

Affinity Density: a Novel Genomic Approach to Identification of Transcription Factor Regulatory Targets

Dennis J. Hazelett^{1,*}, Daniel L. Lakeland² and Joseph B. Weiss^{3*}

¹Integrative Biosciences, 611 SW Campus Dr SOD705, Portland, OR 97239

²Independent Consultant, Los Angeles, CA

³Department of Cardiology, 3303 SW Bond Ave CHH14W, Portland, OR 97239

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Methods:

A new method was developed for identifying novel transcription factor regulatory targets based on calculating Local Affinity Density. Techniques from the signal-processing field were used, in particular the Hann digital filter, to calculate the relative binding affinity of different regions based on previously published *in vitro* binding data. To illustrate this approach, the genomes of *Drosophila melanogaster* and *D. pseudoobscura* were analyzed for binding sites of the homeodomain protein Tinman, an essential heart development gene in both *Drosophila* and Mouse. The significant binding regions were identified relative to genomic background and assigned to putative target genes. Valid candidates common to both species of *Drosophila* were selected as a test of conservation.

Results:

The new method was more sensitive than cluster searches for conserved binding motifs with respect to positive identification of known Tinman targets. Our Local Affinity Density method also identified a significantly greater proportion of Tinman-coexpressed genes than equivalent, optimized cluster searching. In addition, this new method predicted a significantly greater than expected number of genes with previously published RNAi phenotypes in the heart.

Availability:

Algorithms were implemented in python, LISP, R and maxima, using MySQL to access locally mirrored sequence data from Ensembl (*D. melanogaster* release 4.3) and flybase (*D. pseudoobscura*). All code is licensed under GPL and freely available at www.ohsu.edu/cellbio/dev_biol_prog/affinitydensity.html.

Contact: hazelett@ohsu.edu

1 INTRODUCTION

In coming years much effort will be expended to understand the information encoded in the non-protein coding regions of DNA. The physical recruitment of cellular factors to determine what genes are transcribed into RNA is one of the most important functions of these regions. These cellular factors include the class of proteins called transcription factors. Most transcription factors bind to short, degenerate oligonucleotide sequences of 4 to 15 base pairs in length and activate transcription directly or in conjunction with larger

protein complexes. To date, efforts to understand how genomic DNA guides this process have focused on *de novo motif discovery* and *known motif mapping* (Ji and Wong, 2006). Our concern is with the latter, in which previously characterized binding-site motifs are used to predict the genomic targets of a transcription factor. Successful early attempts (Berman *et al.*, 2002; Stathopoulos *et al.*, 2002) at experimentally verified regulatory target predictions have given rise to a modular view of transcriptional regulation. Under this paradigm, islands of regulatory sequences contain clusters of conserved binding sites for two or more transcription factors required for a given process. These sequences, known as *cis*-regulatory modules (CRMs), are thought to direct the timely expression of downstream target genes as part of a regulatory code. This conceptual advance paved the way for searches for novel transcription factor targets of Ftz-F1 (Bowler *et al.*, 2006) and Tinman (Halfon *et al.*, 2002).

In spite of these well-documented attempts at reading regulatory DNA, an in-depth analysis of the distribution of genomic binding sites and its implications is lacking. The rules governing CRM architecture have not been discoverable by current pattern recognition approaches. As a proxy, functionality is typically inferred from the direct conservation of sequence motifs in cross-species alignments. This approach assumes that unconserved sites play little or no role in transcriptional regulation. We decided to take an unbiased look at the genomic distribution of binding sites. Low affinity sites may serve, for example, to increase the local concentration of factors so that they are more available for recruitment by binding partners, or to increase transcription initiation rates by mass action when conditions allow. In order to achieve this, we abandoned the typical search for clusters in favor of a density histogram representing the likely occupancy of transcription factors along the sequence, with all else being equal.

One of the limitations of the most common approach (searching for binding site clusters), is that varying parameters—window size, number of sites, cutoff positional weight matrix score—often results in widely divergent predictions. The quality of these predictions forms the basis for optimization. The cycle of analysis, evaluation and resetting of parameters leads to an arbitrary fitting of parameters to match prior expectation. The process also produces multiple valid prediction sets, whose individual meanings can be difficult to interpret. For example, a smaller window results in higher sensitivity to dense clusters of sites. If a larger cluster window is

*to whom correspondence should be addressed

chosen, specificity decreases but the search is more sensitive to targets whose binding sites are distributed more sparsely. Varying parameters can be biologically revealing, as demonstrated by the example of the dorso-ventral patterning transcription factor Dorsal, where lower numbers of binding sites were found to be correlated with targets of Dorsal repression in the lateral domains of the embryonic blastoderm (Stathopoulos *et al.*, 2002; Stathopoulos and Levine, 2004). However when only one or two binding sites are available to map, competing optimizations are an impediment to discovery.

In light of the difficulty of separating signal from noise using cluster search methods, a new approach is needed that takes into account both the spacing and regional distribution of sites. We developed a method that addresses these requirements using digital-signal processing techniques. We describe here the application of the Hann filter to this problem using the previously characterized binding site for Tinman to illustrate our approach.

