UNDERSTANDING DISTAL TRANSCRIPTIONAL REGULATION FROM SEQUENCE, EXPRESSION AND INTERACTOME PERSPECTIVES

Arvind Rao*, David J. States and Alfred O. Hero, III
Bioinformatics, University of Michigan,
Ann Arbor, MI 48109, USA
Email: [ukarvind, dstates, hero]@umich.edu

James Douglas Engel
Cell and Developmental Biology, University of Michigan,
Ann Arbor, MI 48109, USA
Email: engel@umich.edu

Gene regulation in eukaryotes involves a complex interplay between the proximal promoter and distal genomic elements (such as enhancers) which work in concert to drive precise spatio-temporal gene expression. The experimental localization and characterization of gene regulatory elements is a very complex and resource-intensive process. The computational identification of regulatory regions that confer spatiotemporally specific tissue-restricted expression of a gene is thus an important challenge for computational biology. One of the most popular strategies for enhancer localization from DNA sequence is the use of conservation based prefiltering and more recently, the use of canonical (transcription factor motifs) or de-novo tissue-specific sequence motifs. However, there is an ongoing effort in the computational biology community to further improve the fidelity of enhancer predictions from sequence data by integrating other, complementary genomic modalities.

In this work, we propose a framework that complements existing methodologies for prospective enhancer identification. The methods in this work are derived from two key insights; one, that chromatin modification signatures can discriminate proximal and distally located regulatory regions. Second, that the notion of promoter-enhancer cross-talk (as assayed in 3C/5C experiments) might have implications in the search for regulatory sequences that co-operate with the promoter to yield tissue-restricted, gene-specific expression.

Keywords: Nephrogenesis, Random Forests, Transcriptional regulation, Transcription factor binding sites (TFBS), *GATA* genes, comparative genomics, functional genomics, tissue-specific genes, network analysis, directed information, heterogeneous data integration.

1. INTRODUCTION

Understanding the mechanisms underlying regulation of tissue-specific gene expression remains a challenging question. While all mature cells in the body have a complete copy of the human genome, each cell type only expresses those genes it needs to carry out its assigned task. This includes genes required for basic cellular maintenance (often called "housekeeping genes") and those genes whose function is specific to the particular tissue type that the cell belongs to. Gene expression by way of transcription is the process of generation of messenger RNA (mRNA) from the DNA template representing the gene. It is the intermediate step before the generation of functional protein from messenger RNA. During gene expression, transcription factor (TF) proteins are recruited at the proximal promoter of the gene as well as at sequence elements (enhancers/silencers) which can lie several hundreds of kilobases from the gene's transcriptional start site (Figs. 1 and 2).

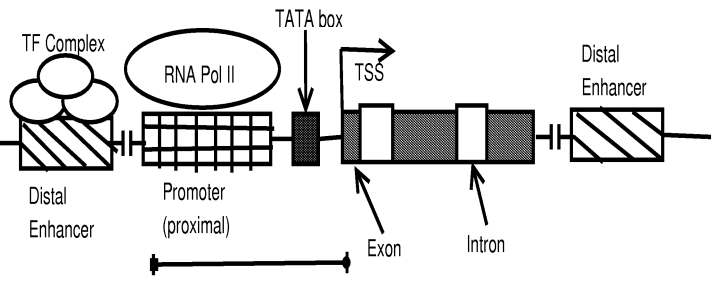

Fig. 1. Schematic of Transcriptional Regulation. Sequence motifs at the promoter and the distal regulatory elements together confer specificity of gene expression via TF binding.

*Corresponding author.

It is hypothesized that the collective set of transcription factors that drive (regulate) expression of a target gene are cell, context and tissue dependent [26, 20]. Some of these TFs are recruited at proximal regions such as the promoter of the gene, while others are recruited at these distal regulatory regions. There are several (hypothesized) mechanisms for promoter-enhancer interaction through protein interactions between TFs recruited at these elements during formation of the transcriptional complex [25]. A commonly accepted mechanism of distal interaction, during regulation, is looping [31, 7, 19], shown in Fig. 2, wherein intervening DNA between the enhancer and promoter is "looped out" to facilitate the interaction between the TFs of the promoter and the enhancer, leading to formation of the transcriptional complex.

An important challenge in current biology is to understand *where* functional regulatory elements (like enhancers) are located for a gene of interest. Given the complexity of the regulatory process, there are several instances wherein the enhancer for a gene is located hundreds of kilobases from the gene it regulates [7, 16, 9]. One of the typical experimental approaches to localize a gene-specific enhancer is via bacterial artificial chromosome (BAC) trap assays [17, 12].Thereafter, using conservation and TFBS based criteria, smaller genomic sequences ($1 - 2kb$) are isolated for subsequent transgenic analysis. However, even short genomic regions can have several conserved sequence elements (CSEs) worthy of experimental testing (e.g. ~ 120 CSEs surpass a 70% conservation in a $45kb$ *human-mouse* aligned region). Since an experimental analysis of each of these several regions is clearly unfeasible, there is a need for the use of principled methodologies that could potentially reduce this large list of enhancer candidates to a much shorter *high-confidence* list for experimental validation.

Since the main problem of interest is the prospective discovery of enhancers in a pre-specified sequence region, it would seem imperative to explore modalities that supplement conservation and TFBS criteria to reduce false positives. In this work, we explore two such modalities that emerge from functional genomic assays (from several recent independent studies as well as from the ENCODE project). These two modalities reveal some interesting new features of regulatory regions that are potentially of great use in discriminating gene-specific enhancers vs. other neutral regions. We note that there are promoter-independent enhancers too, and their computational study has been far more principled [26, 27]; however, their study is outside the scope of this study where we focus on gene-specificity in addition to tissue-specificity. Understanding the characteristics of such regulatory regions entails several aspects:

(1) Do regulatory regions like promoters and enhancers have any interesting *sequence properties* depending on their tissue-specificity of gene expression? Such properties can be examined based on their individual sequences or their epigenetic preferences. A common approach is the identification of canonical or de-novo tissue-specific motif-signatures [20, 15] for such elements, and has been applied quite extensively. In this work, however, we focus on the epigenetic preferences of distal regulatory regions (enhancers) vs. proximal regulatory regions (promoters).

