Chapter 2

Building positional network of transcription factors binding sites from predicted as well as available experimental information about target sites
2.1 Introduction

Biological basis for building positional network of transcription factor binding sites

The regulation of gene expression is fundamental to biological processes such as cell growth, proliferation, differentiation, and response to the extracellular environment. The transcriptional program of a cell is highly dynamic and responsive to intracellular as well as environmental signals. The regulation of transcription is the most important step in control of gene expression. This involves a myriad of mechanisms influencing binding of transcription factors (TFs) to their cognate DNA regulatory elements and activating or repressing transcription. Thus, sequence identity, position, and accessibility of transcription factor binding sites (TFBSs) is a crucial determinant of transcription regulation. Therefore, identification, characterization and validation of these DNA regulatory elements or cis-element is among the most important and challenging tasks for biologists in the post-genome era.

The basic aspects of transcriptional regulation are important to be considered for building these networks of transcription factors. Transcription factors generally bind to short 5–15 base pairs degenerate DNA sequences called consensus binding sequence or motif for the corresponding transcription factors. These binding motifs are generally represented by position-weighted matrices (PWM). Such consensus-binding motifs for hundreds of transcription factors have been identified. However, the binding of a transcription factor is not a simple function of the affinity for its cognate DNA sites, but also of a number of other factors like the chromatin environment and the cooperation or competition with other DNA-binding proteins. As has been observed in higher eukaryotes, transcription factors do not always operate by
themselves, but instead bind to DNA in cooperation with other DNA-binding proteins.

In addition, the regulation of gene transcription depends on interactions among transcription factors and RNA polymerase. This limits the binding locations of transcription factors in vicinity of binding position of RNA polymerase. For example, several cis regulatory elements occur at a specific distance relative to the transcription start site (TSS) (Pilpel et al., 2001). Moreover, several cis elements might occur in the same promoter with restricted spacing between them and alteration of mean distance between these has functional consequences. As an example, increased spacing between the GC-box and the TATA-box in adenovirus 2 E1B promoter diminishes in vivo transcription significantly (Beer and Tavazoie, 2004). There are numerous examples of such positional and spacing restrictions showing importance of positions of transcription factor binding sites (Hannenhalli and Levy, 2003; Thompson et al., 2004).

**Current in silico approaches for building cis-regulatory elements for gene expression and their limitations**

Computational methods for modelling and identification of cis-regulatory elements have been developed over the past two and a half decades (Bulyk, 2003; Brazma et al., 1998b; Brazma et al., 1998a; Wasserman and Sandelin, 2004; Stormo, 2000; Pavesi et al., 2004). Additionally, orthogonal information from comparative genomics or co-regulation at the transcriptional level has also been widely used to identify cis-regulatory sites (Wasserman et al., 2000; Gelfand et al., 2000). Broadly, there are two types of approaches used for identification of cis-regulatory element based on
transcription factor binding sites. The first and most commonly used approach is in silico identification of TF combinations. In this case, interacting TFs are predicted by either enrichment analysis of co-occurring TFBSs on the upstream sequences of genes relative to appropriate background sequences or by comparative genomics using phylogenetically conserved sequences between closely related species. The second approach makes use of available experimental data such as gene expression, chromatin immunoprecipitation (ChIP)-chip, and protein-protein interactions.

It is now well established that regions in the genome that have been conserved over long evolutionary periods are more likely to be functional as regulatory sequences accumulate fewer mutations during evolution than non-regulatory sequences. This is exactly the premise underlying the, now well established, approach of Phylogenetic Footprinting. However, for a genomic region to be functional, evolutionary conservation is important but not sufficient. Besides conservation, identification of shared regulatory signals in the genes that are co-regulated or co-expressed in a specific process further improves the signal-to-noise ratio.

