Computational Identification and Comparative Analysis of Conserved Regulatory Elements Involved in Differentiation and Maturation of Murine Thymocytes

Rebecca L. Fletcher and Narayanan B. Perumal
Indiana University-Purdue University, Indianapolis, School of Informatics, Indianapolis, IN 46202
Email: fletchrl@iupui.edu, nperumal@iupui.edu

ABSTRACT

Thymocytes are precursors of T cells that derive from the thymus, which is the primary site of T cell lymphopoiesis. Proliferation of thymocytes is significant in T cell ontology and ultimately immune response. A comprehensive microarray analysis of the differentiation of thymocytes has provided a gene expression profile of these cells (Puthier, et al 2004). The authors reported a list of 350+ genes as being up- or down- regulated during thymocyte development, which have been categorized either by cell type in which component genes are expressed or the biological process in which component genes function. To determine if there is a common set of transcriptional control elements (and hence similar transcriptional mechanisms) regulating the expression of these genes, we have devised a bioinformatics approach to identify and analyze upstream regulatory regions of subsets of these "thymocyte-specific" genes. We have isolated two sets of sequences relative to the transcription start site (TSS) of the relevant genes (-1,000 to +3000 and -3,000 to +300 with respect to TSS) showing up or down regulation, or no change during thymocyte differentiation. The motif search algorithm, MEME, was employed to derive a definite number of consensus sequences from genes that could be important in transcriptional regulation by acting as clusters of transcription factor binding sites (TFBS). Next, we used the TFBS search algorithm TFSEARCH to look for potential TFBS in the predicted MEME motifs. Although not all of these analyses may identify biologically meaningful transcriptional control elements involved in thymocyte development, we hope that they will provide a guide for the experimental biologist to focus of certain elements for biological validation. In summary, our informatics approach to detect transcriptional control elements many efficiently and effectively aid immunologists in their understanding of transcriptional regulation in T cell ontology.

METHODS

1. Identification of genes expressed during thymocyte differentiation
2. Isolation of promoter sequences
3. Analysis of upstream regulatory regions using MEME (v3.5.4) and TRANSFAC
4. Identification of potentially biologically significant elements

Figure 1. Sample MEME output for Cluster A training set showing relative conservation of each position of Motif 1 among target genes

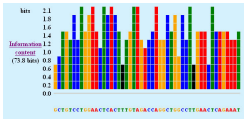


Figure 2. Developmental stages of thymocyte proliferation.

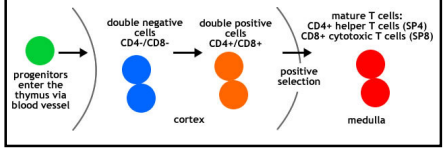


Figure 3. Work Flow

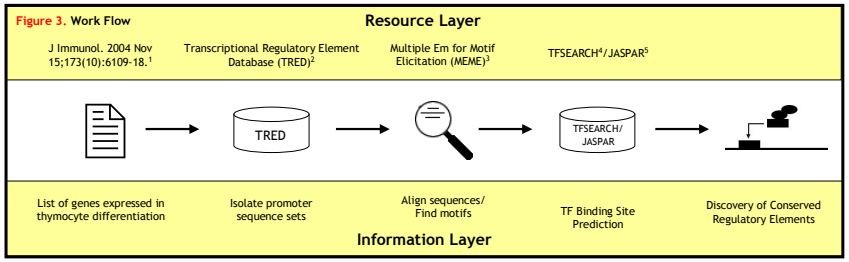
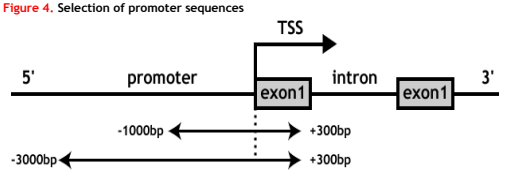
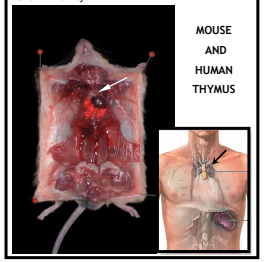


Figure 4. Selection of promoter sequences



Isolated two (2) sets of promoter sequences for each gene involved in differentiation: [-3000 to +300] and [-1000 to +300] relative to TSS

Figure 5. In the thymus gland lymphocytes become specialized. The thymus plays an important role in lymphocyte specialization and immunity.