(2) To reduce the large number of false positives that arise from sequence comparisons alone, we appeal to a mechanistic insight from biology. For long-range transcriptional regulation to be possible, there has to be an enhancer-promoter interaction during formation of the tissue-specific, gene-specific transcriptional machinery. Literature suggests that such interaction is mediated by protein-protein interactions between promoter TFs and enhancer TFs after looping along the chromosomal length [19, 2, 31]. This insight (Fig. 2) leads to two further questions:
   - Which TFs bind the promoter and the putative enhancer(s)?
   - Does this resultant "interaction-graph" between enhancer and promoter TFs have any special *structural* characteristic that can discriminate functional non-coding regulatory regions from non-functional ones?

The primary goal with answering the questions above is to build an enhancer discovery program that can localize tissue-restricted gene-specific enhancers in a given chunk of genome sequence (within a $\sim 200kb$ genomic window, as obtained from BAC trap strategies [12]). These questions will help us understand the nature of distal regulatory regions and provide a way to complement existing approaches in

enhancer localization [27, **?**] to achieve lower false positive rate and higher experimental efficacy.

As a case study to answer these questions, we examine the distal regulation of *Gata2* regulation in the developing kidney. *Gata2* is a gene belonging to the GATA family of transcription factors (*GATA1-6*), and binds the consensus –WGATAR– motif on DNA. It is located on mouse chromosome 6, and plays an important role in mammalian hematopoiesis, nephrogenesis and CNS development, with important phenotypic consequences. The study of long-range regulatory elements that effect *Gata2* expression has been on for several years now.

Recently, [12] reported the characterization of two enhancer elements, conferring urogenital-specific (UG) expression of *Gata2*, between $80-150$kb downstream from the *Gata2* transcription start site, on chromosome 6. In this experiment, 4 regions were selected for transgenic analysis based on sequence identity and TF motif matches. However, only two of these worked *in-vivo*. Based on the insights from the various individual studies since and the ENCODE project, outlined above, we asked if it might now be possible to explain the behavior of these 4 regions along these new modalities (epigenetic signatures and TF-interaction graphs), thereby enabling the proposal of a *framework for promoter-specific enhancer discovery from sequence*.

2. RATIONALE AND DATA SOURCES:

The overall schematic of distal transcriptional regulation via looping is given in Fig. 2. This schematic and the discussion in section. 1 suggests the decomposition of the regulatory process along three main modalities: sequence, expression and interactome. Our main goal in this paper is to understand urogenital enhancer potential of these 4 UG sequence candidates [12] from these three perspectives. These attributes are discussed below:

(1) **Sequence Perspective:** To build motif signatures underlying kidney-specific enhancer activity, it would be ideal to have a database of known, previously characterized, urogenital (UG) enhancers so that we could learn the sequence preferences of such tissue-specific regulatory regions. Here however, due to the unavailability of such data, we take a different approach and examine a public dataset of histone-modified sequences of regulatory regions to find motif-signatures of genomic elements that are potentially enhancers. Though this data source is not kidney specific, we observe that these epigenetic signatures have a strong, discriminative association with distal regulatory regions.

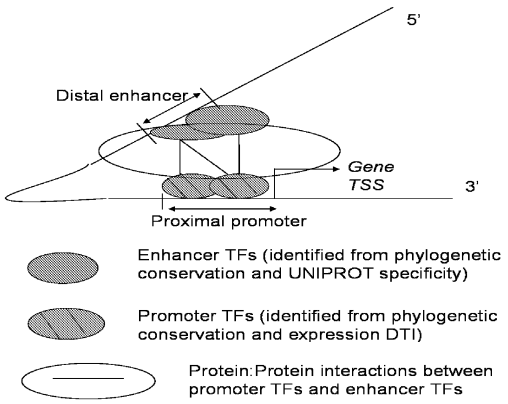

Fig. 2. Distal enhancer-promoter interaction via looping is mediated via protein interactions during TF complex formation. The set of TFs that are putatively recruited at the proximal promoter and distal enhancer can be found from sequence and expression data. Evidence of protein-interaction between proximal and distal TFs can be found from interaction databases.

• Chromatin marks in known regulatory elements: The ENCODE project suggests that monomethylation of the lysine 4 residue of Histone $H3$ is associated with enhancer (or distal regulatory) activity [10] whereas tri-methylation of $H3K4$ and $H3$ acetylation are associated with promoter activity. Using this set of $H3K4me1$, $H3K4me3$ and $H3ac$ sequences, we aim to find sequence motifs that are indicative of such epigenetic preferences during transcription. Though such epigenetic data is available for five different cell lines, we choose the HeLa cell line data because of its widespread use as a model system to understand transcriptional regulation *in-vitro* in the laboratory.

For simplicity, we find the frequencies of six-nucleotide long motifs in the $H3K4me1$ and $H3K4me3/H3ac$ sequences. Then, we build a random forest (RF) classifier to discriminate monomethylated $H3K4$ sequences from trimethylated $H3K4$/acetylated $H3$ sequences based on motif occurrence. We note that even though this data is not kidney cell-specific, it has favorable specificity and sensitivity characteris-

tics. The motifs thus obtained are putatively associated with epigenetic properties of proximal and distally located regulatory regions (such as enhancers), and are predictive of the regulatory potential of new sequences (section: 8).

(2) **Expression Perspective:** There is limited expression data for the developing mouse kidney, mainly due to small tissue yield at such early time points. For this study, we use microarray expression data from a public repository of kidney microarray data (*http://genet.chmcc.org* [32, *http://spring.imb.uq.edu.au/* 4]. Each of these resources contain expression data profiling kidney development from about day 10.5 dpc to the neonate stage. Such expression data can be mined for potential regulatory influence between upstream TF genes and *Gata2* [21, 28].

• *Inference of TF effectors at the promoter region*: The TFs putatively recruited at the proximal promoter are identified using the directed information (DTI) metric, that uses gene-expression (mRNA-level) influence in addition to phylogenetic conservation of the corresponding binding site. We have earlier shown that DTI is a good predictor of gene influence and can be used to infer transcriptional regulatory networks [28].

