Trying on Phylogenetic Footprinting Methods

Rich Sherwood, Department of Biological Sciences, Stanford University

Introduction

Development in multicellular organisms occurs as cells acquire specialized roles. This specialization is accompanied by the differential expression of proteins vital to each task. Arguably the most widely used mechanism for restriction of a cell's protein complement is transcriptional regulation, the use of transcription enhancing or repressing proteins that bind to DNA regions adjacent to genes and that regulate the expression of these genes. Transcription factors, as these regulators are called, bind to short (5-15 bp), highly specific DNA sequences and can regulate large networks of functionally related genes (Bulyk 2003).

Due to the importance of transcriptional regulation, one of the main goals in the post-genomic era is to predict how a gene's expression is regulated based on the presence of transcription factor binding sites (TFBS) in the adjacent genomic regions. Genome-wide knowledge of TFBS could be used to build models of transcriptional regulatory networks that operate in cell fate specification during development (Qiu 2003). Recent advances in genome sequence availability and in high-throughput gene expression analysis technologies have allowed for computational methods of TFBS detection. In vitro assays of transcription factor binding to nucleotide strings have led to the development of databases of position weight matrices (PWMs) such as TRANSFAC that predict the sequences to which transcription factors can bind (Matys *et al* 2003). These PWMs can be used to search genome sequences and even promoter regions determined

from computational promoter-finding programs to identify putative TFBS in the genome. However, only a small fraction of the TFBS predicted by PWMs are functionally relevant, so computational TFBS detection must be combined with some method of enriching for functionally significant TFBS (Qiu 2003).

One method of predicting functionally relevant TFBS involves the use of microarray technology to find genes that are co-expressed or that respond similarly to a stimulus and then to search the promoter/enhancer regions of these genes for conserved motifs that could be TFBS. While this technique has proven effective in simple organisms (e.g. Sinha and Tompa 2003), the identification of regulatory regions of co-expressed genes is a difficult endeavor in higher organisms. To combat this difficulty, cross-species genome comparison, or phylogenetic footprinting, can be used.

Phylogenetic footprinting is based on the premise that TFBS will be highly conserved in comparison to non-regulatory regions in the regions adjacent to genes (Hardison 2000). Furthermore, evidence suggests that transcriptional regulatory regions often occur in modules, so TFBS adjacent to genes will be clustered into regulatory modules that can be distinguished from non-regulatory areas by their high base conservation (Loots *et al* 2000). Phylogenetic footprinting has been used successfully in several test scenarios (Krivan and Wasserman 2001, Oeltjen *et al* 1997, Hardison 2000, Loots *et al* 2000, Wasserman *et al* 2000, Qiu *et al* 2003), and several programs exist that employ phylogenetic footprinting in TFBS detection (rVISTA: Loots *et al* 2002, Footprinter: Blanchette and Tompa 2002, CONSITE: Lenhard *et al* 2003).

I tested how the available phylogenetic footprinting programs, CONSITE, rVISTA and Footprinter, as well as a modified application of motif-finding Bioprospector

(Liu *et al* 2001) would fare in determining the TFBS that allow for the co-expression of a set of five mouse and human genes known to be co-expressed in some but not all tissues. Thus, it would be expected that these genes would share binding sites for some regulatory transcription factors.

Methods and Results

Data Set

I chose five proteins known from multiple experiments to be expressed specifically in the mouse anterior visceral endoderm (AVE) and node—Otx2, Lhx1, FoxA2, Gsc, and Hex (Harland and Gerhart 1997). Some of these proteins such as FoxA2 and Otx2 also are expressed in distinct regions in the developing mouse embryo. I obtained the region 1000 bp upstream of the translation initiation site of each gene as well as of beta-actin and GAPD, two genes expressed ubiquitously, in mouse, rat, and human using the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgGateway). The sequences can be found in the supplemental document. I used four phylogenetic footprinting programs: CONSITE, rVISTA, Footprinter, and Bioprospector. With each program, I attempted to find TFBS preferentially expressed in the five AVE/node sequences. I will present the methods used and results of each program.