However, current approaches for predicting and building such transcriptional regulatory networks are limited in their applicability. Most of these works have exploited the co-occurrence of promoter motifs to predict interacting transcription factors (Grace et al., 2004; Lim et al., 2004) to model expression regulation (Butler and Kadonaga, 2001; Spek et al., 1999) and to detect regulatory modules (Hannenhalli and Levy, 2002; Wu and Berk, 1988) and some of these works impose specific distance constraints between co-occurring motifs. However, very few existing
methods are designed to be applied on a genome-wide scale without prior knowledge about sets of interacting TFs or sets of co-regulated genes.

Despite the success of above approaches, both have notable limitations. While computational approaches can predict TF combinations on a large scale, but they usually lack the ability to functionally annotate interacting TFs. Furthermore, although methods based on phylogenetically conserved sequences can greatly reduce the false prediction rate, these have limitations related to missing potentially significant observations. Moreover, an important technical issue arises if the species analyzed are very closely related. In such a case, non-functional sequences may not have diverged enough to allow functional sequence motifs to be identified. In comparison, if the species are distantly related, short conserved regions may be masked by non-functional background sequences. On the other hand, the approach based on experimental observation needs *a priori* knowledge, such as gene expression patterns in a certain tissue, which restricts interacting TF determination to those tissues or cells studied and thus prevents the discovery of TF combinations from multiple biological conditions.

The individual limitations of the above mentioned methods could be addressed by using them in parallel. Therefore we propose a combinatorial approach to identify regulatory network by analyzing transcription factor binding site position, evolutionary conservation and expression data for transcription factors.
Chapter 2: Building positional network of transcription factors binding sites from predicted as well as available experimental information about target sites

2.2 Results

2.2.1 A strategy for building positional network of transcription factor binding sites

The aim of the present study was to predict, build and characterize the positional networks of transcription factors which might constitute a regulatory module which takes care of inter-distance between TFBS and does not require prior knowledge of gene expression patterns. To identify associated TF pairs, three levels of selection were adopted. First, a pair should show specific inter-distance between their binding sites on common target genes. Second, they should also show evolutionary conservation of enriched inter-distances. Third, their common target genes with enriched inter-distance should show co-expression. Such an approach should greatly facilitate discovery of novel transcriptional regulatory modules in conditions of health and disease e.g. cancer which are marked by dysregulation of gene expression. A schematic representation of the approach is described in Figure 1.

Figure 2.1 A schematic representation of the positional transcription factor binding sites network approach In the first step, we identified target genes common to two transcription factors, which show similar inter-distance between their binding sites across many promoters, or, in other words, could be co-regulating target genes. In the second step, conservation of the inter-distance was checked across organism (human, mouse, rat and chimp). Finally similarity in, expression was checked for target genes common to both transcription factors.
2.2.2 Significant association of TF pairs govern co-expression of target genes

2.2.2.1 Distance specificity of TRANSFAC TF pairs

TRANSFAC (Wingender et al., 1996) database was used to identify Transcription factor Positional Weight Matrices (PWM). We selected 220 TFBS (vertebrate non-redundant PWM) representing 184 TFs to scan human, mouse, rat and chimpanzee promoter sequences to find all possible TFBSs. It was assessed how often a TF pair preferentially occurred at a specific distance from each other. The binding coordinates of each TF were evaluated with rest of TFs generating 184 x 184 TF-TF combinations. Here, the position of individual TF binding sites on promoters were irrelevant and only the distance between binding sites of two TFs was of concern. For each TF pair inter-distances between their binding sites on common target promoters were calculated. Next, for each TF pair, the inter-distance range (1b to 10kb) was divided into 100 bp bins and the frequency of inter-distance occurrence per bin was computed. Fisher’s exact test was used to determine enriched bin by comparing calculated frequency to expected occurrence i.e. random distribution. Expected occurrence per bin was calculated by dividing total occurrence per promoter by total bins analyzed (100 in this study). The TF pairs which showed significant enrichment in at least one bin ($P<0.05$) were chosen. The analysis revealed that out of 33856 TF pairs analyzed, 314 pairs showed significant enrichment at least one bin. Majority of TF pairs showed 100b as enriched inter-distance between their binding sites, however longer inter-distances were also observed in case of few TF pairs. Figure 2.2 shows the distribution of inter-distance for the 314 TF-pairs.
Chapter 2: Building positional network of transcription factors binding sites from predicted as well as available experimental information about target sites