• *Inference of TF effectors at each non-coding region*: At the distal enhancer, it is believed that there is recruitment of tissue-specific transcription factors that co-operate with the basal transcriptional machinery (at the promoter) to direct tissue-specific gene expression [13, 20]. Whereas phylogeny and expression-based influence metrics can yield high confidence candidates for promoter TFs, a similar analysis for enhancers is not possible, because of higher order effects [20, 15]. To this end, the only way to search for putative enhancer TFs is to combine phylogeny with tissue-specific annotation (from UNIPROT or MGI). Hence, every transcription factor, whose motif is conserved at a non-coding (putative enhancer) region and is tissue-specific in annotation is considered a likely candidate TF at that non-coding region.

(3) **Interactome Perspective:** The identification of phylogenetically conserved effector TFs at the promoter (identified via DTI), as also those that are phylogenetically conserved at the putative enhancer candidate regions, lead to the exploration of protein-interactions (PPI) between these TFs, during distal enhancer-promoter interaction (Sec:9). The STRING database (*http://string.embl.de*) integrates various experimental modalities (genomic context, high-throughput experiments such as co-immunoprecipitation, co-expression and literature) to maintain a list of organism-specific functional protein-association networks that is amenable to such exploration.

In this work, the above perspectives are examined in the context of the urogenital enhancers identified in [12]. We aim to show that each of these modalities (epigenetic signatures and TF-interaction graphs) has a predictive value for the identification of enhancers and the integration of these heterogeneous perspectives can lead to potential reduction in false positive rate during large-scale enhancer discovery, genome-wide. To date, there has been no comprehensive study for summarizing these various heterogeneous data sources to understand the characteristics of such regulatory regions.

3. VALIDATION/BIOLOGICAL APPLICATION

As suggested in Sec: 1, we use the recently identified *Gata2* urogenital (UG) enhancers to validate our computational approach. All the data sources (and their analysis) are therefore going to be focused on the kidney.

The experimental characterization of these enhancers was done as follows. Based on BAC transgenic [12] studies, the approximate location of the urogenital enhancer(s) of *Gata2* were localized to a 70 kilobase region on chromosome 6. Using interspecies conservation plots, four elements were selected for transgenic analysis in the mouse. These were designated UG1, 2, 3 and 4. After a lengthy and resource-intensive experimental effort, two out of these four non-coding elements, $UG2$ and $UG4$ were found to be true UG enhancers. Our goal is to find preferences at the sequence, expression and interactome level, that can explain these experimental observations: i.e, that $UG2, 4$ are *Gata2*-specific urogenital enhancers and $UG1, 3$ are not urogenital enhancers for *Gata2*.

It is easy to see the utility of such a *"enhancer discovery"* methodology, since this can be applied

also to other genes of interest. Given the complexity of 1% of the genome, made possible by the ENCODE project, the search for functional elements genome-wide is going to be an important and challenging exercise.

4. ORGANIZATION

With a view to understanding the discriminating characteristics of transcriptional regulatory regions, the first part of this paper (Sections 5-8) addresses identification of motif signatures representative of transcriptional control from epigenetically marked sequences. The second part of this work (Sections 9.1 - 9.2) integrates phylogeny and expression data to find regulatory TFs at the proximal promoter and enhancer(s) of *Gata2*. Using the notion of TF interactions between enhancer and promoter, we examine if protein-interaction data (Sec: 9.3) can offer supporting evidence for the observed *in-vivo* behavior of the four *Gata2* candidate sequences. Classifiers are designed to discriminate regulatory vs. non-regulatory regions based on these two modalities (epigenetic signatures and TF-interaction graphs). Finally, a probabilistic combination of these classifiers is done to obtain a validation (Sec: 10) of the *Gata2* UG enhancer (UGE) candidates ($UG1-4$). Sections: 11 and 12 conclude the paper.

5. Sequence Data Extraction and Pre-processing

Before proceeding to motif identification, a matrix of motif–chromatin-sequence correspondences is created. In this matrix, the counts of hexamer (six-nucleotide) motif occurrence in the *'H3K4me1'* and *'H3K4me3/H3ac'* regions is obtained using sequence parsing (*R package: 'seqinr'*). The motif length of six is not overly restrictive, and can be changed based on biological insight. A Welch t-test is then performed between the relative counts of each hexamer in the two epigenetic-modification categories (*'H3K4me1'* and *'H3K4me3/H3ac'*) and the top 1000 hexamers with $p-value \leq 10^{-6}$ are selected. This set of discriminating hexamers is designated ($\overrightarrow{\mathbf{H}} = H_1, H_2, \ldots, H_{1000}$). This procedure resulted in two hexamer-gene co-occurrence matrices, – one for the *'H3K4me1'* (or $+1$) class of dimension $N_{train,+1} \times 1000$ and the other for the *'H3K4me3/H3ac'* (or -1) class – dimension $N_{train,-1} \times 1000$. Here $N_{train,+1}$ is the matrix of $H3K4me1$ sequences corresponding to distal regulatory regions. $N_{train,-1}$ is the set of *'H3K4me3/H3ac'* sequences that are associated with proximal promoters.

This dataset is obtained from the Sanger ENCODE database (*http://www.sanger.ac.uk/PostGenomics/encode/data-access.shtml*), and contains 298 sequences that undergo modification ($me1/me3/ac$) in histone ChIP assays. 140 of these correspond to $H3K4me1$ (enhancers), and 158 correspond to $H3K4me3/H3ac$ marks (promoters).

Table 1. The 'motif count matrix' for a set of histone-modified sequences. The first column is their genomic locations along the chromosome, the next 2 columns are hexamer quantile labels, and the last column is the corresponding sequence class label ($+1/-1$).