2 APPROACH

We designed a search algorithm that assigns a statistic to each region of the genome and is a continuous function of both binding site density and location. Digital signal processing techniques are best suited to address these requirements (Hamming, 1998). The problem of assigning an accurate measure of binding affinity to each locus of the chromosome can be thought of as a special case of downsampling a digital signal, and digital signal processing techniques for downsampling are well established. This new approach allows us to search for genes in regions containing either dense clusters of sites or a high background of sites, or both, simultaneously. In addition, we used binding affinity data to assign scoring weights for binding sites. Finally, to increase specificity, we compared significant predictions from two species of *Drosophila*, effectively treating binding affinity as a conserved property of chromatin.

3 METHODS

3.1 Scoring

We scored each genomic locus with a value proportional to the measured dissociation constant for each known binding site. We chose this method over a positional weight matrix to characterize the Tinman binding site in order to leverage the wealth of biochemical data available for Tinman and other NK homeodomain proteins. We normalized the sequence score by dividing by the strongest Tinman-monomer/binding-site interaction. Watada *et al.*, 2000 reported a 5-fold higher affinity of the NK homeodomain protein Nkx2.2, a mammalian ortholog of Tinman, for the sequence ‘TCAAGTG’ than for an alternative binding site, ‘TTAAGTG’. Thus, we assigned a relative affinity of 1.0 to the sequence ‘TCAAGTG’. We therefore assigned a relative affinity of 0.2 to the sequence ‘TTAAGTG’.

In order to account for cooperativity that has been reported for Tinman homodimers (Kasahara *et al.*, 2001; Zaffran and Frasch, 2005), we incorporated data from electrophoretic mobility shift assays. Zaffran and Frasch (2005) reported an 8-fold difference in binding affinity between the monomer and dimer binding sites (Kd 430 nM vs 52 nM, respectively). A dimer binding site consists of two binding sites in opposite orientation. Dimer strength varied slightly with the number of intervening nucleotides between the constituent monomers (Zaffran and Frasch, 2005), with the strongest dissociation constant measured at 6 bp (52 nM). We interpolated

the intermediate values from 0 to 15 bp between binding sites and gave a maximum weight of 8.27 to dimers 6 base pairs apart. The dimer function was truncated to 2.0 for spacers sized at < 3 b.p. or > 12 b.p., equivalent to no cooperativity.

3.2 Signal Processing

We calculated the intensity of binding at each nucleotide of the genome based on a combination of binding affinity data and pattern matching of nucleotide sequences. First, the genome was scanned by a scoring function, which assigned a score $\phi(S_i)$ to each base of the chromosome. In our study, this score was based on binding affinity information for fixed sequences starting at the given base pair. Then, these data were reduced to a regional density over a region of size $2N$ by filtering the sequence of scores with a convolution filter, a weighted average of the scores at each base pair. Each $2N$ base pair window overlapped the adjacent windows on either side by N bases. The particular weighting function (or kernel) that we chose is known as the Hann window, a raised cosine curve (Hamming, 1998; see equation 1). The advantage of this type of convolution kernel is that it is a function not only of the number of binding sites within the window, but also how close together the sites are within the window and where they occur. The resulting statistic makes a reasonable tradeoff between measuring the density of sites, and measuring the location where that density occurs. Thus, the affinity density ρ_n , for the n^{th} genomic segment of length N , is given by

$$\rho_n = \frac{1}{K} \sum_{i=0}^{2N} \left(\frac{\cos\left(\frac{(i-N)\pi}{N}\right) + 1}{2} \times \phi(S_i) \right) \quad (1)$$

N is a resolution factor in number of bases, meaning that the information in chromosomes is reduced by a factor of $N : 1$. The normalization constant K is simply the sum of the weights, which ensures that the statistic does not scale with different window sizes (N):

$$K = \sum_{i=0}^{2N} \frac{\cos\left(\frac{(i-N)\pi}{N}\right) + 1}{2} \quad (2)$$

For comparison and interpretability purposes the output was multiplied by 1000 to get a value per kb.

Unless otherwise indicated, $N = 2^{12}$ for this study, dividing the average *Drosophila* gene (~ 10 kb) into two or three overlapping regions. Although we used a particular scoring function defined by a set of fixed patterns and their associated binding affinities, ϕ represents any generic scoring function that assigns a value to the nucleotide sequence S_i beginning at position i of the n th segment for which density ρ is being calculated. Any suitable function, such as a positional weight matrix could be substituted for ϕ .

