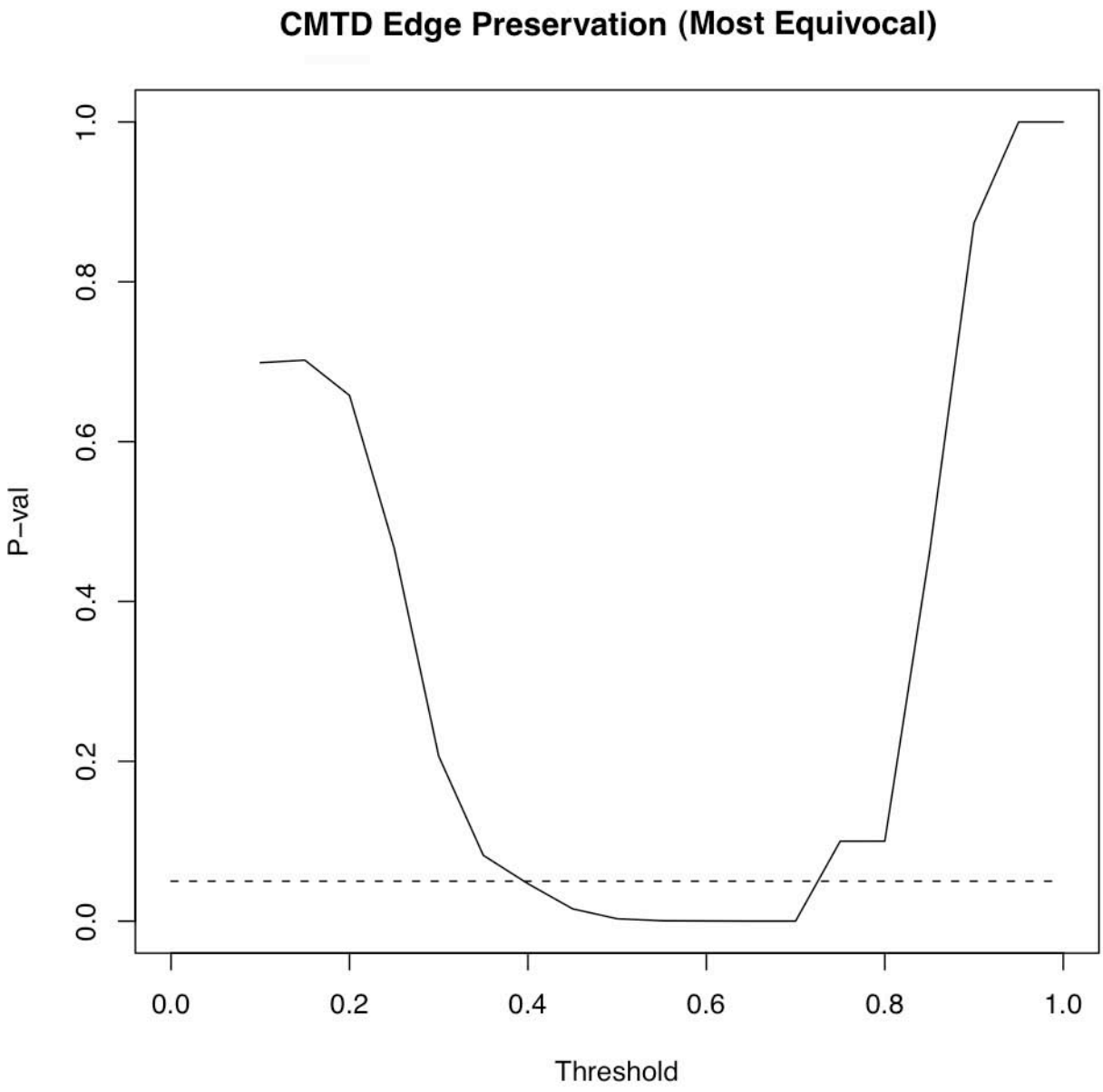
Solid line represents average p-value for 10 classifier variations

Figure 4.3- P-values of connected edge preservation analysis, highest probability site selection criteria (c-Myc)