Sequence	AAAATA	AAACTG	Class
chr2:41410492-41411867	2	1	+1
chr6:41654502-41654782	4	2	+1
chr3:41406971-41408059	1	1	-1
chr2:41665970-41667002	2	3	+1
chr4:41476956-41478365	1	2	-1
chrX:41783327-41784532	1	2	+1

6. MOTIF-CLASS CORRESPONDENCE MATRICES

From the above, $N_{train,+1} \times 1000$ and $N_{train,-1} \times 1000$ dimensional count matrices are available for the chromatin-modified sequences. Before proceeding to the feature (hexamer motif) selection step, the counts of the $M = 1000$ hexamers in each training sample are normalized to account for variable sequence lengths. In the co-occurrence matrix, let $gc_{i,k}$ represent the absolute count of the k^{th} hexamer, $k \in 1, 2, \ldots, M$ in the i^{th} chromatin-sequence. Then, for each sequence g_i, the quantile labeled matrix has $X_{i,k} = l$ if $gc_{i,[\frac{l-1}{K}M]} \leq gc_{i,k} < gc_{i,[\frac{l}{K}M]}, K = 4$. Matrices of dimension $N_{train,+1} \times 1001$, $N_{train,-1} \times 1001$ for the specific and non-specific training samples are now obtained. Each matrix contains the quantile label assignments for the 1000 hexamers ($X_i, i \in (1, 2, \ldots, 1000)$), as stated above, and the last column would have the corresponding class label ($Y = -1/+1$). Having constructed two groups of sequences for analysis, enhancer-associated (*'H3K4me1'*) and promoter-associated (*'H3K4me3/H3ac'*) – we seek to find the

smallest set of hexamer motifs that are most discriminatory between these two classes.Towards this goal, we use random forest classifiers (RF) [3] for finding such a discriminative hexamer subset.

7. RANDOM FOREST CLASSIFIERS

A random forest (RF) is an ensemble of classifiers obtained by aggregating (bagging) several classification trees [3]. Each data point (represented as an input vector) is classified based on the majority vote gained by that vector across all the trees of the forest. Each tree of the forest is grown in the following way:

- A bootstrapped sample (with replacement) of the training data is used to grow each tree. The sampling for bootstrapped data selection is done individually at each tree of the forest.
- For an M-dimensional input vector, a random subspace of m ($\ll M$)-dimensions is selected, and the best split on this subspace is used to split the node. This is done for all nodes of the tree.

During the training step, before sampling by replacement, one-third of the cases is kept "out of the training bag". This OOB (out-of-bag) data is used to obtain an unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance.

Several interesting insights into the data are available using random forest analysis. The variable importance plot yields the variables that are most discriminatory for classification under the 'ensemble of trees' classifier. This importance is based on two measures– 'Gini index' and 'decrease in accuracy'. The Gini index is an entropy based criterion which measures the purity of a node in the tree, while the other metric simply looks at the relative contribution of each variable to the accuracy of the classifier. For our studies, we use the 'randomForest' package for R. The classifier performance on the individual data and the related diagnostics are mentioned under Sec: 8.

8. RFs ON CHROMATIN-MODIFIED SEQUENCES

We train the RF classifier on the set of 298 chromosome sequences that have varying chromatin modifications associated with them (i.e., $H3K4me1/me3$, and $H3ac$), as mentioned in Section: 2.These are derived from the HeLa cell line and are not necessarily context-specific for kidney development. However, given the widespread use of this cell line for transcriptional studies, we aim to find if the motifs associated with regulatory elements are indeed predictive of enhancer activity.

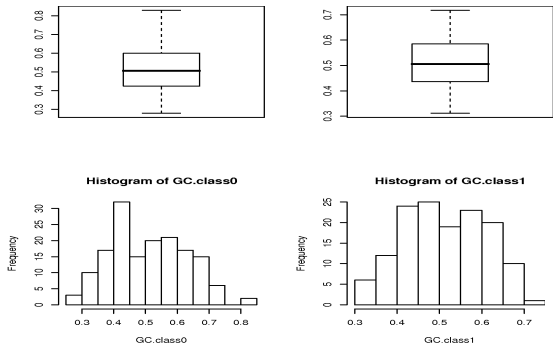

Fig. 3. GC plots for sequence bias in $H3K4me1$ histone sequences vs. $H3K4me3$ and $H3ac$ sequences. We observe that there is no significant bias in GC content.

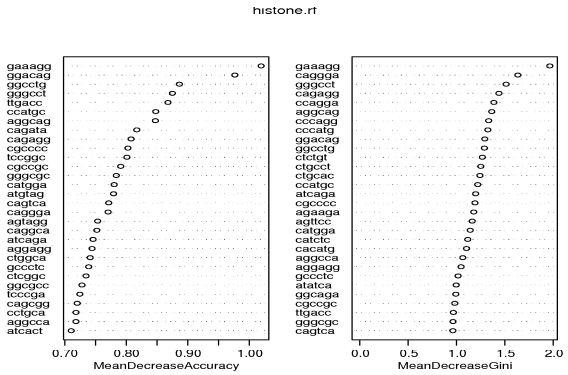

Fig. 4. Top hexamers which can discriminate between $H3K4me1$ histone sequences vs. $H3K4me3$ and $H3ac$ sequences.

Before proceeding to motif identification, we check for possible sequence bias (such as GC-nucleotide composition) between these two classes of chromatin modified sequences. If there is a significant bias, then the motifs turn out to be just GC rich sequences that are not very biologically informative for determination of regulatory potential. The GC composition of these two classes of sequences is represented in Fig. 3. As can be seen, the average GC composition is the same and that there is no

such sequence bias that would skew the discovery and subsequent interpretation of these epigenetic motifs. The performance of the histone-RF classifier is explained in the context of the classifier combination in Section:10.

The motifs obtained from the random forest analysis indicate the "sequence-preferences" of regulatory elements that are nucleosome-free in HeLa cells (Fig. 4). We analyze the performance of these classifiers on the 4 UG candidate regions, mentioned previously. In both cases, $UG2-4$ are classified as enhancers, whereas $UG1$ is correctly classified as not being regulatory. Additionally, a control set of "promoter-independent" enhancers derived from the Mouse Enhancer database [26] was also classified as enhancers based on these chromatin-sequence motif signatures. This high prediction accuracy inspite of non-specificity of cell context (*HeLa* cell line) is very interesting and has potentially high predictive value.

9. PPI BETWEEN PROMOTER AND ENHANCER TFs

In order to understand the nature of interactions between the enhancer and promoter TFs (Fig. 2), we decouple the overall regulation problem into three parts:

(1) Identification of putative TF effectors at the promoter (Section: 9.1),
(2) Identification of enhancer TFs (Section: 9.2), and
(3) Examination of the interaction-graph formed between enhancer-TFs and promoter TFs (Section: 9.3).