3.3 Significance Predictions

In order to filter out the expected background noise in the binding site density data, we multiplied each density by a sigmoidal function of the density (see equation 3). Since the weighted density is a continuous function of position and number of sites, we used this continuous version of a threshold rather than a hard cutoff. We set the 50% point of the sigmoid at the expected density for background, and the transition rate to make the exponent equal to 5 when one additional base over the expected value was detected in the region. The result is that high-scoring regions are reduced by a negligible amount, whereas regions of lower than average density are reduced to nearly zero all in a continuous manner.

$$\gamma = \frac{1}{1 + e^{(-5\left(\frac{\rho_n - \bar{\rho}}{\bar{\rho}_1 - \bar{\rho}}\right))}} \quad (3)$$

In equation 3 the symbol $\bar{\rho}_1$ is the expected density when one additional site above the background noise is found in the region. This sigmoid was changed for each genome and binding site motif by calculating appropriate expected density, and transition rate. A plot of the raw data thus obtained

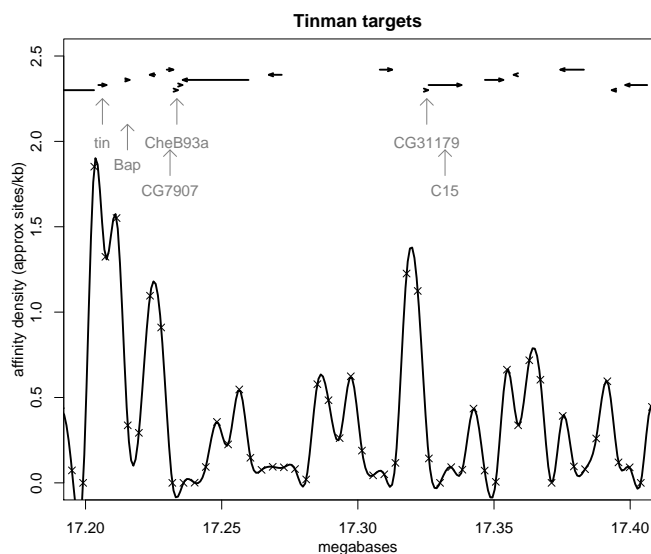


Fig. 1. A region of interest demonstrates the physical location of several predicted Tinman targets and other genes (arrows) relative to regions of high Tinman binding affinity. A spline was fit to the individual affinity data ('x'). Some targets shown not conserved between *D. pseudoobscura* and *D. melanogaster* are therefore not among our predictions (Table 2).

reveals the variation in binding affinity at the level of the whole genome (not shown). At this resolution, peaks are clearly visible overlapping many known Tinman targets and other mesodermally expressed genes, as might be expected. A detailed view of a region of interest containing *tin* and *Bap*, both Tinman targets, reveals clear peaks in the vicinity of these genes as well as other targets from our prediction set, as expected (Figure 1).

The next step is to assign a figure-of-merit to each gene by combining the binding site affinity densities in the regions around the location of the gene in the genome. Our figure of merit was the sum of the sigmoidally filtered binding site affinity density scores for convolution windows that overlap the flybase annotated gene or the region 1kb to either side of the annotation. We explored various buffer sizes for the regions around the gene, and settled on 1 kilobase-pairs. We use this equally weighted centered window around the gene to reflect the lack of prior information about where the regulatory regions are located relative to genes. If better information were available about the relative locations of regulatory regions across a sample of genes, an unequally weighted sum that reflected these relative probabilities could be used. However, since certain promoter sequences have been found far downstream or upstream of their target genes, we were unwilling to use a more narrow, or informed prior distribution at this time.

3.4 Conserved Prediction Sets

Most published accounts of genomic binding site searches to date have leveraged sequence conservation with great success. Therefore, we chose to incorporate a test of conservation between two *Drosophilid* species as a further means of filtering our predictions.

We accomplished this by comparing two independently derived lists of target predictions from parallel searches in two related genomes, *D. melanogaster* and *D. pseudoobscura*. We examined various scoring mechanisms that combine two independently derived scores such as adding the scores, multiplying the scores, and using the first principle component of the two scores. We settled on making a list of genes for each organism with a cutoff score of > 0.7 and taking the intersection of the two lists. This procedure captured the bulk of known Tinman targets (Fig. 2). Our affinity-density model assumes that the number of sites within a region

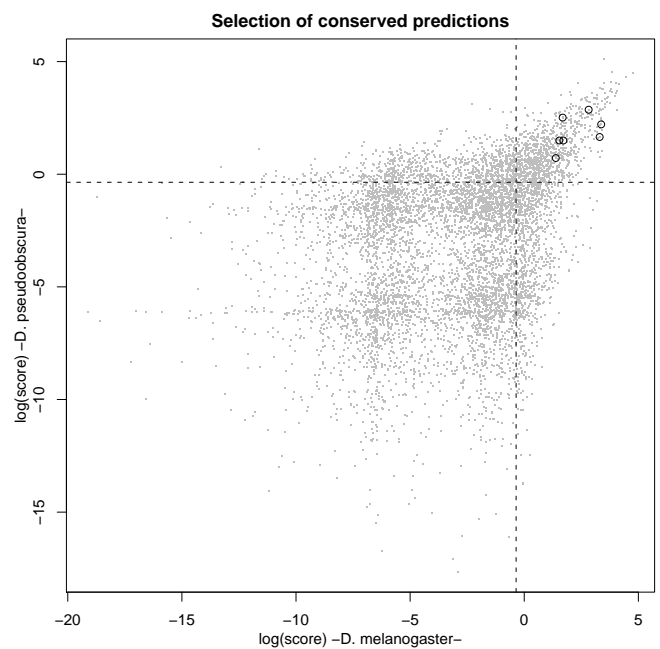


Fig. 2. Target predictions were selected on the basis of conservation in two species of *Drosophila*. Each point (grey) represents the log of scores of a single gene. Dotted lines reveal the cutoff values used for selection. The points corresponding to known targets of Tinman from the literature are highlighted with circles.

and their placement within the region are important for recruitment of transcription factors, and therefore even as evolutionary distance increases the affinity-density should remain similar. Since affinity density is as much a structural property of the chromosome as it is a function of the primary sequence of the DNA, we selected common predictions made on this basis instead of nucleotide-for-nucleotide sequence conservation (see Disc.). Predictions were ranked by the average of the figure-of-merit statistic from *D. melanogaster* and *D. pseudoobscura*.

