Association Based Prioritization of Genes

by

Jang H. Lee

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved March 2011 by the
Graduate Supervisory Committee:

Graciela Gonzalez, Co-Chair
Jieping Ye, Co-Chair
Hasan Davulcu
Amelia Gallitano-Mendel

ARIZONA STATE UNIVERSITY

May 2011

ABSTRACT

Genes have widely different pertinences to the etiology and pathology of diseases. Thus, they can be ranked according to their disease-significance on a genomic scale, which is the subject of gene prioritization. Given a set of genes known to be related to a disease, it is reasonable to use them as a basis to determine the significance of other candidate genes, which will then be ranked based on the association they exhibit with respect to the given set of known genes. Experimental and computational data of various kinds have different reliability and relevance to a disease under study. This work presents a gene prioritization method based on integrated biological networks that incorporates and models the various levels of relevance and reliability of diverse sources. The method is shown to achieve significantly higher performance as compared to two well-known gene prioritization algorithms. Essentially, no bias in the performance was seen as it was applied to diseases of diverse kinds, e.g., monogenic, polygenic and cancer. The method was highly stable and robust against significant levels of noise in the data.

Biological networks are often sparse, which can impede the operation of association-based gene prioritization algorithms such as the one presented here from a computational perspective. As a potential approach to overcome this limitation, we explore the value that transcription factor binding sites can have in elucidating suitable targets. Transcription factors are needed for the expression of most genes, especially in higher organisms and hence genes can be associated via their genetic regulatory properties.

While each transcription factor recognizes specific DNA sequence patterns, such patterns are mostly unknown for many transcription factors. Even those that are known are inconsistently reported in the literature, implying a potentially high level of inaccuracy. We developed computational methods for prediction and improvement of transcription factor binding patterns. Tests performed on the improvement method by employing synthetic patterns under various conditions showed that the method is very robust and the patterns produced invariably converge to nearly identical series of patterns. Preliminary tests were con-

ducted to incorporate knowledge from transcription factor binding sites into our network-based model for prioritization, with encouraging results.

To validate these approaches in a disease-specific context, we built a schizophrenia-specific network based on the inferred associations and performed a comprehensive prioritization of human genes with respect to the disease. These results are expected to be validated empirically, but computational validation using known targets are very positive.

ACKNOWLEDGEMENTS

I really thank my advisor Dr. G. Gonzalez for facilitating this study. Her encouragement and support were vital and my appreciations for them are very deep. I also thank my committee members, Drs. J. Ye, and H. Davulcu for overseeing this study, and especially Dr. A. Gallitano-Mendel for collaboration on schizophrenia research. Dr. Matt Huentelman is much thanked for providing experimental data on Alzheimer's disease. A great appreciation is due to Nate Sutton of Diego lab for processing various data items. Members of Diego lab are acknowledged for creating a stimulatory research environment.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Chapter 1. INTRODUCTION

The chances of developing a disease are often greatly increased with aberrations in certain genes or in accordance with particular genotypes. For example, the $\varepsilon 4$ allele of the APOE gene is known to significantly increase its carrier's susceptibility to Alzheimer's disease. Mutations or deletions of certain genes such as BRCA and TP53 are found in a large part of genomes of cancer patients. Genes have different degrees of implication and significance toward the cause and development of a particular disease. Gene prioritization refers to the problem of assessing gene significance with respect to a disease on a genomic scale. A small number of genes (seed genes) that are already known to be relevant to the disease are typically assumed as an input. Then the rest of genes from a genome are candidates, the significance of which are to be quantified, typically via ranking. Various approaches were proposed [11, 13, 15, 17, 21, 22, 39, 41]. Most utilize associations between genes in translating the significance of seed genes to candidate genes.

Genes can be associated in a number of ways: (1) based on biochemical associations such as protein interactions, (2) based on the similarity of biological processes in which they participate, or (3) based on regulatory associations where a transcription factor regulates the expression of its target genes. These associations can be seen as a biological network relating the genes with each other in a genome. Biological networks based only on one kind of associations are often sparse. From a computational perspective, such sparsity, or absence of necessary associations, is often an impeding factor to the performance of many gene prioritization algorithms. The model introduced here is thus based on an integrated view of such associations.

The biological processes in which genes participate are obtained from the "Gene Ontology" [12]. It provides a system of controlled terms which can be used to describe genes from that perspective, as well as from the perspectives of molecular function and cellular location. An ontological similarity between genes can be quantitatively assessed using over-

laps in the Gene Ontology terms [14, 30] that describe them. Thus, genes with similarity above an appropriate threshold can be associated from an ontological perspective.

The third type of association is more difficult to derive. Expression levels of genes are regulated by transcription factors. The gene encoding a transcription factor and its targets can be associated with one another from a genetic regulatory perspective. Transcription factors comprise large parts of genomes. Nearly all eukaryotic gene expressions are under regulatory controls of transcription factors, hence, their role in the biological processes of organisms is essential. Each transcription factor functions by binding to specific patterns located in the promoter regions of its target genes. Current level of knowledge on the regulatory associations between transcription factors and their targets is still very poor, with most transcription factors still lacking a comprehensive identification of their targets. Thus, efforts for transcription factor target identification are of interest, given the need for understanding the regulatory associations within organisms, and considering their role as a critical performance determinant for gene prioritization algorithms.

Pattern-based identification of potential transcription factor targets is a challenging problem. Since short patterns can occur merely by random chances in genomes, the information content and specificity of binding patterns are, in general, low. Each transcription factor recognizes and binds to multiple similar patterns. The sequence pattern was shown to be a dominant determinant in the interaction of a transcription factor with its targets [122]. Binding sequence patterns of many transcription factors are still unknown. Hence, a reliable computational approach to identifying them is of merit.

This work proposes a computational approach to identify binding patterns using the interaction propensity between residues, which can be expressed in terms of affinity or probability. Transcription factor-DNA interaction is fundamentally an energy-based process, where a low energy combination and spatial arrangement of residues confer a high affinity. Selex [8] and phage display [9] can be used in comprehensively probing interaction energies between protein amino acids and DNA nucleotides. Berg [111] employed Boltzman's statistical mechanics formulation to translate the interaction energy levels to the

probability distributions of DNA bases at binding sites. We utilize Selex data from [87] and the formulation from [111] to develop a method for predicting the binding patterns of zinc finger (ZF) transcription factors, which comprise the largest class of transcription factors in eukaryotes. Then we apply our method to a major subclass of ZF transcription factors, C2H2 (cysteine-2, histidine-2), to predict their target patterns.

Additionally, we developed a method to utilize the ChIP-seq data together with previously known patterns of low accuracy to approximate true patterns as closely as possible. When binding patterns of a transcription factor are reported in the literature, they often exhibit a large degree of inconsistency. Many traditionally available patterns were based on small numbers of transcription factor-bound sequences. The inherently high levels of degeneracy of transcription factor binding patterns imply that the patterns of bound sequences are diverse, which would hardly be represented by the small sets of sequences. Since the presence and conformance of binding patterns are the most critical factors in the transcription factor-target gene interactions, inaccurately represented patterns will render inaccurate the subsequent association inferences that employ the patterns. Considering that the distribution of functional binding patterns would be a unique characteristic of a transcription factor in a genome, it would be desirable to obtain a representation corresponding to such uniqueness. The experimental platform of ChIP-seq [137] involves chromatin immunoprecipitation of transcription factors and sequencing of the bound DNA, and can be used in finding the patterns that are bound by transcription factors in vivo. While useful, the loci reported as bound by transcription factors involve substantial tolerances ranging from 10's to 100's of bases. Traditional approaches addressing ChIP-seq data focused mostly on finding new patterns, without assuming prior knowledge of patterns. Attempts by existing algorithms to find the transcription factor-bound patterns from more prevalent unbound background sequences often produced inaccurate patterns. Furthermore, results often vary each time algorithms are invoked, as they employ a substantial amount of randomness, so as to reduce the problem complexity of finding de novo patterns.

Overall, this work contributes a novel integrative gene prioritization method, an algo-

rithm for the prediction of binding patterns, and an approach for improving the accuracy of binding patterns. We first introduce our method for integrative gene prioritization (IGP) by multiple data integration (MDI) in Ch.2. Then we explain our methods to predict and improve transcription factor binding patterns (IBP) in Chs.3 and 4 respectively. Finally in Ch.5, we apply the developed methods to schizophrenia and EGR3 transcription factor, a gene important in the disease, to obtain a comprehensive prioritization of genes in the human genome with respect to their potential relevance to schizophrenia.

2.1 Abstract

Many methods have been proposed for facilitating the uncovering of genes that underlie the pathology of different diseases. Some are purely statistical, resulting in a (mostly) undifferentiated set of genes that are differentially expressed (or co-expressed), while others seek to prioritize the resulting set of genes through comparison against specific known targets. Most of the recent approaches use either single data or knowledge sources, or combine the independent predictions from each source. However, given that multiple kinds of heterogeneous sources are potentially relevant for gene prioritization, each subject to different levels of noise and of varying reliability, each source bearing information not carried by another, we claim that an ideal prioritization method should provide ways to discern amongst them in a true integrative fashion that captures the subtleties of each, rather than using a simple combination of sources. Integration of multiple data for gene prioritization is thus more challenging than its single data type counterpart. What we propose is a novel, general, and flexible formulation that enables multi-source data integration for gene prioritization that maximizes the complementary nature of different data and knowledge sources in order to make the most use of the information content of aggregate data. Protein-protein interactions and Gene Ontology annotations were used as knowledge sources, together with assay-specific gene expression and genome-wide association data. Leave-one-out testing was performed using a known set of Alzheimer's Disease genes to validate our proposed method. We show that our proposed method performs better than the best multi-source gene prioritization systems currently published.

2.2 Introduction

Of particular relevance to researchers trying to track the molecular basis of disease is to be able to increase the selectivity and sensitivity when predicting the potential association of a phenotype or function with specific genes, an area referred to as "gene prioritization".

Genome sizes of species of interest are typically large, and gene prioritization is an effective means for data reduction. By ranking genes in terms of their relevance to a disease, and with an appropriate thresholidng, a select set of genes can be generated by gene prioritization. Time and cost considerations in disease research usually favor a reduced gene set which enables more focused research and facilitates more effective use of the limited resources.

Over the years, many methods have been proposed for this purpose, with molecular biologists usually favoring those that focus on the statistical analysis and consequent ranking of lists of genes from the output data of high-throughput experiments. Thus, significance analysis of microarrays (SAM), analysis of variance (ANOVA), empirical Bayes t-statistic, between group analysis (BGA), and other methods are used with the help of biostatisticians, and are sometimes provided with commonly used commercial and open-source bioinformatics tools such as Illumina's Genome Studio or caBIG's geWorkbench. Knowledge about the significant genes is sometimes provided by the tools or by sought out separately by researchers only as a way to annotate the genes, but is not used to prioritize them. Researchers have to pick and choose using their own intuition and experience.

Integrating multiple kinds of heterogeneous data and knowledge sources is a challenging problem for which formulation of a flexible and general approach is sought. A number of approaches employing protein interaction as a single knowledge source[28, 17, 31] have been published. Other systems, the best of which are Kohler et al's [21] GeneWanderer and Aerts et al's [11] Endeavour, use heterogeneous knowledge and data sources. GeneWanderer was shown to outperform many existing network-based gene prioritization algorithms [40]. It assumes a set of seed genes known to be disease genes as input and proposes a method where nodes in a protein interaction network are randomly visited (restarting the walk randomly during the process), ranking candidates with respect to their relevance to the given seed gene set. Aerts et al proposed Endeavour, a similarity-based approach that uses heterogeneous data to calculate the similarity between a set of candidate genes and a set of 'training' or seed genes. It was successfully employed in various biological studies. Candidate genes are ranked independently by using a selection of knowledge sources. An

N-dimensional order statistics is used for combining the multiple rankings. de Bie et al [15] used similarity measures and kernels corresponding to each data source and integrated rankings from multiple sources by weighting kernels. Li et al [22] employed Gene Ontology (GO, [12])-derived gene similarity networks and a protein-protein interaction (PPI) network, applied random walk with restart to each and combined the multiple rankings by using a discounted rating system.

Albeit intended on a genomic scale, most of the currently available knowledge sources and experimental platforms have rather low sensitivity. For example, current PPI databases are estimated to capture only 10% of true interactions [18]. Often times data and knowledge sources are orthogonal, with pieces of information absent in one being provided in another. Thus, distinct sources tend to have a complementary nature such that a holistic perspective on genes can be gained by appropriately complementing and integrating distinct sources. Existing approaches for multiple sources take data and knowledge sources separately, whereby their complementarity can be easily lost. Also, many involve rather high computational cost or assume specific types of data and limit the applicability to other data types.

Given a known group of genes associated with a specific disease as a "seed", we hypothesized that the degree of association of a candidate gene with the seed genes signifies its relevance to the disease. All knowledge about the genes was represented in a single network, which can be appropriately configured based on types of data, availability and reliability. Here, we used protein-protein interactions (PPIs), Gene Ontology annotations, gene expression data and SNP data from a Genome-Wide Association Study for validating our approach. Application to a large number of diseases of distinct kinds showed uniform performance level and hence no bias for particular kinds of diseases. We report the results of this general experiment, as well as a more extensive evaluation using genes related to Alzheimer's Disease (AD).

2.3 Material and methods

PPI and Gene Ontology associations were used as knowledge sources in building an integrated gene-gene association network used for gene prioritization. This is what we called the base scheme (BS) for purposes of evaluation. Additionally, gene expression and GWAS data were used as empirical data sources and incorporated in the prioritization by adding a value (level of significance) to each node in the integrated network above. This is what we called the incorporated scheme (IS). In the following subsections, we outline how the associations for each component of the network are defined and integrated, and present two experimental setups (the base scheme and the incorporated scheme) to validate the approach.

Establishing Gene Ontology associations

The Gene Ontology (GO) consists of a directed graph of terms organized under three main categories: biological process, cellular component and molecular function. Genes are annotated with those terms that apply to them. Resnik [30] defined similarity between two GO terms t_0, t_1 under the same category as

$$sim(t_0,t_1) = \mathrm{IC}_{ms}(t_0,t_1) = \max \mathrm{IC}(t_p) \tag{2.1}$$

where $t_p \in parents(t_0,t_1)$, and IC(t) is the information content of term t which is defined as $\mathrm{IC}(t) = -logP(t)$ with $P(t)$ being the probability of occurrence of the term across a genome.

Couto et al [14] defined similarity between two genes g_0, g_1 with respective terms $t_a \in \{terms(g_0)\}$ and $t_b \in \{terms(g_1)\}$ as

$$sim(g_0,g_1) = \max_{a,b} sim(t_a,t_b)\mathrm{IC}(t_a)\mathrm{IC}(t_b) \tag{2.2}$$

Term similarity is a normalized quantity ranging between 0 and 1. We used GO annotations [20] of the human genome, which included a total of 14,685 genes annotated with biological process terms, with a total term occurrence count of 60,792 for an average of 4.140 terms

per gene. In establishing a gene-gene association based on GO annotations, we varied the similarity threshold from 0.30 to 0.70 in increments of 0.10 to retain gene pair similarity only above or equal to the given threshold, obtaining five nested sets of associations.

Protein-protein interactions

Three protein interaction databases were employed, to match those used by Kohler et al in [21] and allow a fair comparison: HPRD[29], STRING[25] and NCBI yeast protein interactions. HPRD is a manually annotated protein interaction data set: the one we used had 2,125 homomeric interactions and 36,631 heteromeric interactions. The STRING database contains information from four sources (genomic context, high-throughput experiments, coexpression, and derived from text), including direct (physical) and indirect (functional) associations. We used version 8.3, which covers 2.6 million proteins from 630 organisms. Each interaction in STRING is assigned a significance score (non-linear) in the range between 150 and 1000. In addition, known protein interactions in yeast were downloaded from NCBI[26]. Each yeast protein was mapped to a human ortholog using InParanoid[27]. Only interactions involving protein pairs that have a 100% match score to human orthologs were retained (a total of 39,665).

Interacting proteins were each mapped to coding genes and then a set of interacting genes were obtained. Some common interactions in the databases derive from single experimental evidence and hence there exists a degree of duplicity among the three databases. The three PPI networks were combined into a single network by counting edges only once irrespective of their duplicity:

$$\{e'(g_1,g_2)\} = \cup\{e_{N_i}(g_1,g_2)\}, 1 \leq i \leq N \tag{2.3}$$

with $e'(g_1,g_2)$ being the edge between nodes g_1 and g_2 in the combined network and N being the total number of PPI networks. Five distinct sets of associations were obtained by using nested sets of interactions with different STRING significance score thresholds (300, 400, 500, 600 and 700).

Gene expression

For this paper, we used microarray expression data sets by Webster et al [32], comprised of control and AD case samples. Genes showing significantly distinct levels between normal and disease cells were identified by using differential expression analysis. Wilcoxon rank sum test was applied to expression levels from the two groups of samples and a P-value of each gene's differential expression was obtained. The P-value threshold was set to 0.05. The significance of a gene G, $S(G)$, from differential expression was calculated as:

$$S(G) = -log(\text{P-value}) \tag{2.4}$$

Genome-wide association study

SNP genotyping is performed on genomes from normal and disease samples. Certain SNP may show distinct presence in one group vs the other e.g., allele A constitutes 80% of disease samples at a certain locus while it constitutes 30% in normal samples. A P-value can be calculated for each SNP and hence for a corresponding gene if the locus of the SNP is within or close to the gene, which would imply the gene is strongly relevant to a specific disease. If a SNP is too distant from genes (more than 20kb away upstream or 5kb downstream), then it was not included in our experiments. Similar to expression data, disease significance P-values were calculated and assigned to genes by using Eq 2.4.

Network representation

To construct the networks used for the base (BS) and incorporated (IS) schemes, the PPI and GO associations described above were used as edges, with genes mapped to nodes. If more than one knowledge source associated two genes g_1 and g_2, then the edge is weighted according to the multiplicity of the number of associating sources. Thus, if N sources were associating the two genes then weight($e(g_1, g_2)$) = N.

Gene g may be completely missing or may not have a P-value above a threshold in the outcome of some experimental data, and have P-values above thresholds only in N_e number of effective sources. Given a significance $S_i(g)$ from empirical data source i ($1 \leq i \leq N_e$)

for a given disease, gene g's overall empirical significance is calculated as

$$S(g) = \sum_{i=1}^{N_e} S_i(g) \tag{2.5}$$

That is, the sum of all significance values is assigned as a combined significance score for the gene (its aggregate experimental significance).

Base scheme

Given a set of training seed genes $\{S_i\}$, candidate gene C was scored as follows:

$$score(C,S) = \sum^{\forall S} e(C,S_i) \tag{2.6}$$

where $e(C,S_i)$ is a non-zero value if an edge exists between C and S and 0 otherwise. Either only the edge presence between C and S can be recognized for scoring, or its weight from the aggregate network can be considered, i.e.,

$$e(C,S)_{BS1} = \mathbf{1}\{e(C,S)\} \tag{2.7a}$$

$$e(C,S)_{BS2} = \text{weight}(e(C,S)) \tag{2.7b}$$

with $\mathbf{1}$ being an indicator function corresponding to edge presence. If only the presence of an edge is considered, then Eq 2.7a is used together with Eq 2.6. This will be referred to as base scheme 1 (BS1). If edge weight is considered, then Eqs 2.7b and 2.6 are used which will be referred to as base scheme 2 (BS2). Candidate genes are ranked according to their scores.

Empirical data incorporation scheme

The network topology used in the empirical data incorporation scheme (IS) is the same as the one in the base scheme. Candidate gene C can have an edge to jth seed gene T_j of an overall empirical significance $S(T_j)$. Then T_j's contribution to the score of C is calculated as

$$e(C,T_j) + kS(T_j) \tag{2.8}$$

where k is a scaling factor, the value of which is to be set according to data reliability. If an edge does not exist between them, then T_j's contribution is 0. The contribution from each training gene T_j, $1 \leq j \leq |T|$, in the training set to candidate gene C is added up for its combined score:

$$score(C) = k_1 S(C) + \sum_{j=1}^{|T|} [e(C,T_j) + k_2 S(T_j)] \tag{2.9}$$

where k_1 and k_2 are scaling factors and $|T|$ the total number of training genes. The ranking of the candidate genes corresponds to the combined scores of the candidate genes. Pseudo-code of the algorithm is shown in Algorithm 1.

Algorithm 1 Pseudo-code of integrative gene prioritization

```
network N = φ
for all sub-networks n_i do
    N∪ = n_i
end for
for all g do
    S(g) = Σ^{∀E_i} S_{E_i}(g), (experimental data, E_i)
end for
for all g do
    score(g)=k_1 S(g) + Σ^{∀T_j} [e(g,T_j) + k_2 S(T_j)], (seed gene T_j)
end for
```

Validation

The disease gene sets from Kohler et al [21] were used. Leave one out testing was performed by holding out one disease gene as a true test gene to be (ideally) recalled from the disease gene set by taking the remainder genes as a training gene set, and this was repeated for each gene over all disease gene sets. Sensitivity and specificity values were calculated as defined in ([11]). Specifically, ranking results were aggregated and the number of true test genes above a given ranking threshold was counted as true positives. The number of test genes below the threshold, non-test genes below the threshold and non-test genes above the threshold were respectively counted as false negatives, true negatives and false positives. As frequently done in literature, a narrowed-down set of genes (e.g., 100) in closest proximity

Table 2.1: Ranking of Alzheimer's disease genes by different algorithms

	Base		Endeavour	GeneWanderer		Incorp.	
Gene	Rank	Rk100	Rk100	Rank	Rk100	Rank	Rk100
APOC1	93	2	5	275	7	1	1
APOE	1	1	4	17	4	1	1
APP	382	1	4	264	1	156	1
CLU	7	1	9	102	2	17	1
CR1	437	2	44	1158	3	352	2
GAB2	202	1	31	496	3	452	2
MSRA	-	100	24	6511	11	-	100
PICALM	444	1	8	978	3	95	1
PSEN1	1	1	2	14	1	1	1
PSEN2	7	1	4	84	1	25	1
PVRL2	7	1	47	67	4	15	2
RELN	439	1	43	957	5	413	1
TOMM40	1261	10	86	3319	18	34	2

to the true test gene along its chromosome is given as a candidate set. We also show the ranking obtained over all genes in the genome.

Current knowledge sources may involve degrees of incompleteness and incorrectness. This would correspond to false positive and negative edges in networks. Facing this, we randomly perturbed 10% of network edges by randomly reassigning them in an experiment. Eight such instances of randomly perturbed networks were generated and the base scheme was applied to each of them.

2.4 Results

Genes implicated in AD were collected from the literature ([10], [19], [24], [23], [33], [34]) (Table 2.1). For comparison of performance, gene prioritization based on random walk with restart (RWR) as described by Kohler et al ([21]) was implemented. In RWR, nodes are navigated in a random fashion starting from a gene randomly selected from a given set of seed genes. Gene ranking in RWR is according to the visit frequency at the conclusion of iteration following a convergence criteria. In addition, Endeavour [11] was downloaded from the authors' website. It randomly selects 99 genes other than true test gene to produce a 100 gene candidate set together with the test gene. Even though the candidate gene sets

used for Endeavour are different from the ones used for base scheme and RWR, we reasoned the set size is sufficiently large from a statistical sense to facilitate sound comparisons and show the rankings under the column name of Rk100.

The base gene prioritization scheme was applied to the AD gene set. The same set was also used for Endeavour and GeneWanderer. When gene APOC1 was left out as a true test gene to be recalled and the other genes were used as a training seed gene set (row 1 in Table 1), there were 92 other genes from the human genome which ranked more significantly (column Base-Rank in Table 1). When the candidate gene set was reduced to the 100 genes of closest proximity (Loc100 set), APOC1 ranked 2nd highest (column Base-Rk100). Endeavour's ranking of the gene was 5th out of 100 genes and RWR's ranking was 275th among entire genome and 7th among Loc100 genes. Each subsequent row can be read in a similar fashion. Thus, the base gene prioritization scheme ranked the AD set genes more significantly than RWR (signed rank test P-value=6.836×10^{-3}.) and Endeavour (P-value= 2.148×10^{-2}).

In order to assess the applicability of the base scheme (BS1) to other diseases besides AD, we applied it to disease gene set of Li et al [22] (Li10) which was derived from the complete Kohler et al set. It includes 36 diseases and genes implicated therein. The receiver

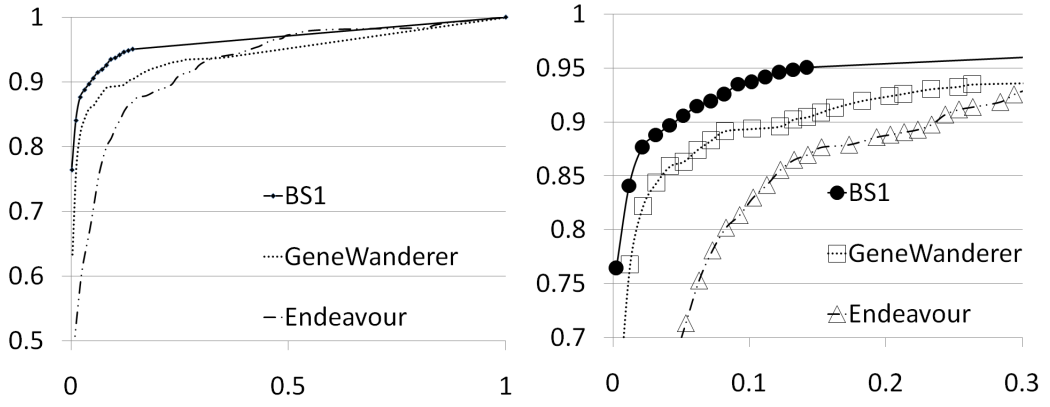


Figure 2.1: ROC curves of specificity vs. 1-sensitivity (a) Base scheme has a larger AUC than Endeavour and RWR. (b) Close-up of higher sensitivity range

Table 2.2: AUC difference between base scheme 1 and base scheme 2 (units in $1e^{-4}$)

GO-PPI	300	400	500	600	700
30	-5.603	-12.101	-12.751	-14.561	-14.451
40	-0.265	+5.767	+9.010	+8.708	+4.035
50	+1.815	+5.567	-0.048	+6.935	+7.971
60	+0.986	+1.065	-0.157	+0.390	+0.000
70	+0.532	+0.464	+0.000	+0.165	+0.398

operating characteristic (ROC) curve of the base scheme BS1 is shown in Fig 2.1 together with the curves of Endeavour and RWR for the same set. AUC value of the base scheme was 0.9655 while, for Endeavour and RWR, the AUC values respectively were 0.9287 and 0.9442. The reasonable AUC value means the base scheme is applicable to other diseases in general as well. Base schemes 1 and 2 were compared over the Li10 set and their AUC values showed a marginal difference possibly suggesting edge multiplicity does not greatly contribute in distinguishing true test gene from the other candidate genes (Table 2.2). Subsequently, we used only base scheme 1 and will refer to it as the base scheme.

Knowledge sources such as PPI or GO may entail some levels of false and missing annotations. In order to evaluate the influence of such noise on the performance of the base scheme, 10% of the edges in the combined network were randomly rewired. Eight such instances of the perturbed networks were generated, and then the base scheme was applied. In all cases, AUC values decreased by small degrees, but consistently from that of the unperturbed network; average AUC value was 0.96070 and standard deviation 0.00223 (Fig. 2.2 and Table 2.3). Only a slight degradation in the AUC of the perturbed network means our base scheme is robust with respect to a noticeable amount of possible mis-curations in the knowledge sources and corresponding noise in the network.

Table 2.3: AUC values from perturbed networks

Instance	1	2	3	4	Average
AUC	0.96119	0.95875	0.96216	0.95869	0.96070
Instance	5	6	7	8	St.dev.
AUC	0.96533	0.95993	0.96101	0.95908	0.00223

Diseases were categorized as belonging to one of three types by Kohler et al: cancer,

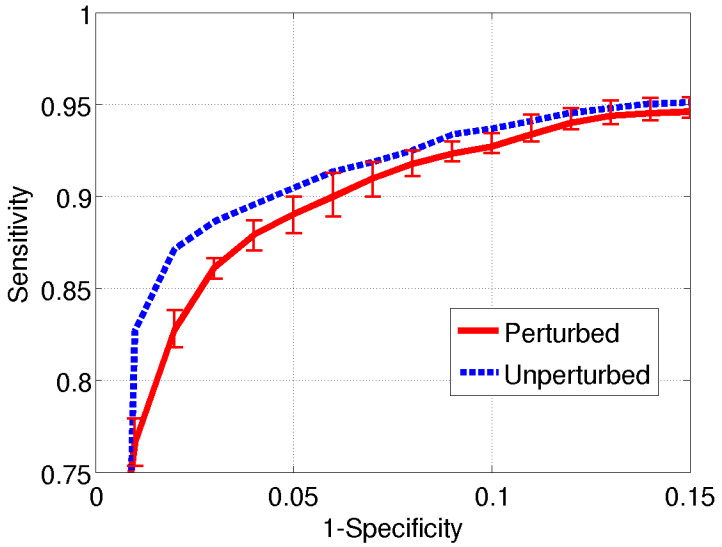


Figure 2.2: ROC's from perturbed and unperturbed networks. Perturbed networks cause slight decreases in the AUC.

Table 2.4: AUC values from application to different disease categories

Type	Cancer	Monogenic	Polygenic	Average
BS1	0.95727	0.96677	0.98025	0.96810
RWR	0.95414	0.90535	0.94978	0.95890
Endeavour	0.87947	0.94471	0.88191	0.90203

monogenic and polygenic. Cancer and polygenic categories each included 12 diseases, and monogenic 86 diseases. We chose the 6 largest disease gene sets from each category to form categories balanced in count and applied the base scheme, Endeavour and RWR to each. AUC values were similar across disease categories (Table 2.4), thus suggesting that the base scheme is not biased to a particular category of diseases. Higher AUC values were produced by BS throughout the different categories.

The contribution of individual knowledge sources was assessed by using either PPI or GO associations alone and by comparing the resulting AUC values with the ones obtained with aggregate sources. Specifically, 5 sets of GO associations were produced with distinct thresholds of 0.30 to 0.70 in increments of 0.10, and also 5 sets of PPIs with thresholds 300 to 700 in increments of 100. A total of 35 networks resulted; 5 with only GO associations

as edges, 5 PPI only, and aggregate networks in 25 different combinations of GO and PPI thresholds. The Base Scheme was applied to the Li10 set for each of the networks. The AUC value monotonically increased as GO or PPI thresholds were lowered (resulting in more network edges) (Figs 2.3(a), 2.3(b)). The highest AUC value was produced with the aggregate network of least stringent threshold combination (PPI 300 and GO 0.30).

The PPI network alone shows reasonable AUC values under varying thresholds (bottom-most curve of Fig. 2.3(a)). Aggregation with GO network consistently improves the AUC values. However, GO networks alone show rather low AUC values especially at high thresholds, but aggregation with PPIs, even at the highest threshold, drastically improves AUC values. Clearly, aggregation of networks from distinct knowledge sources is an effective way of comprehensively utilizing their respective information content, and our base scheme indeed utilizes the higher information content.