The key question that is explored in the following sections is: having identified the set of tissue-specific TFs that might putatively bind the promoter and the candidate regulatory regions, does the *structure of the bipartite TF-interaction graph* (across the promoter TFs and the enhancer TFs) reveal any interesting features that discriminate the functional $UG2,4$ regions from the non-functional $UG1,3$ regions.

9.1. TF effector identification at Promoter and Enhancer

Promoter TF identification: TFs that regulate basal transcription at the promoter can be identified from phylogenetic conservation or co-expression studies. In this approach, the promoter sequence (here, the *Gata2* promoter) is aligned across multiple species and the TFBS motifs that are conserved in the multiple alignment are considered to be putative effectors of gene regulation. Such sequence-based approaches have been examined in literature [20, 15].

Since the list of putative TFs (identified above) that potentially bind at the promoter is still large, there have been efforts to incorporate gene-expression data to reduce the set of potential TF effectors. In this scenario, if the gene corresponding to the conserved TF has a high expression-level influence on *Gata2* expression, then that TF has stronger evidence for being a potential regulator [21]. Recently, we introduced the directed information (DTI) as a metric to infer expression-level influence between any putative transcription factor (TF) gene and a target gene (such as *Gata2*) [28]. This seeks to integrate sequence and expression data into the determination of relationships between transcription factors and their target-genes. All additional details (performance on synthetic data, other biological data and comparison with other metrics) are available in [28]. Information-based measures have enabled the investigation of non-linear gene relationships in the presence of measurement noise [21]. An important point to note is that unlike mutual information, the DTI is a *directed* metric that enables the determination of the strength, significance and direction of gene influence. For *Gata2*, this list of effectors is listed in Fig. 5 below.

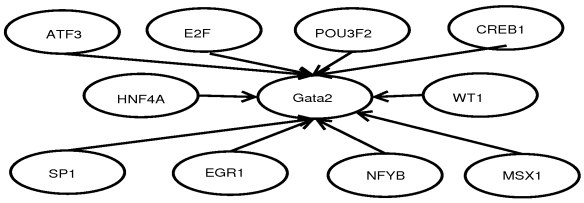

Fig. 5. Putative upstream TFs using DTI for the *Gata2* gene.

9.2. Enhancer TF identification

In section 9.1, we have examined the identification of promoter TFs using phylogenetic sequence conservation of TFBS motifs in conjunction with expression level influence using DTI. The next key step towards determining the structure of promoter-enhancer TF interactions is the identification of enhancer-TFs. As

has been alluded to earlier, there is no method to precisely infer which transcription factors bind a certain regulatory element during long-range gene regulation. Thus, we appeal to a traditional approach of finding tissue-specific transcription factors that are phylogenetically conserved at any potential regulatory region [27, 20](one caveat, however, is that conservation is not a very reliable predictor of TF binding [22, 23]). This is consistent with earlier observations that enhancers recruit tissue-specific transcription factors during the formation of the overall transcriptional machinery during gene expression, whereas promoters recruit components of the basal transcriptional machinery [13, 20, 15, 31].

To ascertain the tissue-specificity of each TF that putatively binds a regulatory element (identified via phylogenetic conservation), we examine that TF's annotation in the UNIPROT or MGI database.

9.3. ENHANCER-PROMOTER DISTAL INTERACTION VIA PROTEIN-PROTEIN INTERACTIONS - A GRAPH BASED ANALYSIS

Using the notion of protein-protein interaction (PPI) mediating long-distance interactions between promoters and enhancers during looping [25, 2, 8], we explore the interactome to look for within-group and between-group interactions in the promoter-TF and the enhancer-TF groups.

The interaction-graphs (e.g: Fig. 6) are obtained in the following manner:

- One part of the graph (hollow circles) corresponds to the TF effector group at the promoter. These V_p TFs are identified based on phylogenetic conservation, tissue-specificity and directed information (section: 9.1).
- The other part of the graph (filled circles) corresponds to the V_e tissue-specific TFs group at the enhancer, identified based on phylogeny and tissue-specificity annotation (section: 9.2).
- The interaction-graph is defined by the vertices $V = (V_p \cup V_e)$, and the edges $E = e_{i,j}$, $i, j \in (1, 2, \ldots, |V_p \cup V_e|)$. Each bidirectional edge $E = (e_{i,j})$ is derived from an annotated interaction between TFs i and j, based on an interaction database. These edges describe both within-group TF interactions as well as between-group interactions. These interactions are obtained from the STRING (*http://string.embl.de/*) and MiMI (*http://mimi.ncibi.org/MiMI/home.jsp*) databases, both of which contain data derived from multiple sources, such as yeast-2-hybrid screens, literature etc.

Though it would be of great value to use a catalog of gene-specific and tissue-specific regulatory regions (with all possible transcription factors) from which to find such interaction characteristics - such a repository does not yet exist. In this section, we use a few examples (*Gata3* OVE, *Gata3* KE, *Fgf* OVE, *Mecp2* F21/F6 , *Shh* FE) of known tissue-specific and gene-specific regulatory elements from literature, as a positive training set. For the negative training set, we consider the set of regions that were reportedly investigated in these transgenic experiments but did not yield gene-specific regulatory activity.

We have presented a preliminary analysis of enhancer-promoter TF interaction-graphs for some genomic elements with known regulatory or non-regulatory activity [19, 18, 9, 24] in Table. 2. The table represents the listing of some of the structural attributes of these interaction-graphs, following analysis methods from literature [1]. A deeper analysis of other graph topology metrics and their relation to functional enhancer activity is a topic of future interest.

• Clustering coefficient: The clustering coefficient of a node is always a number between 0 and 1. The network clustering coefficient is the average of the clustering coefficients for all nodes in the network.

• Characteristic Path length: The characteristic path length denotes the average shortest-path distance of the graph. This gives the expected distance of any two connected nodes in the graph and is a global indicator of network-connectivity.

• Heterogeneity: Network heterogeneity denotes the coefficient of variation of the degree distribution.

• Centralization: This refers to the overall connectivity (cohesion) of the graph. It indicates how strongly the graph is organized around its most central point(s).