4 RESULTS

4.1 Affinity density measurement increases sensitivity to known Tinman targets

We found that affinity density measured by application of the Hann filter to binding site data provided greater sensitivity than searches for clusters of conserved motifs such as those produced in Target Explorer (Sosinsky *et al.*, 2003) or flyenhancer (Markstein *et al.*, 2002) as a predictor of known Tinman regulatory targets (see Figure 1). One method of measuring the sensitivity of target prediction algorithms is to compare expression patterns between the predicted target list and the transcription factor whose binding-site motif was used to derive it.

The Berkeley *Drosophila* Genome Project (<http://www.fruitfly.org>) maintains a database of *in situ* patterns as gene names annotated with a controlled vocabulary of expression terms, plus the images used to assign terms. If our prediction list were valid, we would expect to find associated with these genes the same expression terms associated with *tinman* to a greater degree than would be expected for any random selection of genes of equivalent size from the database. Therefore we compared the

Table 1. Coexpression of Prediction Sets with Tinman

Algorithm	expression data	expression overlap (%)	p-value
Affinity Density	68	30(44%)	0.0205
Target Explorer	55	17(31%)	0.1043

Predicted Tinman targets with equivalent coexpression terms in the BDGP *in situ* expression database.

expression terms associated with *tinman* and our prediction set. We observed a greater fraction of genes with *tinman*-associated expression terms than would be expected in an equivalent-sized random selection of genes from the fly genome (Table 1). We obtained statistically significant enrichment of *tinman* expression overlap using the binomial distribution ($p = 0.0205$, Table 1).

Furthermore, we compared our binding affinity-derived target predictions with the results of our optimized conserved-motif cluster search using Target Explorer. We optimized our cluster search by varying window size and score-cutoff parameters until we obtained a list with the greatest number of expected Tinman targets and other genes whose expression overlapped with *tinman*. We found greater enrichment in *tinman* in the affinity density-derived prediction list than in the conserved-motif cluster list (Table 1).

Next we determined the sensitivity to known Tinman targets. We identified in the literature all published accounts where the authors claim to have evidence for direct binding of Tinman to the promoter or altered transcript or protein levels of the target gene (under conditions of perturbed or misexpressed Tinman). Next we compared how well these genes were predicted by different algorithms.

To compare approaches, we compared the relative ranked scores assigned to the predictions by each. For affinity density, the score is a function of the predicted local affinity for the transcription factor around the target gene. For conserved motif-cluster searches, the score reflects the cumulative scores of the positional-weight matrix within a fixed window relative to the gene of interest. Thus, the score also reflects the relative strength of the prediction from each algorithm, because it is an expression of the likelihood of finding transcription factor at that location.

We therefore ranked the target lists derived by different methods by decreasing score. We evaluated the ranked lists according to two criteria: sensitivity to detect targets that were previously identified by conventional means, and relative strength of predictions as determined by rank. Out of a total of twelve published *tinman* target genes, affinity-density identified 7, including *zfh1*, *biniou* and *jelly belly*. The conserved-motif searches produced only 4 (Table 2), only one of which (*eve*) was not predicted as a target by the affinity algorithm. Together these observations suggest that the local affinity density measurement is more sensitive. In addition, five of the affinity density predictions were ranked higher than 100, compared with only two of the conserved motif cluster predictions that were ranked higher than 200, suggesting higher specificity in the affinity density algorithm.

Table 2. Comparison of Motif-Mapping Methods

Tinman targets	annotation	AD	TE
<i>tinman</i>	CG7895	2	11
<i>bagpipe</i>	CG7902	4	291
<i>zfh1</i>	CG1322	23	
<i>pannier</i>	CG3978	36	14
<i>jelly belly</i>	CG30040	67	
<i>midline</i>	CG6634	147	
<i>biniou</i>	CG18647	192	
<i>eve</i>	CG2328		266
<i>dMef2</i>	CG1429		
<i>Hand</i>	CG18144		
<i>Sur</i>	CG5772		
<i>Six4</i>	CG3871		

AD, affinity density; TE, TargetExplorer cluster identification algorithm. Lower rank (larger number) reflects decreasing score produced by each algorithm. Absence of a rank indicates that the algorithm did not predict the target.

Table 3. Predicted Tinman Targets With Heart Phenotypes

RNAi target	annotation	score	rank
<i>zfh1</i>	CG1322	3.86	23
<i>pannier</i>	CG3978	3.49	36
<i>branchless</i>	CG4608	3.05	52
<i>Traf2</i>	CG10961	2.94	65
<i>scribbled</i>	CG5462	2.73	76
<i>polychaetoid</i>	CG31349	2.75	79
<i>Pdp1</i>	CG17888	2.51	93

105 of 246 predicted targets were also screened in the RNAi mutation project (Kim *et al.*, 2004). Seven were defective in heart development ($p < 0.01$).

