14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

PROBABILISTIC DATA INTEGRATION AND VISUALIZATION FOR UNDERSTANDING TRANSCRIPTIONAL REGULATION

Arvind Rao [b,c,d], *Alfred O. Hero* [a,b,d,f], *David J. States* [b,e], *James Douglas Engel* [c]

Departments of [a]Biomedical Engineering, [b]Bioinformatics, [c]Cell and Developmental Biology,
[d]Electrical Engineering and Computer Science,[e]Human Genetics,[f]Statistics,
The University of Michigan, Ann Arbor, MI

ABSTRACT

In this paper we propose a manifold embedding methodology to integrate heterogeneous sources of genomic data for the purpose of interpretation of transcriptional regulatory phenomena and subsequent visualization. Using the *Gata3* gene as an example, we ask if it is possible to determine which genes (or their products) might be potentially involved in its tissue-specific regulation - based on evidence obtained from various available data sources. Our approach is based on co-embedding of genes onto a manifold wherein the proximity of neighbors is influenced by the probability of their interaction as reported from diverse data sources - i.e. the stronger the evidence for that gene-gene interaction, the closer they are.

1. INTRODUCTION

Below we give a characterization of what we mean by transcriptional regulatory networks. As the name suggests, gene A is connected by a link to gene B if a product of gene A, say protein A, is involved in the transcriptional regulation of gene B. This might mean that protein A is involved in the formation of the transcriptional complex [10] which binds at the promoter or regulatory element of gene B to drive gene B regulation. This is indicated below:

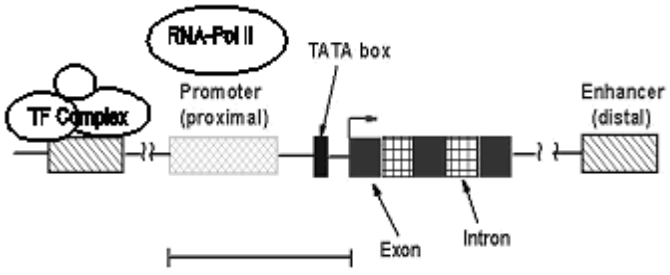

Figure 1: Schematic of Transcriptional Regulation.

As can be seen, the components of the Transcription Factor (TF) Complex, shown in Fig. 1, are the products of several genes. Therefore, the incorrect inference of a transcriptional regulatory network can led to several false hypotheses about the actual set of genes affecting a target gene. Since biologists are increasingly relying on computational tools to guide experiment design, a principled approach to biologically relevant network inference can lead to significant savings in time and resources. To make the inference of these networks relevant for a biologist to design useful experiments it would seem imperative that we incorporate biological knowledge to an extent suitable for making such network inference meaningful. In this paper we try to combine some of the other available data (protein-protein interaction data and phylogenetic conservation of binding sites across genomes) combined with mRNA expression features to build 'proximity maps' such that TF encoding genes lie close to the genes whose transcription they are involved in. These proximity maps not only aid visualization, but also, by construction, integrate interactions from diverse data sources. A straightforward interpretation of such a proximity map is that the stronger the evidence for a true TF gene - target gene interaction, the closer they would lie.

2. WHY BUILD PROXIMITY MAPS ?

As already mentioned above, the mechanism of regulation of a target gene is via the binding site of the corresponding Transcription factor (TF). It is believed that several TF motifs might have appeared over the evolutionary time period due to insertions, mutations, deletions etc in the vertebrate genomes. However, if we are interested in the regulation of a process which is known to be similar between several organisms (say Human, Chimp, Mouse, Rat and Chicken), then we can look for the conservation of functional binding sites over all these genomes. This helps us isolate the functional binding sites, as opposed to those which might have randomly occurred. This however, does not suggest that those other binding sites (TFBS) have no functional role. Since we are interested in the mechanism of regulation of the *Gata2/Gata3* genes (which are known to be implicated in mammalian nephrogenesis), we examine their promoter regions for phylogenetically conserved TFBS (Fig. 2). Such information can be obtained from most genome browsers [6].