Figure 2.2 Distribution of inter-distance for the 314 TF-pairs Distribution of the distance-specificity for the 314 TF pairs in 10kb region is shown by heat map. Patterns of Proximal as well as distant inter-distance were observed among selected TF pairs. Scale shows percentage of co-occurrence (TF binding sites) per 100 base bins. x-axis contains inter-distance.

2.2.2.2 Evolutionary conservation of TRANSFAC TF pairs binding inter-distance

In order to get functional TF pair (cis-regulatory modules), the second level of filtration criteria i.e. conservation of inter-distance across human, mouse, rat and chimpanzee was applied. This involved TF-TF co-occurrence analysis on 19730 chimpanzee, 18785 mouse and 14292 rat promoter sequences and calculation of significantly associated TF pairs. Fisher’s exact test was used to obtain significant inter-distance for each TF pair in individual organism. Following this, significantly enriched inter-distance bin of human TF pairs was compared to their respective chimpanzee, mouse and rat counterparts for conservation. Figure 2.3A shows schematic representation of approach.
Chapter 2: Building positional network of transcription factors binding sites from predicted as well as available experimental information about target sites

This approach identified 300 TF pairs where the significantly enriched inter-distance bin was found to be conserved in all four species. For this analysis, only the TF pairs which had showed significant inter-distance in human were considered. Figure 2.3B shows inter-distance conservation for 184 x 184 TF pairs across human, mouse, rat and chimpanzee. It was evident from the analysis that most of TF pairs shared same inter-distance in human, mouse, rat and chimpanzee. For example, as shown in inset, 100 bp was the enriched inter-distance between the binding sites of MyoD and Ebox on target genes in all the four organisms.

2.2.2.3 Functional relevance of inter-distance specific TF pairs

It was investigated whether the presence of a TF pair at a specific inter-distance range correlated with the expression pattern of their target gene. To assess the correlation between inter-distance specificity and the target gene expression, the transcriptome profile of 79-normal human tissues (Su et al., 2004) was analyzed by comparing target genes with specific inter-distance (TF-TFd) and 20 random set of genes with non-specific inter-distances (TF-TFNod).

The analysis revealed that target genes of TF-TFd showed statistically significant co-expression across 79 tissues in contrast to the genes with nonspecific inter-distance (z-score > 4.0). Strikingly, expression of genes with TF-TFd was concomitant with expression of TF pairs itself (Figure 2.4). This observation also suggests that among the all co-targeting genes, genes with TF-TFd were true targets of TF pair.
Chapter 2: Building positional network of transcription factors binding sites from predicted as well as available experimental information about target sites

Figure 2.3 Inter-distance conservation for 184 x 184 TF pairs across human, mouse, rat and chimpanzee (A) Schematic representation of inter-distance and conservation analysis. (B) Transcription factor pairs with preferred inter-distance show conservation of inter-distance across four species. TF pairs which did not show significant inter-distance in human were excluded in this analysis. A heat map to test whether inter-distance found in human was conserved in other species. Inter-distance distribution of MyoD and Ebox in human and other species is shown below.