4.2 Enrichment of target genes required for heart development

The gene *tinman* is one of the earliest factors required for formation of the visceral and heart mesoderm primordia (Bodmer, 1993; Azpiazu and Frasch, 1993). Downstream targets of Tinman would thus be expected to affect processes required for the patterning and morphogenesis of the heart.

In a screen for cardiogenic genes using an RNAi approach Kim *et al.* (2004) injected embryos with double-stranded RNA representing a large proportion of individual genes in the *Drosophila* genome. RNAi results in the partial or complete knock-down of expression of the gene whose sequence or partial sequence is contained in the double stranded RNA. And therefore, embryos that have been treated in this manner behave as functional hypomorphs for the gene corresponding to the injected sequence (Misquitta and Paterson, 1999). Kim *et al.* (2004) assayed for perturbed heart development by scoring injected embryos as wild-type or mutant with respect to the expression of the d-Mef2-lacZ transgene.

To assess our target list for enrichment of genes functionally required for heart development, we cross-referenced our predictions with these data, which are available from the Fly Embryo RNAi project (<http://flyembryo.nhlbi.nih.gov/>). In our study, out of 246 predicted target genes from the affinity density algorithm, 105 were also screened in the RNAi mutation project (Kim *et al.*, 2004). Seven of these candidates were defective in heart development (Table 3), a significant increase over background ($p < 0.01$, binomial dist.). This suggests that our algorithm efficiently identified putative heart development genes with the Tinman binding motif.

5 DISCUSSION

5.1 Advantages of affinity density measurement over cluster searches.

Many methods exist for analysis of DNA sequence motifs and their distribution. The practice of finding regulatory targets near regions with statistically overrepresented transcription factor binding sites has been referred to as “known motif mapping” (Ji and Wong, 2006). Here we demonstrated that measurement of regional affinity density offers several advantages over traditional cluster searches. In particular, the model for transcription factor activity that affinity density addresses is that local recruitment of proteins to the chromosome binding sites affects the rate of transcription from nearby loci. Therefore the greater the binding affinity, the greater the likelihood of transcription when cellular conditions allow.

In order to address this model of regional recruitment, we needed to score each point on the chromosome in a way that takes into account the local density of binding sites, as well as their relative strength of binding. The strength of binding of a site is related to its sequence, whereas the density of sites is related simultaneously to the number of sites, and how close together those sites are. The convolution filter we employed gives a regional score in such a way that the score is a function of the binding affinity of individual sites, the aggregate number of sites, and the spacing of sites within the window. Because of the overlapping nature of the windows, every site falls within the center region of one window and hence contributes most strongly to that window.

Secondly, our method does not rely on sequence alignment algorithms for assessment of conservation and functional specificity. We observed the population of sites and their distribution relative to each gene as a predictor of a regulatory relationship, and compared this property between species (Fig 2). By comparing the predictions directly instead of aligning individual binding sites and subsequently scoring clusters in which those sites were found, we introduced different assumptions about the functional relevance of unconserved sites. We speculate that the success of our technique relies at least partly on these assumptions. If this view is correct, the benefits of interspecies comparisons of affinity density outweigh the obvious shortcomings from excluding alignments.

In contrast, an equivalent search method that relies on choosing clusters of sites in an arbitrary window with arbitrary site-number cutoff-parameters is problematic. Because these windows are uniformly weighted, the scores are very sensitive to small changes in the window width, since even a single base change in width can increase the total score by the amount associated with one binding site. If we are looking for binding sites where the expected number

Table 4. Tinman regulatory target predictions

I	II	III	IV
fas ^{f^mh}	CG31708	CG14250	scrib ^h
tin ^{f^mvh}	CG12772	bnl ^{m^vh}	tutl
mXr	msi	LpR2	CG31647
Bap ^{m^vh}	fz ^h	CG7196	pyd ^h
Gbeta5	eag	l(3)82Fd	CG5842
CG12607	klar	CG30268	how ^m
pk	CG32048	CG30387	CG6296
hbn	nau ^m	olf413	CG6295
Rgk1 ^{m^v}	CG10301	ptc ^{m^h}	Spn
CG18262 ^m	CG10300	Btk29A ^h	CG6271
sm	pnr ^{m^vh}	knrl	UGP
rols ^{m^v}	Sox21b ^f	CG13862	CG32040
ed ^{f^m}	mspo ^h	CG5391	Irk2
CG33100	cnc ^m	CG14559	nuf ^{m^vh}
Lmpt ^v	Sulf1 ^m	Traf2	CG3599
beat-IIa ^v	aPKC ^{f^m}	dve	faf
dlp ^f	Src64B	jeb ^m	CG8475
eya ^m	heph ^f	pros	Pdp1 ^h
CG8086	fred	robo3	Gr77a
Aats-asn	Fas2 ^m	CG18769	CG7918
CG15336	Dh31	beat-VI ^m	Ero1L
CG10959	CG3502 ^m	Mdr50	wb ^v
zfh1 ^{m^vh}	Doc3 ^{m^h}	CG10882	CG31221
baz	Argk ^{m^v}	unc-5	Ggamma30A
Sema-1a	CheB93a	lqf ^{m^h}	CheA7a