Incorporation of empirical data

Alzheimer's Disease GWAS and differential expression data were incorporated in the gene prioritization process (Table 2.1 column Incorp.) as explained in the incorporated scheme. Improvement over the base scheme was rather marginal (P-value=0.1934). This may be

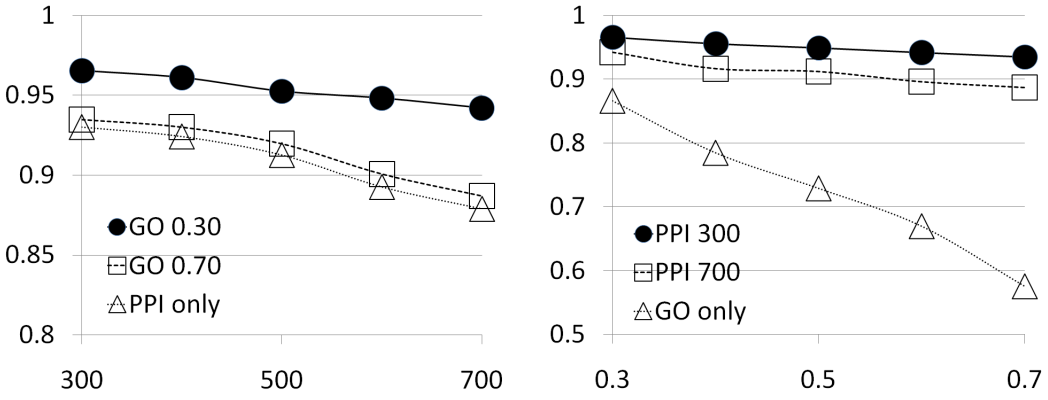


Figure 2.3: AUC values from different knowledge source combinations (a) AUC vs. PPI threshold (b) AUC vs. GO threshold. Using combined networks produces higher AUC values.

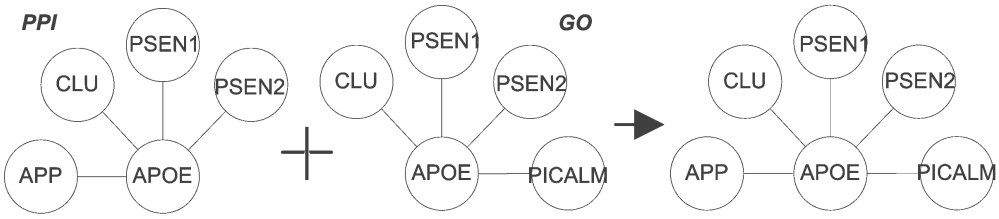


Figure 2.4: Network aggregation. Different networks are complementary to each other, where missing associations in one can be complemented from the other.

attributable to a rather low reproducibility of significant genes between experiments, especially expression data [16, 37, 36]. A number of approaches have been suggested for an appropriate interpretation and extraction of useful information from experimental data including shifting of focus towards groups of genes rather than on individual genes[38]. A new formulation of the incorporated scheme is left as a future work, which considers the difference in nature of experimental data.

2.5 Discussion and conclusion

Two different knowledge sources were each represented in a network and unified in a model that allows for additional sources to be added in a similar fashion. Each independent knowledge source is likely incomplete and missing many associations between genes [18]. The proposed knowledge integration method (base scheme) complements incomplete knowledge sources to produce a more comprehensive view of genes. For example, among well known AD genes, APOE has edges to genes APP, CLU, PSEN1 and PSEN2 in PPI network and lacks an edge to PICALM (Fig 2.4). The GO network does not have the APOE-APP edge but contains the APOE-PICALM edge. We compared our proposed method to two of the best multi-source gene prioritization algorithms. Endeavour utilizes knowledge sources separately and tended to produce the lowest AUC values among the compared algorithms. The method proposed here effectively integrates individual knowledge sources to overcome the incompleteness of each.

The base scheme alone showed better performance than Endeavour and RWR. Rank-

ings based on combined networks were consistently better than rankings based on individual networks. There is a degree of overlap between the two knowledge sources (PPI and GO), since the same information from literature is frequently used to annotate genes. Still there is information content in one source which is not captured in the other. The edge formation by similarity criterion in the GO network can associate genes that are highly related in pathways or from biological perspectives which do not directly interact through their protein products and hence is missed in a PPI network. The described schemes rely on the association between genes to infer disease genes from known genes. The effectiveness of this approach was shown through a series of experiments. The information from knowledge sources and experimental data vary in reliability, degree of curation and level of acceptance. For example, many protein interactions have been verified over time and are well accepted, while high throughput interaction data tends to involve a high rate of false positives.

Our Gene ontology annotation of genes reflects a relatively high level of verification and curation. On the other hand, experimental data is subject to a high level of noise and variance and has not been extensively and thoroughly verified. Hence a network was not directly formed from experimental evidence at this stage, and only node significance was adjusted in accordance with the experimental significance. Our schemes are robust against false positives and missing knowledge as shown in the perturbation experiment. Future work will be directed at incorporating empirical data from experiments in a way that is more consistent with the way knowledge sources are used. While particular knowledge sources and experimental data were used for illustration, the described schemes are sufficiently general to be used with other data types as well. After the preparation of our manuscript, a gene prioritization method [39] was noted for its use of diverse data with a Bayesian approach. While a readily accessible version of their algorithm was unavailable, it will be interesting to perform a comparative study involving it.

2.6 Acknowledgement

Authors are grateful to Dr. Matthew Huentelman for AD GWAS data. This work was partially funded by the NIH/NIA under grant P30 AG019610-09.

Chapter 3. PREDICTING BINDING PATTERNS

3.1 Abstract

Binding patterns are important characterizing features of transcription factors, yet, the patterns of most transcription factors are still unknown. Considering their significance in finding genetic regulatory associations, a computational approach to the prediction of binding patterns deserves research attention. The interactions of many transcription factors with DNA can be complicated, because they are comprised of constituent interactions between multiple residues.

On the other hand, the zinc finger proteins, a specific class of transcription factors, interact with DNA mainly via rather simple one-to-one residue interactions. Hence they permitted the development of a quite a tractable, canonical interaction model. Then the prediction of binding patterns for the zinc finger transcription factors reduces to the determinations of nucleotide probabilities, with respect to given protein residues, which generally require a matrix that provides the interaction propensity between residues. Given a propensity matrix expressed in energy, we explored ways to obtain probability matrices of varying degrees of conservation, by exploiting a degree of freedom conferred by a parameter set that originates from statistical mechanics. Here, we study the performance of predictions that are obtained by using two well-known residue propensity matrices [87, 100] , thereby assessing the feasibility of binding pattern predictions in general. They exhibited a low level of consistency, implying that the set of data underlying at least one was rather small, or that there was an unstability in the algorithm that was used to obtain the matrices. Moderate to large deviations from the reported patterns [136] imply that the propensity matrices are not sufficiently representative of true residue probability distributions, or that the zinc finger-DNA interactions frequently deviate from the canonical model.

3.2 Introduction

Transcription factors comprise an important class of proteins in genomes, which regulate the expression of their target genes by binding to characteristic patterns in their upstream regions. Zinc finger (ZF) transcription factors in turn comprise the largest subclass in mammalian genome transcription factors [92, 91, 98]. Patterns are typically degenerate: multiple similar, yet different patterns exhibit binding activities to a given transcription factor, with varying levels of affinity. The genetic regulatory associations between transcription factors and their targets are mediated via the recognition of and binding to the characteristic DNA patterns by the transcription factors. Identification of such binding patterns is an important goal in computational biology research. In the interaction between a transcription factor and DNA, each participating residue from one molecule may form bondings with multiple residues from the other molecule. The resultant many-to-many interactions require a large number of parameters in models of interactions, thereby weakening the tractability of models. On the other hand, the interactions of a class of transcription factors, C2H2 (cysteine-2, histidine-2) ZFs, are believed to be mainly comprised of one-to-one residue interactions. This greatly eases modeling efforts, and indeed, the transcription factor class has served as a model case of transcription factor-DNA interactions. A unit of zinc finger recognizes 4 DNA bases, and multiple units can be combined in a rather modular fashion. It is believed that the modular flexibility has conferred a rich repertoire to the regulatory controls of organisms [95].

The modularity and the simplicity of element-wise interactions encourages attempts to predict binding patterns solely by using the sequence information of transcription factors. Considering the importance of binding patterns, a reliable computational approach, if available, would greatly aid in the elucidation of genetic regulatory networks, which are still far from a comprehensive understanding. A substantial amount of prior research exists, which try to predict the binding patterns of transcription factors, e.g., [87, 88, 100]. They mainly rely on the collections of DNA and protein sequences that were reported to have

high affinities to bind to each other.

Since it is desired to predict the DNA binding pattern, given an arbitrary zinc finger transcription factor protein sequence, what is called for is a generalized prediction scheme, possibly represented via a matrix, that relates each amino acid of the protein to a probability distribution of nucleotides. The works by Benos et al [87] in 2002 (henceforth to be abbreviated to Benos02) and by Kaplan et al [100] in 2005 (abbreviated to Kaplan05) are attempts to relate arbitrary amino acid residues to their likely nucleotide interaction partners, or vice versa.

Comparative studies involving the two were performed in rather partial manners. Persikov [88] did compare them, while assuming a binary classification of interactions: binding or no binding. Transcription factor-DNA binding events span a wide range of densely occupied energy levels, due to the large number of possible amino acid-nucleotide combinations. Then the binary classification of interactions of binding or no-binding would be rather too coarse to suitably represent the diversity, complexity and the degeneracy of the binding patterns [101, 102]. Here, we performed a comparison study of the two sets of propensity matrices by using a representation scheme of the residue probability distributions of binding patterns.

Patterns obtained by using the two matrix sets were compared against those retrieved from the Jaspar database [136], that were assumed as standards. The residue propensity data of Benos02 are in the units of energy, while what are desired are the probability distributions of residues. Hence, we also explored systematic ways to convert the energy levels to probability distributions. We discovered that the parameters involved in the formulation of Benos02 can serve as a means to control the conservation levels of the patterns produced. Using obtained results, we discuss the biases of the data that underlie Benos02 and Kaplan05 sets, and factors that render the prediction problem hard.

3.3 Material and methods

C2H2 subfamily forms the largest subclass in the zinc finger transcription factor class, accounting for approximately 40% of human transcription factors [98, 91]. Its name derives from two cysteine (Cys) and two histidine (His) residues that coordinate a zinc ion to form a strand-strand-helix protein structure. The motif shows a strong conservation in the sequence pattern: Cys-X2,4-Cys-X12-His-X3,4,5-His, with X being any amino acid. Shown in the list below are three zinc fingers from the EGR3 transcription factor of M.mus, which all conform to the canonical pattern.

1. Finger 1, position 275-299: HACPAEGCDRRFSRSDELTRHLRIH, E=4e-07
2. Finger 2, position 305-327: FQCRICMRSFSRSDHLTTHIRTH, E=1.3e-05
3. Finger 3, position 333-355: FACEFCGRKFARSDERKRHAKIH, E=5.6e-05

The strong conservation stems from a requirement for proper structural shaping of the zinc fingers and their interactions with DNA. The second histidine is located at the 7th residue in an alpha helix, and to be numbered +7. The residues at positions -1,+2,+3,+6 recognize 4 consecutive DNA residues, by respectively forming contacts with nucleotides at positions 3,4,2,1 (Table 3.1). When two zinc fingers are separated by 6 amino acids, they are arranged to contact consecutive DNA bases, with one residue overlap.

Each amino acid-nucleotide interaction is rather independent of other interactions between different residues, in the interactions of zinc fingers with DNA. Additionally, each zinc finger-DNA interaction is independent of others except the one residue overlap between adjacent fingers, which overall gives a high level of modularity to the zinc finger transcription factor-DNA interactions. From a bioengineering perspective, a great flexibility in protein-DNA interaction can be achieved by arranging zinc fingers in a random linear fashion, which would then facilitate a highly specific recognition of long consecutive DNA residues. Experimental techniques such as selex [8] and phage display [9] have been used

for the interrogation of binding affinities between amino acids and nucleotides in the zinc finger-DNA interactions [103, 104, 105, 106].

Conversion of Benos02 matrices

We employ Table 2 of [87] which tabulates binding energies between all possible pairs of amino acids and nucleotides. While the entries are in the unit of energy, we seek to obtain a probability distribution of nucleotides, n's, given an amino acid, a, at position p in a zinc finger: $p(n|a,p)$. Berg [111] used the Boltzman's statistical mechanics formulation to relate the binding energies of nucleotides to their probability values. Let a denote an amino acid out of 20 possible alphabets, $\{a_i\} = \{\mathrm{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}\}$, $1 \leq i \leq 20$, and let n_j denote a nucleotide, $\{n_j\} = \{\mathrm{A,C,G,T}\}$. Given a_i at position p in a finger, we seek to evaluate the probability distribution of the nucleotides, $p(n_j|a_i,p)$. Drawing from statistical mechanics, Berg [111] interpreted each nucleotide as a state, which then enables to relate their energy levels to probability values:

$$p(n_j) = c\exp(\frac{-E_{n_j}}{k_B T}) \tag{3.1}$$

Here, c is a normalization factor specific to each (a_i,p) pair, which ensures that the probability values sum to unity.

$$\sum_j p(n_j) = p(A) + p(C) + p(G) + p(T) = 1 \tag{3.2}$$

$$c\sum_j \exp(\frac{-E_{n_j}}{k_B T}) = 1 \tag{3.3}$$

Inclusion of the Boltzman constant, k_B, and the temperature, T, stems from the original statistical mechanics setting which involved a system of particles. Berg [111] suggested to adopt suitable constant factors in their places, since the macro-molecules of multiple atoms are being addressed in the amino acid-nucleotide interactions. Then the formula corresponding to the new system is:

$$c\sum_i \exp(\frac{E_{a_i}}{kT'}) = 1 \tag{3.4}$$

, with k and T' being the new factors to be determined. Since c is determined only to ensure a proper normalization, one degree of freedom is given by the product, kT'. Its determination would have to reflect the nature of the system: the molecular interactions between amino acids and DNA. There exists a wide range of affinity in the interactions between transcription factors and DNA. The measure of information content [150] can be used in assessing the degrees of conservations of DNA residues. We empirically determine suitable k values, based on the agreements between the patterns produced and typical transcription factor binding patterns, in terms of the levels of conservations as quantified by information contents.

Determination of k

The values of the physical constant, k_B, in different units are: $k_B = 1.3806504e^{-23}JK^{-1} = 8.617343e^{-5}eVK^{-1} = 1.3806504e^{-16}ergK^{-1}$. We used the k value, $3e^{-4}$, that is in the vicinity of $8.617343e^{-5}$, as our initial guess at the k value, and $T' = 300(K)$. The probability values, $p(n_j|a_i, p)$'s that were obtained for positions $p = 1,2,3,4$ are shown in Tables 3.2, 3.3, 3.4 and 3.5. Under $k = 3e^{-4}$, the information content (IC) values are close to 2 in many rows. Considering a substantial range of diversity in typical binding sites of transcription factors, this would imply that the dominant residue frequency values, f^1's are rather too high at $k = 3e^{-4}$. So, the k-value was varied over $\{3e^{-4}, 5e^{-4}, 7e^{-4}, 1e^{-3}, 3e^{-3}, 5e^{-3}\}$ in order to search for more appropriate ones.

The statistics of the information contents, the frequencies of the most frequent residues, f^1's, and the second highest frequencies, f^2's, are shown in Tables 3.8, 3.9, 3.10 and 3.11. There is a clear trend of a decrease in the information contents, or equivalently, an increase in the residue diversity, with increasing k values (Table 3.12). We selected the values, $k = 3e^{-3}, 1e^{-3}, 7e^{-4}$, as approximately good ones that would produce appropriate information content values for most transcription factors.

Probability calculation at the overlapping position

DNA bases at positions 1 and 4 contact two adjacent zinc fingers; the sixth amino acid of the preceding finger (ZF1a6) contacts the fourth base, and the second residue of the following finger (ZF2a2) contacts its complement. We complemented the entries of Table 3.5 to reflect their complementary contact at the overlapping position. The energy contributions of the residues from respective fingers were assumed to be additive, i.e., letting $E_{1,2}$ denote the combined energy from contributions, E_1 and E_2, of the preceding and following fingers at the overlapping position, then $E_{1,2} = k_1E_1 + k_2E_2$. Here, k_1, k_2 are weighting factors that reflect their relative contributions, hence, $k_1 + k_2 = 1$. We used the values, $k_1 = 0.50$ and $k_2 = 0.50$, which corresponds to the case of equal energy contributions. Then, a table similar to Table 3.2 is obtained, consisting of 400 rows which correspond to all possible two-amino acid permutations. Then, we applied Eq.3.1 to obtain nucleotide probability values at the overlapping position.

Impacts of energy level offsetting and of kT product changes

Energy levels can be expressed in reference to an arbitrary level. Given a set of energy values, offsetting them by a constant does not change their relative magnitudes. Berg [111] suggested to offset the energy levels of nucleotides by the lowest one, which would then cause the new minimum value to be 0. We compared the effect of offsetting on the overall distributions of probability values against the one without offsetting (Fig.3.1). They showed very high correlation values across base positions, ranging from 0.9733 to 0.9952, which imply that offsetting would not significantly affect the probability values.

In order to assess the impact of the product kT, its value was set to 1.0, and the resulting probability matrices were compared with the matrices that were obtained by using the values $k = 1e^{-3}$ and $T = 300$ (Fig 3.2). While their correlation coefficients were high across different base positions, their relations were highly non-linear. Altogether, it was concluded that the relative magnitudes of probabilities would be preserved under energy level offsetting and kT product value changes. The impact of offsetting was marginal, and

the impact of using different parameter product values was more significant. We tentatively used k values in the vicinity of $1e^{-3}$ and applied offsetting to energy values.

Conversion of Kaplan05 propensity matrices

Kaplan05 tables were normalized in a manner similar to the Benos02 case. Four tables were obtained, which specify nucleotide probabilities, given an arbitrary amino acid residue. At the overlapping position, two amino acids generally specify different probability values for nucleotides. The arithmetic mean of the two probability values for each nucleotide was taken as the nucleotide probability corresponding to the amino acid pair, which ensures an appropriate normalization without further processing, due to the nature of the arithmetic mean.

3.4 Results

Putative finger sequences from 1574 zinc finger transcription factors of human were obtained from Pfam [90]. The number of fingers possessed by zinc finger transcription factors is widely varying. The distribution of the finger counts of human proteins is shown in Table 3.13. It has been suggested that a small number of fingers have a rather low specificity, since only a small number of DNA residues are specified [88]. Transcription factors with large finger counts frequently have the spacing between their constituent fingers exceeding 6 residues, which would imply that the zinc fingers form separate modules that recognize non-consecutive DNA residues. The fashion in which large finger count transcription factors bind DNA has been suggested to be more complicated than that of the smaller zinc finger count transcription factors [46, 95].

Given a zinc finger sequence set from a transcription factor, $\mathbb{ZF} = \{\mathrm{ZF_i}|\mathrm{ZF_1}, \mathrm{ZF_2}, \ldots, \mathrm{ZF_n}\}$, the characteristic residues that contact DNA bases were extracted. Each amino acid residue specifies a vector of nucleotide probability values. At the overlapping DNA position, the amino acid pair, ZF1a6 and ZF2a2, is used. Concatenating the vectors, which are specified by a single or pairs of residues, then forms a nucleotide residue probability matrix corresponding to the $\mathbb{ZF}$ set.

Application of the method to EGR3 transcription factor with the parameter $k = 3e^{-3}$ produces a pattern with the consensus sequence GCGTGGGCG (Fig.3.8(a), Table3.14), which is the same as one reported in the literature [48]. The first G residue shows the strongest conservation, and the two C's at positions 2 and 8, least conservations. The parameter k was varied to $1e^{-3}$ and $7e^{-4}$ to produce different patterns (Figs.3.8(b) and 3.8(c)). The f^1 residues are seen to be invariant across positions, irrespective of the k value change. The second C residue is very small at $k = 3e^{-3}$ but larger at $k = 7e^{-4}$. Selection of an appropriate k value would depend on the conservation levels of transcription factors under study, e.g., those of a specific transcription factor family from a species.

Since the same residue sets are used as indices to retrieve entries across the probability matrices, by using the entry-wise correspondence among nucleotides and amino acids, correlation coefficients between each pair of matrices can be calculated. The Kaplan05 matrices are highly similar across base positions, as evident from their high correlation coefficients (Table 3.7). To the contrary, such correlations are nearly absent among the Benos02 matrices (Table 3.6).

We then checked the consistency between the Benos02 and Kaplan05 matrices that correspond to the same positions (Figs.3.3 and 3.4), via correlation coefficients. The highest correlation coefficient value was 0.53402511 at the 2nd position, and the lowest was 0.05788589 at the 4th position (Figs.3.3(a) to 3.3(d)). The median was 0.29506800. While the median correlation value is rather significant against the null hypothesis of no correlation (P-value=0.015431), still, the level of overall agreement between the two sets of matrices is fairly low.

The matrices of binding patterns of zinc finger transcription factors were retrieved from Jaspar [136] (Fig.3.7). While the matrices could involve some inaccuracy, in the absence of more readily available alternatives, we used them as standards to compare the predicted patterns against. Patterns were predicted by using the Benos02 and Kaplan05 matrices, which are shown in Figs.3.5 and 3.6. The patterns that result from using the Kaplan05 matrices are seen to possess lower levels of conservation. They overall show moderate to

rather large differences from the Jaspar matrices.

3.5 Discussion

Prediction of transcription factor binding patterns is an important research problem in view of the importance of identifying genetic regulatory associations. Existing approaches commonly rely on the information about the interaction propensities between amino acid and nucleotide residues. Such data typically come from previously known and reported interactions.

While many transcription factors exhibit rather complicated ways of interactions with DNA, the zinc finger transcription factors, especially the C2H2 subfamily, have permitted a rather straightforward interaction model. By resorting to the model, the zinc finger-DNA interactions can be decomposed into constituent, elementary amino acid-nucleotide interactions. The quality and the reliability of predictions critically hinge on those of the propensity data used. If the data are biased or unreliable, then predictions based on the data will consequently be of low quality.

The EGR1 transcription factor (also known as zif268) reliably forms three active finger structures, and has served as a stable structural scaffold in many studies on the residue specificities of general zinc finger-DNA interactions. The resultant data as a whole then represent sequences similar to EGR1 well, while they rather poorly represent highly dissimilar sequences. Other members of the EGR transcription factor family, EGR2, EGR3 and EGR4 have the same amino acid residues as EGR1, at the characteristics positions that interact with DNA. Additionally, quite a number of points need to be addressed before or during an application of the plain zinc finger transcription factor-DNA interaction model. If the number of intervening residues between consecutive fingers exceeds 6, then the zinc fingers are possibly located in different modules [95]. Alternatively, zinc fingers may interact with RNA or protein [95, 46], rather than with DNA, hence, their interaction targets first have to be determined. When a zinc finger transcription factor interacts with DNA, and a large number of fingers are present therein, the modes of interactions of the zinc fingers are

Table 3.1: DNA and protein residue pairs forming contacts

DNA	Protein
3	-1
4	2
2	3
1	6

often very diverse and deviate from the canonical model. Within a zinc finger module, some amino acid residues at the characteristic positions may not participate in the interaction with DNA, and some others from non-canonical positions may. The relative contributions from residues at the overlapping position have to be precisely determined, while they were assumed to make equal contributions in the current study. A full computational model for zinc finger-DNA interactions is required to 1) predict the constituent fingers comprising a module 2) determine which residues are participating in the interactions and 3) determine the contributions of individual participating residues to the overall binding. The rather large discrepancies between the Jaspar and the Benos02-based and Kaplan05-based patterns, in a fairly large number of cases, appear to suggest the insufficient representativeness of the propensity data, or deviations from the canonical interaction model by the transcription factors that were studied. The performed comparison study serves to illustrate many issues to be addressed, in order to realize a fully computational prediction of binding patterns.

We employed the propensity matrices of [87] and explored the implication of a product of parameters that was pertinent in the process of converting the energies of states to their probability values. Varying its value resulted in patterns of different conservation levels. The overrepresentation of the EGR1-like sequences in publicly available data was manifest as a close agreement between the predicted and reported patterns for the transcription factor.

Table 3.2: Nucleotide Probability matrix at the 1st bp, finger position 6

aa\nuc.	A	C	G	T	IC
A	0.90222042	0.00000001	0.09777184	0.00000774	1.53796603
C	0.97855660	0.01284231	0.00823424	0.00036685	1.82750844
D	0.00000000	0.00007912	0.00000022	0.99992066	1.99880260
E	0.13137117	0.00000000	0.00000030	0.86862853	1.43880644
F	0.00000000	1.00000000	0.00000000	0.00000000	2.00000000
G	0.00000421	0.00000421	0.31479637	0.68519521	1.10120940
H	0.00000000	0.00000000	0.98960062	0.01039938	1.91657067
I	0.39068246	0.60931754	0.00000000	0.00000000	1.03476145
K	0.00000000	0.00000000	0.01612527	0.98387473	1.88090630
L	0.03444520	0.00000001	0.96555479	0.00000000	1.78378338
M	0.00000012	0.07204383	0.92781306	0.00014299	1.62447505
N	0.10802155	0.00000006	0.89197838	0.00000000	1.50607901
P	0.00000000	0.00000000	0.99723760	0.00276240	1.97254017
Q	0.99991065	0.00000000	0.00008842	0.00000093	1.99866190
R	0.00000000	0.00000000	1.00000000	0.00000000	2.00000000
S	0.05982310	0.00415672	0.49398394	0.44203625	0.70082355
T	0.99772000	0.00000020	0.00198030	0.00029950	1.97542109
V	0.00056812	0.00000004	0.97167544	0.02775640	1.81006495
W	0.00000000	0.00000000	0.00000000	1.00000000	2.00000000
Y	1.00000000	0.00000000	0.00000000	0.00000000	2.00000000

Table 3.3: Nucleotide Probability matrix at the 2nd bp, finger position 3

aa\nuc.	A	C	G	T	IC
A	0.00000001	0.00000000	0.00000000	0.99999999	1.99999978
C	0.00000000	0.00000000	0.00000000	1.00000000	1.99999999
D	0.00000000	1.00000000	0.00000000	0.00000000	2.00000000
E	0.00006334	0.99963655	0.00000003	0.00030008	1.99507998
F	0.99997914	0.00000000	0.00000000	0.00002086	1.99964559
G	0.00017220	0.00000000	0.00000000	0.99982780	1.99759842
H	0.00000002	0.00000000	0.99999998	0.00000000	1.99999943
I	0.00000000	0.00000031	0.00000000	0.99999969	1.99999294
K	0.00000000	0.00000000	0.95257413	0.04742587	1.72464005
L	0.00000000	0.00072965	0.00000000	0.99927035	1.99134445
M	0.00221318	0.00000000	0.00000000	0.99778682	1.97729113
N	1.00000000	0.00000000	0.00000000	0.00000000	2.00000000
P	0.00000000	0.00000000	0.00000394	0.99999606	1.99992359
Q	0.99885179	0.00000002	0.00001070	0.00113750	1.98704270
R	0.00000000	0.00000017	0.00931596	0.99068387	1.92377236
S	0.00000000	0.00000000	0.00000000	1.00000000	2.00000000
T	0.00000000	0.96555480	0.00000000	0.03444520	1.78378368
V	0.00000000	0.52774924	0.00000000	0.47225076	1.00222295
W	0.00000000	0.00000000	0.00000000	1.00000000	2.00000000
Y	0.00000000	0.00000000	1.00000000	0.00000000	2.00000000

Table 3.4: Nucleotide Probability matrix at the 3rd bp, finger position -1

aa\nuc.	A	C	G	T	IC
A	0.01612308	0.98374125	0.00013567	0.00000000	1.87898370
C	0.00037784	0.09773566	0.00000000	0.90188650	1.53343858
D	0.00000000	0.99999999	0.00000000	0.00000001	1.99999975
E	0.00000012	0.70865402	0.00000948	0.29133638	1.12938863
F	0.00000000	0.00000000	1.00000000	0.00000000	2.00000000
G	0.00000000	0.58226089	0.00053095	0.41720816	1.01374088
H	0.00000000	0.00024031	0.00000000	0.99975969	1.99676412
I	0.01282032	0.00918616	0.97688104	0.00111247	1.81338165
K	0.00000006	0.00014478	0.24762793	0.75222723	1.19050573
L	0.00000001	0.00072965	0.00000000	0.99927034	1.99134411
M	0.00000000	0.00000000	0.00000000	1.00000000	2.00000000
N	0.00000024	0.00000000	0.00000020	0.99999956	1.99998961
P	0.00000000	0.00000000	0.00000000	1.00000000	2.00000000
Q	1.00000000	0.00000000	0.00000000	0.00000000	1.99999999
R	0.00000000	0.00000000	1.00000000	0.00000000	1.99999995
S	0.00000001	0.00000053	0.00000202	0.99999743	1.99994657
T	0.00000002	0.00000000	0.00000000	0.99999998	1.99999947
V	0.00000000	0.00000000	0.00072965	0.99927035	1.99134445
W	1.00000000	0.00000000	0.00000000	0.00000000	2.00000000
Y	0.00000000	0.00000000	1.00000000	0.00000000	2.00000000

Table 3.5: Nucleotide Probability matrix at the 4th bp, finger position 2

aa\nuc.	A	C	G	T	IC
A	0.00000394	0.99999596	0.00000000	0.00000010	1.99992110
C	0.00000000	1.00000000	0.00000000	0.00000000	2.00000000
D	0.00000000	0.00005955	0.84108080	0.15885964	1.36751861
E	0.00000000	1.00000000	0.00000000	0.00000000	1.99999999
F	0.00000000	0.00000000	1.00000000	0.00000000	2.00000000
G	0.00000000	0.99998929	0.00000000	0.00001071	1.99980767
H	0.00000018	0.00015410	0.00000000	0.99984573	1.99782204
I	1.00000000	0.00000000	0.00000000	0.00000000	2.00000000
K	0.00000000	1.00000000	0.00000000	0.00000000	2.00000000
L	0.00000000	0.00000000	0.00000000	1.00000000	2.00000000
M	0.00000000	0.00000000	0.00000000	1.00000000	2.00000000
N	0.00000000	1.00000000	0.00000000	0.00000000	2.00000000
P	0.00000000	1.00000000	0.00000000	0.00000000	2.00000000
Q	0.00000005	0.99999994	0.00000000	0.00000002	1.99999839
R	0.00000000	1.00000000	0.00000000	0.00000000	2.00000000
S	0.00000000	0.99752737	0.00000000	0.00247262	1.97502476
T	0.00000034	0.99999966	0.00000000	0.00000000	1.99999216
V	0.00000000	0.98839266	0.00000002	0.01160732	1.90872982
W	0.00000000	0.10802154	0.00000020	0.89197826	1.50607606
Y	0.00000000	1.00000000	0.00000000	0.00000000	2.00000000

Table 3.6: Correlation coefficients of nucleotide probabilities between matrices of corresponding base positions (Benos02)

Pos	1	2	3	4
1	+1.00000000	+0.08668779	+0.02111910	-0.22734421
2	+0.08668779	+1.00000000	+0.36890810	-0.04949463
3	+0.02111910	+0.36890810	+1.00000000	+0.10822308
4	-0.22734421	-0.04949463	+0.10822308	+1.00000000

Table 3.7: Correlation coefficients of nucleotide probabilities between matrices of corresponding base positions (Kaplan05)

Pos	1	2	3	4
1	+1.00000000	+0.58178560	+0.58323553	+0.51080332
2	+0.58178560	+1.00000000	+0.83030163	+0.78016042
3	+0.58323553	+0.83030163	+1.00000000	+0.79996452
4	+0.51080332	+0.78016042	+0.79996452	+1.00000000

Table 3.8: Statistics of the IC, f^1 and f^2 values of 1st bp matrices as $f(k)$

k	0.000300	0.000500	0.000700	0.001000	0.003000	0.005000
Avg-IC	1.705419	1.523435	1.379125	1.215411	0.725604	0.498666
Med-IC	1.854207	1.538386	1.343155	1.089204	0.586920	0.368370
Stdev-IC	0.379770	0.430004	0.470111	0.517887	0.640184	0.516942
Avg-f^1	0.913159	0.860172	0.813310	0.756941	0.582173	0.507376
Med-f^1	0.981215	0.903376	0.829929	0.751475	0.519760	0.455017
Stdev-f^1	0.145726	0.161950	0.171839	0.182779	0.221780	0.212093
Avg-f^2	0.083160	0.128290	0.164842	0.203425	0.265843	0.269138
Med-f^2	0.014483	0.071691	0.135411	0.214136	0.310171	0.295074
Stdev-f^2	0.136688	0.144987	0.149257	0.151925	0.146935	0.128673