Figure 2.4 Expression of target genes with TF-TF$^d$ across 79 tissues (A) Target genes of TF pair with TF-TF$^d$ show similar expression across 79-normal human tissues compared to target genes of TF-TF$^{NOd}$. Mann-whitney score (z-score) calculated for target genes of all TF pairs across 79-normal human tissues, Heat map showing z-score. (B) An example was highlighted in insert showing correlation between z-score (lower panel), expression of TFAP2A and SP1 (middle panel) and expression of target genes with TF-TF$^d$ (upper panel) across 79-normal human tissues.
The associated TF pairs found in this study are based on TRANSFAC motifs (PWMs) which are built on functional as well as predicted target sites and constitute all possible genomic occurrences. Therefore, the observations made here were validated using genome-wide experimentally determined TFBS. We noted that datasets from chromatin-immunoprecipitation followed by massive parallel sequencing (ChIP-seq) were available for transcription factors; AP2alpha, SP1 and EGR-1 (ENCODE Project). Using these datasets, the co-occurrence of AP2alpha, SP1 and EGR-1, SP1 binding sites was analyzed. Interestingly, the frequency of AP2alpha-SP1 and EGR-1-SP1 co-occurrence distribution as obtained in case of ChIP-seq was strikingly similar to the one observed with TFBS determined from TRANSFAC (Figure 2.5). These results indicated that the significantly associated TF pairs observed using TRANSFAC datasets are likely to be comparable to those inferred from genome-wide experimental analyses.

![Figure 2.5 Comparision between inter-distance distribution pattern of TF pair binding sites determined by TRANSFAC PWM and ChIP-seq experiment](image)

Experimentally determined datasets (ChIP-seq) show similar pattern of inter distance distribution as observed in PWM-based analysis. Distribution of inter-distance between AP2 alpha, SP1 and EGR1, SP1 across target genes showed in upper panel and lower panel respectively. Distribution calculated using ChIP-seq dataset and TRANSFAC PWM were shown in blue and red colored line, respectively.
2.2.3 Analysis of experimentally determined transcription factors binding sites to explore new TF-TF interactive partners

2.2.3.1 Distance specificity in TF pair binding sites determined by ChIP-seq experiments

Understanding the sensitivity of ChIP-seq experiment which provides genome-wide bindings of transcription factor with exquisite specificity in cellular context, we included genome-wide binding profile of 74 transcription factors from five different cell lines (ENCODE project) to predict global as well as cell line specific transcriptional regulatory modules. Specifically, transcription factor binding peaks present ±5kb to TSS of a known gene were chosen for analysis. The binding coordinates of each TF were evaluated with rest of TFs generating 2701 (74X74) TF-TF combinations. The analysis included first, calculation of inter-distances between TF pairs binding sites on common target promoters, second, calculation of a background random occurrence, and third, statistical analysis for significance.

The TF pairs which showed significant enrichment in at least one bin ($P<0.05$) were selected. Interestingly, it was observed that 620 TF pairs significantly co-occurred in at least one inter-distance bin. Surprisingly, no TF pair showed significant inter-distance beyond 300 bp range. Additionally, it was also noted that for many TF pairs, more than 40 percent of co-occurring sites were falling within significantly enriched inter-distance bin (Figure 2.6). These results suggest that most of TF pairs bind in proximate at maximum target genes. Additionally, high percent of co-occurring sites within enriched inter-distance contrary to TRANSFEC analysis were 15-20 percent co-occurring sites were within enriched inter-distance show specificity of chip-seq dataset in finding TF binding sites.
Building positional network of transcription factors binding sites from predicted as well as available experimental information about target sites

Figure 2.6 Inter-distance distribution in TF pair binding sites determined by ChIP-seq experiments High number of target sites show preferred inter-distance for several TF pairs. Pseudo color representing presence and abundance of co-target sites in enriched inter distance bin for TF pairs. Distribution of inter-distance for Four TF pairs is shown in insert (ATF3-SP1, USF1-SP1, USF1-SP1 and ATF3-STAT2).

2.2.3.2 Evolutionary conservation of transcription factor binding sites determined using ChIP-seq approach

Phylogenetic conservation is a powerful approach to identify functional regulatory elements. To compute conservation score for each TF binding site contributed in significant inter-distance, UCSC track for Vertebrate Multiz Alignment & Conservation (16 Species) was used. The first step was to calculate conservation score of each TF in a TF pair by dividing the number of nucleotides conserved across 16 species with total number of nucleotides in binding regions on each target. Second, the final conservation score was calculated by taking average of conservation scores of both TFs in a TF pair on each target site (Figure 2.7A). The binding sites of TF
pairs with final conservation score of more than 60 percent were considered as conserved. Interestingly, the analysis showed that, for 340 TF pairs the percentage of conserved binding sites was found to be significantly higher in TF-TF\textsuperscript{d} sets as compared to TF-TF\textsuperscript{Nod} sets (see section 2.2.1.3) (Figure 2.7B).