CONSITE

I used CONSITE (http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite) to analyze orthologous human:mouse sequences. I searched all vertebrate transcription factor matrices and used a conservation cutoff 10% below the recommended and a TF score

cutoff of 70%. The full results for each orthologous pair are contained in the supplemental document. I then compared results of all seven orthologous pairs to find TFBS expressed in at least four of the AVE/node sequences. There were 10 TFBS that fit this criteria (Table 1). By comparing these results with the results of the ubiquitously expressed sequences, I found that 6/10 of the TFBS were also expressed in at least one of the ubiquitous sequences and thus 4/10 of the TFBS were unique to the AVE/node (Table 1). These were E2F, TEF-1, HFH-2, and HNF-3beta.

rVISTA

I used rVISTA (http://rvista.dcode.org/), also analyzing orthologous human:mouse sequences. I searched all vertebrate transcription factor matrices using default specifications and considered conserved TFBS. Only Otx2, Gsc, Hex and GAPD were sufficiently homologous to qualify for rVISTA, so I compared the results of these sequences. Full results can be found in the supplemental document. Eight TFBS were found in more than one of the AVE/node sequences, and none of these was detected in GAPD (Table 2). Only one TFBS, E2F, was detected in all three sequences.

Footprinter

I used Footprinter (http://abstract.cs.washington.edu/~blanchem/FootPrinterWeb/FootPrinterInput2.pl), analyzing orthologous human:mouse:rat sequences using default specifications. Gsc did not return any results, so it was omitted from analysis. As the results of Footprinter are merely conserved 10-base motifs, they are harder to analyze comparatively. Instead, I

analyzed conserved motifs from two AVE/node sequences with the fewest conserved motifs (Lhx1 with 5 motifs and Otx2 aith 14 motifs) by determining whether they corresponded to known TFBS using Patch (http://www.gene-regulation.com/cgi-bin/pub/programs/patch/bin/patch.cgi?) and then determining whether the other AVE/node and ubiquitous sequence contained identical sequences. The full results of Footprinter are in the supplementary document, and a comparative analysis of Lhx1 and Otx2 motifs is presented (Table 3). This analysis yielded potential binding sites for HNF-3beta, ETS-1, LEF-1 and Sp1 expressed preferentially in two AVE/node sequences and CTCF and PEA3 motifs expressed in one AVE/node sequence.

Bioprospector

I used Bioprospector (http://bioprospector.stanford.edu/), analyzing AVE/node sequences from only human, only mouse, human and mouse, and human, mouse and rat, as well as ubiquitous sequences from human and mouse. I used default specifications and then analyzed the degenerate motif of the top 5 sequences using Patch. Full results are in the supplementary document, and the motifs from the searches that contained potential TFBS are summarized (Table 4). The results were different using only human or only mouse as compared to using combined species. Of TFBS in the AVE/node sequences but not in the ubiquitous sequences, combined species tests found GATA-1, Gbx2, Crx, and c-Myb and one species tests found HNF-3alpha, HNF-3beta, HNF-1, and Gbx2.

Combined Analysis

Comparing the results of the four phylogenetic footprinting approaches, an analysis was constructed of potential TFBS in the AVE/node sequences but not in ubiquitously expressed sequences. Table 5 consists of a summary of the TFBS indicated by at least one method to be regulators of AVE/node gene expression. These results indicate that HNF-3beta, an alternate name for FoxA2, was found by 3 programs, E2F, Gbx2 and GATA-1 were found by 2 programs, and 16 TFBS were unique to a single method.

Discussion

I performed analysis of a set of co-expressed genes using four phylogenetic footprinting techniques. The genes I chose were human, mouse and rat genes known to be co-expressed in the AVE/node during development. I chose these genes for several reasons: 1) I wanted to compare the techniques' performances using eukaryotic sequences; 2) These genes have a very specific expression pattern in the AVE/node, are functionally important in the establishment and maintenance of this region, and are well-documented; 3) Much is known about the factors and signals involved in AVE/nose formation, yet the TFBS are unknown. However, there were several drawbacks to choosing these genes: to allow for a manageable analysis, I limited analysis to a 1000 bp region directly upstream of the translation initiation site, which may not contain the full regulatory regions of these genes; the genes also contain distinct areas of expression so likely contain distinct TFBS as well; and the data for mammalian TFBS is less expansive than for simpler organisms.

The methods examined gave different results. As shown in table 5, only one TFBS was identified in ¾ of the methods, and 3 TFBS were identified in 2/4 of the methods. As to consistency with the literature, the node is known to be transcriptionally regulated by transcription factors downstream of the Wnt, FGF and Nodal pathways and by the 5 members of the group (Harland and Gerhart 1997). TFBS for these regulators were among those identified: LEF-1 is downstream of Wnt (Wodarz and Nusse 1998), ETS-1 and PEA3 are downstream of FGF (Roehl and Nusslein-Volhard 2001), and of those related to proteins on the list, FoxA2 (HNF-3beta) is on the list along with related proteins HFH-2 and HNF-1 (Lehmann *et al* 2003), and Gbx2 and Crx are closely related to Otx2 and Lhx1 (Joyner *et al* 2000, Hodges *et al* 2002). So, the list created is faithful to literature references of AVE/node regulators with the exception of an absence of Nodal downstream transcription factors.