Table 3.9: Statistics of the IC, f^1 and f^2 values of 2nd bp matrices as $f(k)$

k	0.000300	0.000500	0.000700	0.001000	0.003000	0.005000
Avg-IC	1.919116	1.849680	1.769835	1.652969	1.122734	0.786046
Med-IC	1.999784	1.988140	1.938062	1.798150	1.122192	0.854499
Stdev-IC	0.228713	0.265627	0.303998	0.361082	0.491043	0.430636
Avg-f^1	0.971595	0.956201	0.936263	0.903920	0.737707	0.631588
Med-f^1	0.999987	0.998937	0.992692	0.968902	0.723038	0.622714
Stdev-f^1	0.105235	0.111007	0.118860	0.131545	0.177921	0.167690
Avg-f^2	0.028400	0.043590	0.062513	0.091328	0.210398	0.258962
Med-f^2	0.000012	0.001061	0.007285	0.030624	0.222065	0.268338
Stdev-f^2	0.105236	0.111069	0.119040	0.130625	0.148171	0.131009

Table 3.10: Statistics of the IC, f^1 and f^2 values of 3rd bp matrices as $f(k)$

k	0.000300	0.000500	0.000700	0.001000	0.003000	0.005000
Avg-IC	1.826941	1.739105	1.651512	1.528357	0.905378	0.581367
Med-IC	1.999968	1.994685	1.953842	1.778610	0.791565	0.451698
Stdev-IC	0.328815	0.396704	0.470512	0.548707	0.617685	0.521663
Avg-f^1	0.945197	0.918546	0.893937	0.860213	0.684045	0.574983
Med-f^1	0.999998	0.999611	0.995542	0.971768	0.662194	0.536812
Stdev-f^1	0.119459	0.144833	0.163961	0.186896	0.223519	0.207121
Avg-f^2	0.054227	0.076819	0.094210	0.114374	0.187846	0.219924
Med-f^2	0.000001	0.000245	0.002534	0.014527	0.151923	0.208588
Stdev-f^2	0.119521	0.142436	0.152932	0.158360	0.134352	0.111587

Table 3.11: Statistics of the IC, f^1 and f^2 values of 4th bp matrices as $f(k)$

k	0.000300	0.000500	0.000700	0.001000	0.003000	0.005000
Avg-IC	1.937744	1.889097	1.843447	1.767169	1.167887	0.759916
Med-IC	2.000000	1.999975	1.999107	1.987855	1.342409	0.697607
Stdev-IC	0.174029	0.254491	0.303724	0.360790	0.547602	0.475335
Avg-f^1	0.985940	0.970537	0.956672	0.936251	0.783512	0.658848
Med-f^1	1.000000	0.999998	0.999942	0.998926	0.881065	0.705544
Stdev-f^1	0.041745	0.075815	0.097719	0.120289	0.185243	0.183328
Avg-f^2	0.014056	0.029330	0.042536	0.060036	0.147362	0.201326
Med-f^2	0.000000	0.000001	0.000057	0.001070	0.091752	0.193474
Stdev-f^2	0.041734	0.075413	0.095883	0.114430	0.127646	0.099203

Table 3.12: Average IC values of matrices at different base positions as functions of k

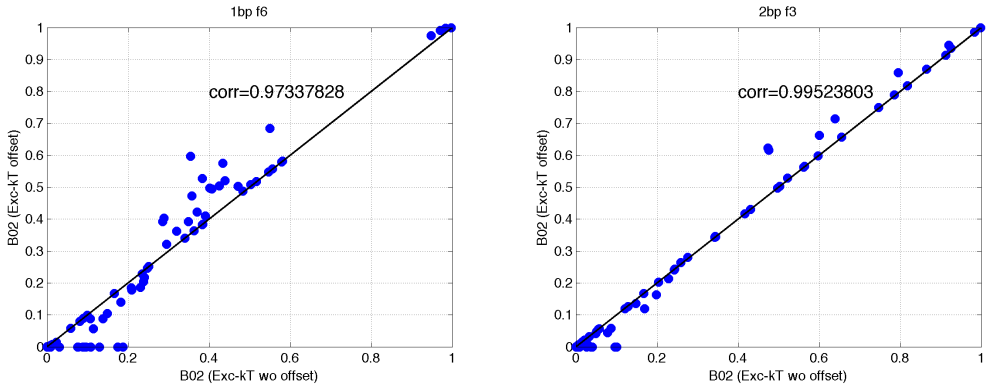
-logk	k	k	bp-1	bp-2	bp-3	bp-4	Max-Min
3.523	3E-4	0.0003	1.7054	1.9191	1.8269	1.9377	0.2323
3.301	5E-4	0.0005	1.5234	1.8497	1.7391	1.8891	0.3657
3.155	7E-4	0.0007	1.3791	1.7698	1.6515	1.8434	0.4643
3.000	1E-3	0.0010	1.2154	1.6530	1.5284	1.7672	0.5518
2.523	3E-3	0.0030	0.7256	1.1227	0.9054	1.1679	0.4423
2.301	5E-3	0.0050	0.4987	0.7860	0.5814	0.7599	0.2874

Table 3.13: Number of zinc fingers in human proteins

N-ZFs	Proteins
1	142
2	109
3	179
4	119
5	103
6	89
7	84
8	88
9	119
10	63
11	99
12	81
13	66
14	67
15	36
16	27
17	27
18	20
19	22
20	5
21	15
22	4
23	2
24	3
25	0
26	2
27	0
28	0
29	0
30	1
31	0
32	1
33	1
34	0

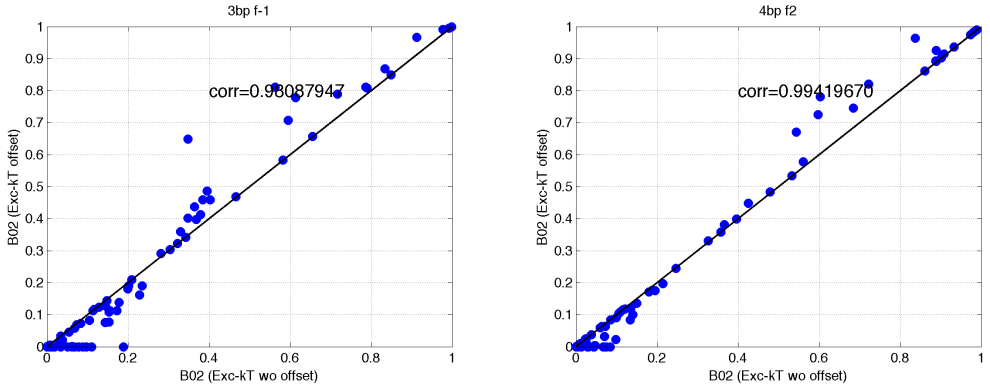
Table 3.14: IC of the calculated EGR3 matrix ($k = 3e^{-03}$)

Position	f^1 residue	f^1	IC
0	g	0.965695	1.733784
1	c	0.499156	0.251648
2	g	0.826380	1.144139
3	t	0.486353	0.730754
4	g	0.844987	1.321063
5	g	0.826380	1.144139
6	g	0.780029	1.084716
7	c	0.499156	0.251648
8	g	0.826380	1.144139
9	g	0.438536	0.395198
Sum	-	6.993053	9.201229



(a) 1st bp

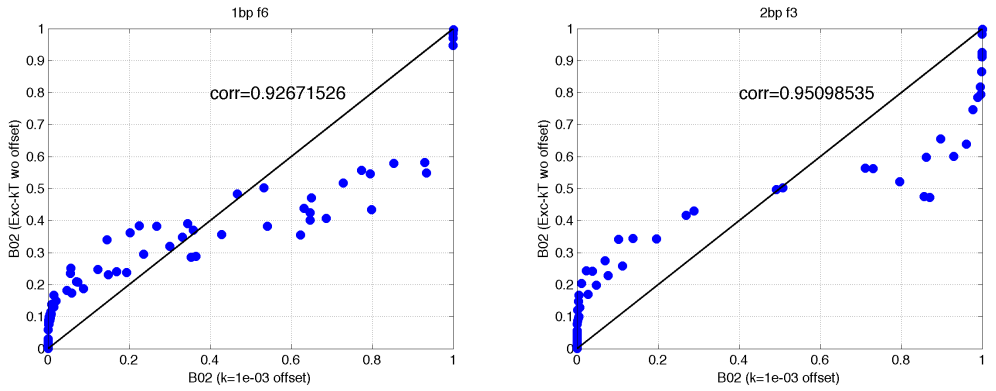
(b) 2nd bp



(c) 3rd bp

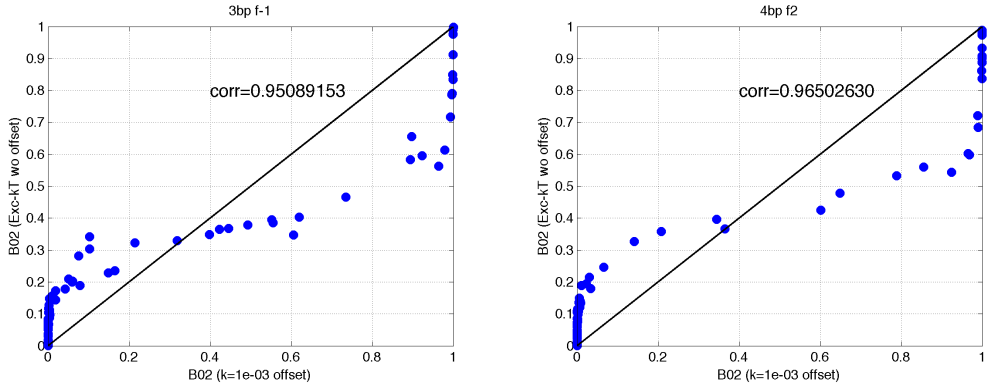
(d) 4th bp

Figure 3.1: Scatter plots of Benos02 matrix entries without energy level offsetting vs. Benos02 with offsetting. Overall correlations are very high.



(a) 1st bp

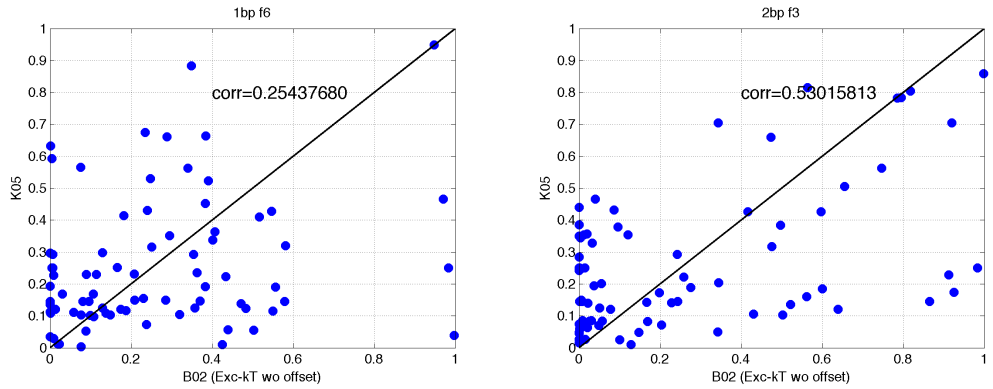
(b) 2nd bp



(c) 3rd bp

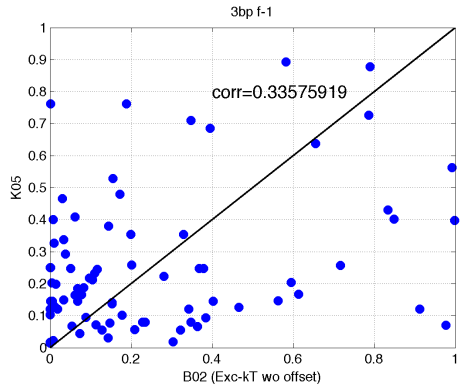
(d) 4th bp

Figure 3.2: Scatter plots of Benos02 matrix entries with $kT' = 1$ vs. Benos02 with $k = 1e^{-3}$ (all with offsetting). Correlations are high, yet the relations are very nonlinear.

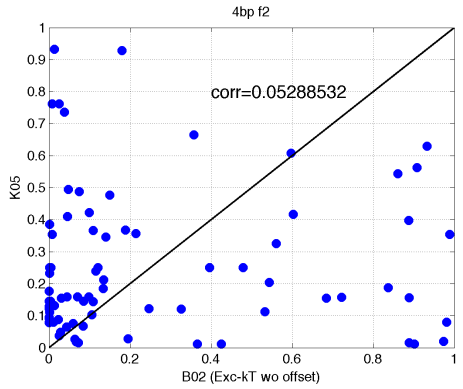


(a) 1st bp

(b) 2nd bp

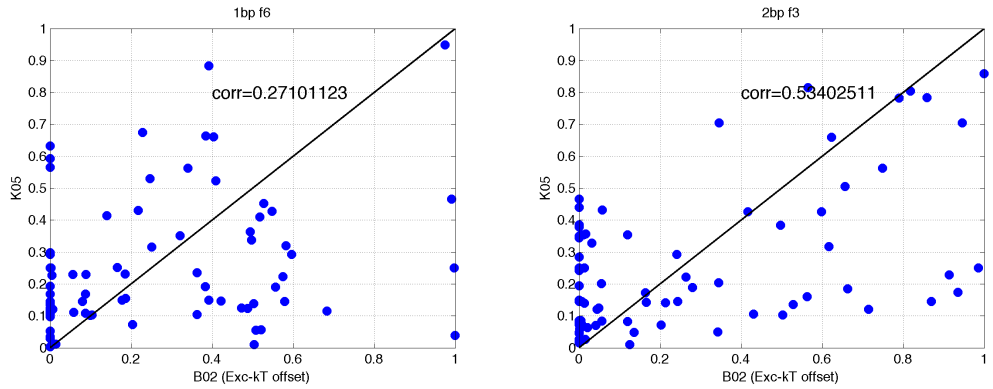


(c) 3rd bp



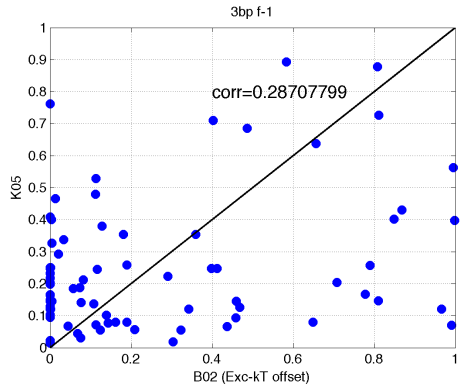
(d) 4th bp

Figure 3.3: Scatter plots of Benos02 matrix entries with $kT' = 1$ and without offsetting vs. Kaplan05. Correlations between matrices from the two sets are rather low.

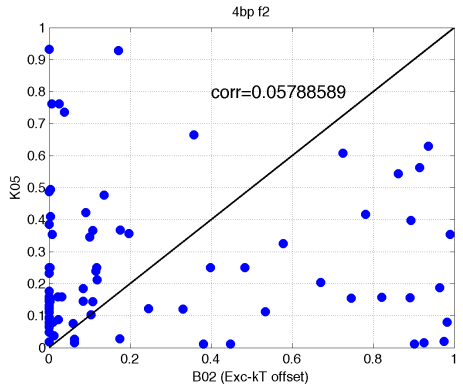


(a) 1st bp

(b) 2nd bp



(c) 3rd bp



(d) 4th bp

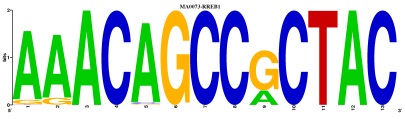
Figure 3.4: Scatter plots of Benos02 matrix entries with $kT' = 1$ and with offsetting vs. Kaplan05. Correlations are low, especially at the 4th position.

(a) EVI1

(b) GFI

(c) MZF1-1-4

(d) MZF1-5-13



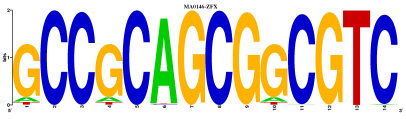
(e) RREB1

(f) SP1

(g) YY1

(h) ZNF354C

(i) REST



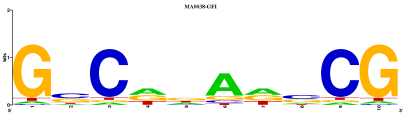
(j) ZFX

(k) INSM1

(l) EGR1

Figure 3.5: Patterns calculated by using Benos02 matrices. The predicted patterns tend to have strong levels of conservations.

(a) EVI1



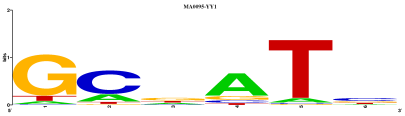
(b) GFI

(c) MZF1-1-4

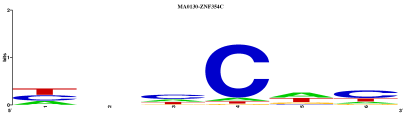
(d) MZF1-5-13

(e) RREB1

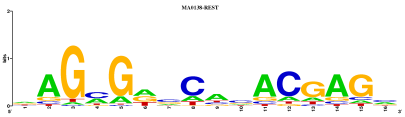
(f) SP1



(g) YY1



(h) ZNF354C



(i) REST

(j) ZFX

(k) INSM1

(l) EGR1

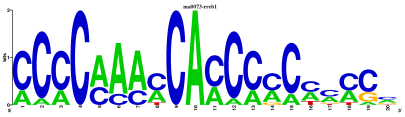
Figure 3.6: Patterns calculated by using Kaplan05 matrices. The predicted patterns have low levels of conservations.

(a) EVI1

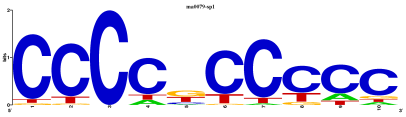
(b) GFI

(c) MZF1-1-4

(d) MZF1-5-13



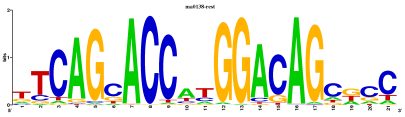
(e) RREB1



(f) SP1

(g) YY1

(h) ZNF354C



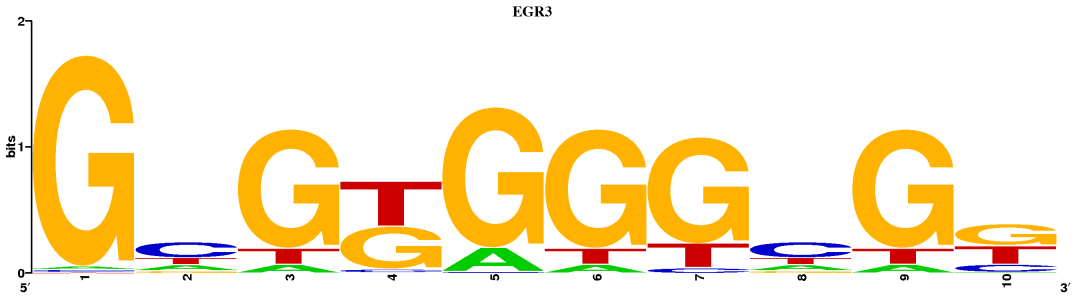
(i) REST

(j) ZFX

(k) INSM1

(l) EGR1

Figure 3.7: Patterns from Jaspar. They were used as reference patterns to compare predicted patterns with.



(a) $k = 3e^{-3}$

(b) $k = 1e^{-3}$

(c) $k = 7e^{-4}$

Figure 3.8: Changes in the calculated EGR3 binding patterns with varying k values. Decreasing the k value results in patterns with higher information contents.

Chapter 4. IMPROVING BINDING PATTERNS

4.1 Abstract

Transcription factors recognize particular DNA sequence patterns and regulate the expression of downstream target genes. ChIP-seq and ChIP-chip data provide information about the loci of bound sequence patterns to a resolution of tens to hundreds of bases. Existing approaches to finding transcription factor-binding patterns from ChIP-seq data have focused mostly on the de novo discovery of patterns, without addressing systematic ways to properly utilize a priori knowledge on the binding patterns. Many existing matrices that represent binding patterns are based on small numbers of sequences, or result from the computationally challenging problem of de novo pattern discovery. Hence, they often exhibit high degrees of deviation from true distributions of bound sequence patterns. Accurate representations of the sequence pattern distributions through matrices are important in understanding the relative contributions of residues to binding energy, and in estimating the significance of putative pattern matches in the course of establishing target genes of transcription factors.

We studied the problem of closely approximating truly representative matrices by utilizing ChIP-seq data, together with matrices that are presumably inaccurate. An iterative approach in combination with a proper parameter selection facilitates to reveal the enrichment of true patterns in the sequence data. Applications to synthetic data demonstrate that widely varying initial matrices stably converge to accurate matrices by using our approach.

4.2 Introduction

Transcription factors regulate the expression of their target genes by binding to specific patterns in the upstream regions of the targets. Their bindings are typically characterized by short DNA sequence patterns, ranging 5-20 base pairs (bp's) in higher organisms. Frequently, different nucleotide residues at a position show nearly identical binding energies, or residues over multiple positions are substitutable without causing significant changes

in binding energies: this has been often referred to as the degeneracy of patterns. Relatively recent experimental techniques such as ChIP-seq [137] can provide important information on the in vivo binding of transcription factors to their cognate patterns. Such patterns have typically been represented via letter probability matrices [150] which specify the expected frequencies of residues at each residue position. Additionally, position-specific scoring matrices can be used, which specify the odds of pattern conformance against random patterns originating from a background nucleotide distribution. While the problem of de novo pattern discovery from ChIP-seq data has been studied by many [114, 115, 116, 124, 127, 129, 133, 130], little attention has been paid to the problem of improving inaccurate patterns. Traditionally available knowledge about patterns has been often based on small scale experiments that involved only small numbers of transcription factor-bound patterns. Such small sample sizes severely weaken the statistical representativeness of the patterns.

ChIP-seq [137] data produces a large number of transcription factor-bound patterns, albeit with a substantial tolerance in locus. Candidate bound regions are 10's to 100 bp long sequence fragments that harbor the patterns bound by transcription factors. Finding the bound patterns can be challenging, given the pattern degeneracy, large sequence count and the substantial sequence lengths.

Most existing algorithms for ChIP-seq data focus on finding new patterns from the data, typically relying on the inherent enrichment of the bound patterns therein, and do not consider ways of systematically utilizing known patterns. Frequently, genes are under a combinatorial regulation. It is especially prevalent in higher organisms, where multiple transcription factors cooperatively regulate the target genes by binding to their respective patterns that are located in close proximity to each other. Multiple patterns jointly enriched in the data can mislead algorithms that mainly rely on pattern enrichments. Low complexity regions in DNA were often implicated to be germane to the births of transcription factor binding patterns [153]. The low complexity regions present another source of noise to the enrichment-driven operation of the algorithms. Together with the computational complex-

ity of the problem, these confounding factors pose a serious challenge to the algorithms, and the resulting algorithm performance can be highly unstable; inferred patterns can be related to ones that are different from the true bound patterns that were sought after. Additionally, they can be highly inaccurate due to the mixed enrichment of heterogeneous patterns. Hence, utilization of prior pattern knowledge, albeit weak and inaccurate, has the potential to produce more specific and accurate patterns corresponding to the transcription factors under study. Using letter probability matrices [150] is still a popular choice for representing binding patterns. The matrices specify frequency vector of nucleotides for each residue position. Those vectors together specify a probability distribution of patterns, rather than a small set of specific patterns.

Given a matrix, methods for the statistical assessments of the conformance of candidate patterns to the pattern represented by the matrix were studied in [118, 121]. In the interactions between a transcription factor and the promoter regions of its target genes, the presence and the conformance levels of sequence patterns to the transcription factor-characteristic binding pattern were shown to be the most critical determinants [122]. Other factors, such as the nucleosome occupancy and chromatin modification signals were far less significant [122].

It is known that binding patterns frequently are highly degenerate. Patterns that are distinct from the one having the highest affinity, can still show moderate to high affinities to bind to the transcription factors and play biological roles [147, 148, 135]. The specificity levels of positions are typically different: the most dominant residue at a position can be strongly conserved, and nearly no other residues may be found in transcription factor-bound patterns. Alternatively, some positions bind less strongly to transcription factors, hence more diverse residues are permitted. Accurate representations of precise residue probabilities would be critical to subsequent inferences that employ the binding pattern information. Identifying the relative specificities of positions facilitates a more precise estimation of the significance of pattern matches that are obtained from searches for patterns in large sequence sets. Hence, it is important to have highly reliable probability matrices (PM) that

are representative of true pattern distributions.

Here we propose a systematic way of utilizing ChIP-seq data together with known patterns that are presumed to be inaccurate, to approximate the true binding pattern distributions as closely as possible. The proposed approach is iterative, and each step consists of scanning sequence fragments from ChIP-seq data. The patterns with P-values more significant than a threshold are retained, which then serve to form a new matrix for the next iteration. Globally, two phases of p-value series are used: 1) a stringency phase where the P-value is gradually decreased in order to successively approximate the true matrix, starting from an initial matrix, and 2) a sensitivity phase where the P-value is progressively relaxed, so as to gain sensitivity and to reach the residue diversity that is characteristic to each of the transcription factors under study. Case studies on a set of transcription factor binding patterns demonstrate that our method, combined with sensible P-value scheduling, facilitates the discovery of true patterns.

4.3 Approach

Assume a set of N sequence patterns $S = \{s_j | j = 1 \ldots N\}$ is given which putatively bind to a transcription factor, all of length L. They can be aligned without permitting gaps and the frequencies of residues can be calculated at each position. Then each aligned position will have a vector of residue frequency values. Denoting by α a letter from a possible alphabet set, each residue frequency at a position i is bounded, i.e., $0 \leq f_i(\alpha) \leq 1$ which sum to unity since they are frequency values: $\sum^{\forall \alpha} f_i(\alpha) = 1$. A matrix can be used to specify the frequency values of residues over multiple positions. Map each position in the alignment to a row of the matrix, and each alphabet to a column of the matrix. Henceforth, we focus our attention on DNA sequences, then the alphabet set $\{\alpha\} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. If a position has nearly identical residue frequency values, then the residue diversity, and consequently entropy are high, and its information content is low. Specifically, the entropy $H_i(\alpha)$ at a position i is defined as

$$H_i(\alpha) = \sum^{\forall \alpha} f_i(\alpha) \log(f_i(\alpha)) \tag{4.1}$$

and the entropy of the entire matrix is defined as:

$$H(\alpha) = \sum_i H_i(\alpha) = \sum_i^L \sum^{\forall\alpha} f_i(\alpha)\log(f_i(\alpha)) \tag{4.2}$$

The information content at a position i is defined as:

$$IC_i = \log|\{\alpha\}| - H_i(\alpha) \tag{4.3}$$

and that of the entire matrix is defined as:

$$IC = \sum_{i=1}^{L}[\log|\{\alpha\}| - H_i(\alpha)] \tag{4.4}$$

with $|\{\alpha\}|$ the size of alphabet set, and is 4 in the case of DNA. The unit of bits can be used by employing log base of 2. A highly divergent position will have a high entropy and low IC, and a highly conserved position will have a low entropy and high IC. In the case of perfect conservation, one of the nucleotides will have $f(\alpha) = 1.0$ for some α, and then $IC = 2(bits)$. If all residues are equally likely, then $f(\mathrm{A}) = f(\mathrm{C}) = f(\mathrm{G}) = f(\mathrm{T}) = 0.25$, and $IC = 0(bits)$.

The letter probability matrix, M, can be interpreted to specify a zero-order Markov chain, i.e., each row of the matrix is a state specifying the emission probability values of nucleotides at the position. Then it can serve as a generative model, where the states are traversed from the first to the last, to generate a sequence. Probabilities of nucleotides follow the frequency vector specified at each position. By iterating the process k times, a total of k sequences can be generated, distribution of which follows M (henceforth referred to as M-sequences or $S(\sim M)$). The k sequences generated in turn can form another matrix M'. Denote the formation of matrix M from a sequence set S by '$S \rightarrow M$', and the generation of a set, S_x, of k sequences from matrix M by '$M \rightarrow_k S_x$'. While an aligned sequence set S uniquely specifies a matrix, the matrices, $M_a(\sim S_a), M_b(\sim S_b), \ldots$, formed from sequence sets, $S_a(\sim M), S_b(\sim M), \ldots, \exists M^*$, will in general not be the same, $M_a \neq M_b \neq \ldots$, due to the inherent stochasticity in the sequence generation following the Markov process. Matrix identity between two matrices M and M' can be defined in terms of the element-wise identity, where all elements are required to be identical: $M = M'$ if and only if $e_{i,j} = e'_{i,j}$ with

$e_{i,j}$ the ith row, jth column element of M, and $e'_{i,j}$ similarly defined. It would be convenient to introduce a measure of difference between two matrices. Simply the absolute difference between each corresponding element of the two matrices, M and M', can be taken and summed for the overall difference between them:

$$\Delta(M,M') = \sum_i \sum_j |e_{i,j} - e'_{i,j}| \tag{4.5}$$

Each position of a matrix specifies a normalized frequency distribution of residues independently of others, which would form a vector of dimension 4 in the case of DNA. Then a matrix is interpreted to be a sequence of vectors. Then a 2-norm vector difference can be applied to each position, and then the differences summed over all positions:

$$\Delta(M,M') = \sum_i \sqrt{\sum_j (e_{i,j} - e'_{i,j})^2} \tag{4.6}$$

They will be respectively termed the 1- and 2-norm differences of matrices: $|M - M'|_1$ and $|M - M'|_2$. Alternatively, the IC difference between two matrices can be calculated. Contrary to the element-wise difference measures, IC is a scalar quantity calculated on individual matrices and would not reflect the vector-wise differences between the matrices. For example, if residue A is of probability 1.0 at the first position of M_1, and T is of probability 1.0 at the same position in M_2, then the two have same IC values while having distinct frequency vectors.

Let M_1 and M_2 denote two matrices generated by first using $M^* \rightarrow_k S_1$, $M^* \rightarrow_k S_2$, two sets each of k sequences generated from M^*, and then using $S_1 \rightarrow M_1$, $S_2 \rightarrow M_2$. While $M_1 \neq M_2$ in general, the matrices M_i's are expected to approach M^* for sufficiently large values of k, i.e., $M_i (\sim S_i)$ will be asymptotically the same as M^*:

$$\lim_{k \rightarrow \infty} |M - M^*| \approx 0 \tag{4.7}$$

$$\lim_{k \rightarrow \infty} M \rightarrow M^* \tag{4.8}$$

Given a sequence pattern of length L, its conformance to a matrix M can be assessed by using likelihood. Assuming the sequence was generated following M, respective prob-

abilities of constituent residues can be multiplied together (assuming position independence) to form a product of likelihoods, l_M: $l_M = p_1(\alpha_1)p_2(\alpha_2)\dots p(\alpha_L) = \Pi_i p_i(\alpha_i)$, given α_i at position i, $1 \leq i \leq L$. An alternative model to explain the sequence is obtained by using random background distribution of nucleotide frequencies. If the nucleotides have frequency values $f(\alpha) = \pi_B(\alpha)$ in a genome under study, then the likelihood that the sequence pattern was generated according to the random background distribution is $l_B = \pi_B(\alpha_1)\pi_B(\alpha_2)\dots\pi_B(\alpha_L)$. The log-odd value can be calculated by taking the ratio of the likelihood values, and then its log: log–odds $= \log(l_M/l_B)$.