It has been proposed that genes comprising a given regulatory module are physically clustered on a chromosome. To test this, the genomic distribution of TF binding sites comprising preferred inter-distances was investigated. A subset of 340 TF pairs (with $P < 0.05$) obtained previously (Figure 2.7B) was taken up for analysis of inter-distances (with $P < 0.05$). For each chromosome, total genes harboring significantly associated TF pairs (TF-TF\textsuperscript{d}) were identified and compared to the chromosomal distribution of genes expected by chance alone. The investigation revealed that chromosomal distribution of TF-TF\textsuperscript{d} genes was different from (mention what was the control set). This suggested that genes regulated by TF pair with significant inter-distance have non-random distribution (Figure 2.8).

### 2.2.3.3 Target genes of TF pairs with specific inter-distance show co-expression

It was investigated whether the presence of a TF pair at a specific inter-distance range was correlated with the expression pattern of target genes. To assess the correlation between inter-distance specificity and the target gene expression, the transcriptome profile of 60 human cancer cell lines representing cancers of nine tissue origins were analyzed. All significantly associated TF pair target genes (TF-TF\textsuperscript{d}) and distantly associated TF pairs target genes (TF-TF\textsuperscript{Nod}) were chosen to examine co-expression pattern in NCI60 cell lines.
Chapter 2: Building positional network of transcription factors binding sites from predicted as well as available experimental information about target sites

Figure 2.7 Conservation analyses of TF pairs with significant inter-distance across 18 species
A) Schematic representation of conservation score analyses by presenting an example of AP2 alpha and ERRA binding on a target gene (details in method). B) Target sites with preferred inter distance (Group A TF-TF$^d$) compared to target sites with random inter distance (Group B TF-TF$^{Nd}$) show higher conservation for more than 300 TF pairs across 18 species; y-axis represents percentage co-targeted sites of TF pair with more than 50% conservation. Percentage of target sites from group A and group B represented by red and blue dots respectively. $P$ value calculated using chi-square test between group A and group B for each TF pairs plotted as line ($P \leq 0.05$ as cut-off).
Chapter 2: Building positional network of transcription factors binding sites from predicted as well as available experimental information about target sites

Figure 2.8 Representation for the chromosomal distribution of genes harboring co-occurring TF pairs with significant inter-distance

Chromosome ideograms with gene density were shown around the outer ring and are oriented pter–qter in a clockwise direction with centromeres indicated in red. Other tracks contain E2F6 binding distribution (Red track): MAX binding sites (blue track); sites where E2F6 and MAX binds within 100 base (100b is significant inter-distance observed in analysis for E2F6 and MAX) (purple track); first and second from inside to outside were track for showing chromosomal distribution of human genes (expected) and E2F6 and MAX target genes with significant inter-distance respectively.

The expression values of TF-TF$^{d}$ set of genes was compared with 10 random set of genes from TF-TF$^{NOd}$ set in NCI60 expression datasets. The results indicated significant co-expression ($z$-score > 4.0) for TF-TF$^{d}$ set in majority of TF pair across NCI60 cell lines (Figure 2.9A).
Chapter 2: Building positional network of transcription factors binding sites from predicted as well as available experimental information about target sites

Figure 2.9 Expression of target genes with TF-TF<sup>d</sup> across NCI-60 cell lines A) Target genes of TF pair with TF-TF<sup>d</sup> show similar expression across 60 cancer cell lines compare to target genes with TF-TF<sup>NOd</sup>, Mann-whitney score (z-score) calculated for target genes of all TF pairs across 60 cell lines, Heat map showing z-score. B) TF interacting network based on experimentally determined binding sites. Each node representing individual TF; black edge shows known interacting partners; red line shows novel interactions based on our positional network approach.