We observe that even for a fairly short stretch of sequence (1 kilobase) upstream of the gene, there are several conserved sequence elements which are potential TFBS (light grey regions). In this work, we are focusing on the TFBS at the promoter upstream of the gene. Since we have data from phylogeny, protein-DNA interactions as reported by ChIP (chromatin-immunoprecipitation) assays, as well as microarray expression data, the presence of a TFBS for a TF which is known to have a DNA binding motif at the promoter as well as whose expression is correlated with *Gata2/Gata3's* expression indicates very strong evidence for that TF to be functional (involved in the target gene's regulation). Given the large number of TFs which are phylogenetically conserved at the promoter, we would need an approach to re-

duce the number of candidates for experimental validation to a much more confident subset. From here onwards, for the purpose of illustration, we continue with the *Gata3* example to demonstrate our methodology.

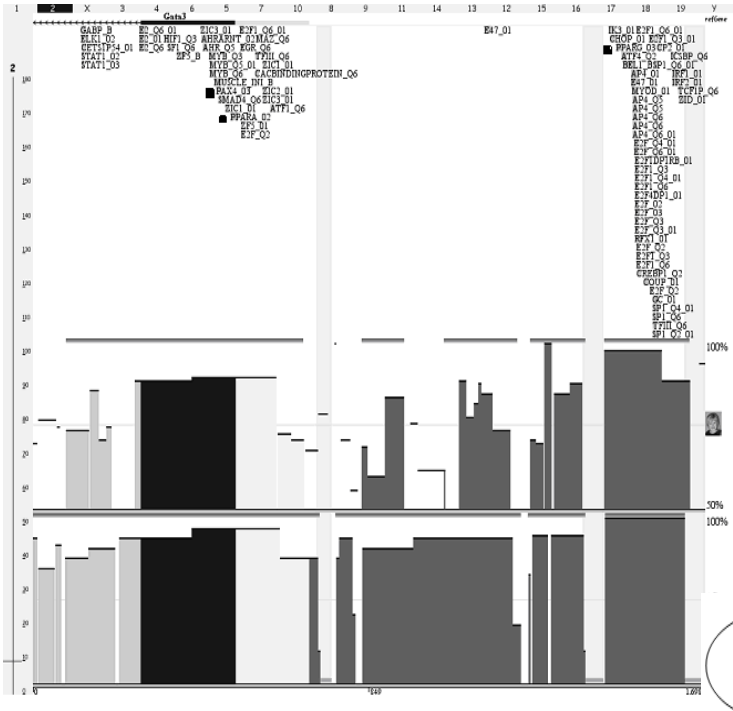

Figure 2: TFBS conservation between Human and Rat, upstream of *Gata3*, the square dots represent TFs which are known to be functional for regulation.

As can be seen above, there are atleast fifty Transcription Factors (TFs) which could possibly bind in the region 1kb upstream of the *Gata3* gene [6]. It would be very useful to find which TFs would bind in the given context of the biological process. In the normal course, every TF is a candidate since its binding site is phylogenetically conserved. Performing a ChIP assay to determine if the TFBS is true is a laborious exercise for this huge set of candidate TFs. But the presence of other sources of data can help us guess which TFs might be really functional to obtain a set of high confidence TF candidates for the assay. This is seen to be extremely useful in understanding tissue specific regulation too, since each tissue provides an environment in which a different set of TFs can possibly bind at the promoter for triggering tissue-specific spatio-temporal expression. Integrating tissue microarray data along with known interactions obtained from other sources can reduce the experimental overhead significantly.

The proximity maps, would place the functional TFs in close vicinity of the target gene (*Gata3*) here. The experimentalist can then look at a small neighborhood around the gene of interest and come up with a list of all the TFs that are possibly functional for transcription, under the evidence provided.

3. SETUP

Computational inference of transcriptional regulatory networks from diverse data has proved to be a bigger challenge than previously imagined. The gold standards for each data source are highly variable, and considering the diversity of interactions that each experimental or computational method aims to recover, their meaningful integration for the purpose of understanding underlying phenomena is a non-trivial task. For this study, we examine three kinds of data sources, two of which are experimentally derived (protein-protein interaction assays, phylogenetic conservation of Transcription Factor Binding sites (TFBS)) and the third is a computational measure (Directed Information) [3] for inferring interactions.