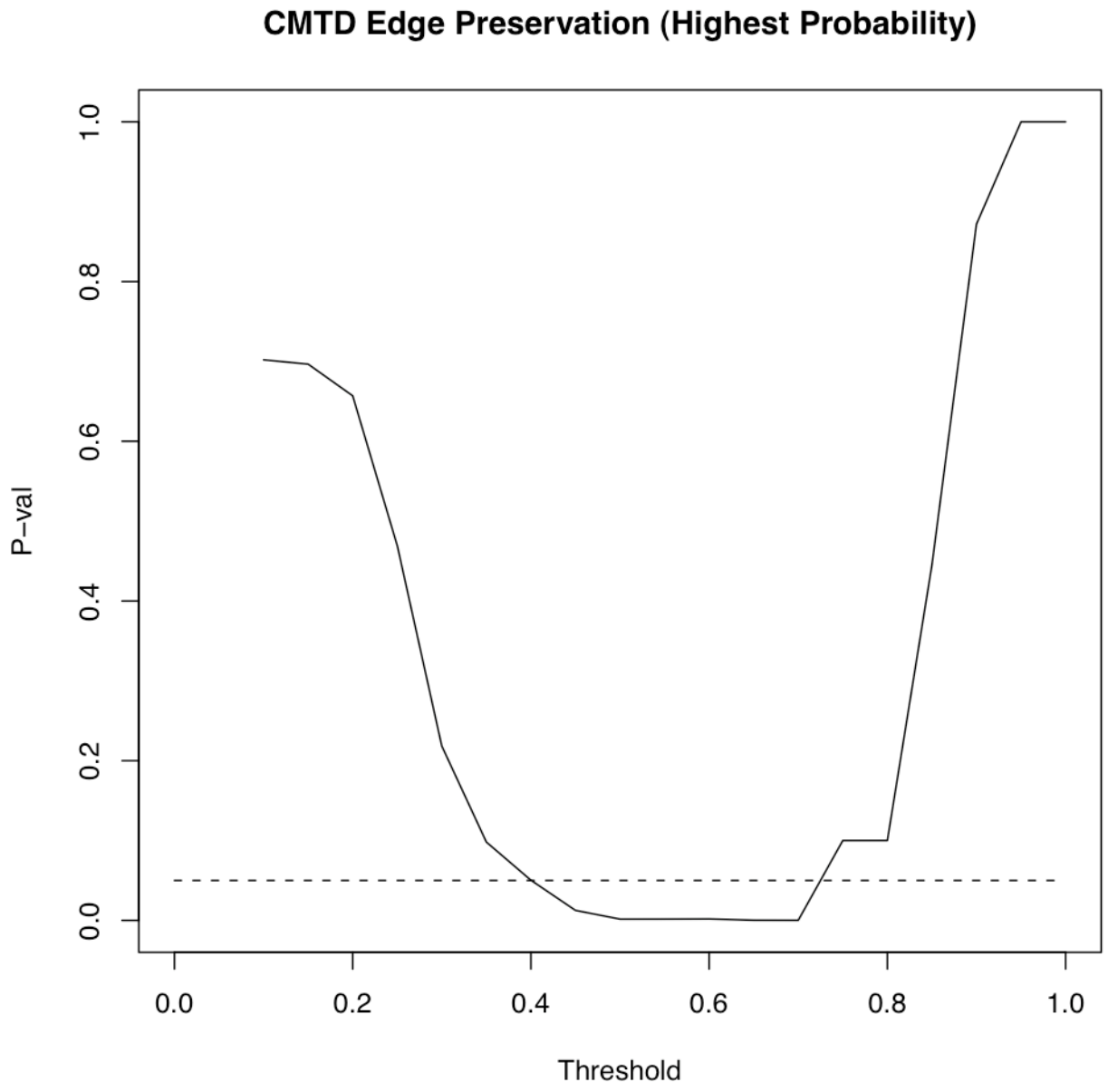
Solid line represents average p-value for 10 classifier variations

Figure 4.4 - P-values of connected edge preservation analysis, most equivocal site selection criteria (c-Myc Target Database)