To identify known interacting partners among predicted TF pairs using positional network approach, we compared our prediction with publicly available databases PubGene (Jenssen et al., 2001) (http://www.pubgene.org/), STRING (von et al., 2007)
Interestingly, we found that out of 340 TF pairs, 96 had been previously shown as interacting partners across different biological process (Figure 2.9B; TF pair list provided as Appendix 1).

2.2.3.4 Transcription factor positional network in K562 cells

To identify TF interaction networks involved in cell type specific manner, we considered K562 leukemia cell line. Several studies used K562 as cellular model for CML. Recently a study showed K562 cell lines a good models for studying drug resistance mediated by P-glycoprotein in chronic myelogenous leukemia (CML) (Hait et al., 1993). Additionally, 28 out of 74 ChIP-seq experiments (dataset available) were performed in K562 cell lines (TF details provided as Appendix 2).

Therefore, we performed our analyses using TFs for which ChIP-seq experiments were performed in K562 cell line. Additionally, we downloaded expression profile for K562 from GEO (http://www.ncbi.nlm.nih.gov/geo/). Interestingly, we found that most of target genes of TF-TF\textsuperscript{d} genes showed co-expression in K562 expression dataset (Figure 2.10A). We found that many TF pairs predicted with positional network approach were previously shown to work together in leukemia For example GATA-1 - c-JUN (Elefanty et al., 1996), MYC-MAX (Si et al., 2010) and c-Jun – c-Fos (Motiwala et al., 2011).

This approach also led to identification of several novel TF interaction partners (Figure 2.10B). Many of the TF partners already implicated in leukemia like TF2B which involved in growth inhibit in leukemia LH-60 cells (Tu et al., 2004). ATF3
promotes proliferation of adult T-cell leukemia cells (Hagiya et al., 2011). YY1 expression suppresses cell proliferation in K562 cells (Shi and Rui, 2011).

Figure 2.9 Transcription factor positional network in K562 cells A) Target genes of TF pair with TF-TF\(^5\) show similar expression across three samples of K562 cell lines compare to target genes with TF-TF\(^{NO}\); Mann-whitney score (z-score) calculated for target genes of K562-specific TF pairs across replicates of K562 cells, Heat map showing z-score. An example was highlighted in insert showing expression of E2F6 and MYC target genes with significant inter-distance across three samples of K549 cells (right panel). B) K562 specific TF interacting network based on experimentally determined binding sites in K562 cells. Each node representing individual TF; black edge shows known interacting partners (searched from STRING, PubGene and INTERPRO); red line shows novel interactions based on our positional network approach.
2.3 Material and Methods

2.3.1 Sequence Retrieval

The ± 5 kb region centered at annotated TSS of 19793 human, 19730 chimpanzee, 18785 mouse and 14292 rat non-redundant promoter sequences were retrieved from UCSC build hg18 for human, PanTro2 for chimpanzee, mm9 for mouse and rn4 for rat.

2.3.2 Analysis of transcription factors binding site

For TRANSFAC PWM based analysis, 220 PWMs were considered which represented potential TFBS for 184 unique TFs. Promoter sequence retrieved for human, chimpanzee, mouse and rat were analyzed for presence of these TFBS using MATCH™ (TRANSFAC® professional 12.1).

For experimental dataset, ChIP-Seq data for 74 TFs were used and downloaded from ENCODE website (http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=wgEncodeYaleChIPseq). Next, peaks for each factor were mapped to nearest gene using UCSC known gene track file for hg18 build. For further studies, only those peaks which were mapped within ±5kb to TSS for known genes were used. To maintain consistent size for binding sites, ±10b from center of peak was considered.