Our objective is to demonstrate that not only is this method scalable to as many kinds of 'relevant' data sources but also encompass both experimental and computational measures of association. Our approach is to construct an interaction probability matrix between K genes under consideration. This matrix is a $K \times K$ matrix with $P(i,j) = P(Z_{i,j} = 1)$, the probability that there is a 'true' functional interaction between the genes i and j, denoted by the event $Z_{i,j} = 1$. This true interaction depends on the probability that the l^{th} data source confirms this interaction (i.e. $Z^l_{i,j} = 1$). If we have L (=3, here) different data sources, we can write this as:

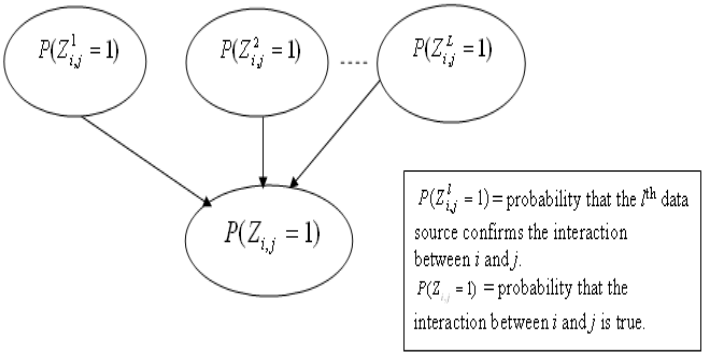

Figure 3: A schematic for how data integration over diverse data sources may be done - each data source is an expert suggesting the probability of functional interaction in its realm, and the bottom node combines the votes confirming or denying that interaction.

$$p_{i,j} = P(Z_{i,j} = 1 | Z^1_{i,j} = 1, Z^2_{i,j} = 1, \ldots, Z^L_{i,j} = 1)$$
$$= \frac{P(Z_{i,j} = 1, Z^1_{i,j} = 1, Z^2_{i,j} = 1, \ldots, Z^L_{i,j} = 1)}{P(Z^1_{i,j} = 1, Z^2_{i,j} = 1, \ldots, Z^L_{i,j} = 1)}$$

By the conditional independence of the various data sources $1, 2, \ldots, L$, the joint distribution (in the denominator) factors into the product of the marginal distributions $P(Z^l_{i,j})$, thus,

$$p_{i,j} = \frac{P(Z_{i,j} = 1) \prod_{l=1}^{L} P(Z^l_{i,j} = 1 | Z_{i,j} = 1)}{\prod_{l=1}^{L} P(Z^l_{i,j} = 1)}$$

$$= P(Z_{i,j}=1)\prod_{l=1}^{L}\frac{P(Z_{i,j}=1|Z^l_{i,j}=1)P(Z^l_{i,j}=1)}{P(Z_{i,j}=1)}\cdot \cdot\frac{1}{\prod_{l=1}^{L}P(Z^l_{i,j}=1)}$$

$$= \frac{\prod_{l=1}^{L}P(Z_{i,j}=1|Z^l_{i,j}=1)}{\mathscr{Z}} \qquad (1)$$

with $\mathscr{Z} = [P(Z_{i,j}=1)]^{(L-1)}$. Also, if the expected number of interactions per protein/gene pair is I, with N entities in the protein/gene universe, then the probability of true interaction, $P(Z_{i,j}=1) = \frac{I}{N}$. Note, we have assumed that the the nature of protein-protein or gene-gene interactions is similar, and that there is no notion of a favorable occurrence of some interactions over the others. A more rigorous analysis would require a knowledge of the interaction network structure of the corresponding proteome/genome.

Thus, from (1), we see that the existence of a 'true' functional relation between two genes i and j, depends on $p_l = P(Z_{i,j}=1|Z^l_{i,j}=1)$ which is computed from a joint histogram of the training data for a particular (l^{th}) data source. This reflects the degree of confidence that biologists have come to associate with the interactions predicted from the l^{th} data source. The multiplication of posterior probabilities is equivalent to the addition of log-likelihoods of generation from each of the various data sources. This is a specific instance of a graphical model formalism [1,5] within such a framework. The expression (1) above decomposes the overall structure of the relationship into a product of marginal conditionals due to the assumed independence of the various data sources.

For the purpose of both visualization and integration of these diverse data sources with a goal to recover biologically relevant relationships, we now explore manifold embedding [2,4] as a method to incorporate the probability weights obtained from the interaction probability matrices to bring those genes closer which have a higher probability of interaction. Manifold embedding helps to understand the local structure of the data points, instead of imposing a global structure on them. This is particularly useful in our scenario since we only have evidence for the immediate upstream affectors of a gene, and not of its global network structure. What we can hope to recover is a fairly accurate picture of the global transcriptional network by piecing together the evidence of the local interactions. For understanding transcriptional regulatory mechanisms, it can be hypothesized that the genes in close vicinity to a gene of interest are either co-regulated or potentially involved in the regulation of the target gene (through its product). A good embedding would use these diverse data sources to reflect such relationships.