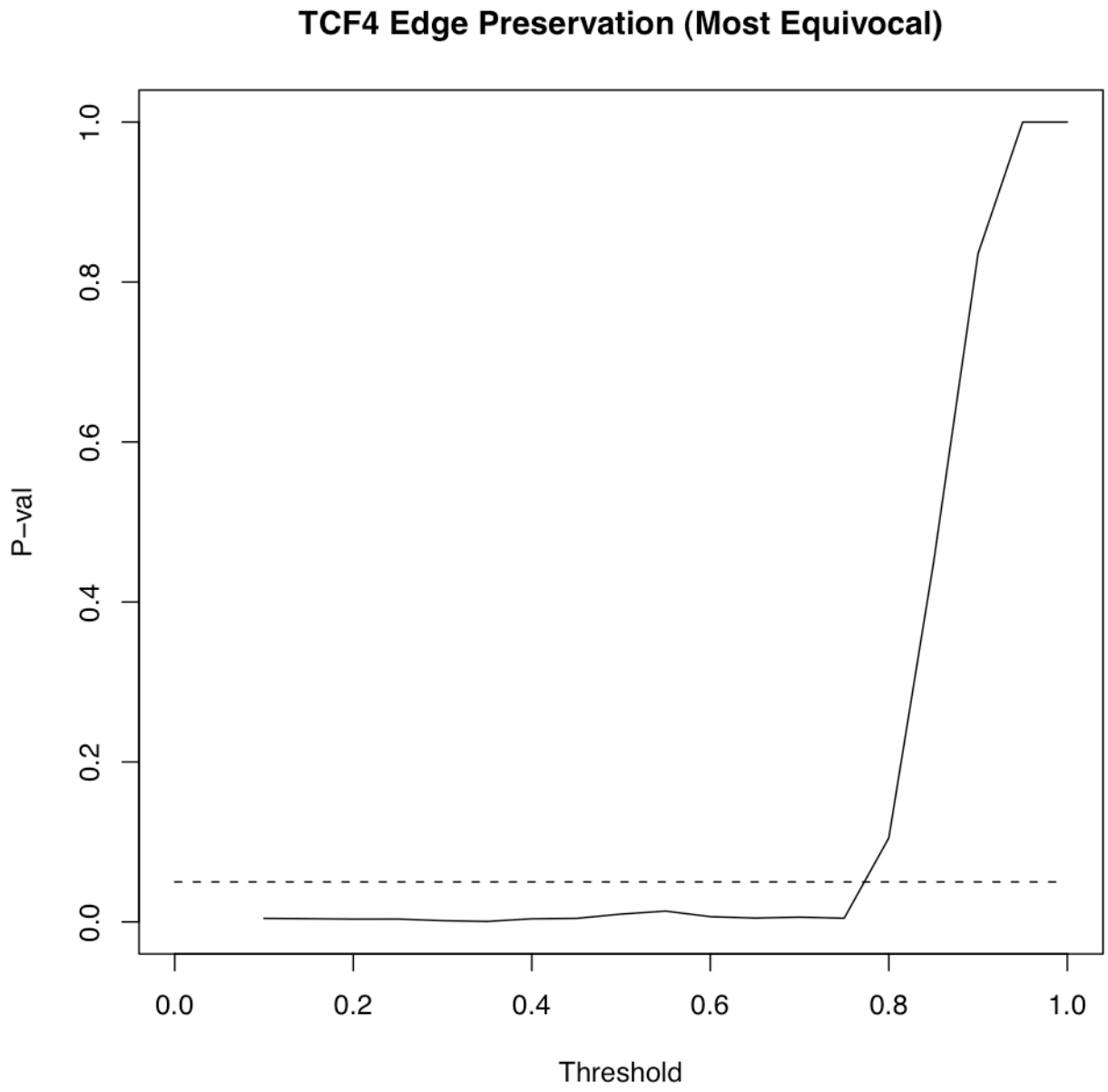
Solid line represents average p-value for 10 classifier variations

Figure 4.5- P-values of connected edge preservation analysis, highest probability site selection criteria (c-Myc Target Database)



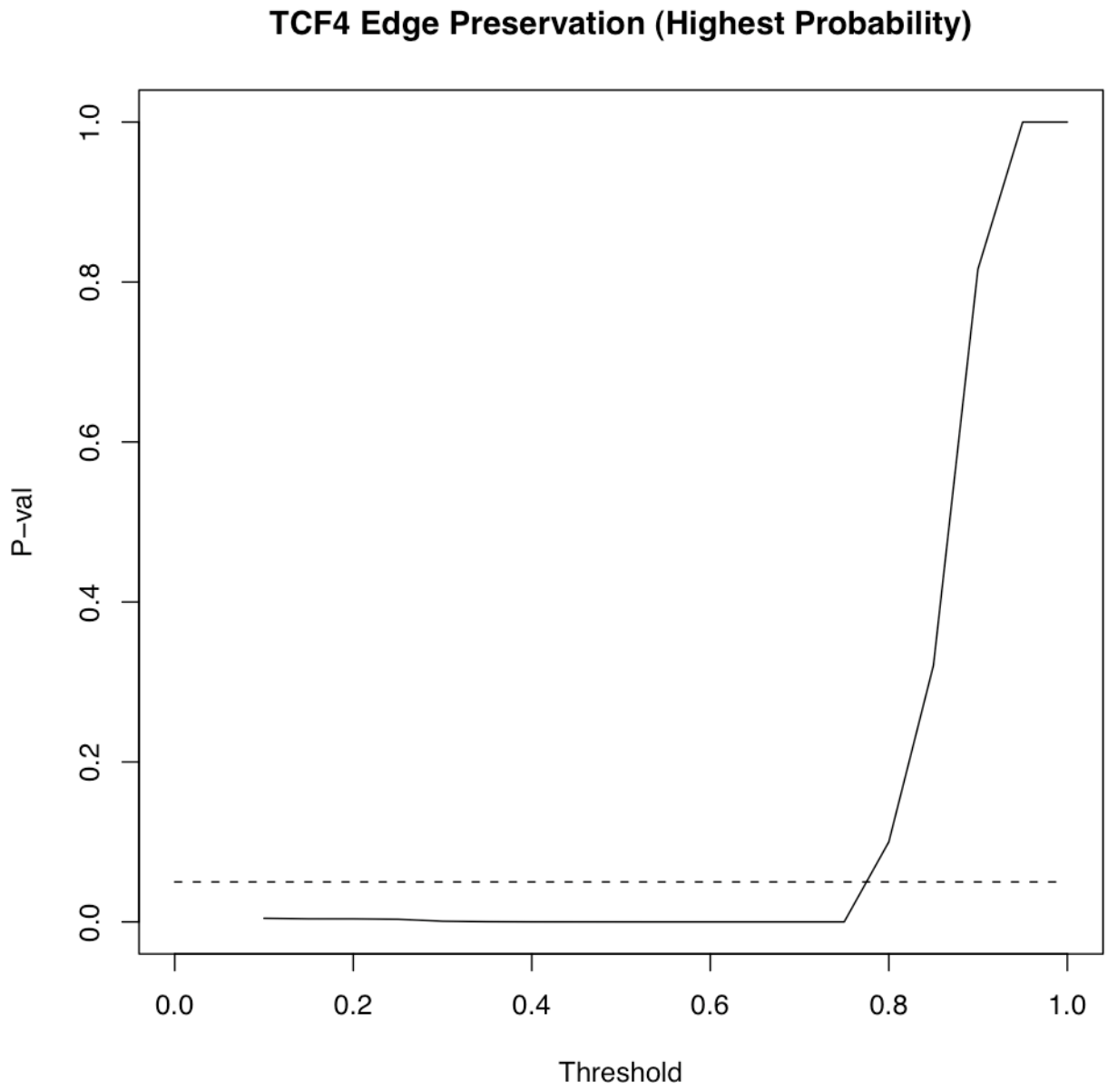
Solid line represents average p-value for 10 classifier variations