• Density: It shows how densely the network is populated with edges (i.e. how "close-knit" an empirical graph is). A network which contains no edges and solely isolated nodes has a density of 0, whereas the

Table 2. The first column is the various regulatory and non-regulatory elements from literature, the next column corresponds to its class label (+1/−1). The subsequent columns correspond to the attributes of the overall TF-interaction graph (both within-group and between-group interactions).

Sequence	Class	Clustering Coefficient	Characteristic path length	Heterogeneity	Centralization	Density
Mecp2 F21 [19]	+1	0.208	2.824	0.668	0.184	0.133
Mecp2 F6 [19]	-1	0	1.75	0.342	0.067	0.145
Gata3 OVE [9]	+1	0.036	2.254	0.779	0.359	0.154
Gata3 KE [9]	+1	0.409	2.0	0.813	0.684	0.216
Gata3 NE1 [9]	-1	0.383	2.131	1.139	0.757	0.15
Gata3 NE2 [9]	-1	0.458	2.013	0.872	0.699	0.203
Fgf10 OVE [24]	+1	0.313	2.433	0.72	0.323	0.133
Shh FE [18]	+1	0.394	2.312	0.797	0.49	0.175

density of a clique (completely connected graph) is 1.

The above mentioned several network properties (as well as clustering coefficients, number of connected components etc.) are examined for the overall interaction-graphs for the reported enhancers from literature. A logistic regression reveals that low values of heterogeneity, characteristic path length and centralization are strong predictors of potential enhancer activity. All of these attributes point to the decentralized, homogenous and somewhat tighter connectivity of the interaction-graphs for true enhancers. We note that the OOB error rate of the RF here is about 20%. The quality of this classifier can be expected to improve as we obtain more data (gene-specific regulatory regions) from which to extract features.

We now examine the interaction-graphs for the test set, i.e. the four *Gata2* UGE candidates. For illustration, we only show the largest connected component of the inter-group edges for each interaction graph (Fig. 6).

This figure indicates a very interesting property of the real enhancers vis-a-vis the other conserved elements. We see that the TF effectors for *Gata2* such as *SP1*, *POU3F2* (identified in the TF effector network above, Fig. 5), are involved in cross-element interactions at the protein level, between the promoter and true enhancers ($UG2/4$). However, the network linkage in the elements that showed no enhancer activity is very sparse suggesting low cross-talk between promoter and enhancer. Also, the TFs at the enhancer nodes (dark circles) have a more uniform degree distribution in the functional elements $UG2/4$ as compared to the non-functional ones. Both these observations suggest lower heterogeneity and centralization of such functional interaction-graphs. Thus, the extent of TF cross-talk is a potential discriminator of possible enhancer function. This shows that superimposing such PPI information along with sequence and expression data helps reduce the number of false positives while integrating various aspects of distal regulation.

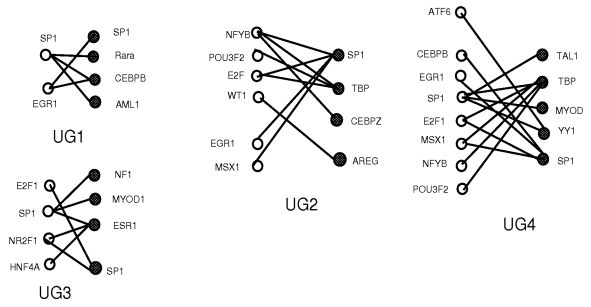

Fig. 6. Protein-protein interaction between putative *Gata2* TFs (hollow circles) and putative UG element TFs (filled circles). Note: This only shows the connections between two groups for one of the connected components. For our analysis, we consider both *intra-* and *inter-group* connections. From *http://string.embl.de/*

10. HETEROGENEOUS DATA INTEGRATION AND VALIDATION ON GATA2 UGE CANDIDATE SEQUENCES

As mentioned previously, the primary goal of the framework developed above is to understand the behavior of known regulatory elements along different genomic modalities. To validate their predictive potential, we demonstrate their application to predicting the behavior of the experimentally-verified *Gata2* UG enhancer candidates (which is our test set). Here we combine the results of the individual classifiers (histone RF and interactome-RF) to obtain an in-

tegrated prediction that a candidate sequence is an enhancer. For combining heterogeneous classifiers, we use a "probabilistic belief fusion" approach.

The framework involves combining the 'beliefs' of the individual classifiers to obtain a combined belief of prediction. To compute the belief of each classifier we start with examining the confusion matrices for each of the classifiers (histone-RF and graph-RF), following ([33]. Since each of the classifiers are random forests, we can obtain their OOB error estimates through these confusion matrices. For the graph-RF, this confusion matrix is as below,

$$\mathbf{CM_{graph-RF}} = \begin{pmatrix} Class & -1 & 1 & class.error \\ -1 & 4 & 1 & 0.20 \\ 1 & 1 & 4 & 0.20 \end{pmatrix},$$

thereby yielding an OOB error estimate of $\sim 20\%$.

Similarly, we have,

$$\mathbf{CM_{histone-RF}} = \begin{pmatrix} Class & -1 & 1 & class.error \\ -1 & 134 & 24 & 0.15 \\ 1 & 21 & 119 & 0.15 \end{pmatrix},$$

yielding an OOB error estimate of $\sim 15\%$.

As can be seen, these classifiers have fairly good sensitivity and specificity characteristics. This is expected to improve as more training data for these classifiers becomes available. Moreover these are two complementary data sources and can be effectively combined to improve detection reliability. Since they are trained on very different modalities, they can be assumed to be independent. It can also be seen that this method of belief combining is applicable to as many modalities (K) as necessary to the biological problem of interest, and hence is truly scalable.

Let each classifier be characterized by its decision function $e_k(x) = j_k$ that maps a data point (x) to the class 'j', for $k = 1, 2, \ldots, K$ and $j_k \in (-1, 1)$. Here, $K = 2$, and $J = 2$ classes.