4. MODIFIED LAPLACIAN EIGENMAP EMBEDDING

Suppose we are investigating the role of $(K-1)$ genes in relation to our target gene (*Gata3*) - we proceed as follows:

- Standardize these K gene expression profiles to 0 mean and unit variance. Notice that the Euclidean distances become the Pearson correlation measure.
- Build the $K \times K$ dimensional weight matrix W from the Hadamard product of the L interaction probability ($P(Z^l_{i,j}=1)^L_{l=1}$) matrices, from each of the L different sources of data.
- Find n Nearest Neighbors using the Euclidean distance (or within some ε-neighborhood). Assign weight $W_{i,j} = p_{i,j}$, from (1) for the pair (i,j), for each of the $\binom{K}{2}$ gene pairs.
- Form the Graph Laplacian [2]:

$$L_{i,j} = \begin{cases} d_i = \sum_k W_{i,k} & \text{if } i = j; \\ -W_{i,j} & \text{if } i \text{ is connected to } j; \\ 0 & \text{otherwise.} \end{cases}$$

- Solve: $min_y y^T L y = \frac{1}{2}\sum_{i,j}(y_i - y_j)^2 W_{i,j}$ (2), subject to:
  - $y^T D y = 1$ and
  - $y^T D \mathbf{1} = 0$

  , where $D_{i,i} = \sum_j W_{j,i}$, a diagonal weight matrix.
- Embed the co-ordinates to a lower dimensional manifold, using the solution (the Laplacian Eigenmap) obtained from the minimization above.
  - The solution to (2) is given by the d generalized eigenvectors associated with the d smallest generalized eigenvalues solving $L\mathbf{y} = \lambda D\mathbf{y}$.
  - If $\mathbf{y} = [y_1, \ldots, y_d]$ is the collection of these eigenvectors, then the embedding is given by : $y_i = (y_{i1}, \ldots, y_{id})^T$, i.e., the d dimensional representation of the i^{th} data point (gene).
- In our representation, we take dimensionality, d = 2 and number of neighbors, n = 5.

5. INTEGRATING DIVERSE DATA SOURCES

We demonstrate the utility of the presented approach to understand the mechanisms underlying transcriptional regulation of the *Gata2/Gata3* genes in the developing kidney [3]. The primary source of data used to obtain distances is the microarray expression profiles of 47 genes known to be co-expressed with *Gata3* in the embryonic kidney. These genes were obtained after screening for differential expression between two tissue types in the embryonic kidney - the Ureteric Bud (UB) and the Metanephric Mesenchyme (MM). This characteristic is a discriminating one for the set of genes which have behavior similar to *Gata2/Gata3*. This gene expression data are obtained from http://genet.chmcc.org [9]. A large amount of data encompassing literature mining, microarrays, protein-protein interactions have been available from the STRING database (http://string.embl.de/) - for most of the

$K = 48$ genes selected above, a lot of functional information from several experiments is available. For our purpose, we find the strength of association between any two genes i and j using significance scores from three different sources:

- Phylogenetic conservation of protein i's binding site in the upstream region of gene j.
- Interaction of Protein i with Protein j OR ChIP evidence for protein i's interaction with a DNA-domain in gene j's promoter.
- Directed Information [3], measuring causality in expression of gene j due to gene i - based on microarray expression.

Most of the data sources report significance scores as log-likelihood scores or p-values. These scores are then standardized for true positives, for example, a p-value from ChIP of 0.15 can be seen to be significant and predictive of a true interaction, whereas in a computational measure a more stringent p-value like 0.05 might be necessary to infer true interaction. From here, a joint histogram of true vs. predicted interactions, under each data source (l), can be obtained and used for the evaluation of the probability p_l as pointed to above. We note that the scores derived from the three sources are not symmetric - hence for our purpose we take the events $(Z_{i,j} = 1)$ and $(Z_{j,i} = 1)$ to be equivalent. However, inference needs to be done with regard to the biological process to avoid misinterpretation.