Figure 4.6- P-values of connected edge preservation analysis, most equivocal site selection criteria (TCF4)



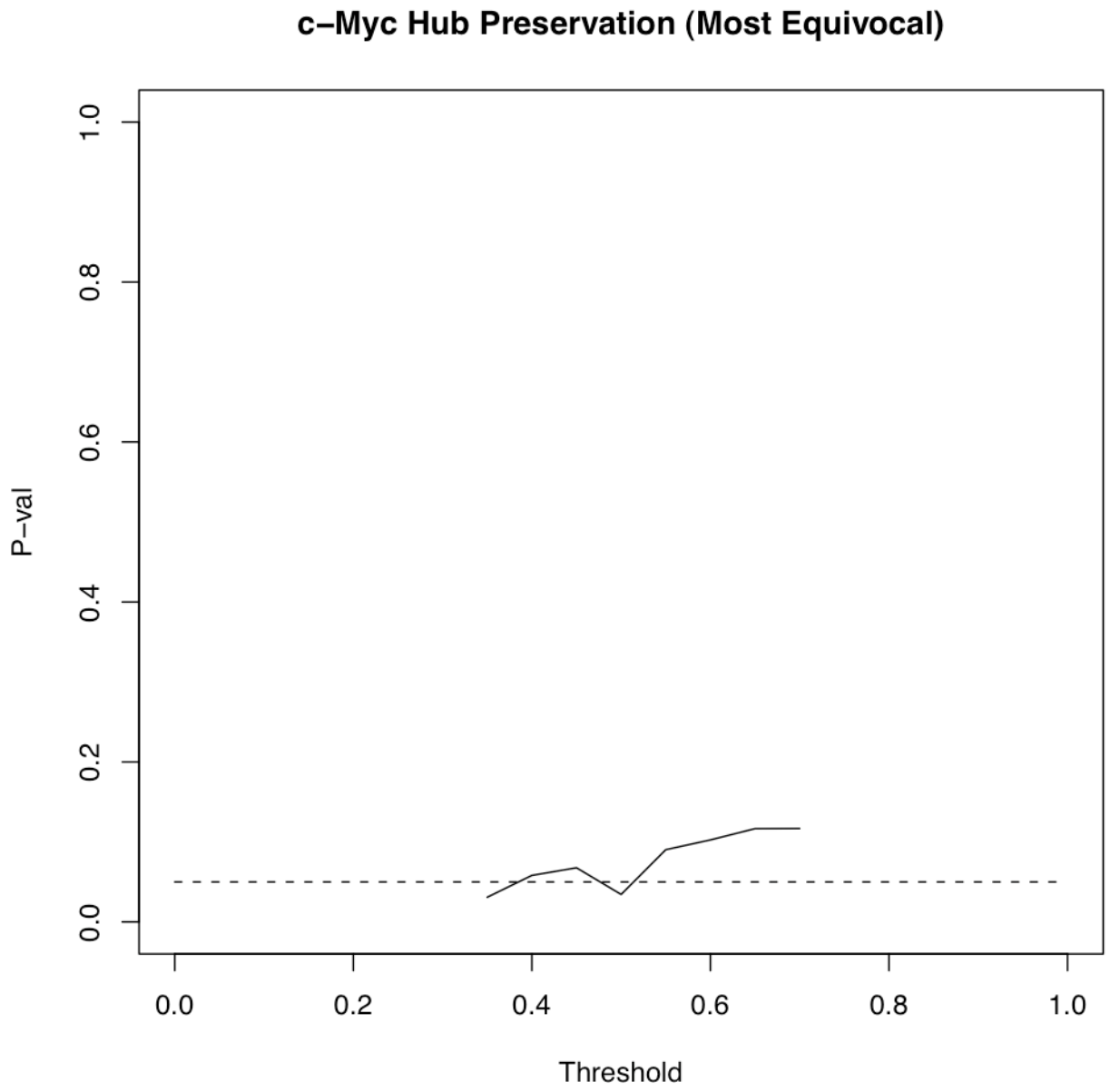
Solid line represents average p-value for 10 classifier variations