2.3.3 TF-TF inter-distance Analysis

For both PWM based binding sites and ChIP-seq binding sites, analysis for inter-distance was performed in similar fashion. The binding coordinate of each TF was compared with rest of TFs generating n*n TF combinations. Next for each TF pair inter-distances between their binding sites on common target promoters were calculated. For each TF pair, the inter-distance range (1b to 10kb) was divided into 100 base bins and computed.
frequency of inter-distances per bin. To calculate statistical significance of inter-distance enrichment, observed frequency of each inter-distance bin was compared to expected frequency. Expected frequency was calculated by total co-occurrence of TF pair divided by number of bin considered i.e. 100 (10,000 base (total region consider)/ 100 base (size of bin considered)) in our case. Then, for each bin significance of difference between actual and expected frequency was evaluated using Fisher’s exact test with a cut-off of \( P < 0.05 \). A TF pair showing significant enrichment in at least one bin was used for the next step of approach i.e. conservation of that inter-distance across other organisms. In case of PWM based approach similar inter-distance analysis was performed on chimpanzee, mouse and rat TFBS.

2.3.4 **Evolutionary conservation of TF-TF inter-distance**

Regulatory elements are often characterized by evolutionary conservation. Therefore, this step was used to reduce false positives in the present approach for identification of regulatory modules. For TRANSFAC approach, three additional genomes (chimpanzee, mouse and rat) along with human were considered. For each organism, promoter sequence extraction, identification of TFBS and analysis of inter-distance were done independently. To check conservation of human TF pair inter-distance, same TF pair was considered from other organism and check whether they shows same inter-distance as preferred distance between there binding site on other organisms also. A TF pair which showed same inter-distance in all four genomes was considered for further study.

For ChIP-seq dataset, conservation of TF pair inter-distance was checked across 16 genomes. For each TF pair all co-occurred sites were distributed into two groups, first
where TF pair binding sites were separated by specific inter-distance (TF-TF\textsuperscript{d}), second where sites were separated by non-specific inter-distances (TF-TF\textsuperscript{NOd}). To check for conservation, data representing multi-alignments of 16 vertebrate genomes with human was extracted from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/). Then, for each occurrence of TF pair, binding site of both partners (e.g. TF1 and TF2) was checked individually for conservation across 16 species. Using alignment data, a percentage score for both TF1 and TF2 binding site was prepared by looking for base by base alignment. A final conservation score was calculated by taking average of percentage score of TF1 and TF2. Finally, for each TF pair, number of sites which showing more than 60 of conservation score (i.e. 60 percent total bases from site conserved across species which termed as conserved site) was compared between the two groups (TF-TF\textsuperscript{d} and TF-TF\textsuperscript{NOd}).

### 2.3.5 Expression data, normalization and significance analysis

For TRNASFAC TF pair analysis, tissue-specificity was checked in 79 human tissues. Analysis was done based on a previously described method (Xie et al., 2005). The expression data of each gene across all tissues was first normalized to be mean 0 and variance 1 before ranking them as per their normalized expression level in each tissue, hence generating 79 tissue-specific ranked gene lists.

For ChIP-seq TF pair analysis, NCI60 cell line expression data were used. Expression data were downloaded from NCI website using Cellminer program (Shankavaram et al., 2009). The NCI60 is a set of 60 human cancer cell lines derived from diverse tissues: brain, blood and bone marrow, breast, colon, kidney, lung, ovary, prostate and skin. The expression data of each gene across all cell lines was normalized by z-score
transformation method (Cheadle et al., 2003). For cell line specific positional network, K562 expression data was used. Data was downloaded from GEO (http://www.ncbi.nlm.nih.gov/geo/) under GEO ID GSE12056. Expression data was normalized by Z-score transformation.

In Z-score transformation method, raw intensity data for each experiment is log10 transformed. Z scores are calculated by subtracting the overall average gene intensity (within a single experiment) from the raw intensity data for each gene, and dividing that result by the SD of all of the measured intensities, according to the formula:

\[ Z \text{ score} = \frac{\text{intensity}_G - \text{mean intensity}_{G1...Gn}}{\text{SD}_{G1...Gn}} \]

Where \( G \) is any gene on the microarray and \( G1...Gn \) represents the aggregate measure of all of the genes.