Given a sequence pattern and π_B, distinct matrices will clearly assign different log-odd values to the sequence. If a normalized statistic is desired such as P-value, which would facilitate comparison of significance of pattern matches on a uniform scale across different matrices, then the log-odd values have to be converted to P-values. Exhaustive enumerative generation of all possible L-mers, calculation of their log-odds and then mapping to p-values is one way to do it. Its computational complexity is exponential (4^L) which would be prohibitively computationally expensive for large values of L. Efficient algorithms based on function generation were studied in [118] and [121]. We used the FIMO implementation from meme tool suite [109] for the calculation of P-values of pattern conformance.

Given a sequence set S and a matrix M, the p-values representing the significance of conformance to M of sequences therein will form a distribution, $D = D_P(S,M)$. If $M \rightarrow S$, then being generated from M as a template, the sequences in $S(\sim M)$ are expected to tend to have significant P-values when they are scored against M. It is the probabilistic nature of the M-based sequence generation, $M \rightarrow S$ that will inherently produce some insignificant p-values which can be best comprehended by using the distribution of p-values. If the set of entire sequence patterns S^* that are targeted by a transcription factor (TF) is given, then a corresponding matrix M^* can be formed, $S^* \rightarrow M^*$. Assume another matrix, M_i is given with a substantial difference from M^*, $\delta \leq |M_i - M^*|$, for some constant $\delta > 0$. The matrix M_i is presumably formed from a small number of M^*-sequences mixed with some random sequence patterns of length L. Now using M_i as a generative model, generate a

sequence set S using M_i, calculate their p-values against M_i and form a distribution of p-values, $D_P(S,M_i)$. If the M^*-sequences are scored against M_i, then they are likely to form a distribution $D_p(S(\sim M^*),M_i)$ distinct from $D_p(S(\sim M_i),M_i)$. Now assume there are a total of N^* of M^*-sequences, N_i of M_i sequences, and $N^* \gg N_i$. Form a matrix M_{i+1} from a concatenation of $S(M^*)$ and $S(M_i)$: $S(M^*) \cup S(M_i) \rightarrow M_{i+1}$. Then intuitively the $S(M_{i+1})$ will form a distribution located closer to $D(M^*)$ as scored against M_i. By iterating the above steps, each successive M_i is expected to be closer to M^* than previous M_{i-1}'s:

$$|M_i - M^*| > |M_{i+1} - M^*| > \ldots > |M_\infty - M^*| \approx 0 \tag{4.9}$$

ChIP-seq data are inherently enriched in M^*-sequences since they are obtained from the in vivo binding of TFs to their target patterns. A sequence of length L'' ($L'' > L$) has $N' = L'' - L + 1$ of L-mers. Then a ChIP-seq sequence set, C, of N sequences each of length L' is a set of $N(L' - L + 1)$ of L-mers. Given a matrix M_i, the p-values of the L-mers from C are calculated and a threshold p-value θ_p is applied. Only L-mers with p-values $p < \theta_p$ can be retained, and the L-mers with higher p-values are discarded. This step will be iteratively applied to the set C with an appropriate sequence of threshold p-values. The skeleton of the method is shown in Algorithm listing 2.

Algorithm 2 Binding pattern improvement algorithm

```
while true do
  for all sequences s_j do
    scan s_j with M_i and retain patterns with p-values < θ_p
  end for
  form M_{i+1} with the retained patterns
  if an appropriate condition then
    break;
  end if
  i = i + 1;
end while
```

Given NL number of random sequence patterns, scanning them with a matrix with a threshold θ_p will produce a total of $NL\theta_p$ matches on average. A very lenient p-value will produce an excessive number of matches, which randomly originate from background and

do not reflect the presence of true binding patterns. Using a stricter p-value will produce less number of random matches. On the other hand, applying a too strict p-value to M^*-sequences will likely even discard a substantial sub-population of the M^*-sequences of lower conformance levels, and iterating the $S \rightarrow M$ steps will no longer approximate M^* due to such loss. Hence it is suggested to start the process with an appropriate p-value which finds a balance between the rejection of random noise and the retention of M^*-sequences. The relative proximity of M_{i+1} to M^* over M_i suggests to use a sequence of successively stricter p-values. Once an M_i obtained during iteration is sufficiently close to M^* such that it approximates M^* well while rejecting the random noisy matches, then the p-value is successively relaxed so that the search is sufficiently sensitive to the inherent diversity of M^*-sequences.

We used ChIP-seq data from [107] which studied the TFs involved in the embryonic stem cell development in mouse. Three TFs were chosen, binding patterns of which were suspected to have a reverse complementary symmetry where one side is reverse complement of the other. Through iterations over data, reverse complementarity is expected to be restored . The binding patterns were assumed to mostly reside in the promoter regions of the genome. Loci of the 3' boundary of promoter regions that are farthest from the transcription start sites of genes were obtained from [123]. The 5' upstream sequences 10kbp long or spanning up to the boundary of the upstream gene were extracted, and the shorter of the two was retained as the promoter sequence for each gene. The frequency values of the mono-nucleotides of promoter sequences were as follows: A-0.256629, C-0.243552, G-0.243132 and T-0.256288.

4.4 Results

We used various patterns, synthetic and real, to assess the performance of our approach.

Synthetic pattern

The synthetically generated pattern is shown in Fig.4.5). Specifically, the consensus of the first half of the pattern was AGTC, and that of the latter half was GACT, which was a re-

verse complement of the first. Hence, the pattern has a reverse complementary symmetry. The most frequent residue value (henceforth referred to as f^1, and the second most frequent f^2, and so on) was 0.90, and values from a geometric progression were assigned as frequencies to remaining residues: i.e., $G = \{g_1, g_2, g_3, g_4\} = \{0.9000, 0.0900, 0.0090, 0.0009\} = \{f^1, f^2, f^3, f^4\}$. Residues were taken in a circular manner, i.e., if value f^1 was assigned to G then the value f^2 was assigned to T, f^3 to A, and so on. A gap of length 1 was placed in the middle where residues had equal frequency values of 0.250. The information content (IC) of each conserved position was 1.480001, and the overall IC of the pattern (to be referred to as the pattern P_s) was 11.840011. Scaling factors were randomly chosen among $S_1 = \{s_j | 1 \leq j \leq 2\} = \{0.70, 0.80\}$, and the most frequent residue of each position was scaled down by multiplying by the chosen factor: $f^{1'} = s_j f^1$. Resulting distortion could be similar to the ones typically seen in disagreeing matrices of a given transcription factor (TF) found across the literature. The frequency values of the remaining residues were increased by equal portions from the difference, $f^1 - f^{1'}$, so that they sum to one, i.e., $f^{i'} = f^i + \delta/3$ with $\delta = f^1 - f^{1'}$ and $2 \leq i \leq 4$. The resulting frequency and the IC values after the perturbation were as shown in Table 4.1 and Fig.4.2(b) (referred to as P'_{S_1}). Some f^1 values decreased down to 0.630000, and correspondingly IC to 0.489850, which is a large decrease from the *IC* of 1.480001. Note the overall IC decreased to 4.411192 which is a drastic decrease from the initial IC of 11.840011. The extreme loss of the IC would make searches for patterns based on the perturbed one hard, and hence the case is quite challenging to discover the M^*-sequences from.

In order to mimic ChIP-seq data, $N = 1000$ random sequences of length 141 were generated following the background mono- and di-nucleotide frequency distributions of the mouse promoter sequences. By using the matrix M^* corresponding to P_s as a generative model, a total of $N = 1000$ patterns each of length 9 were generated, and inserted into random positions. Then each of the resulting sequences was of length 150bp ($= 141 + 9$). The perturbed matrix was used as an initial matrix, M_0 for iterative improvement. The series

of threshold P-values, Φ^1, that was employed was

$$\Phi^1 = \Phi_d + \Phi_i \tag{4.10}$$

$$\Phi_d = \{1.0e^{-3}, 7.5e^{-4}, 5.0e^{-4}, 2.5e^{-4}, 1.0e^{-4}\} \tag{4.11}$$

$$\Phi_i = \{2.5e^{-4}, 5.0e^{-4}, 7.5e^{-4}, 1.0e^{-3}\} \tag{4.12}$$

which consists of a decrease and then an increase phase: $\Phi^1 = \Phi_d + \Phi_i$. Figures 4.2(c) to 4.2(k) show the change in the IC values of patterns formed along the iteration, and Table 4.2, the IC of each iteration step.

If the f^1 residues are the same between corresponding positions of two matrices, then the standard deviation of the f^1 values can be used to measure their conformance, i.e., f^1 values are taken over all positions excluding gap. Standard deviation (stdev) of the initial matrix M_0 was 0.04657943, and stdev of the M_1 matrix obtained after the first iteration was 0.01465860 (Table 4.3). Nearly constant f^1 and IC values across positions (excluding the gap at position 4, Table 4.5) means that the pattern from the retained sequences with p-value thresholding quickly converged to M^* after the first iteration.

Given a matrix which is supposed to possess a reverse complementary symmetry, we measure the degree of its deviation from such symmetry, i.e., how asymmetric frequency vector sequence of one side is with respect to the other after reverse complementing it: under 1-norm, $AS_1 = \sum_{i=1}^{L/2} |e_i - e_{L-i}|_1$, and under 2-norm, $AS_2 = \sum_{i=1}^{L/2} |e_i - e_{L-i}|_2$. Large reduction in the asymmetry in M_0 is seen in matrices obtained from the iteration. At the 1st iteration with p-value $p = 1.0e^{-3}$, the IC value, 11.859641 is close to the $IC = 11.840011$ of M^*. This suggests that the p-value may be interpreted to correspond to the inherent specificity level of the matrix which, when searching for bound patterns, would enable close approximation to M^*, given the mixture of random and the M^*- sequences. Similar test was performed a total of 10 times. The results with each perturbed matrix instance as an initial matrix were overall similar to the above.

Another scaling factor set, $S_2 = \{s_j | 1 \leq j \leq 3\} = \{0.60, 0.70, 0.80\}$ was used to obtain P'_{S_2} by perturbing P_s, and then patterns were iteratively obtained (Figs.4.3(b) to 4.3(k)).

A pattern sufficiently close to M^* was obtained through iteration, while the early patterns are somewhat more deviant from M^* than the corresponding ones from the iteration that resulted from the perturbation with S_1 (tables 4.6 and 4.7). Results from a total of 10 random initial matrix instances were overall similar.

Swapping residue frequency

In another test, we randomly selected $k = 2$ positions and applied random derangement to their chosen frequency vectors. By derangement of a set where no single element is same as the other, frequency of each residue is assigned a value different from the previous one, i.e., $f'(\alpha) \neq f(\alpha)$, for all α, with $f(\alpha)$ and $f'(\alpha)$ respectively denoting the frequencies of α before and after derangement. Figure 4.4(b) shows an instance of the derangement, $P'_{k=2}$, where the first and the last positions were selected (A to C and T to A) from P_s. After the first iteration, each of the most frequent residues after derangement, $f^{1'}$, is still dominant over the f^1 residue at the deranged position. At the second iteration with $p = 7.5e^{-4}$, the f^1 residues are restored, albeit with significantly lower IC values than those of M^*. By the iteration with $p = 5.0e^{-4}$, IC heights at deranged positions are restored to around 1.5 bits, which are approximately the same as the per-position IC of M^* (Table 4.8). A total of 10 such instances of perturbed matrices were generated, and iterations were performed with respective matrices as initial ones. Results were nearly constant in that patterns all converged to M^*, regardless of particular positions deranged.

Additionally, $k = 1$ position was randomly selected and deranged. A pattern nearly the same as M^* was attained by iteration (Fig.4.5(e)). When $k = 3$ residues were selected and deranged, patterns did not converge to M^*, which would suggest that the initial matrix is too distant from M^* and would instead converge to a different pattern (Figs. 4.6(b) to 4.6(k)).

Real ChIP-seq data

We applied our method to the data set from [107]. Four TF's (Oct4, Sox2, Nanog, Smad1) had highly homologous binding patterns according to [107] which would imply rather promiscuous bindings and make the determination of patterns specific to each TF hard and

were excluded. Zinc finger TF's of large finger counts are known to have highly promiscuous binding patterns [95], then the representativeness of the data set can be limited. Two such TF's, Ctcf and Zfx, were excluded.

Binding patterns of 3 TF's were investigated first: Nmyc, Stat3 and Tcfcp2l1 (hereafter abbreviated to Tcfcp). Matrices from Jaspar [136] were retrieved to be used as initial matrices which are shown in Figs.4.7(a),4.7(b),4.7(c). They were suspected of having reverse complementary symmetry. IC of the matrices were: Nmyc - 11.104085, Stat3 - 13.600680 and Tcfcp - 11.650291 (bits). We padded extra positions at the beginnings and the ends of the matrices with equal residue frequency values so as to check for any conservations at the fringes. Top $N = 1000$ sequences of the highest intensities were used, and the value of pseudo-count parameter of FIMO was set to 2.0.

While the initial Stat3 TF pattern is highly asymmetric (asymmetry coefficient, $AS_1 = 1.83523400$ and $AS_2 = 1.10530225$), the asymmetry decreases by a large margin down to $AS_1 = 0.25608466$, or to $AS_2 = 0.16666693$ by iteration 4 (Fig.4.8(e), Table 4.9). At each iteration step, it was checked how many sequences out of $N = 1000$ were found to have matching patterns with p-values $< \theta_p$ of the step. The counts were compared with the ones obtained by using the initial Jaspar matrices, M_0's, at the same p-values without iteration (Figs.4.13(a),4.11(a),4.12(a), Table 4.17). Matrices from iteration, M_i's, $1 \leq i \leq 9$, consistently produced larger number of sequences than raw Jaspar matrices (signed-rank test significance p-value=0.0027). Also checked was the total number of sites that were matched out of the total sequence fragments, with the iteratively obtained matrices, M_i's, and with the raw M_0's: $|S|_{iter}$ and $|S|_{raw}$ (Figs.4.13(b), 4.11(b), 4.12(b)). Again $|S|_{iter}$'s were consistently larger than $|S|_{raw}$'s of corresponding p-values. A similar trend was seen in the total number of unique sequence patterns matched (Figs.4.13(c), 4.11(c), 4.12(c)). The higher recall count at an identical p-value means that the sensitivity was improved at identical levels of specificities, hence attesting to a true improvement of patterns. The f^2 residues in the first half site (CGAT) form a reverse complement of the other half f^2 residues, as well as the f^1 residues do with the corresponding ones. This is expected since the Stat3 TF is

known to bind DNA as a homo-dimer [138], hence the matrix truly representative of the distribution of its binding sequence patterns is expected to have a reverse complementary symmetry. At the p-value, $1e^{-4}$, 6 out of 44 matrix entries were zero (Table 4.9) which may be indicative of a rather stringent p-value or slight overfitting. The lowest AS_1 value attained was 0.06208051, and the lowest AS_2, 0.04130808 at p-value $p = 5.0e^{-4}$ during Φ_i (table 4.11), which is more than a decade reduction from those of M_0. Corresponding logo (4.8(h)) reveals that the outer-most positions excluding fringes are the most conserved ($f^1 = 0.936242$ and $IC = 1.562637$ at position 1, and $f^1 = 0.940436$ and $IC = 1.588085$ at position 9). This possibly suggests the mode of the interaction between Stat3 TF and DNA, where the end residues are in a stronger contact than the rest. Residues at the fringe positions were only marginally conserved (IC=0.144503 and 0.149006 bits, Table 4.11), and the middle position showed nearly no conservation. It was seen that, while each half site is overall conserved, there exists a discernible variation in the degrees of conservations therein (f^1 ranges from 0.795302 at position 2 to 0.936242 at position 1).

With more stringent, smaller p-values, the frequencies of non-dominant residues (f^2, f^3 and f^4) decrease, whereas lenient p-values are more permissive of diverse patterns. While it can be asked which p-value is the right one to choose, selection of an appropriate value appears to be specific to each TF and to particular ChIP-seq data. Then another study on a comprehensive scale would be called for its elucidation. Here we leave the problem of picking an appropriate value to users facing specific TF's, and instead focus on the problem of approximating M^* starting from a presumably inaccurate M_0, which entails the rejection of noisy patterns and the reflection of true distribution of bound patterns.

When applied to Tcfcp, the recall count statistics obtained with the M_i matrices from iteration, and with raw M_0 matrix, under varying p-values were as in Table 4.18. Sequences and sites found with M_i's consistently outnumbered those with M_0 matrix (Figs.4.12(a),4.12(b)). The lowest asymmetry values attained were $AS_1 = 0.41343279$ and $AS_2 = 0.23105012$ at iteration 9 with $p = 1.0e^{-3}$ (Figs.4.9(b) to 4.9(j)). Highest conservations in the first half site were at the second and fifth positions ($IC = 1.585289$ and $IC = 1.621942$, Table 4.5).

The third position had a significant fashion of frequency bipartition between C and T, and the fourth position had quite an even frequency partition between A and G (Table 4.5). Nucleotides C and T are pyrimidines of single ring structures, and A and G are purines of double rings. This suggests another modes of conservations at the levels of purines or pyrimidines, instead of conservations at single residue levels. From the initial matrix (Figure 4.9(a)), such modes of conservations could have been easily regarded as spurious and missed out due to similar heights that are attained by positions of little conservations (e.g., positions 7 and 10).

Nmyc forms a bHLH leucine zipper structure and binds DNA as a dimer. At $p = 5.0e^{-4}$, the position 4 strongly favors G, and alternatively favors A to somewhat less degree, and the reverse complementarily corresponding position 7 had a similar preference to their complements.

1. Position 4: A-0.35966387 C-0.02521008 G-0.61512605 T-0.00000000

2. Position 7: A-0.00000000 C-0.65042017 G-0.00000000 T-0.34957983

This mode of purine or pyrimidine preference is similar to what was seen in tcfcp. With increasing p-values during Φ_i, the dinucleotide pattern GC becomes more prevalent through patterns. This would be attributable to the enrichment of clustered CpG's from CpG islands ([154, 155]) in the ChIP-seq set which would be predominantly found in promoter regions. It is well known that mammalian promoter regions often harbor enriched CpG dinucleotides, and particular attention is called for, whenever a TF pattern contains many CpG subpatterns. Although it attests to the tracking for patterns enriched in the data by the algorithm, clearly the pattern was not sought after. The number of sites found with the iteration, $|S|_{iter}$, far exceeded those without, $|S|_{raw}$, especially during the increase phase, possibly due to spurious matchings with CpG (Fig.4.13(b)). During Φ_d where the CpG subpattern had little impact, the numbers of sequences recalled with iteration still exceeded by significant margins those without. In addition, the binding pattern changes of Esrrb and Klf4 TFs through iterations are shown in Figs.4.21(f), 4.22(f). The numbers of sequences,

sites and unique patterns detected, with the matrices from iteration and with the raw Jaspar matrices were as shown in Figs.4.14(c), 4.15(c). During Φ_d, sequences recalled by using M_i's consistently exceeded those that were recalled by using M_0. Our approach is intended to be a supervised approach, and due discretion is called for, in selecting an appropriate pattern and rejecting patterns which were not sought for, possibly by referencing the initial matrices that were used.

Alternatively, we used a starting p-value $p_1 = 1.0e^{-2}$, and ran another iteration with the p-value series:

$$\Phi^2 = \{\Phi_{d'}, \Phi^1, \Phi_{i'}\} \tag{4.13}$$

$$\Phi_{d'} = \{1.0e^{-2}, 7.5e^{-3}, 5.0e^{-3}, 2.5e^{-3}\} \tag{4.14}$$

$$\Phi_{i'} = \{2.5e^{-3}, 5.0e^{-3}, 7.5e^{-3}, 1.0e^{-2}\} \tag{4.15}$$

, with $\Phi^1 = \Phi_d + \Phi_i$ previously defined. When the series was applied to symmetric TF's, resulting patterns were too divergent or had far lower levels of symmetry. This would imply the starting p-value $1e^{-2}$ does not have sufficient specificity to reject noisy patterns in the data. As for the pivoting p-value, p_p, where phase changes from Φ_d to Φ_i, many matrix entries were observed to turn into zeroes by the p-value $p = 1.0e^{-4}$, across the TF's. Given appreciable levels of diversity of many positions in TF's in general, this appeared to signify a large loss of such diversity information. Continued iteration with decreasing p-values would then incur further loss of sensitivity. So we used the p-value $p = 1.0e^{-4}$ as p_p.

Taking the raw matrix of each TF, we decreased the f^1 frequency values of each position randomly by using one of the scaling factors from $S_1 = \{s_1, s_2\} = \{0.70, 0.80\}$. For each TF, 10 such perturbed matrices, M_0^p's, were generated, each to serve as an initial matrix, and the iterative method was applied. Given the series of matrices, $\{M_1^p, M_2^p, \ldots, M_9^p\}$, resulting from the iteration with a perturbed initial matrix, and the series of matrices, $\{M_1, M_2, \ldots, M_9\}$, resulting from the iteration with an unperturbed matrix, we measured the differences between the matrices of corresponding iteration steps, $|M_i^p - M_i|$. Differences in terms of 1-norm (Figs.4.19(a),4.19(c),4.19(e)), 2-norm (Figs.4.19(b),4.19(d),4.19(f)) and

IC (Figs.4.20(a),4.20(b),4.20(c)) show that series all converge to within very small tolerances, while the initial matrices, M_0^p's, exhibit large differences from M_0. Overall, it is implied that the space of initial matrices that converge to M^* is quite large. The stable convergences to series of highly similar matrices from both the substantially perturbed matrices and the unperturbed ones as initial matrices imply that the matrices detected along the iterations would indeed correspond to true M^*'s representing the true distributions of the M^*-sequences.

4.5 Discussion

In the current study, we developed a method to find approximations to the true binding pattern distributions of transcription factors, using ChIP-seq data together with presumably inaccurate letter probability matrices. The two kinds of data are highly complementary in nature; the ChIP-seq data are enriched in the true patterns, albeit mixed with random patterns, and the initial matrix for a transcription factor is often inaccurate while having a potential to guide the search for the accurate pattern along a right path. Existing algorithms did not address formal ways to approximate the correct pattern, when an inaccurate matrix representation for the pattern is given. Due to the computational complexity, patterns found by the de novo discovery algorithms were often inaccurate. Our proposed method can be used to find precise patterns from ChIP-seq data, when an initial matrix of a rather low accuracy is given. Alternatively, it can be employed during a post-processing stage to improve or validate the patterns found by using the de novo algorithms. Residues in the in vivo transcription factor-bound patterns typically exhibit substantial levels of variability and diversity. Then using the most frequent residues alone to represent the binding patterns, as done in the consensus representation, would incur a large loss of information. It is believed that a substantial fraction of the binding patterns have affinities that are far lower than the highest possible affinity. Elucidating the true distribution of patterns would be an important step in statistically assessing the significance of patterns targeted by a transcription factor, and further using them to find out the genes that are regulated by the transcription

factor. An important observation was that some conservations were at the levels of purines or pyrimidines, instead of the levels of single residues, in some transcription factor binding patterns. In conserved regions of multiple, consecutive residues, the most dominant residues frequently showed substantial variability in the level of conservation. They would result from the differences in the structural and chemical interactions between constituent molecules, and imply that the relative contributions of residues to the overall binding are different.

The iterative nature of the algorithm is somewhat similar to that of the expectation maximization [145]. A sensible p-value schedule has been devised so that random, noisy background patterns are maximally rejected, while the in vivo transcription factor-bound patterns are maximally retained, so as to enable approximations to truly representative matrices. Starting from moderately or highly perturbed initial matrices, stable convergences to nearly accurate matrices were achieved, over the synthetic and real transcription factor cases. Hence, our approach was shown to be robust against significant degrees of perturbation to the initial patterns.

In the current work, we assumed an independence between positions, and employed the letter probability matrices that reflect a zero order Markov model. While the positional independence assumption is believed to provide a fairly good modeling power, still substantial correlations can exist between residues in some transcription factors, especially among the neighboring ones [146]. While utilizing such inter-residue dependency was suggested not to significantly improve the accuracy of transcription factor-binding pattern models [149], it would be still interesting to use such higher order interactions so as to obtain even more realistic representations of patterns, and see if they can make any contributions to the problems such as predicting the target genes of transcription factors.

Table 4.1: Perturbed matrix that was obtained by applying S_1 to synthetic pattern, P_s

Position	A	C	G	T
0	0.720000	0.150081	0.069016	0.060902
1	0.069016	0.060902	0.720000	0.150081
2	0.150081	0.069016	0.060902	0.720000
3	0.090902	0.630000	0.180081	0.099016
4	0.250000	0.250000	0.250000	0.250000
5	0.099016	0.180081	0.630000	0.090902
6	0.630000	0.090902	0.099016	0.180081
7	0.180081	0.630000	0.090902	0.099016
8	0.090902	0.099016	0.180081	0.630000

Table 4.2: IC of the matrix perturbed with S_1

position	f^1 residue	f^1	IC
0	a	0.720000	0.736047
1	g	0.720000	0.736047
2	t	0.720000	0.736047
3	c	0.630000	0.489850
4	a	0.250000	0.000000
5	g	0.630000	0.489850
6	a	0.630000	0.489850
7	c	0.630000	0.489850
8	t	0.630000	0.489850

Table 4.3: IC and other measure changes along iteration (initial matrix was obtained by applying S_1 to P_s)

$\sum f^1$	IC	Zeros	AS_1	AS_2	Stdev-max	Iter.
5.560000	4.657389	0	0.54000000	0.31176915	0.04657943	00
7.482307	11.859641	0	0.09791565	0.06057295	0.01465860	01
7.524837	12.075690	0	0.05017562	0.03547952	0.01657152	02
7.639758	12.503471	0	0.02855575	0.02019196	0.02106235	03
7.767442	13.098560	6	0.00000000	0.00000000	0.01105833	04
7.892319	14.019979	18	0.00000000	0.00000000	0.03489020	05
7.841004	13.653412	10	0.00000000	0.00000000	0.03416863	06
7.667785	12.586329	0	0.00000000	0.00000000	0.01803142	07
7.531282	12.095856	0	0.00000000	0.00000000	0.01532486	08
7.469183	11.796485	0	0.00191112	0.00135137	0.01467024	09

Table 4.4: M_1 from iteration 1

Position	A	C	G	T
0	0.92244304	0.06980126	0.00484731	0.00290839
1	0.00920989	0.00436258	0.92050412	0.06592341
2	0.09646146	0.00533204	0.00290839	0.89529811
3	0.00145419	0.89626757	0.09597673	0.00630150
4	0.25399903	0.24236549	0.24672807	0.25690742
5	0.00678623	0.09306835	0.89869123	0.00145419
6	0.87736306	0.00678623	0.01211827	0.10373243
7	0.07658749	0.90790111	0.00484731	0.01066408
8	0.00290839	0.01017935	0.07998061	0.90693165

Table 4.5: IC of M_1

Position	f^1 residue	f^1	IC
0	a	0.922443	1.562712
1	g	0.920504	1.534886
2	t	0.895298	1.466927
3	c	0.896268	1.474106
4	t	0.256907	0.000383
5	g	0.898691	1.480109
6	a	0.877363	1.369251
7	c	0.907901	1.482424
8	t	0.906932	1.488843

Table 4.6: Matrix perturbed with the scaling factor set S_2

Position	f^1 residue	f_1'	IC
0	a	0.720000	0.736047
1	g	0.540000	0.297386
2	t	0.540000	0.297386
3	c	0.630000	0.489850
4	a	0.250000	0.000000
5	g	0.630000	0.489850
6	a	0.720000	0.736047
7	c	0.540000	0.297386
8	t	0.720000	0.736047
Sum	-	-	4.079997

Table 4.7: IC values from iteration (initial matrix obtained by applying S_2 to P_s)

$\sum f^1$	IC	Zeros	AS_1	AS_2	Stdev-max	Iter.
5.290000	4.079997	0	0.09000000	0.10392305	0.08332381	00
7.508885	11.871491	0	0.03455084	0.04272715	0.02531132	01
7.523262	12.056275	0	0.01350676	0.01910144	0.01433933	02
7.658699	12.562435	0	0.00722624	0.01021944	0.02051380	03
7.769753	13.159174	8	0.00000000	0.00000000	0.01225274	04
7.892319	14.019979	18	0.00000000	0.00000000	0.03489020	05
7.841004	13.653412	10	0.00000000	0.00000000	0.03416863	06
7.667785	12.586329	0	0.00000000	0.00000000	0.01803142	07
7.531282	12.095856	0	0.00000000	0.00000000	0.01532486	08
7.469183	11.796485	0	0.00047778	0.00067568	0.01467024	09

Table 4.8: Changes along iteration in various measures, given the initial matrix of $k = 2$ derangements

$\sum f^1$	IC	Zeros	AS_1	AS_2	Stdev-max	Iter.
7.450000	11.840011	0	1.79819512	1.20741005	0.00000000	00
6.857143	9.866572	5	1.71428571	1.00749068	0.14696018	01
7.814414	13.481942	6	0.45765764	0.28912339	0.07804965	02
7.687429	12.628391	1	0.14722537	0.09934551	0.01561927	03
7.777641	13.144441	9	0.10073708	0.06545160	0.01272891	04
7.904271	13.914492	17	0.17378498	0.12288454	0.03166972	05
7.827676	13.405019	7	0.16710184	0.11319936	0.02305679	06
7.685746	12.620194	1	0.18181817	0.11776472	0.02008365	07
7.516385	12.040598	1	0.16683219	0.10847425	0.01895917	08
7.471770	11.825734	1	0.17224878	0.10850121	0.01879911	09

Table 4.9: IC values along the iteration of Stat3 TF

$\sum f^1$	IC	Zeros	AS_1	AS_2	Stdev-max	Iter.
8.741026	13.600680	3	1.83523400	1.10530225	0.22929127	00
7.959973	10.074737	0	1.23202173	0.74477111	0.19084167	01
8.016487	10.468826	0	0.75555553	0.45669236	0.18688752	02
8.103766	11.085177	0	0.46861924	0.28998010	0.19383665	03
8.274074	11.858616	0	0.25608466	0.16666693	0.20016995	04
8.481982	12.965770	6	0.18918922	0.12814643	0.20491349	05
8.279616	11.911774	0	0.09605119	0.06490790	0.19998989	06
8.103188	11.119593	0	0.06208051	0.04130808	0.19488000	07
7.971616	10.635525	0	0.07569143	0.04887949	0.19227716	08
7.839695	10.176923	0	0.07506361	0.04757993	0.18654629	09

Table 4.10: IC of Stat3 TF (iteration=3)

Position	f^1 residue	f^1	IC
0	c	0.418410	0.126405
1	t	0.905439	1.402181
2	t	0.759833	1.023485
3	c	0.875314	1.287309
4	c	0.841004	1.294243
5	a	0.285356	0.017800
6	g	0.896234	1.480910
7	g	0.898745	1.396351
8	a	0.835983	1.229791
9	a	0.951464	1.660001
10	g	0.435983	0.166701
Sum	-	-	11.085177

Table 4.11: IC of Stat3 TF (iteration=7)

Position	f^1 residue	f^1	IC
0	c	0.423658	0.144503
1	t	0.936242	1.562637
2	t	0.795302	1.115277
3	c	0.885906	1.342946
4	c	0.859899	1.343904
5	t	0.275168	0.007066
6	g	0.865772	1.366794
7	g	0.888423	1.353798
8	a	0.808725	1.145577
9	a	0.940436	1.588085
10	g	0.423658	0.149006
Sum	-	8.103188	11.119593

Table 4.12: IC values over the iteration steps of Tcfcp TF

$\sum f^1$	IC	Zeros	AS_1	AS_2	Stdev-max	Iter.
9.824523	11.650291	0	2.90228100	1.87883315	0.25512101	00
9.030560	9.850817	0	1.86247879	1.18468062	0.23884510	01
8.962775	10.194990	0	1.26903552	0.79887896	0.23929739	02
9.012322	10.690545	1	0.90236965	0.56474865	0.24200261	03
9.270992	11.566697	5	0.66921120	0.41822132	0.24383441	04
9.485380	12.743679	11	0.63157892	0.39453408	0.24676533	05
9.292181	11.776022	4	0.55967079	0.33198134	0.24581214	06
9.083588	10.911586	1	0.57492353	0.34724792	0.24093670	07
8.907191	10.339027	0	0.42809362	0.24994807	0.23890038	08
8.809702	9.980596	0	0.41343279	0.23105012	0.23775316	09