As to a comparison of the phylogenetic footprinting methods, it is important to understand the rationale for phylogenetic footprinting: cross-species comparison should highlight regions that are conserved because they contain regulatory modules. Identifying TFBS only in these conserved regions should greatly increase the functional relevance of TFBS motifs found. Keeping this rationale in mind, Bioprospector is ill-suited for phylogenetic footprinting. Bioprospector works by finding the 8 bp motifs most highly expressed in the sequences entered. So, there is no benefit to applying Bioprospector to multiple species versus applying it to one species. Its method of enriching for functional motifs focuses on the idea that functional motifs will be overrepresented in the data set, which is independent of species used. In fact, using sequences of different species should

lower the efficacy, as the transcription factors from each species may have incurred mutations that slightly change the optimal binding site for that transcription factor.

The three remaining algorithms, CONSITE, rVISTA and Footprinter, all use a method of TFBS recognition, sequence alignment, and conservation threshold determination to filter functionally irrelevant TFBS. Among these programs, Footprinter theoretically maximizes the advantage created by phylogenetic footprinting, as it allows for multispecies comparison as opposed to dual sequence comparison. Footprinter takes into account the phylogenetic by assigning more weight to sequences that are evolutionarily less related but maintain similarity—thus, Footprinter takes into account phylogeny and does not unduly weight the sequences from highly related species. Footprinter did find biologically relevant TFBS, so these sites must be conserved among rat, mouse and human.

CONSITE and rVISTA are virtually identical with regards to algorithm. Both align the orthologous sequences using BLAST-like algorithms, set a threshold of conservation based on the average conservation of the sequences used, and find TFBS based on PWM. Between them, CONSITE generated potential TFBS that were highly enriched for biologically relevant factors, likely due to an improved TFBS database. CONSITE's interface is also more user-friendly, and its use of non-repetitive TFBS allows for easy identification of the binding factor, whereas rVISTA sometimes does not give the transcription factor that binds to the site.

In the future, algorithms for phylogenetic footprinting could be improved by several factors. First, better biological knowledge of how TFBS occur would be of great import. Are they clustered, do they occur in the regions of highest conservation only, is

there a way to predict eukaryotic promoter/enhancer regions, are there regions of DNA outside of the exact DNA-binding site that affect PWM of transcription factors? These facts could be integrated into an algorithm to increase concentration of functional TFBS. Second, the algorithms could be improved by genomic comparisons. If genomes were more globally aligned, then the regions of highest conservation could be more easily determined, as the alignments derived by the programs currently are likely erroneous. Thus, the sequencing of more genomes, which could allow for the tracing of the mutations and transformations of genomic segments, would allow for more accurate alignments and would vastly increase the benefit of phylogenetic footprinting as a tool for TFBS detection.

Table 1: CONSITE Results

Number of predicted TFBS for each factor in the upstream region of each gene. Factors in italics are preferentially expressed in AVE/node.

Factor	Lhx1	FoxA2	Hex	Otx2	Gsc	ActB	GAPD
AML-1	0	1	5	6	2	1	4
Bsap	0	1	4	2	2	1	3
E2F	2	0	2	4	1	0	0
HFH-2	2	0	1	19	11	0	0
Hnf-3beta	1	0	1	2	6	0	0
Myf	0	2	1	10	6	2	3
NRF-2	0	2	1	2	2	0	0
Spz1	2	3	0	5	7	1	2
TEF-1	0	1	2	2	2	0	0
Thing1-E47	3	1	0	8	13	1	4

Table 2: rVISTA Results

Binding sites recognized by these factors were identified in more than one sequence. Sequences containing conserved motifs for this factor are denoted +.

Factor	Otx2	Gsc	Hex	GAPD
AREB6	+	+	-	-
Sp1	+	+	-	-
S8	+	+	-	-
USF	+	-	+	-
XVENT	+	+	-	-
LDSPOLYA	+	-	+	-
CAAT	+	-	+	-
E2F	+	+	+	-

Table 3: Footprinter Results

The following motifs were found to be highly conserved in sequences denoted by +. Predicted TFBS for the motifs are listed.

Motif	Predicted TFBS	Lhx1	FoxA2	Otx2	Hex	ActB	GAPD
CCGCA	ETS-1, LEF-1	+	-	-	+	-	-
GCTGC	CTCF	+	-	-	-	-	-
AGGAAA	PEA3	+	-	-	-	-	-
CCCCC	Sp1	-	+	+	-	-	-
TGTTT	HNF-3beta	-	+	+	-	-	-

Table 4: Bioprospector Results

The following motifs were found to be the most highly expressed in the sequence set noted. TFBS that bind to the motifs are listed.