The top 100 predicted targets from affinity density algorithm. Superscripts indicate expression data from all available flybase sources; ^f foregut/clypeolabrum primordium, ^m mesoderm or somatic muscle, ^v visceral mesoderm, ^h heart (dorsal vessel). Known targets (Table 2) are highlighted in boldface.

of sites in a region of interest is small (perhaps 2 or 3) then a single extra score at the edge of our window can change the total score by 33 to 50 percent. This is the essence of the problem of aliasing (artifact that arises in digital signal processing). This problem can only be addressed by using a convolution kernel designed to reduce the effect of aliasing, such as the one we have used here.

Our method resulted in a significant increase in the representation of known targets from the literature, in addition to likely targets determined by RNAi, over a standard cluster search. Our findings demonstrate a marked improvement of local affinity density over cluster searching as a module detection algorithm. These data give us increased confidence that our prediction set includes a large number of novel true targets of Tinman (see Table 4). Many of these are coexpressed with tinman or in tissues derived from Tinman-expressing precursors. Ultimately, the goal is to apply this approach to the targets of other transcription factors or groups of factors.

5.2 Disadvantages of motif-mapping in general

All motif-mapping studies face several challenges that are not addressed by our method. First among these is that they treat a three dimensional object, the genome of interest, as two dimensional. Linear distances in DNA sequence do not accurately represent spatial distances between sites on transcriptionally active DNA.

Another limitation is lack of information about transcription-factor/DNA interactions for many transcription factors. Also, many transcription factors—especially those with short, frequently occurring binding sites—are not by themselves sufficient to predict gene expression. In such cases a search for enrichment of binding sites will not yield a specific list of predictions. The comparison of techniques presented in this study were facilitated by the high information content of the Tinman binding site. Tinman motifs occur relatively infrequently (< 1 motif per kb in the fly genome), ideal for separation of signal from noise.

In addition, many transcription factor binding sites are shared among families of transcription factors, complicating analysis. For example, our choice of Tinman potentially overlaps homeodomain proteins that share the core NK binding motif. Three of the most important NK homeodomain proteins in development, Tinman, Bagpipe, and Vnd, are known to bind the consensus sequence ‘TCAAGTG’ (Zaffran and Frasch, 2005; Gehring, 1987) with high affinity. Since Bagpipe acts in concert with Tinman to induce visceral mesoderm (Azpiazu and Frasch, 1993), we assume its targets are a subset of Tinman’s. However, Vnd is required for cell-fate specification in the developing CNS (White *et al.*, 1983; Skeath *et al.*, 1994). We anticipate that any prediction set exemplified by the one produced in this study necessarily includes targets of other transcription factors in addition to false positives. Indeed, our predictions for Tinman include a number of genes known to be involved in nervous system development such as *Semaphorin-1a* (Yu *et al.*, 1998) and *robo3* (Simpson *et al.*, 2000) (4), consistent with the expression of Vnd in the central nervous system. To complicate matters further, we can not rule out overlap of Tinman and Vnd target sets. For example the NK homeodomain gene *ladybird early* is involved in both cardiac development (Jagla *et al.*, 1997, 2002; Zikova *et al.*, 2003) and neuronal specification (De Graeve *et al.*, 2004), making it a potential candidate target of either Tinman or Vnd. An additional complication is that some homeobox genes act as repressors. It is almost surprising, given these caveats, that there is any specificity to computational predictions at all.

Studies of well-characterized pathways in which the authors analyzed several transcription factors simultaneously (Stathopoulos *et al.*, 2002; Berman *et al.*, 2002) have been able to profitably sidestep around these issues. More recently (Segal *et al.*, 2008) predicted the spatial distribution of segmentation genes from the distribution of transcription factors along the antero-posterior axis of the *Drosophila* embryo by calculating the free energy of transcription factor binding in enhancer regions. This “thermodynamic” model for expression of gene regulation is conceptually related to our approach, but the emphasis on spatial prediction and the analysis of promoter regions of preselected genes make it difficult to compare. To identify enhancers, we experimented with different window sizes and relative positioning. We tried scoring only the regions 5’ or 3’ of genes, excluding coding regions, widening the gene region, or using fixed windows at the center of the gene. None of these alterations in protocol enriched for known targets, in fact in some cases they resulted in loss of specificity (data not shown), suggesting that functional enhancer sequences are found anywhere within a gene, and the probability of binding sites affecting gene expression decreases with distance from the coding region.

In our study we used binding-affinity weighted pattern-matching, however the signal processing approach used here may theoretically be used in combination with any of the available motif-detection scoring algorithms, represented as ϕ in equation 1. If we were to investigate two or more transcription factor binding motifs using this approach, it would be necessary to know more about the relative binding affinities of each, or to assume equal affinity, for better or worse.