Table 4.13: The letter probability matrix of Tcfcp TF (iteration=9)

Position	A	C	G	T
0	0.35754190	0.08119181	0.29757914	0.26368715
1	0.00260708	0.92662942	0.06331471	0.00744879
2	0.12551210	0.54562384	0.00335196	0.32551210
3	0.40335196	0.00707635	0.46443203	0.12513966
4	0.00446927	0.05139665	0.94264432	0.00148976
5	0.20260708	0.34078212	0.07821229	0.37839851
6	0.11098696	0.32290503	0.10800745	0.45810056
7	0.17690875	0.33333333	0.19217877	0.29757914
8	0.29757914	0.19180633	0.33370577	0.17690875
9	0.45735568	0.10689013	0.32476723	0.11098696
10	0.37839851	0.07970205	0.34078212	0.20111732
11	0.00148976	0.93966480	0.05437616	0.00446927
12	0.12253259	0.46964618	0.00744879	0.40037244
13	0.32774674	0.00335196	0.54189944	0.12700186
14	0.00744879	0.06108007	0.92886406	0.00260708
15	0.26219739	0.30018622	0.08119181	0.35642458

Table 4.14: IC of Tcfcp TF matrix (iteration=9)

Position	f^1 residue	f^1	IC
0	a	0.357542	0.147894
1	c	0.926629	1.571023
2	c	0.545624	0.592687
3	g	0.464432	0.532016
4	g	0.942644	1.650707
5	t	0.378399	0.186018
6	t	0.458101	0.258669
7	c	0.333333	0.051949
8	g	0.333706	0.052246
9	a	0.457356	0.260076
10	a	0.378399	0.184000
11	c	0.939665	1.638331
12	c	0.469646	0.535418
13	g	0.541899	0.587905
14	g	0.928864	1.579736
15	t	0.356425	0.147883

Table 4.15: IC change along the iteration of Nmyc TF

$\sum f^1$	IC	Zeros	AS_1	AS_2	Iter.	P-value
7.123932	9.485116	0	0.36431624	0.46610597	01	$1.0e^{-3}$
6.946809	9.706180	0	0.22841051	0.29306899	02	$7.5e^{-4}$
6.795440	10.215353	2	0.15244300	0.19035724	03	$5.0e^{-4}$
6.934996	11.163767	12	0.09349955	0.12499055	04	$2.5e^{-4}$
7.374790	12.641752	19	0.07647059	0.10106718	05	$1.0e^{-4}$
7.190993	10.913690	0	0.06672227	0.08620041	06	$2.5e^{-4}$
7.105651	9.742728	0	0.04766584	0.06115612	07	$5.0e^{-4}$
6.984073	9.155316	0	0.04015588	0.05134455	08	$7.5e^{-4}$
6.846623	8.730967	0	0.02717115	0.03405713	09	$1.0e^{-3}$

Table 4.16: Numbers of sites, sequences and patterns recalled, by using iteratively obtained matrices, and by using Jaspar matrices (Nmyc)

N-sites	N-sequences	N-patterns	Seq./Patterns	Iter.	P-value	iter/Jaspar
744	549	457	-	00	-	-
970	608	543	1.1197	01	1.0e-3	iter
938	569	494	1.1518	02	7.5e-4	iter
816	487	395	1.2329	03	5.0e-4	iter
591	397	250	1.5880	04	2.5e-4	iter
301	218	121	1.8017	05	1.0e-4	iter
597	376	253	1.4862	06	2.5e-4	iter
1122	495	494	1.0020	07	5.0e-4	iter
1645	539	734	0.7343	08	7.5e-4	iter
2127	553	973	0.5683	09	1.0e-3	iter
722	540	448	-	01	1.0e-3	Jaspar
608	477	373	-	02	7.5e-4	Jaspar
492	408	286	-	03	5.0e-4	Jaspar
313	282	172	-	04	2.5e-4	Jaspar
163	158	81	-	05	1.0e-4	Jaspar

Table 4.17: Numbers of sites, sequences and patterns recalled, by using iteratively obtained matrices, and by using Jaspar matrices (stat3)

N-sites	N-sequences	N-patterns	Seq./Patterns	Iter.	P-value	iter/Jaspar
779	607	560	1.0839	01	1.0e-3	iter
701	576	492	1.1707	02	7.5e-4	iter
610	531	409	1.2983	03	5.0e-4	iter
469	416	288	1.4444	04	2.5e-4	iter
345	320	184	1.7391	05	1.0e-4	iter
459	407	277	1.4693	06	2.5e-4	iter
608	529	406	1.3030	07	5.0e-4	iter
710	587	494	1.1883	08	7.5e-4	iter
793	626	569	1.1002	09	1.0e-3	iter
717	567	517	-	01	1.0e-3	Jaspar
651	532	455	-	02	7.5e-4	Jaspar
538	462	374	-	03	5.0e-4	Jaspar
396	361	257	-	04	2.5e-4	Jaspar
257	237	152	-	05	1.0e-4	Jaspar

Table 4.18: Numbers of sites, sequences and patterns recalled, by using iteratively obtained matrices, and by using Jaspar matrices (Tcfcp)

N-sites	N-sequences	N-patterns	Seq./Patterns	Iter.	P-value	iter./Jaspar
1333	804	1295	0.6208	01	1.0e-3	iter
1212	782	1174	0.6661	02	7.5e-4	iter
1070	748	1035	0.7227	03	5.0e-4	iter
792	615	763	0.8060	04	2.5e-4	iter
500	427	473	0.9027	05	1.0e-4	iter
760	601	731	0.8222	06	2.5e-4	iter
1054	743	1021	0.7277	07	5.0e-4	iter
1231	810	1192	0.6795	08	7.5e-4	iter
1383	840	1340	0.6269	09	1.0e-3	iter
1178	749	1150	-	01	1.0e-3	Jaspar
1013	685	988	-	02	7.5e-4	Jaspar
822	593	802	-	03	5.0e-4	Jaspar
596	473	581	-	04	2.5e-4	Jaspar
375	326	367	-	05	1.0e-4	Jaspar

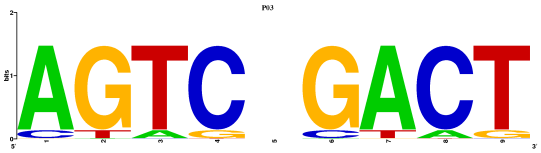
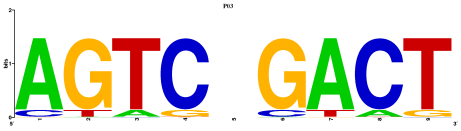
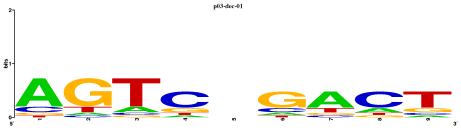


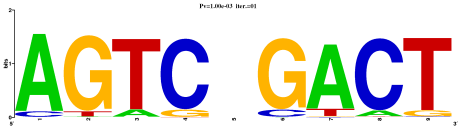
Figure 4.1: A synthetic pattern with reverse complementarity. A gap with no conservation is present at the middle position.



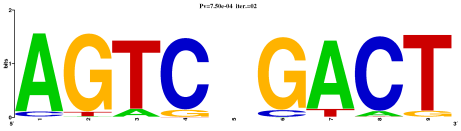
(a) P_s



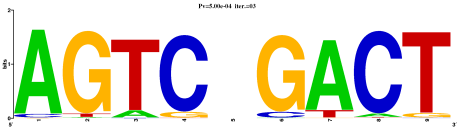
(b) P'_{S_1}



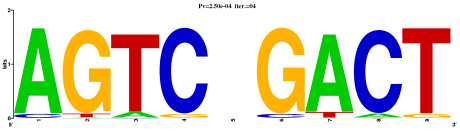
(c) Iter.=1



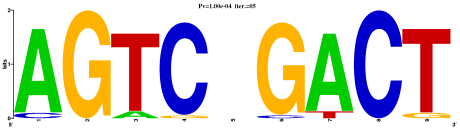
(d) Iter.=2



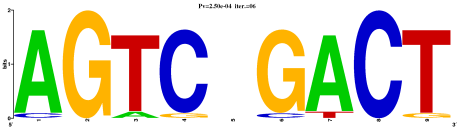
(e) Iter.=3



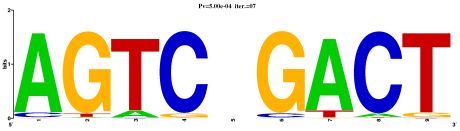
(f) Iter.=4



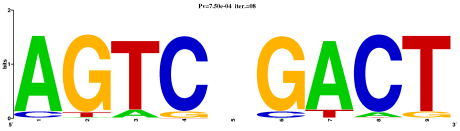
(g) Iter.=5



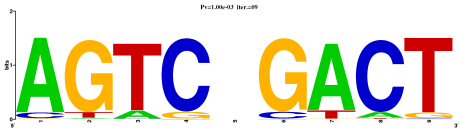
(h) Iter.=6



(i) Iter.=7

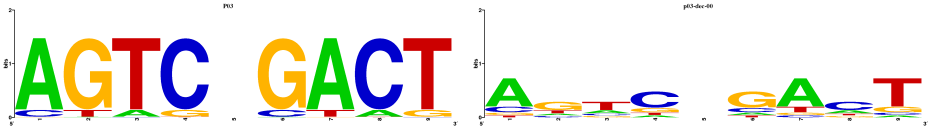


(j) Iter.=8

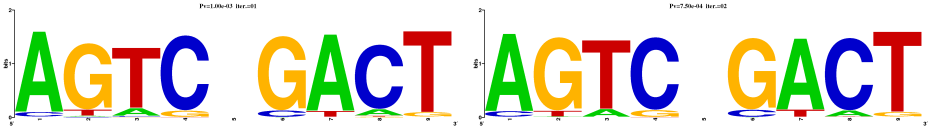


(k) Iter.=9

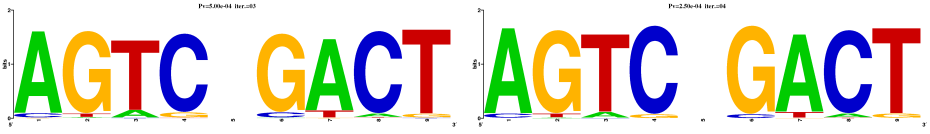
Figure 4.2: Iteration from an initial pattern obtained by perturbing P_s with S_1. Patterns stably converge to the unperturbed pattern, P_s.



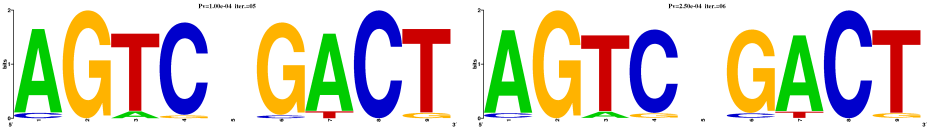
(a) P_s (b) P'_{S_2}



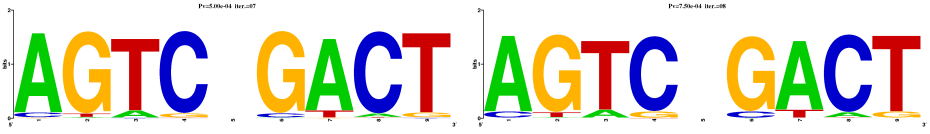
(c) Iter.=1 (d) Iter.=2



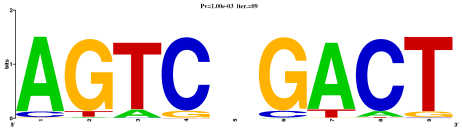
(e) Iter.=3 (f) Iter.=4



(g) Iter.=5 (h) Iter.=6

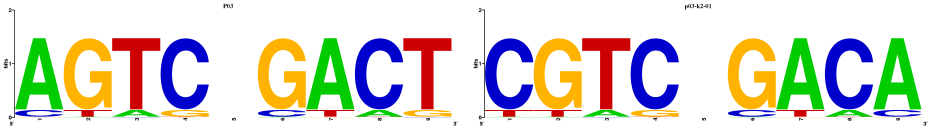


(i) Iter.=7 (j) Iter.=8



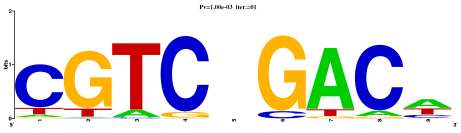
(k) Iter.=9

Figure 4.3: Iteration from an initial pattern obtained by perturbing P_s with scaling factor set, S_2. Patterns converge to the unperturbed P_s.

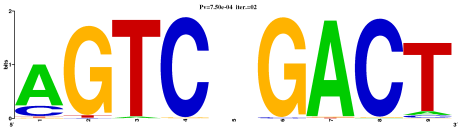


(a) P_s

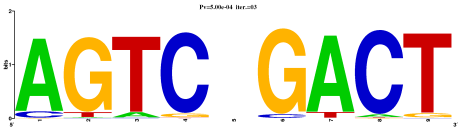
(b) $P'_{k=2}$



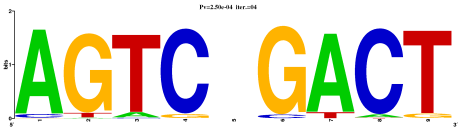
(c) Iter.=1



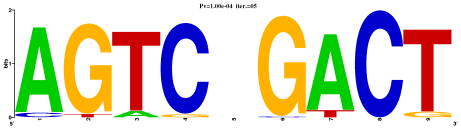
(d) Iter.=2



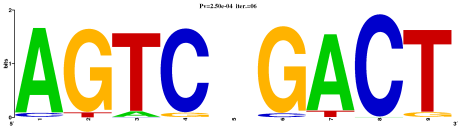
(e) Iter.=3



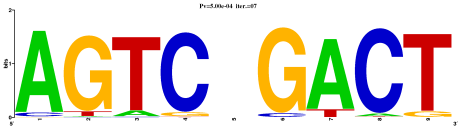
(f) Iter.=4



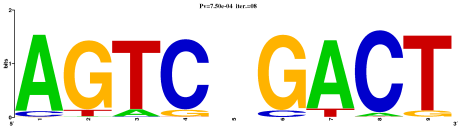
(g) Iter.=5



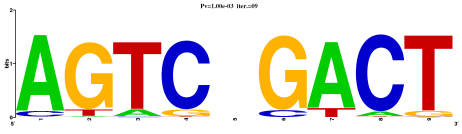
(h) Iter.=6



(i) Iter.=7

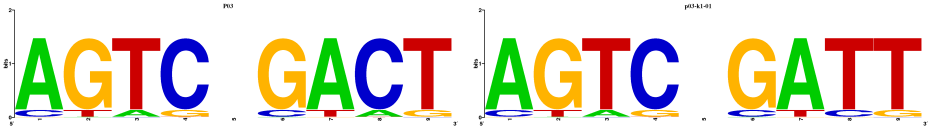


(j) Iter.=8



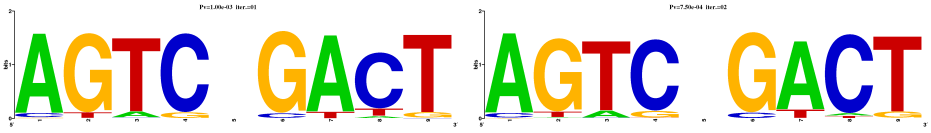
(k) Iter.=9

Figure 4.4: Iteration from an initial pattern with two residues deranged. While the degree of deviation from P_s of the initial pattern is substantial, patterns invariably converged.



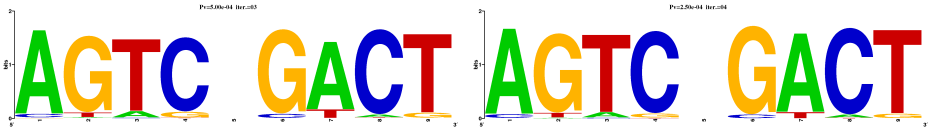
(a) P_s

(b) $P'_{k=1}$



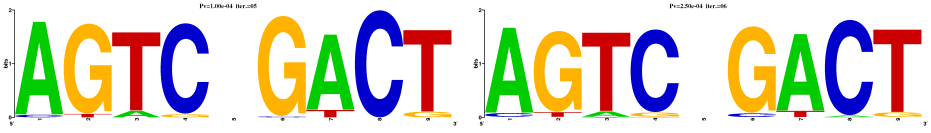
(c) Iter.=1

(d) Iter.=2



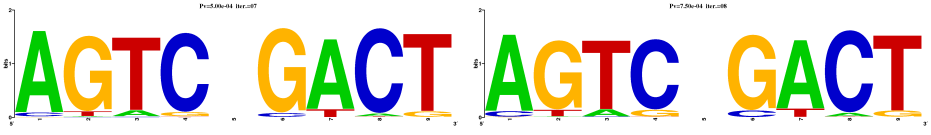
(e) Iter.=3

(f) Iter.=4



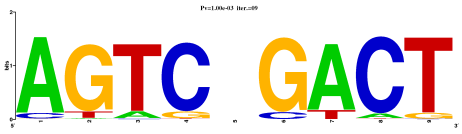
(g) Iter.=5

(h) Iter.=6



(i) Iter.=7

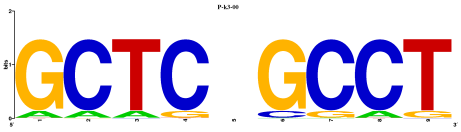
(j) Iter.=8



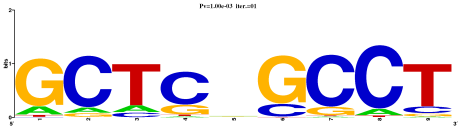
(k) Iter.=9

Figure 4.5: Initial pattern with one residue deranged. Convergence behavior is similar to that of the two residue perturbation case.

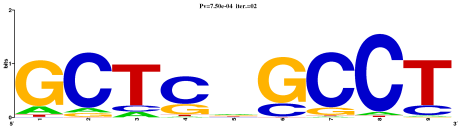
(a) P_s



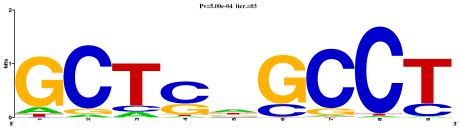
(b) $P'_{k=3}$



(c) Iter.=1



(d) Iter.=2

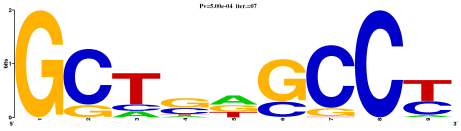


(e) Iter.=3

(f) Iter.=4

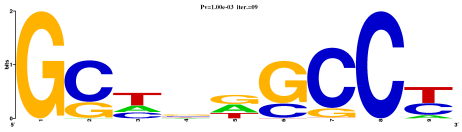
(g) Iter.=5

(h) Iter.=6



(i) Iter.=7

(j) Iter.=8

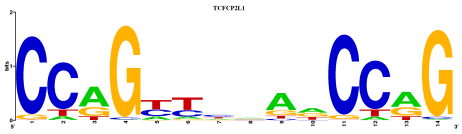


(k) Iter.=9

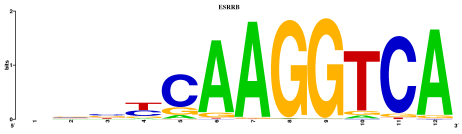
Figure 4.6: Initial pattern with three residues deranged. Patterns do not converge to the synthetic pattern. It implies the degree of deviation of the initial pattern from the synthetic template pattern P_s is too large.

(a) Nmyc

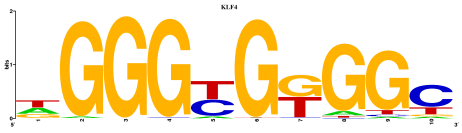
(b) Stat3



(c) Tcfcp

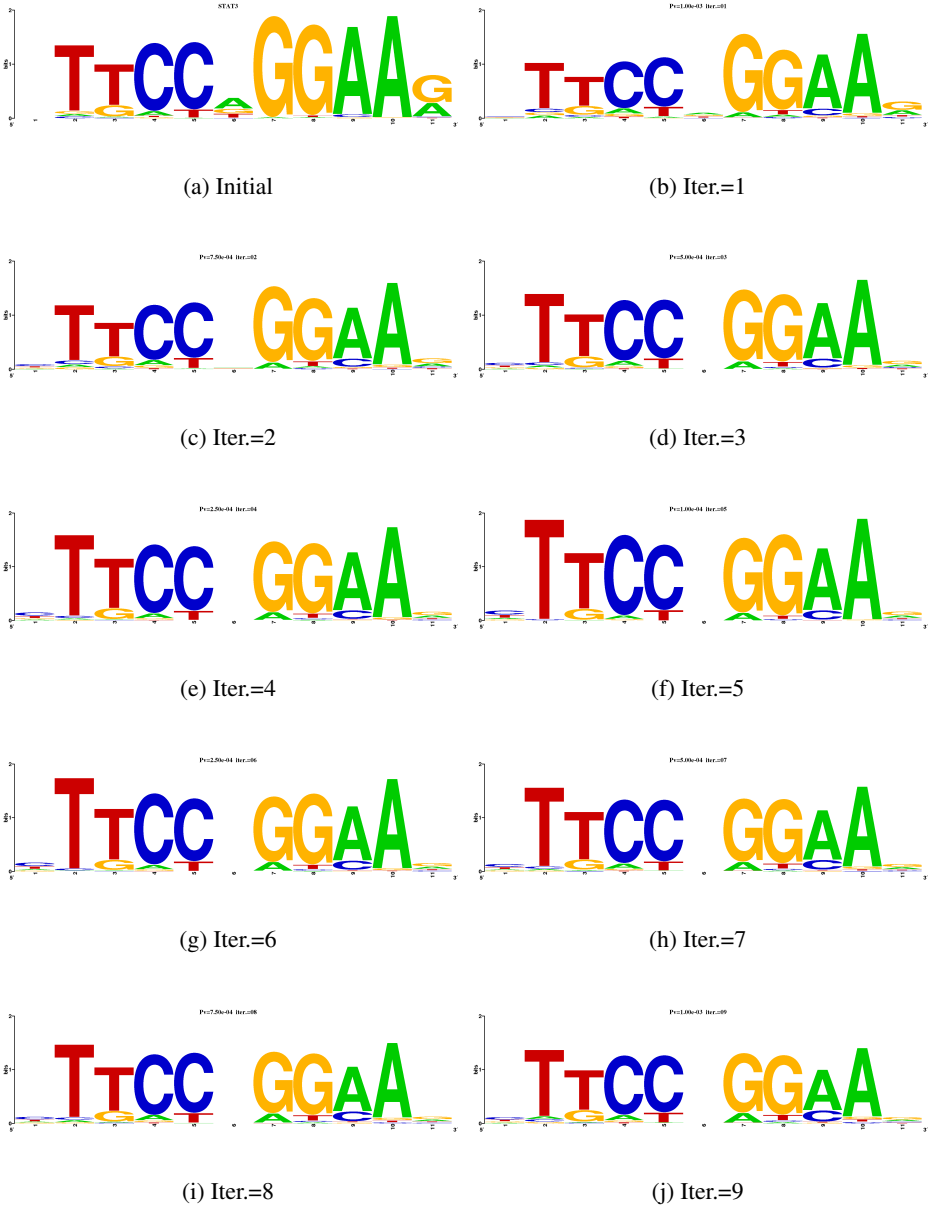


(d) Esrrb



(e) Klf4

Figure 4.7: Initial patterns from Jaspar that were used in case studies. The first three patterns have potential reverse complementary symmetries, to be discovered through iterative refinements.



(a) Initial (b) Iter.=1

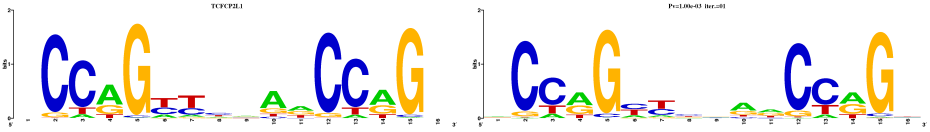
(c) Iter.=2 (d) Iter.=3

(e) Iter.=4 (f) Iter.=5

(g) Iter.=6 (h) Iter.=7

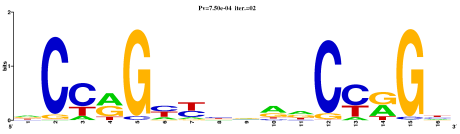
(i) Iter.=8 (j) Iter.=9

Figure 4.8: Stat3 TF binding pattern changes. Symmetry in the binding pattern is discovered through iteration.

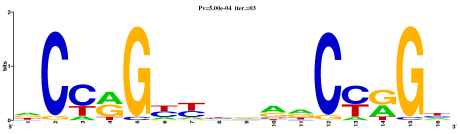


(a) Initial

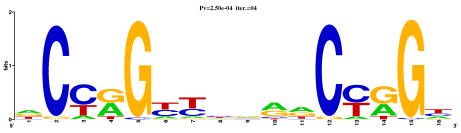
(b) Iter.=1



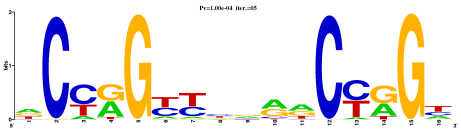
(c) Iter.=2



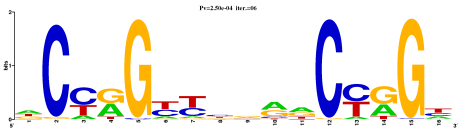
(d) Iter.=3



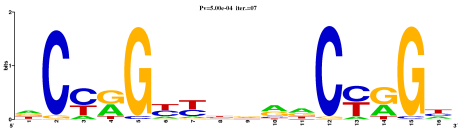
(e) Iter.=4



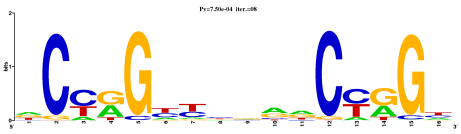
(f) Iter.=5



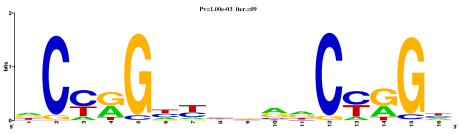
(g) Iter.=6



(h) Iter.=7

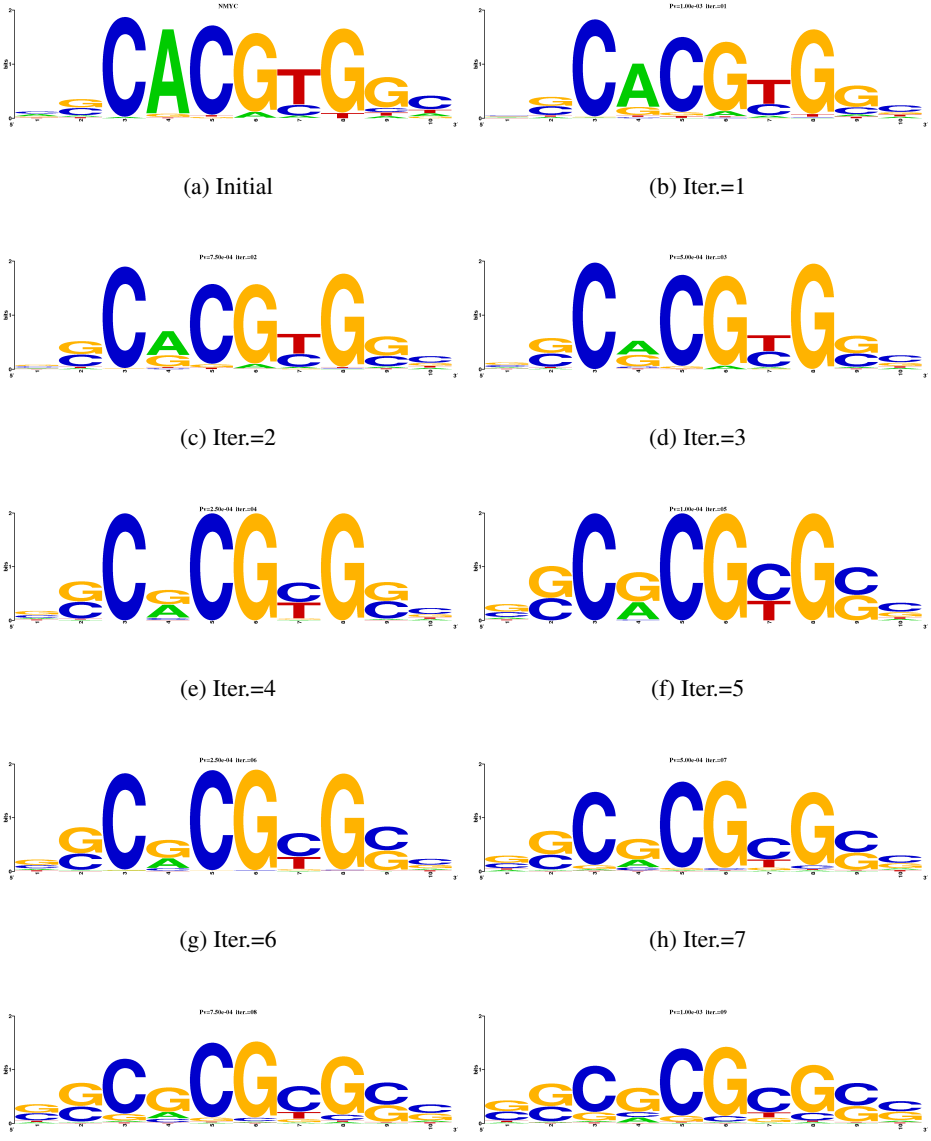


(i) Iter.=8



(j) Iter.=9

Figure 4.9: Tfcfcp TF binding pattern changes. Units of conservation are purines or pyrimidines rather than single nucleotides, at positions 3 and 4.



(a) Initial (b) Iter.=1

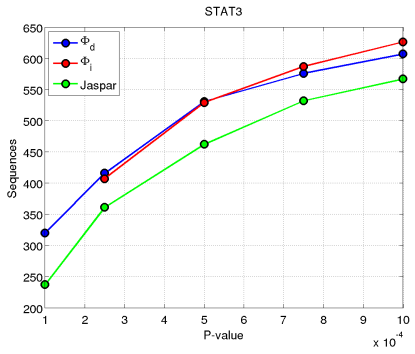
(c) Iter.=2 (d) Iter.=3

(e) Iter.=4 (f) Iter.=5

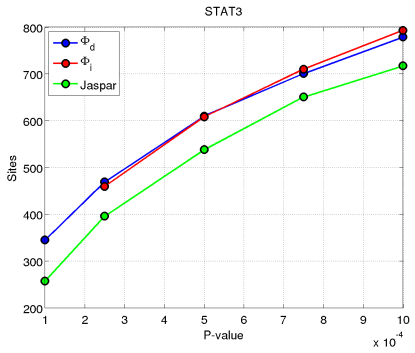
(g) Iter.=6 (h) Iter.=7

(i) Iter.=8 (j) Iter.=9

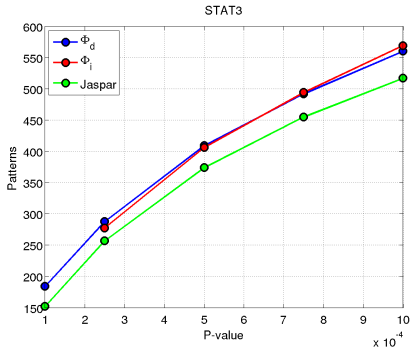
Figure 4.10: Nmyc TF binding pattern changes. During the Φ_i phase, patterns converge to CpG dinucleotides that are abundant in mammalian promoters.