CONCEPTS

Thymocyte
A lymphocyte that derives from the thymus and is the precursor of a T cell, which is responsible for cell-mediated immunity.

Promoter
DNA sequence that enables a gene to be transcribed

Motif
Highly conserved region that suggest similar biological function

TFBS
The DNA sequence that is recognized and bound by a transcription factor (TF); TF is a protein that controls gene expression

Figure 5. Comparison of Motif 1 with Known Motifs in JASPAR[5]

Position Specific Probability Matrix

The JASPAR output file displays the TF binding profile for Pax4.

JASPAR Score (%)
Similarity of MEME motif compared to each of the motifs in the JASPAR database of known TFBS motifs.

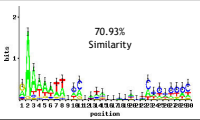


Table 1. Gene Expression Data

Six different transcriptional signatures were named by either (a) cell type in which they are expressed or (b) biological process in which they function.

A. Proliferation status (36 genes)
B. Early blockade
C. TCR rearrangement process
D. Double Positive stage
E. Single Positive stage
F. Stromal Cells

Table 2. Cluster A Corresponds to Genes Involved in Cell Cycle Regulation

Proliferation status of developing thymocytes.

Cdc451, Cdc2a, Ccna2, Cdca1, Cdc20, Cdca3, Hist2, H2afx, H2afz, Nasp, S1bp, Birc5, Bub3, Ube2c, Prim1, Top2a, Pcna, Smc211, Srebf2, Zhx1, Hmgb2, Sfrs3, Rrm2, Hnrpab, No15a, Dutp, Ddx39, Silg41, Ptprv, Hpgd, Topk, Lrpap1, Dnclc1, 2310061F22Rik, 2410017I18Rik, 2810417H13Rik.

EXPLANATION OF RESULTS

MEME *E*-value
MEME finds the most statistically significant (low *E*-value) motifs first. The E-value is an estimate of the expected number of motifs that one would find in a similarly sized set of random sequences.

MEME Information Content Diagram
Shows which positions in the motif are most highly conserved (measured in bits). Higher information content indicated greater conservation of that particular motif sequence among the target genes.

DISCUSSION & FUTURE WORK

We have demonstrated that specific functional clusters of thymocyte genes responsible for functions such as proliferation carry conserved motifs that indicate potential coordinate transcriptional regulation. In the future, analyses will be extended to include the remaining gene clusters, namely those genes involved in early blockade, TCR rearrangement, thymocyte education as well as clusters specific to the double positive and single positive cell stages. We would also like to perform statistical significance tests to validate our predictions. Finally, we hope that research immunologists will attempt biological validation of these findings.

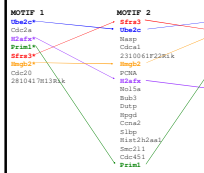
MEME RESULTS

The proliferation genes (cluster A)* contain three motifs:

	Width	E-value	Info Content	Width
MOTIF 1		8.34-061	73.8	50
MOTIF 2		2.8e-039	32.2	41
MOTIF 3		7.8e-036	46.2	36

*Total dataset size (residues) = **36400**

Genes from cluster A that carry each motif:

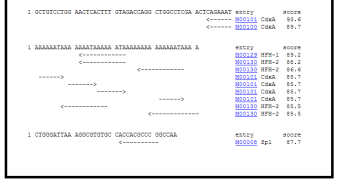


*Denotes genes that carry all three motifs.

Interestingly, MOTIF 2 was present in several genes (4 out of 5) having been identified via bibliographic searches as functionally significant in chromatin structure (i.e., Hist2, H2afx, Nasp and Slbp). Likewise, half of the genes from cluster A that have been classified as cell-cycle regulators carry the same MOTIF 2, as well as MOTIF 3.

TFSEARCH Results

TFMATRIX (TFBS) entries with High Scoring Sequences:



REFERENCES

[1] A General Survey of Thymocyte Differentiation by Transcriptional Analysis of Knockout Mouse Models. Denis Puthier et al . (2004) J Immunol, 15;173(10):6109-18.
[2] TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. Zhao F, Xuan Z, Liu L, Zhang MQ. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D103-7.
[3] Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Timothy L. Bailey and Charles Elkan, Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.
[4] TFSEARCH: Searching Transcription Factor Binding Sites. Yutaka Akiyama, http://www.rwcp.or.jp/papia/ .
[5] A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D95-7. JASPAR