5.3 Implications for sequence analysis

For any motif-scanning exercise it is necessary to determine which motifs are likely to be biologically functional. To gauge this property researchers have focused on direct nucleotide conservation because it is one of the most well-researched fields in bioinformatics (Kumar and Filipinski, 2007). However, intergenic sequences tend to diverge faster than transcript-encoding sequences, so the study of direct conservation of motifs in non-coding DNA is limited to comparisons among closely related species. Still, since some binding sites are always conserved in closely related species, it makes sense that preselection of conserved sites increases specificity. We submit that regional enrichment of binding sites is a better predictor of regulatory targeting than clusters of individually conserved sites. When such conservation is observable it is likely the result of strong selection pressures from which the majority of functioning binding sites are exempted.

A test for conservation of aligned nucleotide sequence motifs acts as a filter to remove all other functional binding sites. But factors may bind any reasonably high-affinity site regardless of its conservation between any arbitrary two species. Bowler *et al.* (2006) described pervasive low-level, non-specific transcription from low affinity sites of the transcription factor Ftz-F1, although the biological significance of this activity remains unclear. Perhaps more significantly, Berman *et al.* (2004) proposed the existence of “preserved” binding sites, close enough in one genome to the analogous position in a sister genome to substitute functionally, but not close enough in a DNA sequence alignment to be considered conserved. Such sites are critical in distinguishing true *cis*-regulatory modules from false positive ones (Berman *et al.*, 2004). These observations led us to consider binding affinity as a thermodynamic and spatial property of the chromosome and to try to measure it using signal processing, in a manner similar to the one that inspired (Segal *et al.*, 2008).

Since it is thought that the core regulatory networks that govern axis formation and subdivision of the germ layers are preserved across animal phyla (Levine and Davidson, 2005), the nucleotide-for-nucleotide conservation constraint seems too narrow. Current search methods assume that critical individual sites must remain intact in a population of sites. This assumption is valid on a microevolutionary scale. However, since non-protein-coding regions probably have fewer constraints on the physical order and number of motifs required to make a functional element, sequence alignment may give an incomplete picture of conservation in regulatory regions.

Mutation of individual sites probably has a pronounced effect on fitness and viability only in rare cases. Some of this robustness may be attributed to the larger pool of binding sites (Berman *et al.*, 2004; Segal *et al.*, 2008). For example, deleterious mutations in one site might produce little net effect and are likely to be compensated for

by selection pressures that stabilize novel binding sites or variants of pre-existing sites with weaker affinity. Our finding that affinity density has a stronger predictive value compared with searches for conserved clusters of motifs suggests that encoded information in DNA presents itself in ways that are detectable by means other than sequence alignments.

By assigning a statistic to binding affinity and selecting shared predictions between *D. melanogaster* and *D. pseudoobscura*, we treated this feature as a potentially conserved structural character of chromatin. If this view is correct it implies that using binding site affinity to approximate the target set of a transcription factor might aid in revealing the evolutionary relationships of regulatory networks amongst closely related families of organisms that are otherwise too distantly related to be analyzed by sequence alignment. We would like to see the application of this principle within *Drosophilidae*. In conclusion, the measurement and analysis of the distribution of transcription factor affinities is a promising novel approach for analyzing the distribution of transcription factor binding sites and their targets.

FUNDING

Dennis Hazelett was supported by a grant from the National Institutes of Health 1T32HD049309 to the Developmental Biology Training Program at OHSU. Joseph Weiss was supported by RO1-HL075498-01A1.

ACKNOWLEDGEMENTS

We thank Drs. Francesca Mariani and Rachel Dresbeck for editing this manuscript and helpful discussions.