(a) Number of sequences recalled-Stat3

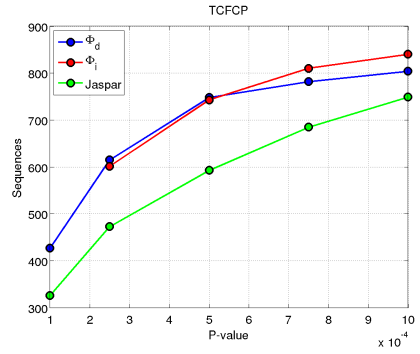


(b) Number of sites recalled-Stat3

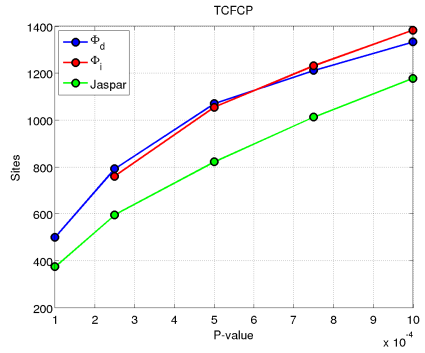


(c) Number of patterns recalled-Stat3

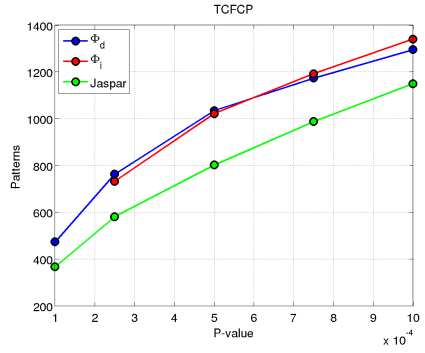
Figure 4.11: Number of sequences, sites and unique patterns found with iteratively obtained matrices and with raw Jaspar matrices without iteration-Stat3. Substantially larger number of sequences are recalled in a consistent manner, with the iteratively obtained matrices.



(a) Number of sequences recalled-Tcfcp

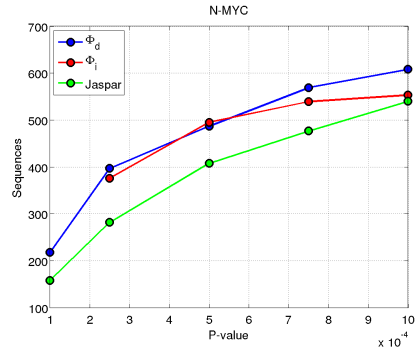


(b) Number of sites recalled-Tcfcp

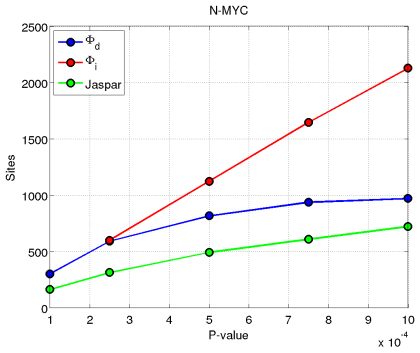


(c) Number of patterns recalled-Tcfcp

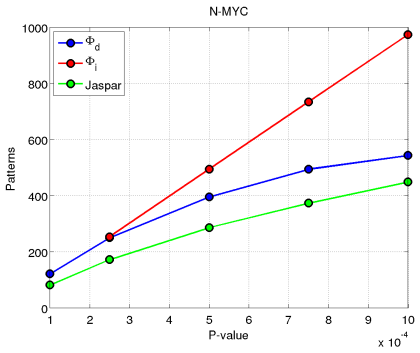
Figure 4.12: Number of sequences, sites and unique patterns found with iteratively obtained matrices and with raw Jaspar matrices without iteration-Tcfcp. Substantially larger number of sequences are recalled in a consistent manner, with the iteratively obtained matrices.



(a) Number of sequences recalled-Nmyc

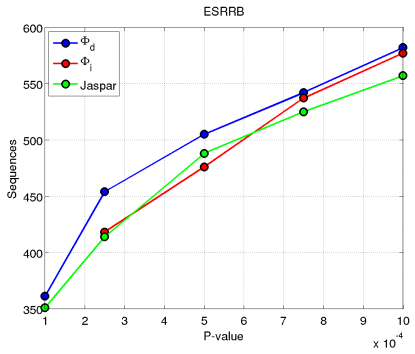


(b) Number of sites recalled-Nmyc

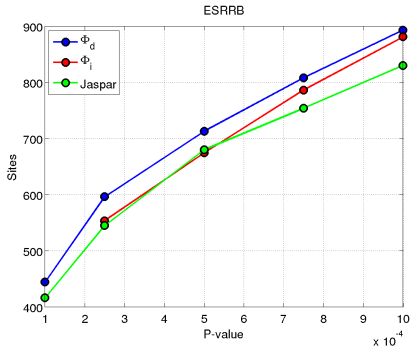


(c) Number of patterns recalled-Nmyc

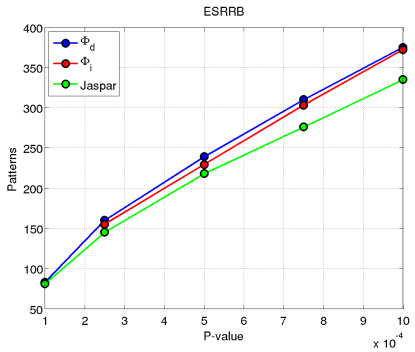
Figure 4.13: Number of sequences, sites and unique patterns found with iteratively obtained matrices and with raw Jaspar matrices without iteration-Nmyc. Recall counts resulting from using iteratively obtained matrices are larger than those resulting from using raw matrices.



(a) Number of sequences recalled-Esrrb

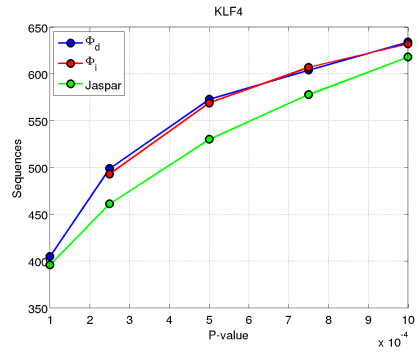


(b) Number of sites recalled-Esrrb

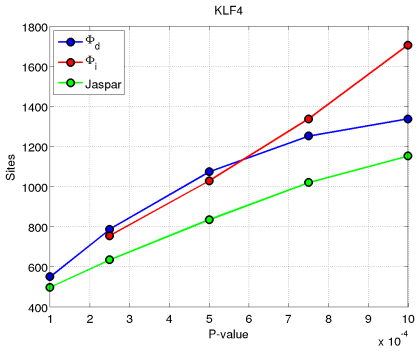


(c) Number of patterns recalled-Esrrb

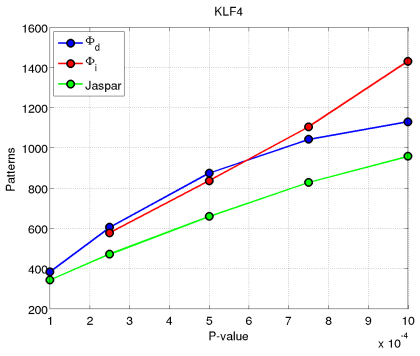
Figure 4.14: Number of sequences, sites and unique patterns found with iteratively obtained matrices and with raw Jaspar matrices without iteration-Esrrb. Larger number of sequences are recalled by using the iteratively obtained matrices than by using raw matrices, especially during the Φ_d phase.



(a) Number of sequences recalled-Klf4

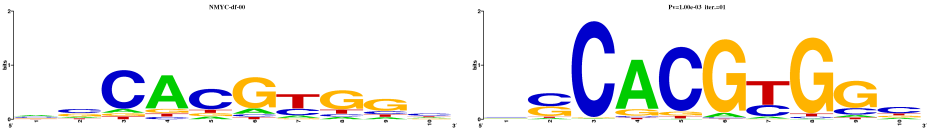


(b) Number of sites recalled-Klf4

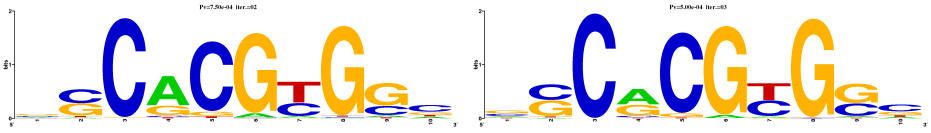


(c) Number of patterns recalled-Klf4

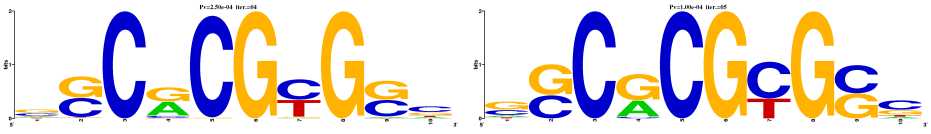
Figure 4.15: Number of sequences, sites and patterns found through iteratively obtained matrices and with raw Jaspar matrices without iteration-Klf4. The overall trends of higher recall counts are similar to those of other TFs.



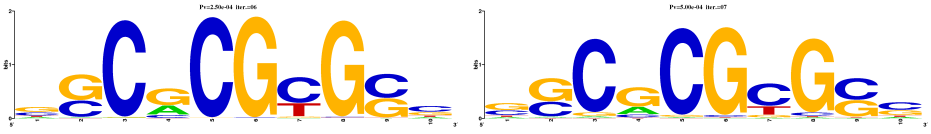
(a) P'_{S_1} (b) Iter.=1



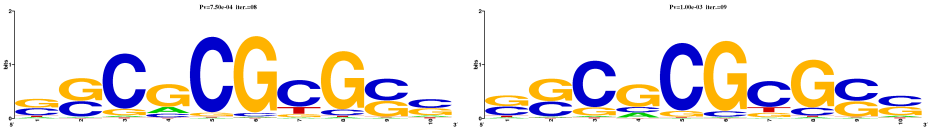
(c) Iter.=2 (d) Iter.=3



(e) Iter.=4 (f) Iter.=5

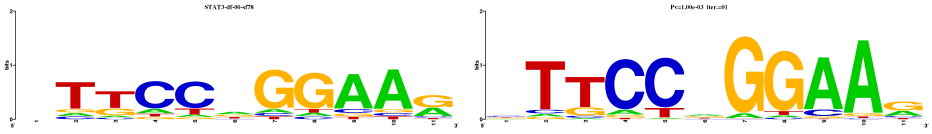


(g) Iter.=6 (h) Iter.=7

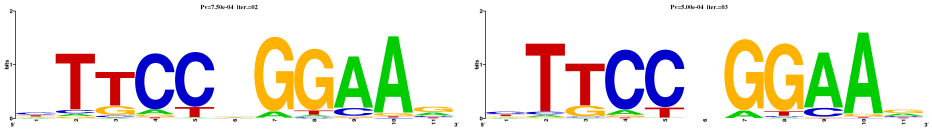


(i) Iter.=8 (j) Iter.=9

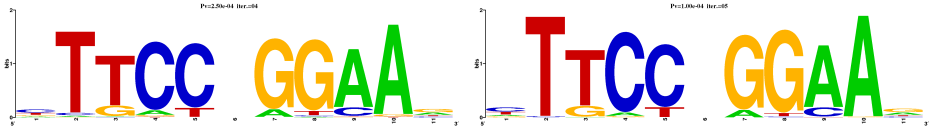
Figure 4.16: Nmyc binding pattern changes with a perturbed initial matrix. The series of patterns obtained are nearly identical to the ones that were obtained without perturbation.



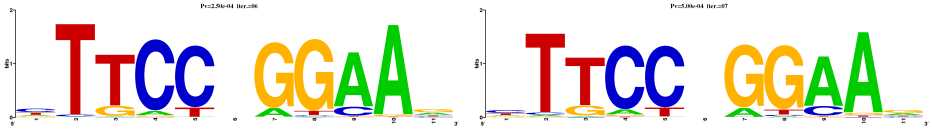
(a) P'_{S_1} (b) Iter.=1



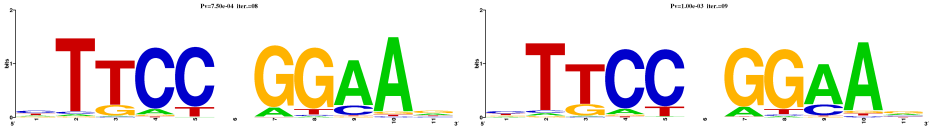
(c) Iter.=2 (d) Iter.=3



(e) Iter.=4 (f) Iter.=5

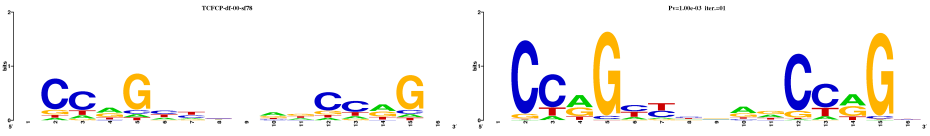


(g) Iter.=6 (h) Iter.=7



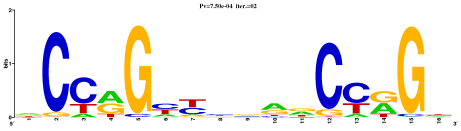
(i) Iter.=8 (j) Iter.=9

Figure 4.17: Stat3 binding pattern changes with a perturbed initial matrix. Stable convergence is exhibited through iteration.

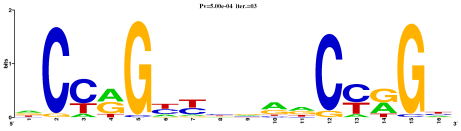


(a) P'_{S_1}

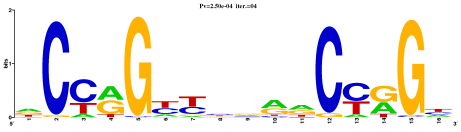
(b) Iter.=1



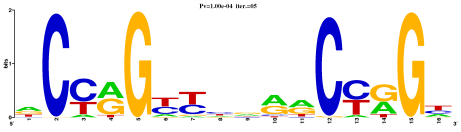
(c) Iter.=2



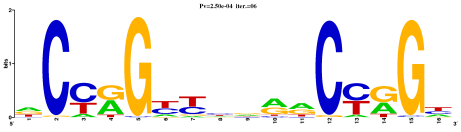
(d) Iter.=3



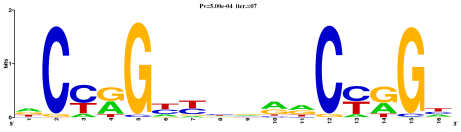
(e) Iter.=4



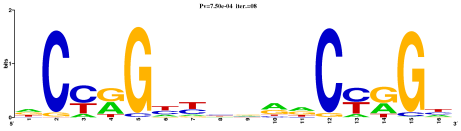
(f) Iter.=5



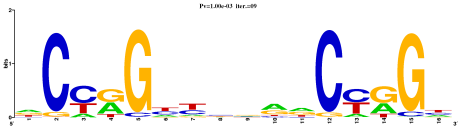
(g) Iter.=6



(h) Iter.=7

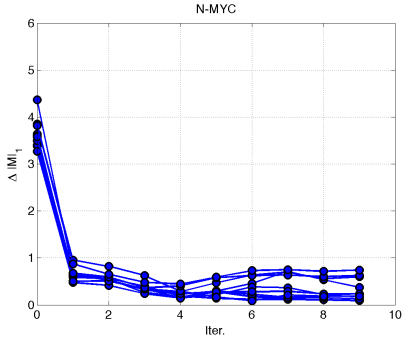


(i) Iter.=8

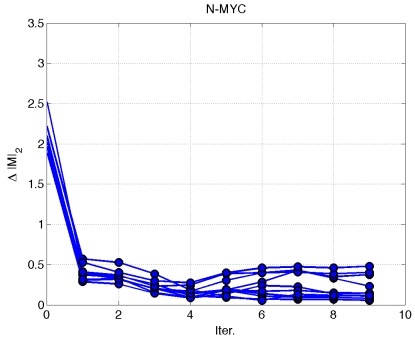


(j) Iter.=9

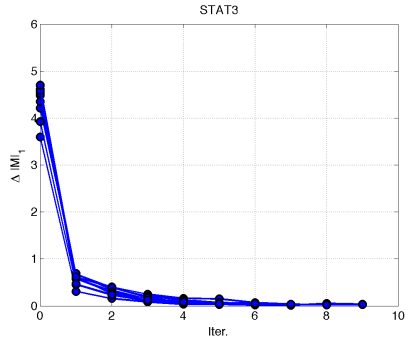
Figure 4.18: Tcfcp binding pattern changes with a perturbed initial matrix. Pattern enrichment in the ChIP-seq data attracts the initial pattern to the series of patterns that are similar to those of the unperturbed case.



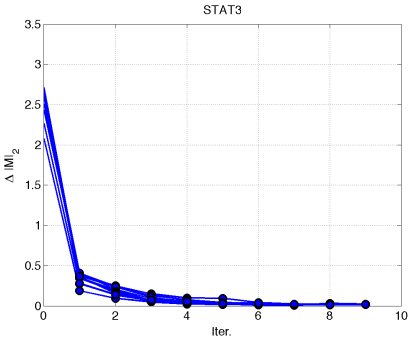
(a) Nmyc-ΔM_1



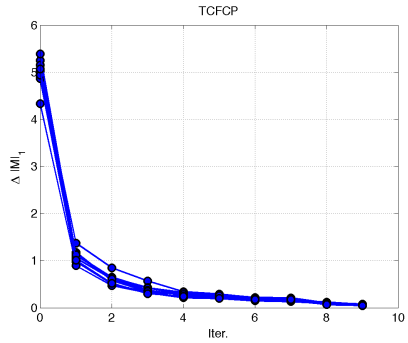
(b) Nmyc-ΔM_2



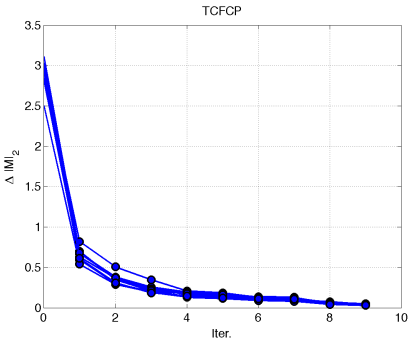
(c) Stat3-ΔM_1



(d) Stat3-ΔM_2

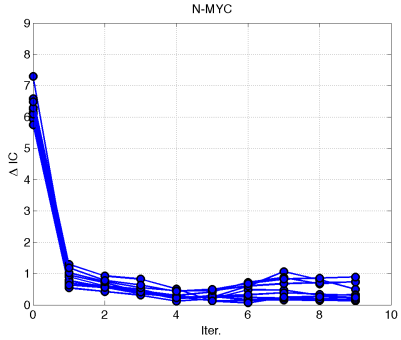


(e) Tcfcp-ΔM_1

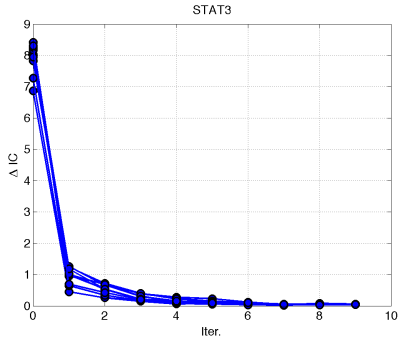


(f) Tcfcp-ΔM_2

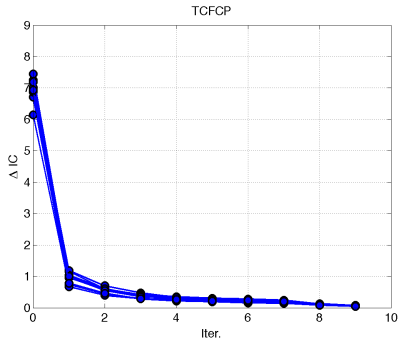
Figure 4.19: Convergence from perturbed Jaspar matrices-$\Delta M_1, \Delta M_2$. Large differences of the initial matrices from the unperturbed matrix are greatly reduced through iteration.



(a) Nmyc

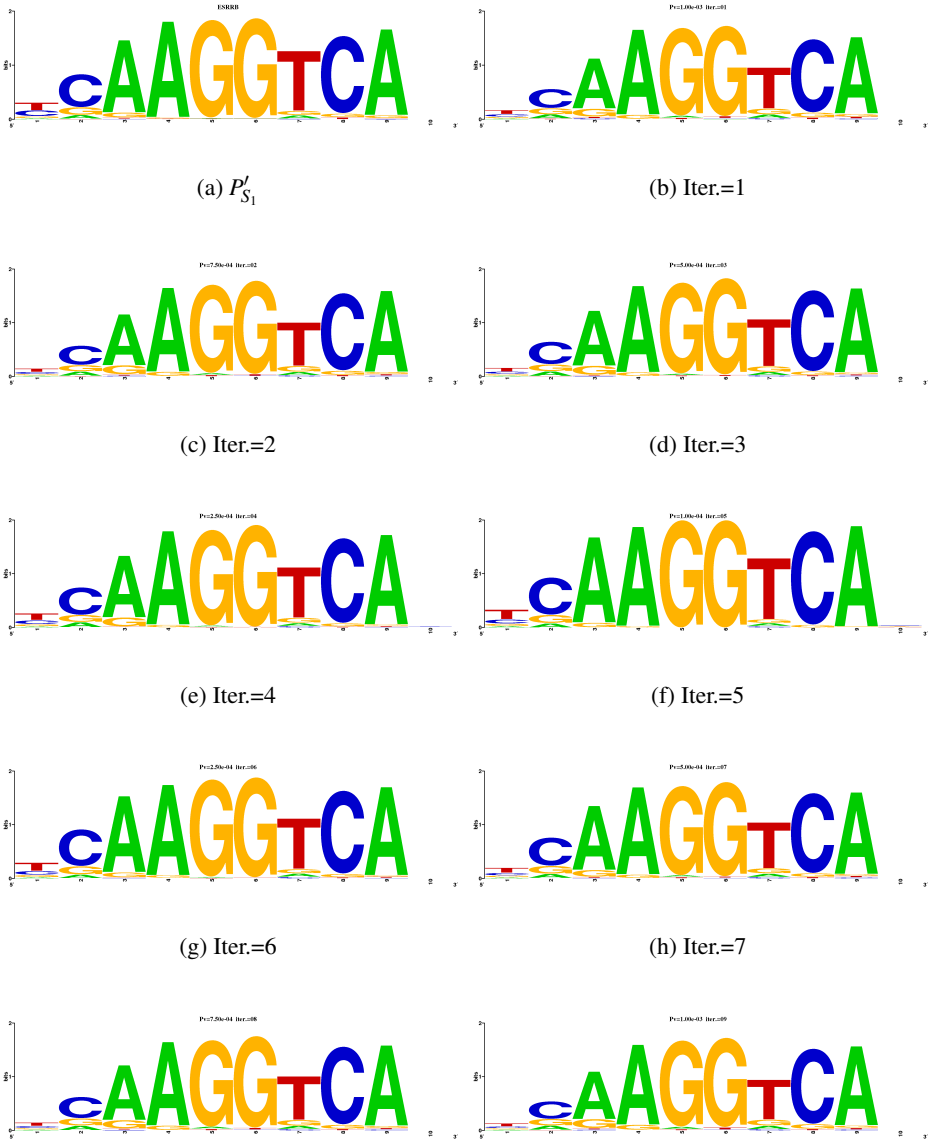


(b) Stat3



(c) Tcfcp

Figure 4.20: Convergence from perturbed Jaspar matrices-ΔIC. Overall decreases in deviation with respect to the unperturbed cases are evident.



(a) P'_{S_1} (b) Iter.=1

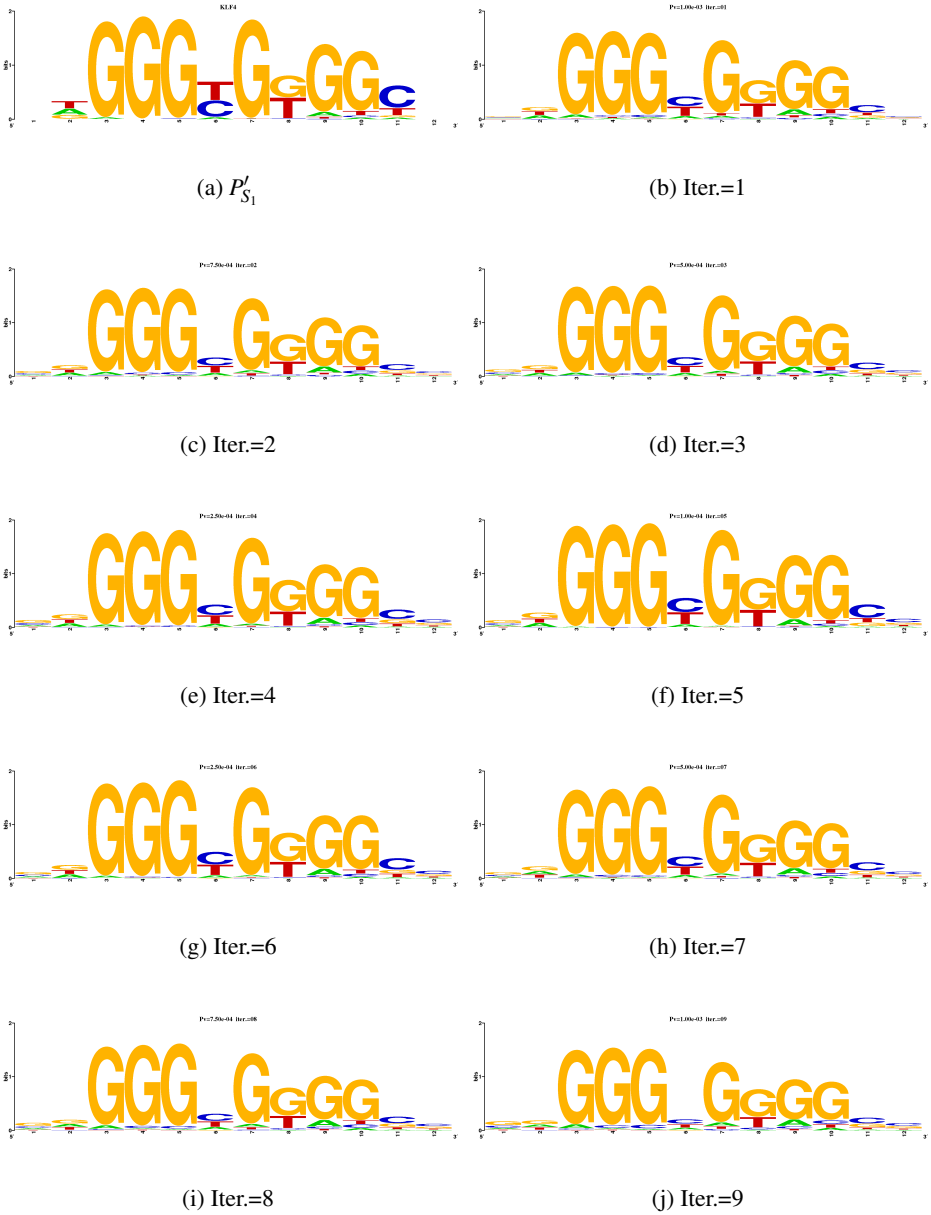
(c) Iter.=2 (d) Iter.=3

(e) Iter.=4 (f) Iter.=5

(g) Iter.=6 (h) Iter.=7

(i) Iter.=8 (j) Iter.=9

Figure 4.21: Esrrb TF binding pattern changes. Using different P-values causes changes in the degree of residue conservations.



(a) P'_{S_1} (b) Iter.=1

(c) Iter.=2 (d) Iter.=3

(e) Iter.=4 (f) Iter.=5

(g) Iter.=6 (h) Iter.=7

(i) Iter.=8 (j) Iter.=9

Figure 4.22: Klf4 TF binding pattern changes. Residues at different positions exhibit different degrees of conservations.

Chapter 5. GENE PRIORITIZATION WITH NEW ASSOCIATIONS

5.1 Introduction

Genes have different levels of relevance and significance in the etiology and pathology of diseases. Gene prioritization algorithms attempt to computationally predict the disease relevance of genes, typically on a genomic scale. Genes that are implicated to be significant by experimental platforms, such as genome wide association study, are often poorly characterized, and their functions and protein interaction partners are unknown. Associations between genes in functional or genetic terms provide important clues in elucidating their ontologies, hence, finding such associations is an important research problem. Most gene prioritization algorithms base their inference on genetic associations. This provides another significance to the problem of finding associations. Genes can be associated with each other in various ways. Expressions of most genes in cells are controlled by transcription factors. Regulatory associations, where a transcription factor as a gene product regulates the expression of its target genes, are an example of such associations. The gene EGR3 is gaining attention in the study of schizophrenia [171]. Regulatory associations involving the gene are inferred in the current study through a computational approach. It is a transcription factor of three C2H2 zinc fingers, and plays important roles in the expressions of early response genes upon environmental stimuli [160]. Inferred associations are then incorporated in the gene prioritization scheme that was introduced in Ch.2, to provide a comprehensive prioritization of the human genes with respect to schizophrenia.

5.2 Approach

Binding patterns of a large number of transcription factors are unknown, which is one of the most limiting factors in the inference of regulatory associations between transcription factors and their targets. Patterns that are reported for a transcription factor in the literature, as represented via letter probability matrices or consensus [150], often exhibit a large discrepancy. Then the challenge of selecting or determining the correct pattern is posed.

While binding patterns typically exhibit a high level of degeneracy and variability, it is still sought to obtain accurate representations of the distributions of functional binding patterns, which are unique characteristics of transcription factors. Given genome sequences, such representations will then facilitate searches for the loci of functional binding patterns with higher levels of sensitivity and specificity. We first use the methods developed in Chs.3 and 4 together to infer putative binding patterns of the EGR3 transcription factor.

Inference of transcription factor binding patterns

The letter probability matrices of EGR3 that were previously obtained in Ch.3 with varying pseudo-Boltzman constants, k's, are shown in Tables 5.1,5.2 and 5.3. Their logo representations are shown in Figs.5.1(a), 5.1(b) and 5.1(c). Decrease in the k value coincides with an increase in the information contents of the patterns.

Binding patterns typically reside in the 5' upstream (promoter) regions of target genes. Upstream region sequences of length 1kbp's of the entire genes from the human genome were obtained. A sequence set enriched in the EGR3 binding patterns is needed to apply the iterative binding pattern improvement method (Ch.4). It is known that the EGR family transcription factors 1,2,3 and 4 all have the same characteristic residues that recognize DNA bases [90]. Given a C2H2 transcription factor, our approach from Ch.3 mainly utilizes the characteristic protein residues of the factor to infer its binding patterns. This implies that the target genes of other EGR members can be used for the training of a classifier for EGR3 targets. The list of known target genes of the EGR transcription factors (EGR+ genes) is shown in Table 5.4, which would harbor the binding patterns that are specific to the EGR family in their promoters. The set of their promoter sequences is far more enriched in the binding patterns of the EGR transcription factors, compared with the sequences that are randomly selected from the genome. The set, then, is also enriched in the binding patterns of the EGR3 transcription factor. This encourages an application of the iterative binding pattern improvement method that was shown in Ch.4. We first use our binding pattern prediction method (Ch.3) in order to obtain a putative EGR3 binding pattern, and apply

the iterative binding pattern improvement method so as to refine its accuracy. Since the search region length of 1kbp's per sequence is rather large, measures against the matches randomly originating from background patterns are necessary, which means a matrix with a rather high information content is desired. So we used the predicted pattern corresponding to the k value, $7e^{-4}$.

The matrix was then used as an initial matrix in the iterative improvement with the EGR promoter set. Figures 5.2(b) to 5.2(j) show the iteration steps. The numbers of the matching sites, genes with a match and the detected unique patterns, varied in accordance with the pattern match significance p-value along the iteration (Table 5.5). We tentatively chose the matrix from iteration 1 corresponding to the p-value $1e^{-4}$. Information contents (IC) of the different positions were as shown in Table 5.6.