Figure 4.7 - P-values of connected edge preservation analysis, highest probability site selection criteria (TCF4)



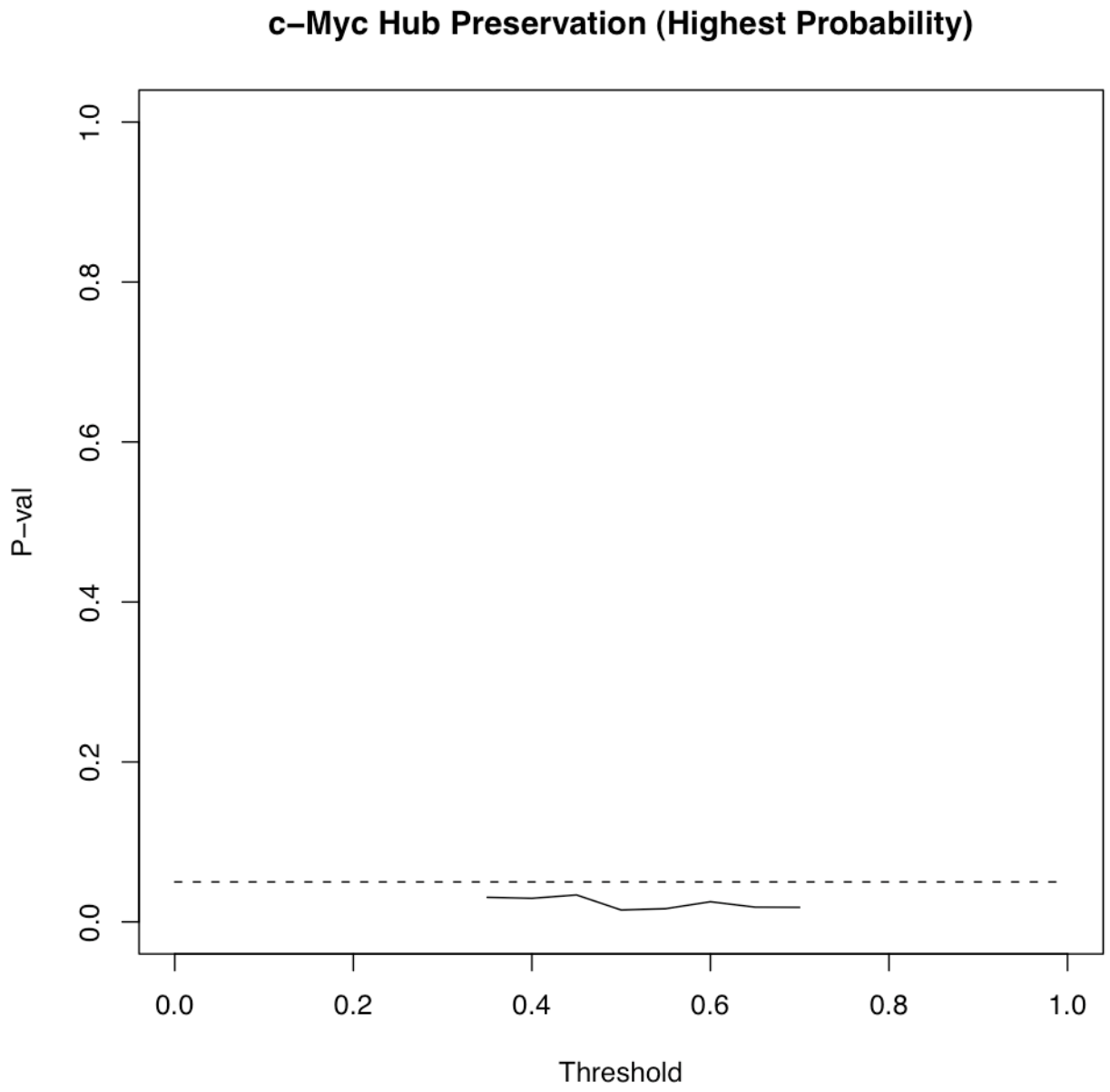
Solid line represents average p-value for 10 classifier variations

Figure 4.8 - P-values of hub preservation analysis, most equivocal site selection criteria (c-Myc)