REFERENCES

- Azpiazu, N. and Frasch, M. (1993). *tinman* and *bagpipe*: two homeobox genes that determine cell fate in the dorsal mesoderm of *Drosophila*. *Genes & Development*, **7**, 1325–1340.
- Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., and Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proceedings of the National Academy of Science of the U.S.A.*, **99**(2), 757–762.
- Berman, B. P., Pfeiffer, B. D., Laverty, T. R., Salzberg, S. L., Rubin, G. M., Eisen, M. B., and Celniker, S. E. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biology*, **5**(9), R61.1–R61.24.
- Bodmer, R. (1993). The gene *tinman* is required for specification of the heart and visceral muscles in *Drosophila*. *Development*, **118**, 719–729.
- Bowler, T., Kosman, D., Licht, J., and Pick, L. (2006). Computational identification of ftz/ftz-f1 downstream target genes. *Developmental Biology*, **299**, 78–90.
- De Graeve, F., Jagla, T., Da Ponte, J.-P., Rickert, C., Dastugue, B., Urban, J., and Jagla, K. (2004). The *ladybird* homeobox genes are essential for the specification of a subpopulation of neural cells. *Developmental Biology*, **270**, 122–134.
- Gehring, W. (1987). Homeoboxes in the study of development. *Science*, **246**, 1245–1252.
- Halfon, M. S., Grad, Y., Church, G. M., and Michelson, A. M. (2002). Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Research*, **12**, 1019–1028.
- Hamming, R. W. (1998). *Digital filters*. General Publishing Company, Ltd., 30 Desmill Road, Don Mills, Toronto, Ontario, 3rd edition.
- Jagla, K., Frasch, M., Jagla, T., Dretzen, G., Bellard, F., and Bellard, M. (1997). *ladybird*, a new component of the cardiogenic pathway in *Drosophila* required for diversification of heart precursors. *Development*, **124**, 3471–3479.
- Jagla, T., Bidet, Y., Da Ponte, J.-P., Dastugue, B., and Jagla, K. (2002). Cross-repressive interactions of identity genes are essential for proper specification of cardiac and muscular fates in *Drosophila*. *Development*, **129**, 1037–1047.
- Ji, H. and Wong, W. H. (2006). Computational biology: Toward deciphering gene regulatory information in mammalian genomes. *Biometrics*, **62**(3), 645–663.
- Kasahara, H., Usheva, A., Ueyama, T., Aoki, H., Horikoshi, N., and Izumo, S. (2001). Characterization of homo- and heterodimerization of cardiac CsxNkx2.5 homeoprotein. *Journal of Biological Chemistry*, **276**(7), 4570–4580.
- Kim, Y.-O., Park, S.-J., Balaban, R. S., Nirenberg, M., and Kim, Y. (2004). A functional genomic screen for cardiogenic genes by rna interference in developing *drosophila* embryos. *Proceedings of the National Academy of Science of the U. S. A.*, **101**(1), 159–164.
- Kumar, S. and Filipinski, A. (2007). Multiple sequence alignment: In pursuit of homologous dna positions. *Genome Research*, **17**, 127–135.
- Levine, M. and Davidson, E. H. (2005). Gene regulatory networks for development. *Proceedings of the National Academy of Science of the U. S. A.*, **102**(14), 4936–4942.
- Markstein, M., Markstein, P., Markstein, V., and Levine, M. S. (2002). Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proceedings of the National Academy of Science of the U. S. A.*, **99**(2), 763–768.
- Misquitta, L. and Paterson, B. M. (1999). Targeted disruption of gene function in *Drosophila* by rna interference (rna-i) a role for *nautilus* in embryonic somatic muscle formation. *Proceedings of the National Academy of Science of the U.S.A.*, **96**(4), 1451–1456.
- Ocorr, K., Akasaka, T., and Bodmer, R. (2006). Age-related cardiac disease model of *Drosophila*. *Mechanisms of Ageing and Development*, **128**, 112–116.
- Prall, O. W. J., Elliott, D. A., and Harvey, R. P. (2002). Developmental paradigms in heart disease: insights from *tinman*. *Annals of Medicine*, **34**, 148–156.
- Schott, J.-J., Benson, D. W., Basson, C. T., Pease, W., Silberbach, G. M., Moak, J. P., Maron, B. J., Seidman, C. E., and Seidman, J. G. (1998). Congenital heart disease caused by mutations in the transcription factor nkx2-5. *Science*, **281**, 108–111.
- Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., and Gall, U. (2008). Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature*, **451**, 535–541.
- Simpson, J. H., Bland, K. S., Fetter, R. D., and Goodman, C. S. (2000). Short-range and long-range guidance by slit and its robo receptors: a combinatorial code of robo receptors controls lateral position. *Cell*, **103**(7), 1019–1032.
- Skeath, J. B., Panganiban, G., and Carroll, S. B. (1994). The ventral nervous system defective gene controls proneural gene expression at two distinct steps during neuroblast formation in *Drosophila*. *Development*, **120**(6), 1517–1524.
- Sosinsky, A., Bonin, C. P., Mann, R. S., and Honig, B. (2003). Target Explorer: an automated tool for the identification of new target genes for a specified set of transcription factors. *Nucleic Acids Research*, **31**(13), 3589–3592.
- Stathopoulos, A. and Levine, M. (2004). Whole-genome analysis of *Drosophila* gastrulation. *Current Opinion In Genetics & Development*, **14**(5), 477–484.
- Stathopoulos, A., Van Drenth, M., Erives, A., Markstein, M., and Levine, M. (2002). Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo. *Cell*, **111**(5), 687–701.
- Watada, H., Mirmira, R. G., Kalamaras, J., and German, M. S. (2000). Intramolecular control of transcriptional activity by the NK2-specific domain in NK-2 homeodomain proteins. *Proceedings of the National Academy of Sciences of the U. S. A.*, **97**(17), 9443–9448.
- White, K., DeCelles, N. L., and Enlow, T. C. (1983). Genetic and developmental analysis of the locus *vnd* in *Drosophila melanogaster*. *Genetics*, **104**, 433–448.
- Yu, H.-H., Araj, H. H., Ralls, S. A., and Kolodkin, A. L. (1998). The transmembrane semaphorin Sema I is required in *Drosophila* for embryonic motor and CNS axon guidance. *Neuron*, **20**(2), 207–220.
- Zaffran, S. and Frasch, M. (2005). The homeodomain of Tinman mediates homo- and heterodimerization of nk proteins. *Biochemical and Biophysical Research Communications*, **334**(2), 361–369.
- Zikova, M., Da Ponte, J.-P., Dastugue, B., and Jagla, K. (2003). Patterning of the cardiac outflow region in *Drosophila*. *Proceedings of the National Academy of Science of the U. S. A.*, **100**(21), 12189–12194.