Classification of EGR +/- sets

Given the EGR+ gene set, a pseudo negative EGR- gene set was formed by excluding the EGR+ genes from the human genome and then randomly selecting from the remainder the same number of genes as the EGR+ genes. Then, the respective +/- promoter sequence sets were searched for the presence of EGR binding patterns. Distinct scanning results were obtained under different P-values (Table 5.7). Frequently, more than one matching site were found in the promoters of genes. Remarkably, the promoters of 57 EGR- genes have a total of 146 matching sites, while there were 84 EGR+ promoters having 234 sites, at p-value $5e^{-5}$. Hence, it would be very challenging to distinguish the EGR+/- sets apart, solely based on the criteria of binding pattern conformance. Transcription factor binding patterns are mostly located in proximity to the transcription start sites of genes, hence their precise coordinates potentially bear a level of signal that is pertinent to the target gene classification problem. Hence we introduced additional features: the locus of minimum distance to transcription start site (min-loc) out of matching sites of a gene, the farthest locus from transcription start site (max-loc), and the number of matches (sites). We used the median p-value of the matches of each gene as its summary p-value. Scatter plots of possible two

feature combinations are shown in Figs. 5.3(a) to 5.3(f). Support vector machine [159] (SVM) is known to be a method that shows good performance in the classification problems from diverse domains. When the SVM was applied with different kernels, the training performance levels were as in Table 5.8. Considering that the sample size was rather small, and in order to avoid potential overfitting that could result from using complex kernels, the quadratic kernel was chosen.

The promoter sequences from the human genome (S_h) were searched in for the presence of the obtained EGR binding patterns, under different p-values (Table 5.9). At the threshold p-value $5e^{-5}$, there were 8342 genes with promoters harboring one or more matches with the significance levels exceeding the threshold (to be referred to as S'_h set).

When the SVM classifier formed above was applied to S'_h, 6394 genes were classified as EGR+ (the initially classified EGR+ set, or EGR_i^+), and 1948 genes as EGR-. Since the set of genes that were classified to be positives, EGR_i^+, is supposed to have a rather low level of specificity, further measures have to be applied to it, in order to increase the level of the enrichment of EGR3 targets therein. Among the EGR_i^+ genes, some genes showed extremely large numbers of matching sites in their promoters (Table 5.10), which is notable considering that a rather moderate length of promoter region was used together with a strict p-value threshold. Genes with matching sites ≥ 10 were retained to obtain a set of 183 putative target genes (G_t, Table 5.11). The genes have strong signals to bind to the transcription factors of the EGR family, hence they would be highly enriched in the EGR3 targets.

Schizophrenia gene prioritization

Exhaustive associations were formed between every two genes from G_t to form a schizophrenia disease-specific subnetwork. This was merged with the GO0.30 and HNS300 networks from Ch.2, resulting in a single aggregated network. Genome wide association studies (GWAS) for the schizophrenia disease were performed in [156], [157] and [158]. We used the GWAS data from [157] and reflected them to the significance levels of genes using

Eq.2.4 from Ch.2. Allen et al. [7] provides a list of genes implicated in the schizophrenia, which was compiled from the literature. We use the set of 42 genes that were graded A,B and C, as a seed gene set. Then the integrative gene prioritization method of Ch.2 was applied to the human genome with the parameter $k = 1/3$ in Eq.2.9. Out of a total of 9061 genes that had positive scores, the 100 genes with the highest scores are shown in Table 5.12. The top 50 and 100 most significant genes were respectively characterized in terms of over-represented Gene Ontology terms (Tables 5.13 and 5.14) by using [175]. Especially prevalent were terms that are specific to neural cells such as the neural signal transmission.

5.3 Discussion

Here we performed a comprehensive prioritization of human genes with respect to the schizophrenia disease. Notably, a subnetwork specific to the schizophrenia disease was formed by drawing associations between putative target genes of the EGR3, a gene highly relevant to the disease. When the search for the loci of binding patterns for a transcription factor is on a genomic scale, a large number of patterns conforming to a given pattern can occur by random chances. This results in a large number of genes that are not truly regulated by the transcription factor, yet are classified as such. Measures are needed to reduce such a high rate of false positives. Additional features such as the loci of binding patterns were employed, together with a kernel of adequate complexity in the classification scheme, so as to model the inherent structure within the features. Significantly-ranked genes from the prioritization result were highly enriched in Gene Ontology terms that are specific to neural cells, such as neural signal transmissions, and may serve well as interesting genes for further research in the schizophrenia disease. The procedure for target gene prediction was purely computational and maximally utilized the existing data. The data from [86] was derived from a fairly large set of experiments, and encourages computational prediction of the binding patterns for transcription factors. The procedure that was adopted realizes, to a great extent, the potential of such data.

Table 5.1: Calculated EGR binding pattern with $k = 3e^{-3}$

Position	A	C	G	T
0	0.01565242	0.01295830	0.96569470	0.00569458
1	0.18985339	0.49915606	0.08918355	0.22180701
2	0.05936601	0.00487306	0.82638039	0.10938054
3	0.01150165	0.03693489	0.46521054	0.48635292
4	0.14440979	0.01049011	0.84498739	0.00011271
5	0.05936601	0.00487306	0.82638039	0.10938054
6	0.00003541	0.04197646	0.78002908	0.17795904
7	0.18985339	0.49915606	0.08918355	0.22180701
8	0.05936601	0.00487306	0.82638039	0.10938054
9	0.02159215	0.16866006	0.43853550	0.37121229

Table 5.2: Calculated EGR binding pattern with $k = 1e^{-3}$

Position	A	C	G	T
0	0.00000426	0.00000242	0.99999312	0.00000021
1	0.04790999	0.87072318	0.00496621	0.07640062
2	0.00036975	0.00000020	0.99731738	0.00231267
3	0.00000705	0.00023351	0.46660369	0.53315575
4	0.00496679	0.00000190	0.99503130	0.00000000
5	0.00036975	0.00000020	0.99731738	0.00231267
6	0.00000000	0.00015399	0.98811233	0.01173368
7	0.04790999	0.87072318	0.00496621	0.07640062
8	0.00036975	0.00000020	0.99731738	0.00231267
9	0.00007175	0.03419713	0.60112834	0.36460277

Table 5.3: Calculated EGR binding pattern with $k = 7e^{-4}$

Position	A	C	G	T
0	0.00000002	0.00000001	0.99999997	0.00000000
1	0.01515852	0.95472191	0.00059480	0.02952477
2	0.00001255	0.00000000	0.99981525	0.00017220
3	0.00000006	0.00000872	0.45252043	0.54747079
4	0.00051469	0.00000001	0.99948530	0.00000000
5	0.00001255	0.00000000	0.99981525	0.00017220
6	0.00000000	0.00000363	0.99822342	0.00177295
7	0.01515852	0.95472191	0.00059480	0.02952477
8	0.00001255	0.00000000	0.99981525	0.00017220
9	0.00000165	0.01105552	0.66392425	0.32501858

Table 5.4: Target genes of EGR transcription factor family

ID	Name	EGR TF	Source
25	ABL1	1	[162]
207	AKT1	1	[162]
355	FAS	1	[162]
387	RHOA	1	[161]
388	RHOB	1	[161]
467	ATF3	1	[161]
468	ATF4	1	[162]
672	BRCA1	1	[162]
677	ZFP36F1	1	[161]
811	CALR	1	[162]
819	CAMLG	1	[162]
867	CBL	1	[162]
928	CD9	1	[161]
1019	CDK4	1	[162]
1021	CDK6	1	[162]
1326	MAP3K8	1	[162]
1388	TNXB	1	[161]
1397	CSRP2	1	[161]
1445	SRC	1	[161]
1525	CXADR	1	[161]
1612	DAPK1	1	[162]

Continued on next page...

Table 5.4 – Continued

ID	Name	EGR TF	Source
1662	DDX10	1	[162]
1674	DES	1	[161]
1786	DNMT1	1	[161]
1855	DVL1	1	[161]
1938	EEF2	1	[161]
1958	EGR1	1	[161, 162]
2004	ELK3	1	[161]
2073	ERCC5	1	[161]
2119	ETV5	1	[162]
2152	F3	1	[161]
2202	EFEMP1	1	[161]
2253	FGF8	1	[161]
2289	FKBP5	1	[162]
2353	FOS	1	[161]
2523	FUT1	1	[162]
2551	GABPA	1	[162]
2885	GRB2	1	[162]
3065	HDAC1	1	[162]
3066	HDAC2	1	[162]
3075	CFH	1	[161]
3087	HHEX	1	[162]
3159	HMGY	1	[161]

Continued on next page...

Table 5.4 – Continued

ID	Name	EGR TF	Source
3161	HMMR	1	[162]
3339	HSPG2	1	[161]
3397	ID1	1	[161]
3399	ID3	1	[162]
3481	IGF2	1	[161]
3491	CYR61	1	[161]
3725	JUN	1	[161, 162]
3880	KRT19	1	[161]
3945	LDHB	1	[161]
3953	LEPR	1	[161]
4088	SMAD3	1	[162]
4254	KITLG	1	[162]
4299	AFF1	1	[162]
4615	MYD88	1	[162]
4616	GADD45B	1	[162]
4869	NPM1	1	[162]
4881	NPR1	1	[161]
5054	SERPINE1	1	[161]
5108	PCM1	1	[162]
5154	PDGFA	1	[161]
5155	PDGFB	1	[161]
5329	PLAUR	1	[162]

Continued on next page...

Table 5.4 – Continued

ID	Name	EGR TF	Source
5359	PLSCR1	1	[162]
5573	PRKAR1A	1	[161]
5591	PRKDC	1	[162]
5594	MAPK1	1	[162]
5605	MAP2K2	1	[162]
5728	PTEN	1	[165]
5730	PTGDS	1	[161]
5734	PTGER3	1	[161]
5829	PXN	1	[161]
5880	RAC2	1	[161]
5888	RAD51	1	[162]
5889	RAD51C	1	[161]
6095	RORA	1	[161]
6125	RPL5	1	[161]
6134	RPL10	1	[161]
6193	RPS5	1	[161]
6464	SHC1	1	[162]
6598	SMARCB1	1	[162]
6609	SMPD1	1	[162]
6647	SOD1	1	[162]
6667	SP1	1	[162]
6774	STAT3	1	[162]

Continued on next page...

Table 5.4 – Continued

ID	Name	EGR TF	Source
7043	TGFB3	1	[161]
7056	THBD	1	[161]
7057	THBS1	1	[161]
7178	TPT1	1	[162]
7409	VAV1	1	[162]
7422	VEGF	1	[161]
7431	VIM	1	[161]
7533	YWHAH	1	[162]
7538	ZFP36	1	[162]
7803	PTP4A1	1	[161]
8503	PIK3R3	1	[162]
8517	IKBKG	1	[162]
8635	RNASET2	1	[161]
8835	SOCS2	1	[161]
8864	PER2	1	[162]
8887	TAX1BP1	1	[161]
9590	AKAP12	1	[161]
9757	MLL4	1	[162]
9988	DMTF1	1	[162]
10456	HAX1	1	[162]
10915	TCERG1	1	[162]
10957	PNRC1	1	[161]

Continued on next page...

Table 5.4 – Continued

ID	Name	EGR TF	Source
10962	MLLT11	1	[162]
11170	FAM107A	1	[161]
23157	6-Sep	1	[162]
26013	L3MBTL	1	[162]
26959	HBP1	1	[162]
54806	AHI1	1	[162]
55904	MLL5	1	[162]
57591	MKL1	1	[162]
64857	PLEKHG2	1	[162]
79870	BAALC	1	[162]
84324	CIP29	1	[162]
91663	MYADM	1	[162]
114034	TOE1	1	[169]
117178	SSX2IP	1	[162]
139285	FAM123B	1	[162]
64783	RBM15	1	[162]
4099	MAG	2	[172]
4155	MBP	2	[172]
50846	DHH	2	[172]
356	FASL	3	[163]
1959	EGR2	3	[167]
2098	ESD	3	[168]

Continued on next page...

Table 5.4 – Continued

ID	Name	EGR TF	Source
2247	BFGF	3	[170]
2557	GABRA4	3	[166]
4804	P75NTR	3	[164]
9956	HS3ST2	3	[168]
23237	ARC	3	[174]
27074	TSC403	3	[163]
7124	TNF	1,3	[173, 162]
4665	NAB2	1,3	[162, 167]

Table 5.5: Recall counts from IBP (Ch.4) iterations over the EGR+ gene promoter set

N-sites	N-genes	N-patterns	Iter	P-value
261	95	130	01	$1.0e^{-4}$
282	96	131	02	$7.5e^{-5}$
255	91	107	03	$5.0e^{-5}$
190	78	66	04	$2.5e^{-5}$
120	60	30	05	$1.0e^{-5}$
181	73	63	06	$2.5e^{-5}$
248	85	104	07	$5.0e^{-5}$
352	98	154	08	$7.5e^{-5}$
429	102	198	09	$1.0e^{-4}$

Table 5.6: IC values of the EGR3 matrix obtained by applying IBP (Ch.4), first iteration

Position	f^1 residue	f^1	IC
0	g	0.777778	0.922372
1	g	0.950192	1.643711
2	c	0.770115	0.892723
3	g	0.946360	1.624380
4	g	0.942529	1.611300
5	g	0.923372	1.522369
6	g	0.693487	0.760520
7	g	0.934866	1.562416
8	c	0.704981	0.746882
9	g	0.900383	1.425357
Sum	-	8.544061	12.712031

Table 5.7: Numbers of genes matched from EGR+/- sets with varying P-values

N-sites	N-genes	N-patterns	$\pm$	p-value
5579	136	2870	pos	5e-3
3810	135	2184	neg	5e-3
357	105	114	pos	1e-4
210	69	88	neg	1e-4
234	84	61	pos	5e-5
146	57	50	neg	5e-5
94	55	14	pos	1e-5
56	28	11	neg	1e-5

Table 5.8: Training performance of SVM with different kernels

Kernel	Err.rate%	FN	FP	TP	TN	Sensitivity%	Specificity%	F-measure%
Linear	36.1702	2	49	82	8	97.6190	14.0351	24.5417
Quad.	29.0780	16	25	68	32	80.9524	56.1404	66.3010
Poly.	22.6950	9	23	75	34	89.2857	59.6491	71.5187
RBF	22.6950	10	22	74	35	88.0952	61.4035	72.3666

Table 5.9: Number of genes with EGR family TF-matching patterns from the human genome

N-sites	N-genes	N-patterns	p-value
21824	8342	71	5e-5
8280	4276	14	1e-5
4018	2919	7	5e-6
1233	997	1	1e-6

Table 5.10: Distribution of matching sites of genes with p-values $\leq$ threshold

Sites	Genes
1	3739
2	1762
3	1065
4	587
5	374
6	246
7	160
8	96
9	77
10	69
11	38
12	29
13	18
14	15
15	23
16	11
17	4
18	10
19	7
20	2
21	1
22	3
23	4
24	0
25	0
26	1
27	0

Table 5.11: Putative EGR transcription factor family target genes

ID	Name	ID	Name
10815	CPLX1	27087	B3GAT1
5527	PPP2R5C	162494	RHBDL3
26053	AUTS2	375790	AGRN
57338	JPH3	108	ADCY2
7392	USF2	2063	NR2F6
5455	POU3F3	147657	ZNF480
80816	ASXL3	91461	PKDCC
11044	PAPD7	65265	C8orf33
144699	FBXL14	1960	EGR3
51305	KCNK9	8324	FZD7
79832	QSER1	6929	TCF3
56666	PANX2	26040	SETBP1
10846	PDE10A	140688	C20orf112
399664	MEX3D	57666	FBRSL1
64067	NPAS3	79047	KCTD15
196528	ARID2	284207	METRNL
23389	MED13L	63926	ANKRD5
222389	BEND7	26173	INTS1
57134	MAN1C1	8239	USP9X
4858	NOVA2	6314	ATXN7
27161	EIF2C2	80145	THOC7

Continued on next page...

Table 5.11 – Continued

ID	Name	ID	Name
359948	IRF2BP2	51317	PHF21A
8912	CACNA1H	6874	TAF4
10000	AKT3	2850	GPR27
4010	LMX1B	3363	HTR7
2776	GNAQ	1454	CSNK1E
170394	PWWP2B	83855	KLF16
140862	ISM1	727800	RNF208
23774	BRD1	9394	HS6ST1
55061	SUSD4	23170	TTLL12
84733	CBX2	9693	RAPGEF2
255783	PRR24	11193	WBP4
140730	RIMS4	58489	FAM108C1
340529	PABPC1L2A	90	ACVR1
375056	MIA3	221061	FAM171A1
9382	COG1	246175	CNOT6L
3784	KCNQ1	57446	NDRG3
23118	TAB2	64864	RFX7
388336	SHISA6	84894	LINGO1
89853	FAM125B	154215	NKAIN2
4784	NFIX	79718	TBL1XR1
222553	SLC35F1	79145	CHCHD7
9612	NCOR2	5324	PLAG1

Continued on next page...

Table 5.11 – Continued

ID	Name	ID	Name
57524	CASKIN1	729830	FAM160A1
4150	MAZ	100290519	LOC100290519
1002	CDH4	148479	PHF13
84133	ZNRF3	343472	BARHL2
6256	RXRA	1459	CSNK2A2
109	ADCY3	161882	ZFPM1
157922	CAMSAP1	64976	MRPL40
100133142	LOC100133142	23152	CIC
7528	YY1	266722	HS6ST3
23023	TMCC1	9672	SDC3
79774	GRTP1	114815	SORCS1
202018	TAPT1	23129	PLXND1
9734	HDAC9	2887	GRB10
152	ADRA2C	23543	RBM9
353116	RILPL1	92714	ARRDC1
84961	FBXL20	134957	STXBP5
6497	SKI	27092	CACNG4
23462	HEY1	2736	GLI2
100132074	FOXO6	79789	CLMN
9715	FAM131B	10801	SEPT9
5782	PTPN12	25817	FAM19A5
94032	CAMK2N2	60678	EEFSEC

Continued on next page...

Table 5.11 – Continued

ID	Name	ID	Name
56899	ANKS1B	113000	RPUSD1
3767	KCNJ11	3953	LEPR
5522	PPP2R2C	8863	PER3
23313	C22orf9	55160	ARHGEF10L
57479	PRR12	10498	CARM1
4644	MYO5A	3064	HTT
3480	IGF1R	10320	IKZF1
3749	KCNC4	644246	LOC644246
79364	ZXDC	284058	KIAA1267
23359	FAM189A1	342865	VSTM2B
11122	PTPRT	678	ZFP36L2
2817	GPC1	64109	CRLF2
4248	MGAT3	57118	CAMK1D
4756	NEO1	2894	GRID1
22880	MORC2	23513	SCRIB
23284	LPHN3	2010	EMD
4325	MMP16	8535	CBX4
9969	MED13	10273	STUB1
57621	ZBTB2	55274	PHF10
1979	EIF4EBP2	126567	C2CD4C
57584	ARHGAP21	1000	CDH2
84376	HOOK3	79772	MCTP1

Continued on next page...

Table 5.11 – Continued

ID	Name	ID	Name
92	ACVR2A	57593	EBF4
29072	SETD2	55323	LARP6
9874	TLK1	7468	WHSC1
6324	SCN1B	286	ANK1
57554	LRRC7	-	-

Table 5.12: Most significantly ranked 100 genes in schizophrenia disease

ID	Name	Score	N-edges	$\sum GWAS$
627	BDNF	40.5482	28	37.645
773	CACNA1A	38.0438	27	33.131
3553	IL1B	37.5542	27	31.662
1812	DRD1	37.3631	26	34.089
7124	TNF	37.0937	27	30.281
1813	DRD2	36.9797	26	32.939
6622	SNCA	36.5542	26	31.662
1141	CHRNB2	36.2898	26	30.869
1814	DRD3	36.1641	26	30.492
3458	IFNG	35.9281	26	29.784
5663	PSEN1	35.3427	25	31.028
3064	HTT	35.2898	25	30.869
3717	JAK2	35.1683	25	30.505
3479	IGF1	34.9281	25	29.784
324	APC	34.9280	25	29.784
857	CAV1	34.9151	25	29.745
538	ATP7A	34.9151	25	29.745
1815	DRD4	34.5808	24	31.743
6531	SLC6A3	34.2485	24	30.746
9370	ADIPOQ	33.9511	24	29.853

Continued on next page...

Table 5.12 – Continued

ID	Name	Score	N-edges	$\sum GWAS$
5027	P2RX7	33.9280	24	29.784
7042	TGFB2	33.7719	24	29.316
1020	CDK5	33.6304	24	28.891
1906	EDN1	33.6175	24	28.852
1136	CHRNA3	33.6174	24	28.852
7040	TGFB1	33.2889	24	27.867
596	BCL2	32.6636	23	28.991
135	ADORA2A	32.6174	23	28.852
1956	EGFR	32.4613	23	28.384
186	AGTR2	32.4613	23	28.384
5743	PTGS2	32.4063	23	28.219
6647	SOD1	32.3930	23	28.179
183	AGT	32.3429	23	28.029
5021	OXTR	32.2889	23	27.867
100	ADA	32.2659	23	27.798
154	ADRB2	32.2498	23	27.750
6868	ADAM17	32.1327	23	27.398
5468	PPARG	32.1002	23	27.301
100133941	CD24	32.0155	23	27.046
5578	PRKCA	31.9913	23	26.974
2185	PTK2B	31.9783	23	26.935
6648	SOD2	31.9560	23	26.868

Continued on next page...

Table 5.12 – Continued

ID	Name	Score	N-edges	$\sum GWAS$
2903	GRIN2A	31.9013	23	26.704
7054	TH	31.7998	23	26.399
3643	INSR	31.7719	22	29.316
207	AKT1	31.7719	22	29.316
595	CCND1	31.5766	22	28.730
2064	ERBB2	31.5257	22	28.577
7533	YWHAH	31.5245	22	28.573
7157	TP53	31.3501	22	28.050
7248	TSC1	31.3198	22	27.959
4092	SMAD7	31.3198	22	27.959
7043	TGFB3	31.2498	22	27.750
27185	DISC1	31.2375	22	27.713
3630	INS	31.2186	22	27.656
351	APP	31.0812	21	30.243
1855	DVL1	31.0623	22	27.187
1312	COMT	31.0558	22	27.167
3611	ILK	30.9913	22	26.974
2898	GRIK2	30.9913	22	26.974
2066	ERBB4	30.9047	21	29.714
4803	NGF	30.8683	22	26.605
2149	F2R	30.7049	22	26.115
811	CALR	30.6807	22	26.042

Continued on next page...

Table 5.12 – Continued

ID	Name	Score	N-edges	$\sum GWAS$
3570	IL6R	30.6807	22	26.042
4035	LRP1	30.6407	21	28.922
5467	PPARD	30.6239	22	25.872
2247	FGF2	30.4742	21	28.423
2690	GHR	30.4613	21	28.384
4846	NOS3	30.4493	22	25.348
3162	HMOX1	30.3867	22	25.160
4842	NOS1	30.3541	22	25.062
10371	SEMA3A	30.2997	20	30.899
9463	PICK1	30.2776	22	24.833
1816	DRD5	30.2531	21	27.759
3667	IRS1	30.1867	21	27.560
6934	TCF7L2	30.1636	21	27.491
23411	SIRT1	30.1636	21	27.491
5194	PEX13	30.1625	21	27.488
5970	RELA	30.1327	21	27.398
6507	SLC1A3	30.1199	21	27.360
348	APOE	30.0954	21	27.286
6777	STAT5B	30.0772	21	27.232
43	ACHE	30.0625	21	27.187
6532	SLC6A4	30.0156	21	27.047
2730	GCLM	30.0078	22	24.023

Continued on next page...

Table 5.12 – Continued

ID	Name	Score	N-edges	$\sum GWAS$
960	CD44	29.9913	21	26.974
51738	GHRL	29.9913	21	26.974
4763	NF1	29.9913	21	26.974
5025	P2RX4	29.9542	21	26.863
4929	NR4A2	29.9433	21	26.830
5590	PRKCZ	29.9209	21	26.763
552	AVPR1A	29.9209	21	26.763
7046	TGFBR1	29.8351	21	26.505
8651	SOCS1	29.8237	20	29.471
7057	THBS1	29.8221	21	26.466
6011	GRK1	29.8221	21	26.466
6850	SYK	29.7962	21	26.388
8877	SPHK1	29.7666	21	26.300
3356	HTR2A	29.6942	21	26.083

Table 5.13: Gene Ontology term enrichment in top N=50 significant genes

Term	Description	Pvalue
GO:0051952	Regulation of amine transport	3.08E-4
GO:0014059	Regulation of dopamine secretion	3.63E-4
GO:0051969	Regulation of transmission of nerve impulse	7.48E-4
GO:0050804	Regulation of synaptic transmission	7.48E-4
GO:0032225	Regulation of synaptic transmission,dopaminergic	9.35E-4

Table 5.14: Gene Ontology term enrichment in top N=100 significant genes

Term	Description	Pvalue
GO:0042391	Regulation of membrane potential	3.62E-6
GO:0051952	Regulation of amine transport	2.15E-5
GO:0019725	Cellular homeostasis	2.61E-5
GO:0032225	Regulation of synaptic transmission, dopaminergic	8.87E-5
GO:0006873	Cellular ion homeostasis	9.62E-5
GO:0055082	Cellular chemical homeostasis	9.62E-5
GO:0050801	Ion homeostasis	9.62E-5
GO:0051940	Regulation of catecholamine uptake involved in synaptic transmission	2.14E-4
GO:0051584	Regulation of dopamine uptake	2.14E-4
GO:0051580	Regulation of neurotransmitter uptake	2.14E-4
GO:0050804	Regulation of synaptic transmission	2.19E-4
GO:0014059	Regulation of dopamine secretion	2.45E-4
GO:0032880	Regulation of protein localization	2.85E-4
GO:0051588	Regulation of neurotransmitter transport	3.31E-4
GO:0051969	Regulation of transmission of nerve impulse	3.55E-4
GO:0007628	Adult walking behavior	3.71E-4
GO:0031644	Regulation of neurological system process	5.57E-4
GO:0010638	Positive regulation of organelle organization	8.27E-4
GO:0048148	Behavioral response to cocaine	8.97E-4

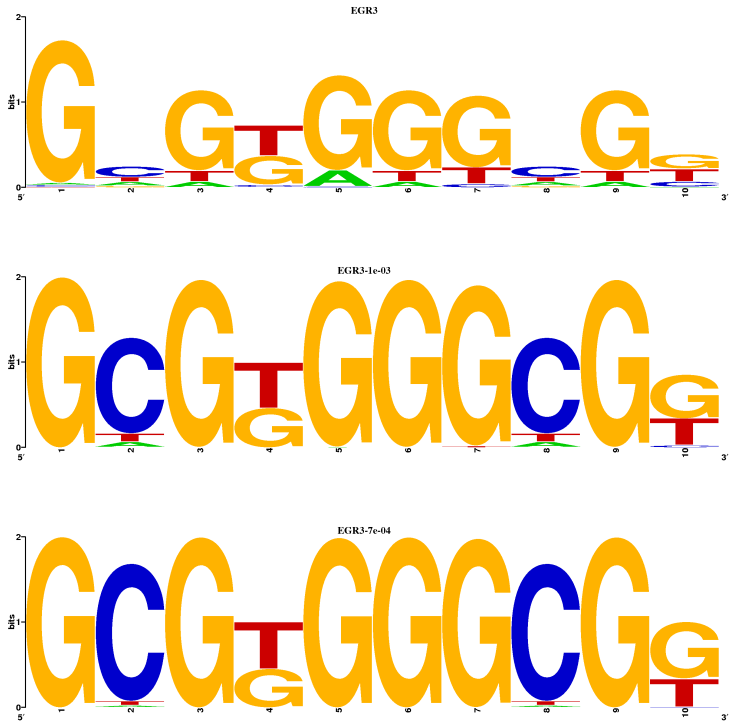
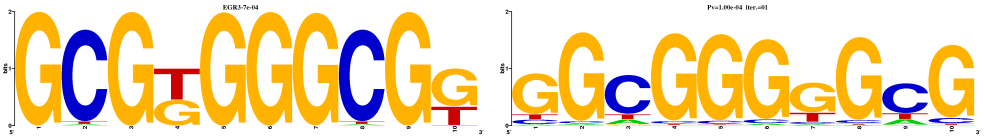
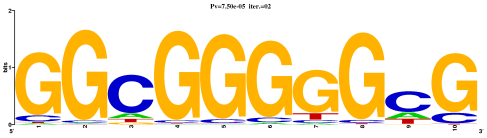


Figure 5.1: Changes in the calculated EGR3 binding patterns with varying k values. Smaller k value coincides with higher information contents of the patterns.

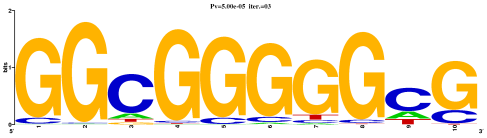


(a) Iter.=0

(b) Iter.=1

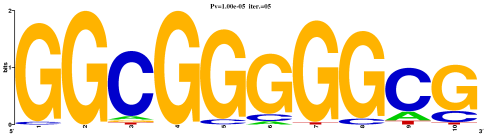


(c) Iter.=2



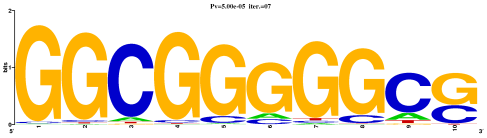
(d) Iter.=3

(e) Iter.=4

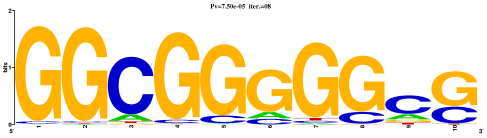


(f) Iter.=5

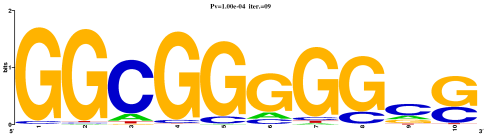
(g) Iter.=6



(h) Iter.=7



(i) Iter.=8



(j) Iter.=9

Figure 5.2: Application of IBP method (Ch.4) to EGR binding pattern. Initial pattern and the patterns resulting from iteration. Sequence set was the 1kbp promoter regions of the known target genes.