6. RESULTS AND DISCUSSION

A common approach used for studying transcriptional regulatory mechanisms is by association. The hypothesis underlying this is that if genes are co-clustered/correlated, they are co-regulated, i.e. have a common set of controls. Since we are interested in the transcriptional regulatory mechanisms of *Gata3*, we look for genes which are in a ε-neighborhood of *Gata3*. Several investigations have hypothesized the existence of regulatory modules [7] governing the co-expression of a gene set to hypothesize for the existence of a common control mechanism underlying their co-ordinated expression. The obtained embedding allows us to identify such gene groups of interest - which are within some neighborhood around *Gata2* or *Gata3* and whose expression profiles are co-ordinated with *Gata2/Gata3* expression. From the embedded manifold in two-dimensions as shown in Fig.4, we observe that the *PPAR*, *Lamc2*, *Pax2* genes are among several which are 'in close proximity' (and possibly functionally relevant in transcriptional regulation) to the *Gata3* gene. This is interesting since each of these have phylogenetically conserved TF binding sites (the three black squares in Fig.2) in the *Gata3* promoter. Recently, a mechanism for the regulation of *Gata3* by *Pax2* has been reported during kidney formation [8]. It is to be noted that *Gata3*'s family member *Gata2* is in the cluster on the top left, indicating that though it is expressed, the influence of the TF genes is more pronounced and hence, they are closer in the network. We note that this embedding has integrated information from three very different data sources to build this 'proximity map' of genes. These findings are currently being verified in the laboratory. It can be seen that not only can this approach combine known interactions graphically, but we can look for interactions not previously identified, from an arbitrary neighborhood surrounding any gene of interest. Thus we can move up or down the transcriptional regulatory network amongst these genes. We note again that these interactions are not truly symmetric, and so experiments have to be designed to confirm these interactions. However, the search space (to look for potential regulators) is reduced significantly.

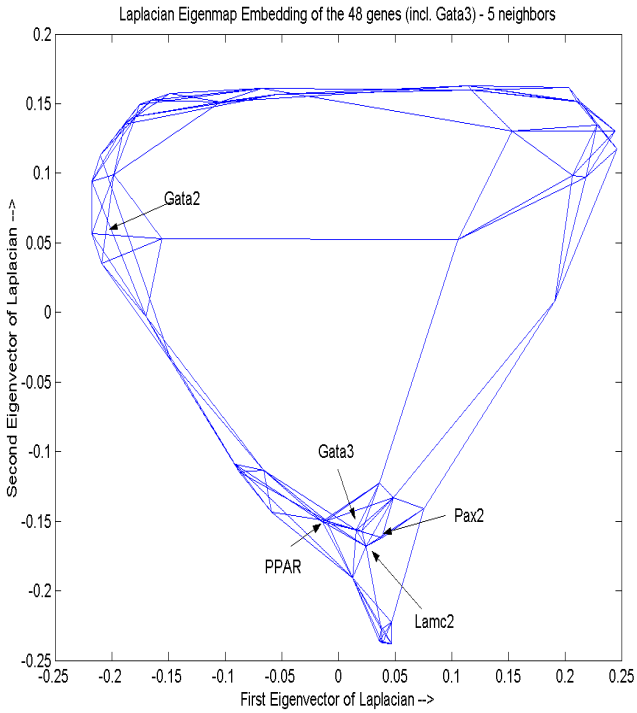

Figure 4: Laplacian Eigenmap Embedding

7. CONCLUSIONS

We have presented a methodology to understand the mechanisms underlying transcriptional regulation of a gene by combining various available data sources via a *modified Laplacian Eigenmap* technique. This framework provides a common ground both for the integration and visualization of diverse data sources for understanding physiological processes. Also, the 'single layer' graphical model formalism can be extended to include conditional interactions among various experiments relating to a particular data source.

Some of the extensions to this work that we are currently pursuing is the integration of data sources that 'build' on each other - for example, integrating two different kinds of experiments reporting the same functional interaction. One such pair is the prediction of protein-protein interactions via ChIP or through Yeast-2-Hybrid (Y2H) screen assays. In this case, the protein-protein interaction expert of Fig.3 will have two parents - one reporting the presence of interaction under the ChIP assay and the other parent for the Y2H assay (Fig. 5).