Set tested and rank	8bp motif	Potential TFBS
Human only #1	ATTTWTTT	HNF-3beta, HNF-3alpha
Human only #2	WTAAATAW	HNF-3beta
Human only #3	GCGGCSCG	Sp1
Human only #5	TGTYAATC	HNF-1, Gbx2
Mouse only #1	MTTWATAC	GATA-1, Gbx2, TBP
Mouse only #2	ATWAATGW	Gbx2
Human and mouse #1 and #4	YGATKGAC	GATA-1
Human and mouse #3	GCTMATCA	GATA-1
Human, mouse and rat #1 and #2	TKGATTKA	GATA-1, Gbx2
Human, mouse and rat #4	TMAATTCA	Gbx2, Crx
Human, mouse and rat #5	GATTGAMR	c-Myb
ActB/GAPD #2	RCGTGCRC	ER-alpha, ER-beta, Sp1
ActB/GAPD #3	TGTGCACS	Sp1
ActB/GAPD #4	TTTTTYTT	TBP
ActB/GAPD #5	YGCACGYA	HIF-1

Table 5: Data Summary

TFBS	CONSITE	rVISTA	Footprinter	Bioprospector Multispecies	Bioprospector single species	Total # methods containing this TFBS
HNF-3beta	+	-	+	-	+	3
E2F	+	+	-	-	-	2
Gbx2	-	-	-	+	+	2
GATA-1	-	-	-	+	+	2
HFH-2	+	-	-	-	-	1
TEF-1	+	-	-	-	-	1
AREB6	-	+	-	-	-	1
Sp1	-	+	-	-	-	1
S8	-	+	-	-	-	1
USF	-	+	-	-	-	1
XVENT	-	+	-	-	-	1
LDSPOLYA	-	+	-	-	-	1
CAAT	-	+	-	-	-	1
ETS-1	-	-	+	-	-	1
CTCF	-	-	+	-	-	1
PEA3	-	-	+	-	-	1
LEF-1	-	-	+	-	-	1
HNF-1	-	-	-	-	+	1
Crx	-	-	-	+	-	1
c-Myb	-	-	-	+	-	1

References

Blanchette M and Tompa M. 2002. *Genome Res.* **12**, 739.

Bulyk ML. 2003. *Genome Biology* **5**, 201.

Hardison R. 2000. *Trends in Genet.* **16**, 369.

Harland R and Gerhart J. 1997. *Annu. Rev. Cell Dev. Biol.* **13**, 611.

Hodges MD, Vieira H, Gregory-Evans K, and Gregory-Evans CY. 2002. *Genomics* **80**, 531.

Joyner AL, Liu A and Millet S. 2000. *Curr. Opin. Cell Biol*. **12**, 736.

Krivan W and Wasserman WW. 2001. *Genome Res.* **11**, 1559.

Lehmann OJ, Sowden JC, Carlsson P, Jordan T and Bhattacharya SS. 2003. *Trends Genet.* **19**, 339.

Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, and Wasserman WW. 2003. *J. Biol.* **2**, 13.

Liu X, Brutlag DL, Liu JS. 2001. *Pac Symp Biocomput*. ,127-38.

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, and Frazer KA. 2000. *Science* **288**, 136.

Loots GG, Ovcharenko I, Pachter L, Dubchak I, and Rubin EM. 2002. *Genome Res.* **12**, 832.

Matys V, Fricke E, Geffers E, Gossling M, Haubrock M, Hehl K, Hornischer D, Karas D, Kel AE, Kel-Margoulis OV *et al*. 2003. *Nucleic Acids Res.* **31**, 374.

Oeltjen JC, Malley TM, Muzny DM, Miller W, Gibbs RA, and Belmont JW. 1997. *Genome Res*. **7**, 315.

Qiu P. 2003. Biochem. *Biophys. Res. Commun.* **309**, 495.

Qiu P, Qin L, Sorrentino RP, Greene J, Partridge NC, and Wang L. 2003. *J. Mol. Biol.* **278**, 167.

Roehl H and Nusslein-Volhard C. 2001. *Curr. Biol.* **11**, 503.

Sinha S and Tompa M. 2003. *Nucleic Acids Res.* **31**, 3586.

Wasserman WW, Palumbo M, Thompson W, Fickett JW, and Lawrence CE. 2000. *Nat.*

Genet. **26**, 225.

Wodarz A and Nusse R. 1998. *Annu. Rev. Cell. Dev. Biol.* **14**, 59.