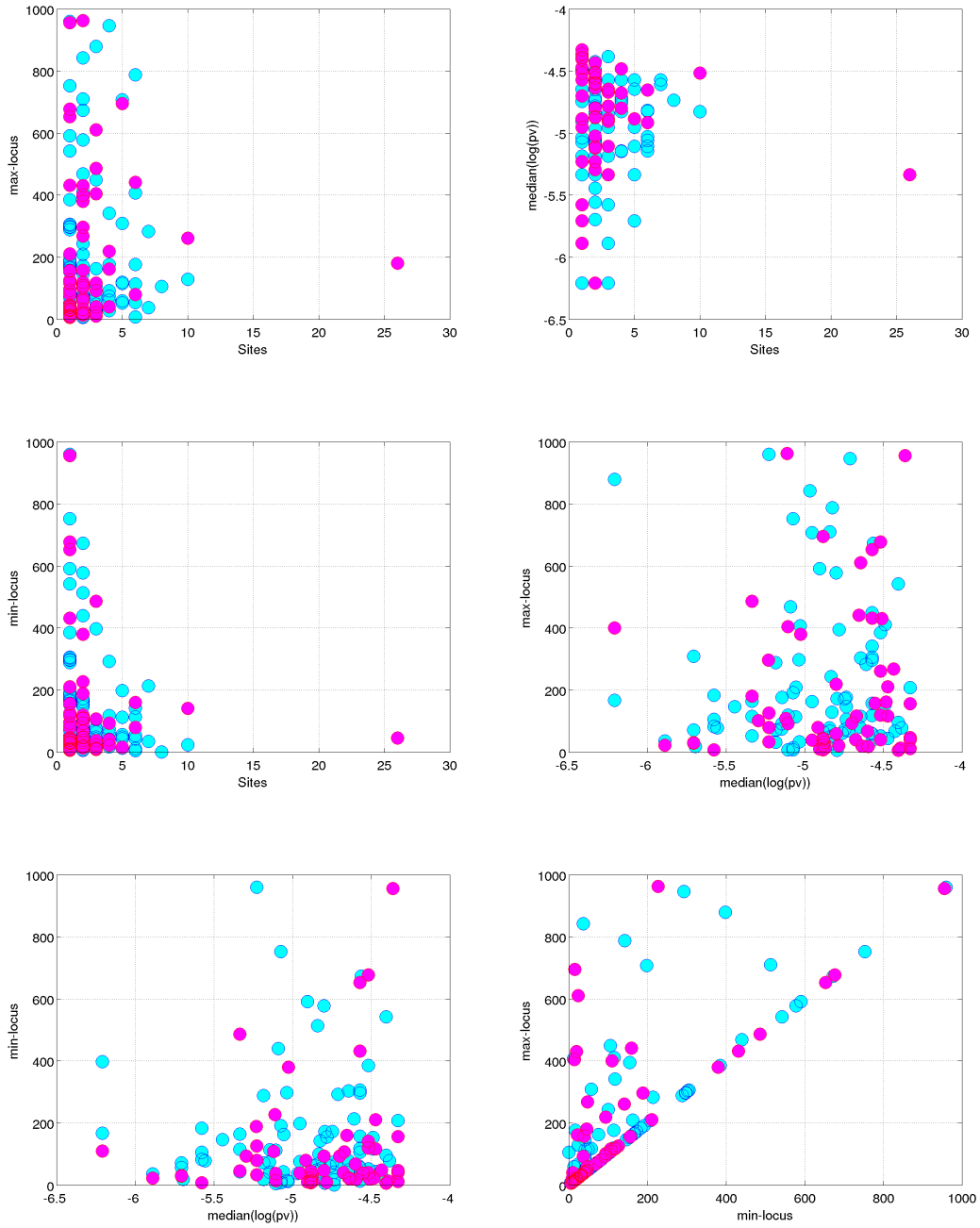


Figure 5.3: Scatter plots of feature combinations. The positive examples (cyan) do not clearly separate from the negative ones (magenta circles).

The levels of relevance and significance of genes in the causes and developments of diseases are widely varying. Focusing on the most relevant ones first helps economize the efforts to prevent and cure diseases. Gene prioritization algorithms rank genes by their disease relevance and provide a means for data reduction, i.e., only the genes above a certain threshold rank can be retained for a more focused research. When the gene prioritization is formulated as a computational problem, algorithms rely, to great extents, on the associations between genes in translating the significance of seed genes to the candidate genes to be ranked.

Distinct data have different levels of reliability and relevance to a disease. In Ch.2, we developed a method to discern and model such aspects of data. Comparisons with two well-known algorithms demonstrated that our method exceeds them in performance. The method was shown to be robust to small inaccuracies that can be present in data, via random perturbation experiments.

Transcription factor-DNA interactions typically consist of many-to-many residue interactions. The C2H2 zinc finger transcription factors, on the other hand, mainly use one-to-one amino acid-nucleotide interactions in their interactions with DNA, which yielded quite a tractable interaction model. The amino acid-nucleotide residue propensity tables are critical components in putting the model to practical use, such as predicting the binding patterns of transcription factors. The sets of tables from [87] and [100] are two well-known examples. In Ch.3, we compared the predicted patterns that were obtained by using them, against the set of patterns from Jaspar [136] that were presumed as standards. Overall, the patterns obtained by using [100] showed lower levels of conservations across positions and transcription factors. Those obtained by using [87] were closer to the Jaspar patterns. While it is important to obtain high quality data of propensities between residues, it was seen to be equally important to accurately group the zinc fingers of a transcription factor into ap-

propriate DNA-interacting modules. Additionally, it was seen to be necessary to predict the deviations from the canonical model in the participation of residues. We studied the problem of improving transcription factor binding patterns, that are quite inaccurate, by using the enrichment of true bound patterns in particular types of data such as ChIP-seq, in Ch.4. The reliability and robustness of the developed method were shown via the convergence of widely varying initial patterns to nearly identical series of patterns. Patterns obtained with the method showed far higher levels of sensitivity at identical levels of specificity, compared with those obtained with raw initial patterns, as they were applied to the transcription factor-bound data from [107].

Biological networks are far sparser than necessary to achieve a reliable performance level for most gene prioritization algorithms. In Ch.5, we comprehensively utilized the methods for gene prioritization, target pattern prediction and improvement, to prioritize human genes with respect to the schizophrenia disease, on a genomic scale. A genetic association network specific to the disease was built. Genes with significant ranks showed enrichment of the Gene Ontology [12] terms that are specific to neural cells, which can serve as interesting lead genes in schizophrenia research.

There often exist, in the literature, different sets of genes that are believed to be significant in a disease. It would be interesting to explore how different gene prioritization results can be, if different input gene sets are used. Frequently, different transcription factors have the same or similar set of characteristic residues that mediate their interactions with DNA. Then naturally arising is the question on the specificity of their target recognition - whether they would bind the same targets or different ones. If targets are different, then a number of potentially relevant factors may be cited, such as nucleosome modification signals or combinatorial regulations of transcription factors. Combinatorial regulations are quite common, especially in higher organisms, where multiple transcription factors cooperatively regulate the expression of their targets. If specificity is conferred by the difference in the combination of transcription factors, it would be reflected via the presence of patterns in the promoter regions of targets, which correspond to the different factors. Such difference

in the presence of patterns may be utilized in further refining computationally predicted, putative target sets.

The significance of pattern matching was mostly assessed by using the measure of p-value, which was assumed to have a strong correlation with the binding energy. Choosing an appropriate threshold p-value in classifying patterns as matching or not matching typically entails a tradeoff between sensitivity and specificity. The problem of the p-value selection was especially conspicuous when multiple threshold p-values had to be selected in a series, where the impact of selection at preceding steps would propagate down the iterative process. While they were determined on a rather empirical basis in the present study, a more systematic method for determining the threshold p-values would be interesting to pursue in the future.

BIBLIOGRAPHY

[1] Jia P, Zheng S, Long J, Zheng W, and Zhao Z. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics, November 2, 2010.

[2] Sun J, Jia P, Fanous AH, van den Oord EJCG, Chen X, Riley BP, Amdur RL, Kendler KS, Zhao Z. Schizophrenia gene networks and pathways and their applications for novel candidate gene selection. PLoS ONE 5(6):e11351 (2010).

[3] Guo AY, Sun J, Jia P, Zhao Z. A novel microRNA and transcription factor mediated regulatory network in schizophrenia. BMC Systems Biology 4:10 (2010).

[4] Sun J, Jia P, Fanous AH, Webb BT, van den Oord EJCG, Kendler KS, Zhao Z. A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases Schizophrenia as a case. Bioinformatics. 25(19):2595-2602 (2009).

[5] Sun J, Kuo PH, Riley BP, Kendler KS, Zhao Z (2008) Candidate genes for schizophrenia: a survey of association studies and gene ranking. American Journal of Medical Genetics B: Neuropsychiatric Genetics 147B:1173-1181.

[6] Raychaudhuri, S., Plenge, R. M., Rossin, E. J., Ng, A. C., Purcell, S. M., Sklar, P., Scolnick, E. M., Xavier, R. J., Altshuler, D. and Daly, M. J. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet 5, e1000534 (2009).

[7] Allen NC, Bagade S, McQueen MB, Ioannidis JPA, Kavvoura FK, Khoury MJ, Tanzi RE, Bertram L "Systematic Meta-Analyses and Field Synopsis of Genetic Association Studies in Schizophrenia: The SzGene Database" Nat Genet 40(7): 827-34 (2008).

[8] Tuerk, C. and Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science, 249(4968):505-510 (1990).

[9] Smith, GP. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science, 228 (4705), 1315-1317 (1985).

[10] R. Abraham et al. A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. BMC Med Genomics, 1:44 (2008).

[11] S. Aerts et al. Gene prioritization through genomic data fusion. Nature Biotechnology, 24(5), 537-44 (2006).

[12] M. Ashburner et al. Gene ontology: tool for the unification of biology. Nat. Genet, 25, 25-29 (2000).

[13] J. Chen, B. J. Aronow and A. G. Jegga. Disease candidate gene identification and prioritization using protein interaction networks. BMC bioinformatics, 10 (2008).

[14] Couto et al. Implementation of a functional similarity measure between gene-products. Technical report DI/FCUL TR 03-29 (2003).

[15] T. de Bie et al. Kernel-based data fusion for gene prioritization. Bioinformatics, 23(13), i25-32 (2007).

[16] L. Ein-Dor et al. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics, 21, 171-178 (2005).

[17] G. Gonzalez et al. Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. Pacific symposium on biocomputing (2007).

[18] G.T. Hart et al. How complete are current yeast and human protein-interaction networks? Genome Biol., 7, 120 (2006).

[19] D. Harold et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet, 41(10), 1088-93 (2009).

[20] ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/.

[21] S. Kohler et al. Walking the interactome for prioritization of candidate disease genes. the American journal of human genetics, 82(4), 949-58 (2008).

[22] Y. Li et al. Integration of multiple data sources to prioritize candidate genes using discounted rating system. BMC Bioinformatics, 11 Suppl 1:S20 (2010).

[23] H. Li et al. Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. Arch Neurol, 65(1), 45-53 (2008).

[24] J. Lambert et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet, 41(10), 1094-9 (2009).

[25] C. von Mering et al. String 7 - recent developments in he integration and prediction of protein interactions. Nucleic acids research, 35, D358-62 (2003).

[26] NCBI ftp://ftp.ncbi.nih.gov/gene/GeneRIF/interactions.gz (2010).

[27] G. Ostlund et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res., 38, D196-203 (2010).

[28] M. Oti et al. Predicting disease genes using protein-protein interactions. Journal of medical genetics, 43(8), 691-8 (2006).

[29] T.S.K. Prasad et al. Human protein reference database - 2009 update. Nucleic acids research, 37, D767-72 (2009).

[30] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. IJCAI (1995).

[31] A. Sharma et al. Gene prioritization in type 2 diabetes using domain interactions and network analysis. BMC genomics, 11 :84 (2010).

[32] J.A. Webster et al. Genetic control of human brain transcript expression in Alzheimer disease. American journal of human genetics, 84(4), 445-58 (2009).

[33] E. Reiman et al. GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. Neuron, 54(5), 713-20 (2007).

[34] P. Kramer et al. Alzheimer disease pathology in cognitively healthy elderly: A genome-wide study. Neurobiol Aging, May 6 (2010).

[35] S. Frantz. An array of problems. Nat. Rev. Drug Discov., 4, 362-363 (2005).

[36] G. L. Miklos and R. Maleszka. Microarray reality checks in the context of a complex disease. Nat. Biotechnol., 22, 615-621 (2004).

[37] E. Marshall. Getting the noise out of gene arrays. Science, 306, 630-631 (2004).

[38] D. Yang. Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. Bioinformatics, 24, 265-271 (2008).

[39] B. Linghu et al. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol. 10 (9):R91 (2009).

[40] S. Navlakha and C. Kingsford. The power of protein interaction networks for associating genes with diseases, Bioinformatics, 26(8), 1057-63 (2010).

[41] S. Karni et al. (2009) A network-based method for predicting disease-causing genes, Journal of computational biolgy, 16(2), 181-9.

[42] L. Bertram et al. (2007) Systematic meta-analyses of Alzheimer disease genetic association studies: the Alzgene database, Nature genetics, 39(1), 17-23.

[43] A. Maayan et al. (2008) Network integration and graph analysis in mammalian molecular systems biology, IET systems biology, 2(5), 206-21.

[44] O. Vanunu et al. (2010) Associating gens and protein complexes with disease via network propagation, Plos computational biology, 6(1).

[45] X. Ma et al. (2007) CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data, Bioinformatics, 23(2), 215-21.

[46] NOLTE, R., et al. Differing roles for zinc fingers in DNA recognition: Structure of a six-finger transcription factor IIIA complex, Proc. Natl. Acad. Sci. Vol. 95, pp. 2938-2943, March 1998.

[47] Itai Yanai. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21, 5: 650-659 (2005).

[48] R. Poirier. Distinct functions of egr gene family members in cognitive processes. Front Neurosci. 2008 Jul;2(1):47-55. Epub 2008 Jul 7.

[49] AH Swirnoff and J Milbrandt. DNA-binding specificity of NGFI-A and related zinc finger transcription factors. Mol. Cell. Biol., Apr 1995, 2275-2287, Vol 15, No. 4.

[50] S. Collins. Opposing regulation of T cell function by Egr-1/NAB2 and Egr-2/Egr-3. European Journal of Immunology, Volume 38, Issue 2, pages 528-536, February 2008.

[51] P. Zipfel et al. The human zinc finger protein EGR-4 acts as autoregulatory transcriptional repressor. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression. Volume 1354, Issue 2, 1 November 1997, Pages 134-144.

[52] P. Chavrier et al. EMBO J. 9 (1990), pp. 1209-1218.

[53] P. Lemaire et al. Mol. Cell. Biol. 10 (1990), pp. 3456-3467.

[54] Alia Al-Sarraj. Specificity of transcriptional regulation by the zinc finger transcription factors Sp1, Sp3, and Egr-1. J Cell Biochem. 2005 Jan 1;94(1):153-67.

[55] G.D. Wieland. Early growth response proteins EGR-4 and EGR-3 interact with immune inflammatory mediators NF-kappaB p50 and p65. J. Cell. Sci. (2005).

[56] Kumbrink J. EGR1, EGR2, and EGR3 activate the expression of their coregulator NAB2 establishing a negative feedback loop in cells of neuroectodermal and epithelial origin. J Cell Biochem. 2010 May 12.

[57] Yamagata K. Egr3/Pilot, a zinc finger transcription factor, is rapidly regulated by activity in brain neurons and colocalizes with Egr1/zif268. Learn Mem. 1994 Jul-Aug;1(2):140-52.

[58] Safford M et al. Egr-2 and Egr-3 are negative regulators of T cell activation. Nat Immunol 2005; 6: 472-480.

[59] X. Gao. Regulation of low affinity neurotrophin receptor (p75NTR) by early growth response (Egr) transcriptional regulators. Mol Cell Neurosci. 2007 December ; 36(4): 501-514.

[60] W. Tourtellotte et al. Functional Compensation by Egr4 in Egr1-Dependent Luteinizing Hormone Regulation and Leydig Cell Steroidogenesis. Mol Cell Biol. 2000 July; 20(14): 52615268.

[61] L. Li et al. Egr3, a synaptic activity regulated transcription factor that is essential for learning and memory. Mol Cell Neurosci. 2007 May;35(1):76-88.

[62] K. Yamada et al. Genetic analysis of the calcineurin pathway identifies members of the EGR gene family, specifically EGR3, as potential susceptibility candidates in schizophrenia. Proc Natl Acad Sci U S A. 2007 Feb 20;104(8):2815-20.

[63] A. Inoue. Transcription factor EGR3 is involved in the estrogen-signaling pathway in breast cancer cells. J Mol Endocrinol. 2004 Jun;32(3):649-61.

[64] L. Eldredge et al. Abnormal sympathetic nervous system development and physiological dysautonomia in Egr3-deficient mice. Development 135, 2949-2957. 2008.

[65] D. S. Roberts. Egr3 stimulation of GABRA4 promoter activity as a mechanism for seizure-induced up-regulation of GABAA receptor 4 subunit expression. PNAS August 16, 2005 vol. 102 no. 33 11894-11899.

[66] V. Lazarevic et al. The gene encoding early growth response 2, a target of the transcription factor NFAT, is required for the development and maturation of natural killer T cells. Nature Immunology 10, 306 - 313 (2009).

[67] Briani C. Adult onset Charcot-Marie-Tooth disease type 1D with an Arg381Cys mutation of EGR2. Muscle Nerve. 2010 Jun;41(6):888-9.

[68] Q. Wu et al. MiR-150 promotes gastric cancer proliferation by negatively regulating the pro-apoptotic gene EGR2. Biochem Biophys Res Commun. 2010 Feb 12;392(3):340-5.

[69] Mager GM. Active gene repression by the Egr2.NAB complex during peripheral nerve myelination. J Biol Chem. 2008 Jun 27;283(26):18187-97

[70] M. Fu et al. Egr1 target genes in human endothelial cells identified by microarray analysis. Gene. 2003 Oct 2;315:33-41.

[71] Virolle T., et al. The Egr-1 transcription factor directly activates PTEN during irradiation-induced signalling. Nat Cell Biol., 2001 Dec;3(12):1124-8.

[72] M. U. Ehrengruber et al. Modulation of early growth response (EGR) transcription factor-dependent gene expression by using recombinant adenovirus. Gene. 2000 Nov 27;258(1-2):63-9.

[73] de Belle I et al, In vivo cloning and characterization of a new growth suppressor protein TOE1 as a direct target gene of Egr1 , Journal of Biological Chemistry , 2003

[74] James et al. Genomic profiling of the neuronal target genes of the plasticity-related transcription factor-Zif268. J Neurochem 95:796-810 (2005) .

[75] Ishikawa , Early growth response gene-1 plays a pivotal role in down-regulation of a cohort of genes in uterine leiomyoma , Journal of Molecular Endocrinology (2007) 39 333-341

[76] A. Kubosaki et al. Genome-wide investigation of in vivo EGR-1 binding sites in monocytic differentiation. Genome Biol. 2009;10(4):R41.

[77] J. M. Fustin et al. Egr1 involvement in evening gene regulation by melatonin. FASEB J. 2009 Mar;23(3):764-73.

[78] T Deguchi. Expression patterns of the Egr1 and Egr3 genes during medaka embryonic development. Gene Expr Patterns. 2009 Apr;9(4):209-14.

[79] F. Mora-Lpez et al. Transcription of PRDM1, the master regulator for plasma cell differentiation, depends on an SP1/SP3/EGR-1 GC-box. Eur J Immunol 2008, 38:2316-2324.

[80] Christy B, Nathans D (1989). DNA binding site of the growth factor-inducible protein Zif268. Proc. Natl. Acad. Sci. U.S.A. 86 (22): 8737-41.

[81] K Nose. Functional activation of the egr-1 (early growth response-1) gene by hydrogen peroxide. Biochem. J. (1996) 316, 381 - 383.

[82] Raychowdhury et al. Interaction of Early Growth Response Protein 1 (Egr-1), Specificity Protein 1 (Sp1), and Cyclic Adenosine 3'5'-Monophosphate Response Element Binding Protein (CREB) at a Proximal Response Element Is Critical for Gastrin-Dependent Activation of the Chromogranin A Promoter. Mol Endocrinol. 2002 Dec;16(12):2802-18.

[83] G. Bai et al. Cloning and analysis of the 5 flanking sequence of the rat N-methyl-d-aspartate receptor 1 (NMDAR1) genestar , Biochimica et Biophysica Acta (BBA) - Biomembranes , Volume 1152, Issue 1, 10 October 1993, Pages 197-200.

[84] A Dinkel, Transcription factor Egr-1 activity down-regulates Fas and CD23 expression in B cells. The Journal of Immunology, Vol 159, Issue 6 2678-2684, 1997.

[85] A. B. James. Transfected Zif268 reduced the activity of psmb9, SGK, and Tap1 promoterreporter constructs. Regulation of the Neuronal Proteasome by Zif268 (Egr1)

[86] P.V. Benos et al. Additivity in protein-DNA interactions: how good an approximation is it?. Nucleic Acids Res., 30, 4442-4451 (2002).

[87] P.V. Benos et al. Probabilistic code for DNArecognition by proteins of the EGR family. J. Mol. Biol., 323, 701-727(2002).

[88] Persikov AV, Osada R, Singh M. Predicting DNA recognition by Cys2His2 zinc finger proteins. Bioinformatics. Jan 1;25(1):22-9 (2009).

[89] Suzuki,M. et al. DNA recognition code of transcription factors. Protein Eng., 8, 319-328 (1995).

[90] Alex Bateman et al. The Pfam protein families database. Nucl. Acids Res. (2004) 32 (suppl 1): D138-D141.

[91] Tupler R. et al. Expressing the human genome. Nature, 409, 832-833 (2001).

[92] Messina, David N., et al. An ORFeome-based Analysis of Human Transcription Factor Genes and the Construction of a Microarray to Interrogate Their Expression. Genome Res., 14: 2041-2047 (2004).

[93] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860-921 (2001).

[94] Benos P. et al. SAMIE: statistical algorithm for modeling interaction energies. Pac Symp Biocomput. 2001:115-26.

[95] Iuchi S. Three classes of C2H2 zinc finger proteins. Cell Mol Life Sci, Apr;58(4):625-35 (2001).

[96] Siggers T and Honig B. Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. Nucleic Acids Res, 35(4): 1085-1097 (2007).

[97] Lam K et al. Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. Nucl. Acids Res. Feb 14, (2011).

[98] Ding G et al. SysZNF: the C2H2 zinc finger gene database. Nucleic Acids Res, Jan; 37: D267-D273 (2009).

[99] Ramirez CL et al. Unexpected failure rates for modular assembly of engineered zinc fingers. Nat. Methods; 5:374-375 (2008).

[100] Kaplan et al. Ab initio prediction of transcription factor targets using structural knowledge. PLoS Comput Biol. 2005 Jun;1(1):e1.

[101] Gwenael Badis et al. Diversity and Complexity in DNA Recognition by Transcription Factors. Science. 2009 June 26; 324(5935): 17201723.

[102] Stormo GD and Zhao Y. Determining the specificity of protein-dna interactions. Nat Rev Genet. 2010 Nov;11(11):751-60.

[103] Segal,D.J. Evaluation of a modular strategy for the construction of novel polydactyl zinc finger DNA-binding proteins. Biochemistry, 42, 2137-2148 (2003).

[104] Meng,X. et al. Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. Nat. Biotechnol., 26, 695-701(2008) .

[105] Perez,E.E. Establishment of HIV-1 resistance in CD4+T cells by genome editing using zinc-finger nucleases. Nat. Biotechnol., 26, 808-816 (2008).

[106] Shukla,V.K. Precise genome modification in the crop species Zea mays using zinc-finger nucleases. Nature, 459, 437-441 (2009).

[107] Chen et al (2008), Integration of external signaling pathways with the core transcriptional network in embryonic stem cells *C*ell, 133(6):1106-17 (2008).

[108] Sridharan et al (2009), Role of the Murine Reprogramming Factors in the Induction of Pluripotency *C*ell, 136(2), 364-77.

[109] Bailey et al (2009), MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research, 37, W202-W208.

[110] Crooks et al (2004), WebLogo: A sequence logo generator, Genome Research, 14:1188-1190.

[111] Berg O. et al (1987), selection of dna binding sites by regulatory proteins: statistical mechanical theory and application to operators and promoters. J Mol Biol, 193 723-750 (1987).

[112] Hu. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. Nucleic Acids Research Volume38, Issue7Pp. 2154-2167.

[113] , Pengyu Hong. A boosting approach for motif modeling using ChIP-chip data. Bioinformatics Volume21, Issue11Pp. 2636-2643.

[114] Ji. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nature Biotechnology 26, 1293 - 1300 (2008).

[115] Victor X. Jin. W-ChIPMotifs: a web application tool for de novo motif discovery from ChIP-based high-throughput data. Bioinformatics Volume25, Issue23Pp. 3191-3193.

[116] Ettwiller L et al (2007), Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. Nature Methods, 4, 563-565 (2007).

[117] Kuttippurathu, L et al. CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments. Bioinformatics, 2011 Mar 1;27(5):715-7.

[118] Rodger Staden, Methods for calculating the probabilities of finding patterns in sequences. Bioinformatics Volume5, Issue2Pp. 89-96 (1989).

[119] Huang, H., et al. Determination of Local Statistical Significance of Patterns in Markov Sequences with Application to Promoter Element Identification. J Comput Biol. 2004;11(1):1-14.

[120] Hertz, G. Z. and Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15, 563-77.

[121] Bailey, T. Details of mcast statistics. Technical report IMB-tr0001.

[122] Holloway D., et al. Machine learning for regulatory analysis and transcription factor target prediction in yeast. Syst Synth Biol, 1(1):25-46 (2006).

[123] Fujita PA, et al. The UCSC Genome Browser database: update 2011. Nucleic Acids Res. 2010 Oct 18.

[124] Ali Mortazavi et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods; 5(7):621-8, (2008).

[125] Fejes AP, FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics, 2008 Aug 1;24(15):1729-30.

[126] Alan P. Boyle, et al. F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics, 2008 Nov 1;24(21):2537-8.

[127] Geetu Tuteja, Extracting transcription factor targets from ChIP-Seq data. Nucleic Acids Res., 2009 Sep;37(17):e113.

[128] Zhang Y., et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

[129] Rozowsky, J. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls, Nature Biotechnology 27, 66 - 75 (2009).

[130] Anton Valouev, Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods. 2008 Sep;5(9):829-34.

[131] Zang C., et al. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics, Aug 1;25(15):1952-8 (2009).

[132] Jothi R, et al. Genome-Wide Identification of in Vivo Protein-DNA Binding Sites From ChIP-Seq Data. Nucleic Acids Res. Sep;36(16):5221-31 (2008).

[133] Kharchenko, P., et al. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. Dec;26(12):1351-9 (2008).

[134] Nix, Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. BMC Bioinformatics, 9:523 (2008).

[135] Ganapathi et al. Extensive role of the general regulatory factors, Abf1 and Rap1, in determining genome-wide chromatin structure in budding yeast. Nucl. Acids Res. November 16, (2010).

[136] Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D91-4.

[137] Johnson, D.S., et al. Genome-wide mapping of in vivo protein-DNA interactions. Science 316, 1497-1502 (2007).

[138] Park OK, et al. Dimer stability as a determinant of differential DNA binding activity of Stat3 isoforms. J Biol Chem. 2000 Oct 13;275(41):32244-9.

[139] Ma A et al. DNA binding by N- and L-Myc proteins. Oncogene. 1993 Apr;8(4):1093-8.

[140] Sarah To et al. Modulation of CP2 Family Transcriptional Activity by CRTR-1 and Sumoylation. PLoS One. 2010; 5(7): e11702.

[141] Yoon JB, Li G, Roeder RG. Characterization of a family of related cellular transcription factors which can modulate human immunodeficiency virus type 1 transcription in vitro. Mol Cell Biol. 1994 Mar;14(3):1776-85.

[142] Murata T, et al. Transcription factor CP2 is essential for lens-specific expression of the chicken alphaA-crystallin gene. Genes Cells. 1998 Jul;3(7):443-57.

[143] Kang HC et al. Erythroid cell-specific alpha-globin gene regulation by the CP2 transcription factor family. Mol Cell Biol. 2005 Jul;25(14):6005-20.

[144] Shirra, MK and Hansen, U. LSF and NTF-1 share a conserved DNA recognition motif yet require different oligomerization states to form a stable protein-DNA complex. J Biol Chem. 1998 Jul 24;273(30):19260-8.

[145] Dempster, A.P. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39 (1): 138 (1977).

[146] Bulyk ML, et al. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Research, Mar 1; 30(5):1255-1261 (2002).

[147] A. Tanay. et al. Extensive low-affinity transcriptional interactions in the yeast genome. Genome res, 16, 962 (2006).

[148] E. Segal, et al. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature, 451, 535-40 (2008).

[149] Barash Y. et al. Modeling dependencies in protein-dna binding sites. Proceedings of the seventh internat ional conference on research in computational molecular biology. ACM press, pp. 28-37 (2003).

[150] Stormo GD. DNA binding sites: representation and discovery. Bioinformatics, 16, 16-23 (2000).

[151] Erill, I. and O'neill, M. C. A reexamination of information theory-based methods for DNA-binding site identification. BMC Bioinformatics, Feb 11;10:57 (2009).

[152] Solomon, M.J., et al. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell 53, 937-947 (1988).

[153] Balmer JE and Blomhoff R. Evolution of transcription factor binding sites in mammalian gene regulatory regions: handling counterintuitive results. J Mol Evol.,68(6):654-64 (2009).

[154] Scarano, E. et al. The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos. Proc. Natl. Acad. Sci. 57 (5): 1394-400 (1967).

[155] Jabbari K and Bernardi G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. Gene 333: 143-9 (2004).

[156] Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature . 2009 Aug 6 ; 460(7256):748-52.

[157] Shi J, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature . 2009 Aug 6 ; 460(7256):753-7.

[158] Stefansson H et al. Common variants conferring risk of schizophrenia. Nature . 2009 Aug 6 ; 460(7256):744-7.

[159] C. Cortes and V. Vapnik. Support-Vector Networks. Machine Learning, 20, 1995.

[160] Patwardhan, S., et al. EGR3, a novel member of the Egr family of genes encoding immediate-early transcription factors. Oncogene, 6(6):917-28 (1991).

[161] Ishikawa , Early growth response gene-1 plays a pivotal role in down-regulation of a cohort of genes in uterine leiomyoma , Journal of Molecular Endocrinology (2007) 39 333-341

[162] A. Kubosaki et al. Genome-wide investigation of in vivo EGR-1 binding sites in monocytic differentiation. Genome Biol. 2009;10(4):R41. Epub 2009 Apr 19.

[163] A. Inoue. Transcription factor EGR3 is involved in the estrogen-signaling pathway in breast cancer cells. J. Mol. Endocrinol, 32(3):649-61 (2004)

[164] X. Gao. Regulation of low affinity neurotrophin receptor (p75NTR) by early growth response (Egr) transcriptional regulators. Mol Cell Neurosci. 2007 December ; 36(4): 501514.

[165] Virolle et al. The Egr-1 transcription factor directly activates PTEN during irradiation-induced signalling. Nature Cell Biology , Yr: 2001

[166] D. S. Roberts. Egr3 stimulation of GABRA4 promoter activity as a mechanism for seizure-induced up-regulation of GABAA receptor 4 subunit expression. PNAS August 16, 2005 vol. 102 no. 33 11894-11899.

[167] Kumbrink J. EGR1, EGR2, and EGR3 activate the expression of their coregulator NAB2 establishing a negative feedback loop in cells of neuroectodermal and epithelial origin. J Cell Biochem. 2010 May 12.

[168] Quach, D., et al. Egr3 is a transcriptional regulator required for normal target tissue innervation during sympathetic nervous system development. FASEB J. 24(Meeting Abstract Supplement) 568.2 (2010).

[169] de Belle, I., et al, In vivo cloning and characterization of a new growth suppressor protein TOE1 as a direct target gene of Egr1. Journal of Biological Chemistry, 18;278(16):14306-12 (2003).

[170] Mayer, SI, et al. Epidermal-growth-factor-induced proliferation of astrocytes requires Egr transcription factors. J Cell Sci,122(Pt 18):3340-50. (2009).

[171] Kim, SH, et al. EGR3 as a potential susceptibility gene for schizophrenia in Korea. Am J Med Genet B Neuropsychiatr Genet. 2010;153B(7):1355-60.

[172] Jang, S. et al. In vivo detection of Egr2 binding to target genes during peripheral nerve myelination. Journal of Neurochemistry, 98, 5, 1678-1687 (2006)

[173] Kishore R et al.ERK1/2 and Egr-1 contribute to increased TNF-alpha production in rat Kupffer cells after chronic ethanol feeding. Am J Physiol Gastrointest Liver Physiol.282(1):G6-15 (2002).

[174] Li, L., et al. The Neuroplasticity-Associated Arc Gene Is a Direct Transcriptional Target of Early Growth Response (Egr) Transcription Factors. Molecular and Cellular Biology, 25(23), 10286-10300, (2005).

[175] Eden, E., et al. GOrilla: A Tool For Discovery And Visualization of Enriched GO Terms in Ranked Gene Lists. BMC Bioinformatics, 10:48 (2009).