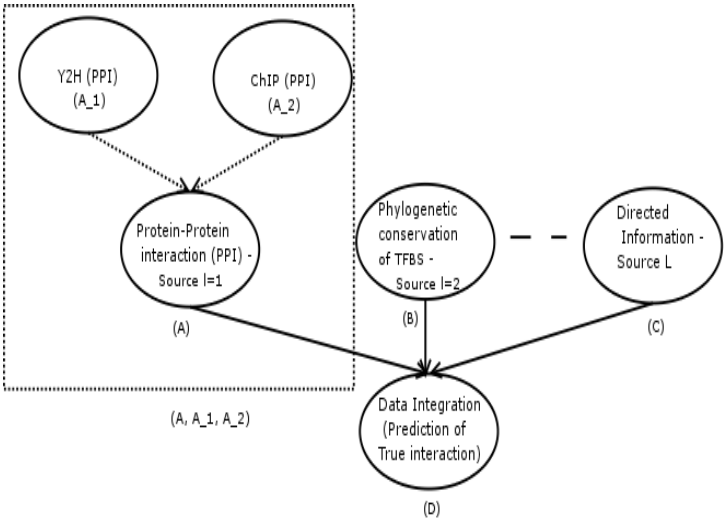

Figure 5: A schematic for diverse data integration under a more sophisticated framework , wherein we see the protein interaction derived from Y2H and ChIP experiments - this can be extended to an arbitrarily complex graphical topology.

Using a graphical model formalism here yields,

$$\begin{aligned} P(D,A_1,A_2,A,B,C) &= P(D|(A,A_1,A_2),B,C).P(A,A_1,A_2) \\ &\quad .P(B).P(C) \\ &= P(D|(A,A_1,A_2),B,C).P(A|A_1,A_2).P(A_1).P(A_2) \\ &\quad .P(B).P(C) \end{aligned}$$

Here, $A = \{Z_{i,j}^{PPI} = 1\}$, $A_1 = \{Z_{i,j}^{Y2H} = 1\}$, $A_2 = \{Z_{i,j}^{ChIP} = 1\}$, $B = \{Z_{i,j}^{Phylogenetic \quad TFBS} = 1\}$, $C = \{Z_{i,j}^{DTI} = 1\}$, $D = \{Z_{i,j}^{overall \quad data} = 1\}$. These probabilities can be obtained from a joint histogram of the corresponding data sets as suggested in Section 3.

We are also examining techniques to 'visualize' the non-symmetric Laplacians which arise due to the non-symmetric nature of the probabilities $p_{i,j}$ and $p_{j,i}$. One way would be to symmetrize this non-symmetric proximity matrix by taking the (geometric) mean of the interaction probability matrix and its transpose.

Finally, to reduce the number of candidate TFs or affectors even further , knowledge of the biophysics of transcriptional regulation needs to be incorporated. This is primarily because steric hindrance factors, kinetics of binding etc. have a very important role for the regulation of transcription.

REFERENCES

[1] Troyanskaya OG, Dolinski K, Owen AB, Altman RB, and Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in S. cerevisiae). Proc Natl Acad Sci USA 100(14): 8348-53, 2003.

[2] M. Belkin, P. Niyogi , Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, Neural Computation, June 2003; 15 (6):1373-1396.

[3] A.Rao, A.O. Hero, D.J. States, J.D. Engel, Inference of Biologically Relevant Regulatory networks using Directed Information, accepted to ICASSP 2006.

[4] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux and M. Ouimet. Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps,and Spectral Clustering, In Advances in Neural Information Processing Systems, 2004.

[5] Lee I, Date SV, Adai AT, Marcotte EM.: A Probabilistic functional network of yeast genes. Science 306(5701):1555-8 (2004).

[6] G.G. Loots and I. Ovcharenko, rVISTA 2.0: evolutionary analysis of transcription factor binding sites, Nucleic Acids Research, 32(Web Server Issue), W217-W221 (2004)

[7] I. Ovcharenko and M.A. Nobrega, Identifying synonymous regulatory elements in vertebrate genomes, Nucleic Acids Research, 33, W403-7, (2005).

[8] Grote D, Souabni A, Busslinger M, Bouchard M, Pax 2/8-regulated Gata 3 expression is necessary for morphogenesis and guidance of the nephric duct in the developing kidney, Development. 2006, Jan;133(1):53-61.

[9] Schwab K, Patterson LT, Aronow BJ, Luckas R, Liang HC, Potter SS., A catalogue of gene expression in the developing kidney, Kidney Int. 2004 Aug;66(2):867-8.

[10] Alberts, Bruce; Johnson, Alexander; Lewis, Julian; Raff, Martin; Roberts, Keith; Walter, Peter ,"Molecular Biology of the Cell", New York: Garland Publishing; 2002.