To assess the correlation between inter-distance specificity and the target gene expression different sets was generated. Target genes with specific inter-distance (TF-TF\text Superscript d) were compared with 20 random set of genes with non-specific inter-distances (TF-TF\text Superscript NOd). Significance was calculated by Mann-Whitney rank sum statistic where, enrichment of expression of each gene set \( S \) in a particular tissue, and its significance was analyzed from the ranked list of genes \( T \), specific for that tissue after evaluating the non-randomness of ranks of \( S \) within \( T \), using the Mann-Whitney rank sum statistic. After summing the ranks of \( S \) in list \( T \), the significance of this rank sum was tested against the rank sum of control set (all genes in \( T \), excluding \( S \)). If \( \mu \) and \( \sigma^2 \) are the mean and variance of the control set then enrichment (Z-score) of \( S \) is given by \( (\mu - S)/\sigma \), which measures enrichment in terms of number of standard deviations away from the mean of the control set. A z-score of 4.0 is considered to be significant in the present study.
2.4 Discussion

An integrative approach for identification of novel TF interactions

Previous methods to identify TF pairs used TFBS enrichment in co-expressed genes or evolutionary conservation. In comparison, this study used a strategy of three step filtrations to predict interacting TF pairs: (1) distance constraint between TF binding sites, (2) evolutionary conservation of inter-distance, and (3) co-expression of genes containing significant inter-distance. In distance constraint, the importance of similar spacing between binding sites of interacting TFs on target promoters was assessed. Previous works showed cis-regulatory motifs to be distance constrained. A study in *Escherichia coli* ranked cis-regulatory motifs based on the enrichment of specific spacing between them, and experimentally validated their functionality in terms of binding sites (Bulyk et al., 2004). Several other studies, exploited the co-occurrence of promoter motifs to predict TF interactions and TF modules (Hannenhalli and Levy, 2003; Thompson et al., 2004). In evolutionary conservation, functional conservation for the identification of TF combinations was considered, and a computational approach was developed to implement this strategy. These approaches included selection of TF combinations based on the similar pattern of binding site arrangement on promoters across organisms. In expression data analysis, a statistical method to identify co-expression patterns for genes harboring TF-TF\textsuperscript{d} was used, and Mann-Whitney rank sum statistic was used to quantify expression pattern difference between TF-TF\textsuperscript{d} and TF-TF\textsuperscript{Nod}.

The proposed model of ‘positional network of TFBS’ was tested using TRANSFAC database, which contained PWMs based on functional as well as predicted target sites for TFs and constitutes all possible genomic occurrences. However, the binding of a
transcription factor is not a simple function of the affinity for its cognate DNA sites, but also of a number of other factors like chromatin environment. Therefore, available TF binding sites determined by ChIP sequencing (ChIP-seq) was also used. ChIP-seq is a new method for genome wide mapping of protein binding sites on DNA, where chromatin immunoprecipitation (ChIP) is performed using TF specific antibodies which is followed by next-generation sequencing of enriched DNA fragments to identify transcription factor binding sites in vivo on the whole-genome scale. ChIP–seq has higher resolution, fewer artifacts, greater coverage and a larger dynamic range than any other method to determine TF binding sites. Thus, the inclusion of ChIP-seq data was expected to identify additional insights into transcriptional control by interacting TFs. In current study, 74 TFs were analyzed. However, it is anticipated that with advancement and availability of ChIP-seq data for more transcription factors, the analyses would improve further and generate many more novel interactions.

A major criticism of this approach could be combining TF binding sites determined from different cell type which came with assumption that TF binding sites determined in one cell type might also be same in others. To overcome this problem, a filter was applied in which genes bound by TF were checked for correlation with the expression TF itself in each cell type. Subsequently, the genes which showed significant correlation were used to build positional network. With information for each cell type becoming gradually available, construction of cell type specific networks should be possible.