The belief of the k^{th} classifier is defined as,

$$bel_k(x \in C_i | e_k(x) = j_k) = P(x \in C_i | e_k(x) = j_k)$$

The overall belief, *bel(i)*, is computed using Bayes rule,

$$bel(i) = P(x \in C_i).\frac{\prod_{k=1}^{K} P(x \in C_i | e_k(x) = j_k)}{\prod_{k=1}^{K} P(x \in C_i)}$$

$$bel(C_i) = \frac{\prod_{k=1}^{K} P(x \in C_i | e_k(x) = j_k)}{\sum_{i=1}^{J} \prod_{k=1}^{K} P(x \in C_i | e_k(x) = j_k)}.$$

Note: $J = 2$ and $K = 2$. Depending on the belief value $bel(i)$, the decision rule ($E(x)$) for classifying data point x is,

$$E(x) = j, \text{if } bel(j) = max_i\, bel(i),$$
$$\text{or, } E(x) = j, \text{if } bel(j) = max_i\, bel(i), \text{and, } bel(j) \geq \alpha,$$

where $0 < \alpha \leq 1$, with α being a threshold.

We now show the output classes of each of the 2 classifiers as well as the combined belief on the *Gata2* UG enhancer candidates in Table. 3. More specifically, for the first row in Table. 3, the overall belief equation above becomes,

$$bel(ug1 = +1) = \frac{\prod_{k=1}^{K} P(ug1 = +1 | e_k(x) = j_k)}{\prod_{k=1}^{K} [P(ug1 = +1 | e_k(x) = j_k)] + \prod_{k=1}^{K} [P(ug1 = -1 | e_k(x) = j_k)]}$$
$$= \frac{\prod_{k=1}^{K} (1 - prec_{n,k})}{[\prod_{k=1}^{K} (1 - prec_{n,k}) + \prod_{k=1}^{K} prec_{n,k}]}$$

Here, $prec_{n,k} = \frac{TN_k}{TN_k + FN_k}$. Similarly, $prec_{p,k} = \frac{TP_k}{TP_k + FP_k}$. These are the negative and positive precision values respectively, for the k^{th} classifier. These rates are obtained from the corresponding confusion matrices shown above. This approach is followed for each of the $UG1 - 4$ elements (Table. 3).

If we set a threshold of $\alpha = 0.60$ or 0.90, we would get $UG2$ and $UG4$ to be the true enhancers (100% accuracy). However, for a choice of $\alpha = 0.50$, $UG3$ is predicted to be an enhancer in spite of being declared a member of the (-1) class by the graph-RF. This choice of threshold thus determines the performance of the combined classifier (just like in any other hypothesis-testing scenario). We note that at the present time, there is no known repository of promoter-specific regulatory elements to carry out such graph-analysis on each element.

Under the $\alpha = 0.50$ case, however, the results are not to be interpreted as a 25% error rate since the nature of the test set (*Gata2* UG enhancers) are very different from the training data of each modality (histone sequences are for a different cell-context; and interaction-graphs are obtained over different genes). The fact that we are getting such good prediction in spite of the training sets being so different is a strong point in favor of examining and integrating these data sources. The test-error rates are given by the OOB error estimates of the individual classifiers.

Table 3. Combined belief generation during heterogeneous classifier integration. The last column represents the combined belief (probability that the UG candidate sequence is an enhancer) as a result of integrating the histone-RF and graph-RF predictions.

Sequence	True Class	Histone RF prediction $e_1(x)$	Interaction-graph RF prediction $e_2(x)$	P(Class=+1) (Overall Belief)
Gata2 UG1	-1	-1	-1	0.0377
Gata2 UG2	+1	+1	+1	0.9520
Gata2 UG3	-1	+1	-1	0.5535
Gata2 UG4	+1	+1	+1	0.9520

11. SUMMARY OF APPROACH

In this work, we have shown that,

- Chromatin modification motif signatures are predictive of regulatory element location. These point to the cell-specific epigenetic preferences of distally located regulatory regions.
- Promoter and enhancer TFs that are putatively recruited during gene (*Gata2*) regulation can be identified using a combination of phylogenetic conservation, expression data, and tissue-specificity annotation.
- Effector TFs at the gene proximal promoter have high network linkage with enhancer TFs in case of functional enhancers. The TF interaction-graphs of truly functional elements are seen to be have a lower centralization, characteristic path length and heterogeneity suggesting higher cross-talk during formation of the transcription factor complex.

These diverse perspectives (based on sequence, expression and interactome data) shed some light on the sequence and mechanistic preferences of true regulatory regions interspersed genome-wide. It is to be noted that this model is data-driven and needs further validation to correspond directly with the biology of transcription.

12. CONCLUSIONS

The novelty of the proposed work spans several areas. Firstly, data sources that are relevant to understand the mechanism of gene regulation (with *Gata2* as an example) have been identified. We have developed methods that reconcile the behavior of known regulatory elements along each of these modalities. The utilization of histone-modified sequences and their exploration for sequence motifs are indicative of epigenetic-preferences and nucleosome-occupancy patterns. This has not been explored before for the characterization of distal regulatory regions. The use of DTI as a metric to infer putative TF to target-gene influence is a recent one that serves to integrate phylogenetic TFBS conservation along with expression data. Finally, the utilization of graph-based analysis techniques to understand the "structure" of the TF interaction-graph between enhancer and promoter helps us understand true enhancer behavior from a mechanistic viewpoint. The probabilistic combination of multiple classifiers (each deriving from a unique data resource) aims to reconcile the behavior of existing enhancers along multiple modalities. We hope to demonstrate that a principled integration of non-overlapping genomic modalities can be used to interpret the context and specificity of gene regulation.

ACKNOWLEDGEMENTS

We thank Ms. Swapnaa Jayaraman for useful discussions about network analysis. We are also very grateful to the two anonymous reviewers for their help in revising the manuscript.

References

1. Bader GD, Hogue CW.,"An automated method for finding molecular complexes in large protein interaction networks".,*BMC Bioinformatics*. 2003 Jan 13;4:2.
2. E Blackwood and J Kadonaga, Going the distance: a current view of enhancer action. Science 281 (1998), pp. 6063.
3. L. Breiman., "Random forests"., *Machine Learning*, 45(1): 5.32, 2001.
4. Challen G, Gardiner B, Caruana G, Kostoulias X, Martinez G, Crowe M, Taylor DF, Bertram J, Little M, Grimmond SM.,"Temporal and spatial transcriptional programs in murine kidney development".,*Physiol Genomics*. 2005 Oct 17;23(2):159-71.
5. Dong J, Horvath S (2007) "Understanding Network Concepts in Modules", *BMC Systems Biology* 2007, 1:24
6. ENCODE Project Consortium, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project"., *Nature*. 2007 Jun 14;447(7146):799-816.
7. Fraser P., "Transcriptional control thrown for a loop"., *Curr Opin Genet Dev*. 2006 Oct;16(5):490-5.