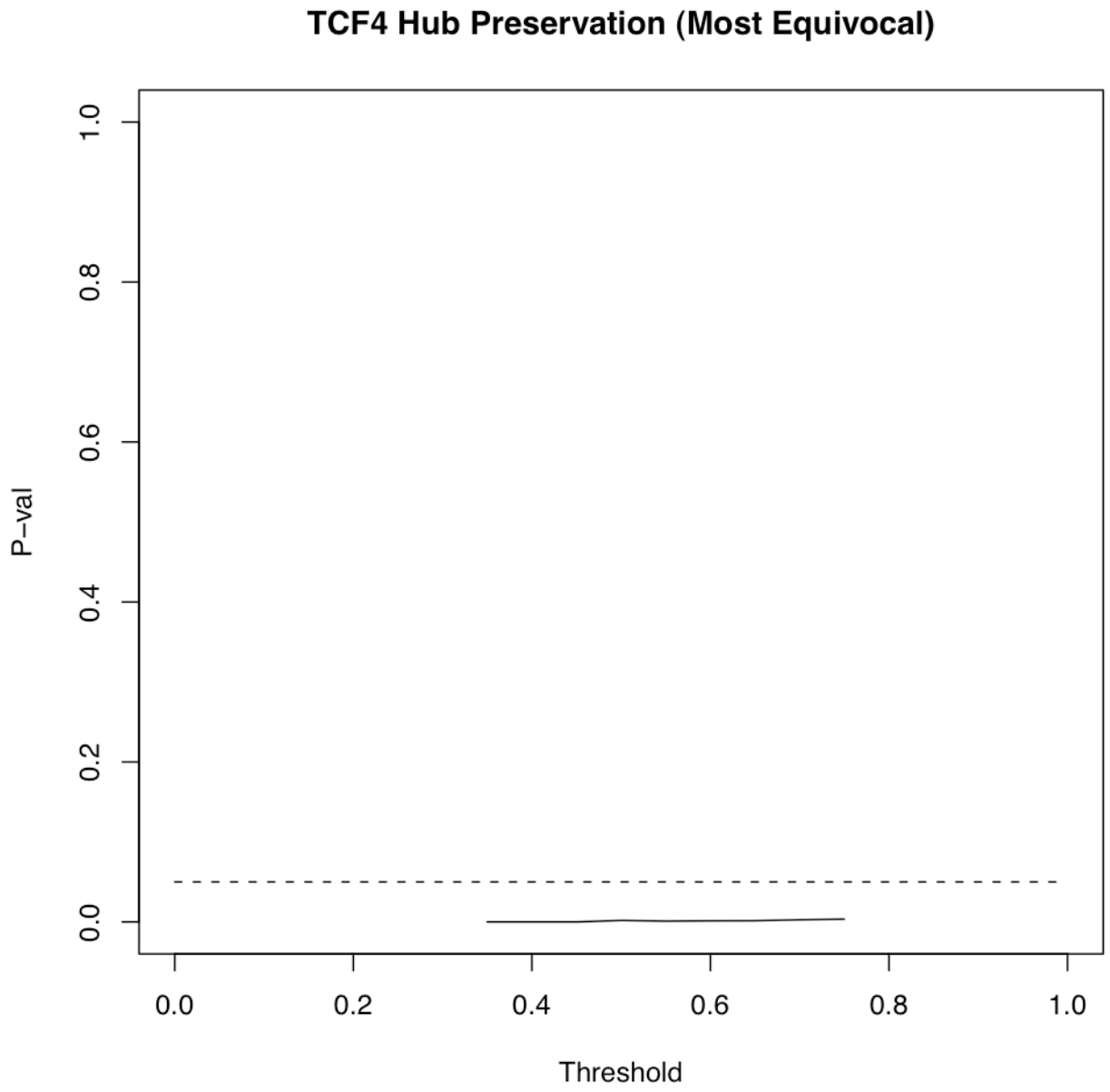
Solid line represents average p-value for 10 classifier variations

Figure 4.9 - P-values of hub preservation analysis, highest probability site selection criteria (c-Myc)