Computational Neuroscience Unit Technical Report, 2003.
8. Gilbert S.F, "Developmental Biology", Sinauer Associates Inc., Publishers Sunderland, Massachusetts, 1997.
9. Hasegawa SL, Moriguchi T, Rao A, Kuroha T, Engel JD, Lim KC.,"Dosage-dependent rescue of definitive nephrogenesis by a distant Gata3 enhancer".,*Dev Biol.* 2007 Jan 15;301(2):568-77.
10. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B.,"Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome"., *Nat Genet. 2007* Mar;39(3):311-8.
11. Hutchinson GB., "The prediction of vertebrate promoter regions using differential hexamer frequency analysis".,*Comput Appl Biosci.* 1996 Oct;12(5):391-8.
12. Khandekar M, Suzuki N, Lewton J, Yamamoto M, Engel JD., "Multiple, distant Gata2 enhancers specify temporally and tissue-specific patterning in the developing urogenital system".,*Mol Cell Biol.* 2004 Dec;24(23):10263-76.
13. Kleinjan DA, van Heyningen V., "Long-range control of gene expression: emerging mechanisms and disruption in disease"., *Am J Hum Genet.* 2005 Jan;76(1):8-32.
14. Koch CM, Andrews RM, Flicek P, Dillon SC, Karaz U, Clelland GK, Wilcox S, Beare DM, Fowler JC, Couttet P, James KD, Lefebvre GC, Bruce AW, Dovey OM, Ellis PD, Dhami P, Langford CF, Weng Z, Birney E, Carter NP, Vetrie D, Dunham I.,"The landscape of histone modifications across 1% of the human genome in five human cell lines".,*Genome Res.* 2007 Jun;17(6):691-707.
15. Kreiman G., "Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes".,*Nucleic Acids Res.* 2004 May 20;32(9):2889-900.
16. Lakshmanan, G., K. H. Lieuw, K. C. Lim, Y. Gu, F. Grosveld, J. D. Engel, and A. Karis. 1999. "Localization of distant urogenital system-, central nervous system-, and endocardium-specific transcriptional regulatory elements in the GATA-3 locus". *Mol. Cell. Biol.* 19:1558-1568.
17. Lee, E. C., D. Yu, J. Martinez de Velasco, L. Tessarollo, D. A. Swing, D. L. Court, N. A. Jenkins, and N. G. Copeland. 2001. A highly efficient Escherichia coli-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA. Genomics 73:56-65.
18. Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E., "A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly".,*Hum Mol Genet.* 2003 Jul 15;12(14):1725-35.
19. Liu J, Francke U.,"Identification of cis-regulatory elements for MECP2 expression".,*Hum Mol Genet.* 2006 Jun 1;15(11):1769-82.
20. MacIsaac KD, Fraenkel E., "Practical strategies for discovering regulatory DNA sequence motifs".,*PLoS Comput Biol.* 2006 Apr;2(4):e36.
21. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A., "ARACNE: an algorithm for the reconstruction of gene regulatory networks in amammalian cellular context"., *BMC Bioinformatics.* 2006 Mar 20;7 Suppl 1:S7.
22. McGaughey DM, Vinton RM, Huynh J, Al-Saif A, Beer MA, McCallion AS., "Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b"., Genome Res. 2008 Feb;18(2):252-60.
23. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E., "Tissue-specific transcriptional regulation has diverged significantly between human and mouse"',.Nat Genet. 2007 Jun;39(6):730-2.
24. Ohuchi H, Yasue A, Ono K, Sasaoka S, Tomonari S, Takagi A, Itakura M, Moriyama K, Noji S, Nohno T., "Identification of cis-element regulating expression of the mouse Fgf10 gene during inner ear development"., *Dev Dyn.* 2005 May;233(1):177-87.
25. Petrascheck M, Escher D, Mahmoudi T, Verrijzer CP, Schaffner W, Barberis A.,"DNA looping induced by a transcriptional enhancer in vivo"., *Nucleic Acids Res.* 2005 Jul 7;33(12):3743-50.
26. Pennacchio, L. A., Ahituv, N., Moses, A., Prabhakar, S., Nobrega, M., Shoukry, M., Minovitsky, A., Dubchak, I., Holt, A., Lewis, K., Plazer-Frick, I., Akiyama, J., DeVal, S., Afzal, V., Black, B., Couronne, O., Eisen, M., Visel, A., and Rubin, E.M. 2006., "In vivo enhancer analysis of human conserved non-coding sequences", *Nature*, 444(7118):499-502.
27. L.A. Pennacchio, G.G. Loots, M.A. Nobrega, and I. Ovcharenko, "Predicting tissue-specific enhancers in the human genome", *Genome Research*, 17(2), 201-11 (2007)
28. Rao A, Hero AO, States DJ, Engel JD, "Using Directed Information to build Biologically Relevant Influence Networks", *Proc. Computational Systems Bioinformatics (CSB)*, 2007.
29. Romer KA, Kayombya GR, Fraenkel E., WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches., Nucleic Acids Res. 2007 Jul;35(Web Server issue):W217-20.
30. Royce TE, Rozowsky JS, Gerstein MB., "Assessing the need for sequence-based normalization in tiling microarray experiments"., *Bioinformatics.* 2007 Apr 15;23(8):988-97.
31. Simonis M, Kooren J,de Laat W (2007) "An evaluation of 3C-based methods to capture DNA interactions"., *Nature Methods* 4(11): 895.
32. Stuart RO, Bush KT, Nigam SK, "Changes in gene expression patterns in the ureteric bud and metanephric mesenchyme in models of kidney development", *Kidney International*,64(6),1997-2008,December 2003.
33. Xu, L.; Krzyzak, A.; Suen, C.Y., Methods of combining multiple classifiers and their applications to handwriting recognition, *Systems, Man and Cybernetics, IEEE Transactions on* Volume 22, Issue 3, May-June 1992 Page(s):418 - 435.