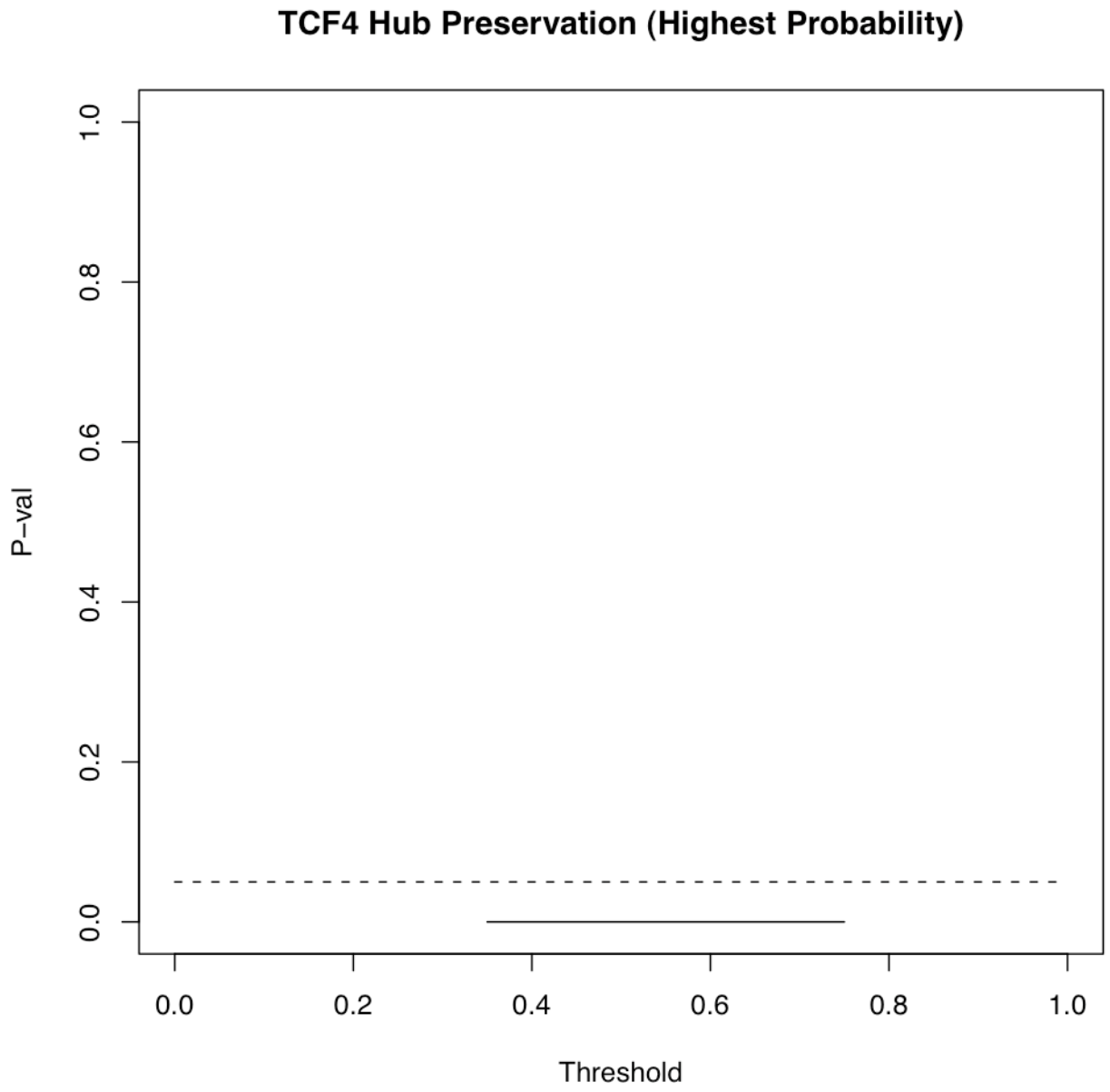
Solid line represents average p-value for 10 classifier variations

Figure 4.10 – P-values of hub preservation analysis, most equivocal site selection criteria (TCF4)



Solid line represents average p-value for 10 classifier variations

Figure 4.11 - P-values of hub preservation analysis, highest probability site selection criteria (TCF4)



Solid line represents average p-value for 10 classifier variations

Figure 5.1 – Illustration of Protocol Workflow for Prioritizing Shared Hubs for Biochemical Confirmation

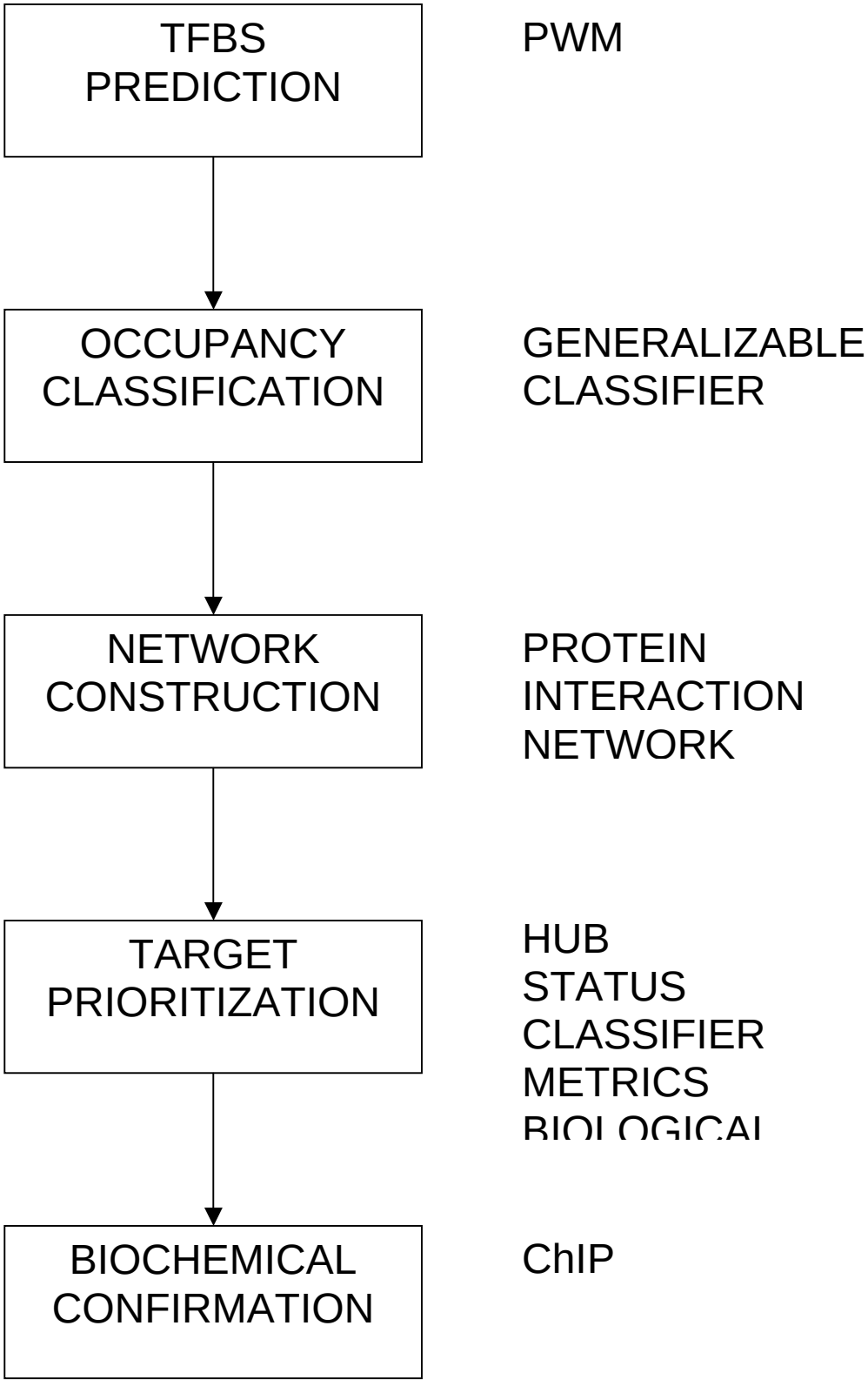


Figure 5.2 – Chromatin Immunoprecipitation Confirmation of Binding at Predicted TCF4 Target Genes

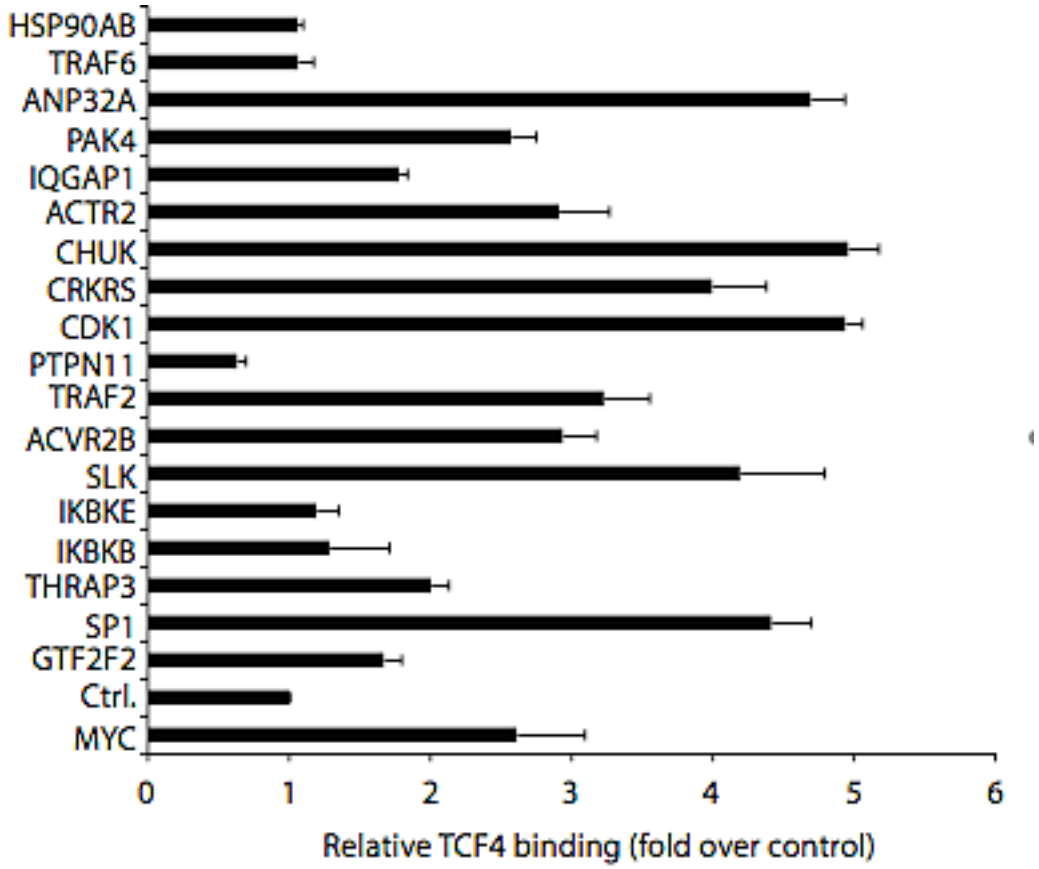


Figure 5.3 – Chromatin Immunoprecipitation Confirmation of Binding at Predicted c-Myc Target Genes