BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 13 2009, pages 1617–1624
doi:10.1093/bioinformatics/btp282

Sequence analysis

Affinity Density: a novel genomic approach to the identification of transcription factor regulatory targets

Dennis J. Hazelett[1,*], Daniel L. Lakeland[2] and Joseph B. Weiss[3]

[1]Integrative Biosciences, Oregon Health and Science University (OHSU), 611 SW Campus Dr SOD705, Portland, OR 97239, [2]Independent Consultant, Los Angeles, CA and [3]Department of Cardiology, Oregon Health and Science University (OHSU), 3303 SW Bond Ave CHH14W, Portland, OR 97239, USA

Received on February 6, 2009; revised on April 21, 2009; accepted on April 22, 2009

Advance Access publication April 28, 2009

Associate Editor: Alex Bateman

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 complete genomes of *Drosophila melanogaster* and *D.pseudoobscura* were analyzed for binding sites of the homeodomain proteinc 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 http://www.ohsu.edu/cellbio/dev_biol_prog/affinitydensity/.

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 4–5 bp 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 indepth 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 a fresh look at the distribution of binding sites across an entire genome. 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 short-range clusters in favor of a density map representing the likely occupancy of transcription factors along the sequence. Our approach bears some resemblance to prior analyses (Frith *et al.*, 2002; Ward and Bussemaker, 2008).

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 chosen, specificity decreases but the search is more sensitive to targets whose

*To whom correspondence should be addressed.

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 and Levine, 2004; Stathopoulos *et al.*, 2002). 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 short-range 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 a computationally efficient manner across the entire genome. 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 versus 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 bp to 15 bp between binding sites and gave a maximum weight of 8.27 to dimers 6 bp apart. The dimer function was truncated to 2.0 for spacers sized at <3 bp or >12 bp, 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 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 statistics makes a reasonable tradeoff between measuring the density of sites, and measuring the location where that density occurs. Thus, the affinity density ρ, for the n-th genomic segment of length N, is given by

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

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):

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

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

Unless otherwise indicated, $N = 2^{12}$ for this study, dividing the average *Drosophila* gene ($\sim$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 to reduce irrelevant background rather than a hard cutoff. The sigmoid can be thought of as a logistic regression curve with two important parameters, one for the location of the 50% transition point, and one for the transition rate. We set the 50% point of the sigmoid at a reference density ρ_{ref} selected to be approximately three times the expected score, and the transition rate so that the SD of the implied density was equal to the expected score for one binding site 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 (e.g. $\rho_{\text{ref}} \approx 0.541$, and $s \approx 0.110$ for *D.melanogaster* using the Tinman motif).

$$\gamma = \frac{1}{1+e^{(-(\rho_n-\rho_{\text{ref}}/s))}} \quad (3)$$

This sigmoid was changed for each genome and binding site motif by calculating appropriate expected density and transition rate. Due to the shape, high scoring regions are not sensitive to the parameters of the sigmoid, but the sigmoid parameters do affect how much of the low scoring regions contribute to the score for each gene.

A plot of the sigmoidally filtered Affinity Density data reveals the variation in binding affinity at the level of the whole genome (data not shown). At this

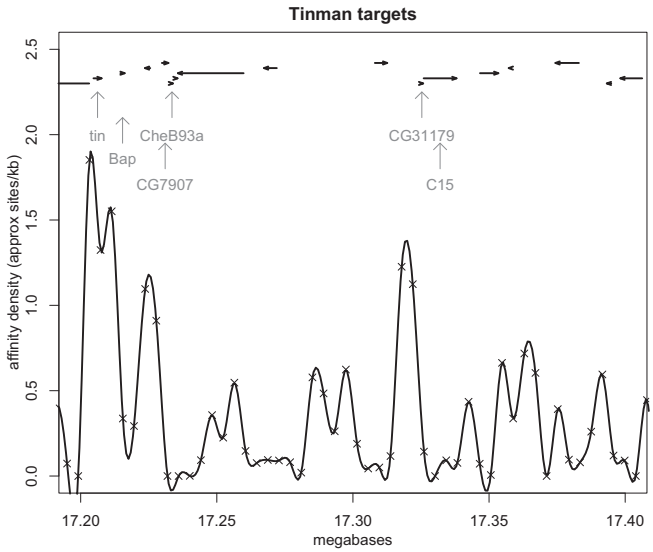

Fig. 1. A region of interest demonstrates the physical location of several predicted Tinman targets and other genes (arrows) relative to regions of high Tinman binding affinity. A spline was fit to the individual affinity data ('cross'). Some of the displayed targets are not conserved between *D.pseudoobscura*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

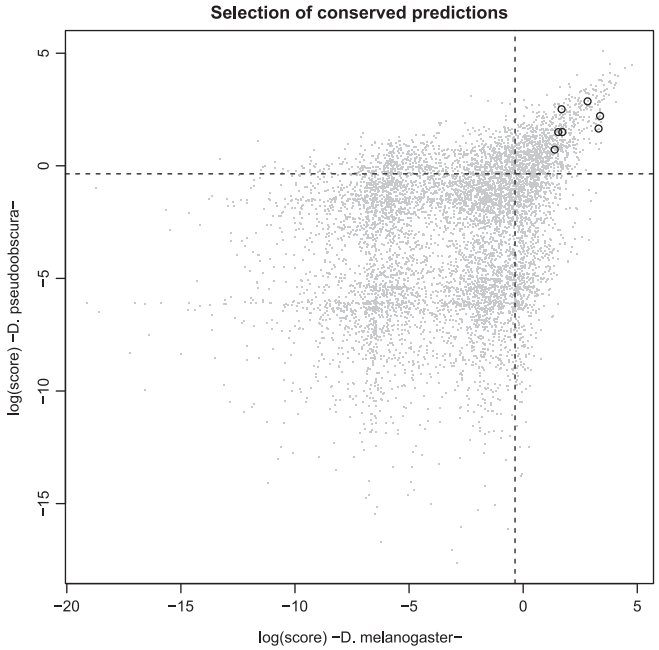

Fig. 2. Target predictions were selected on the basis of conservation in two species of *Drosophila*. Each point (gray) represents the log of scores of a single gene. Dotted lines reveal the cutoff values used for selection. The points corresponding to known targets of Tinman from the literature are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

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 (Fig. 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 1 kb to either side of the annotation. We explored various buffer sizes for the regions around the gene, and settled on 1 kb. 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 promotor 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). This *ad hoc* method was useful for our purposes, but should be replaced with a method based on a systematic approach that optimizes the tradeoff between longer prediction sets, and the researcher's perceived cost of missing an additional valid prediction.

Our Affinity Density model assumes that the number of sites within a region 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 Section 5). 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 GenomeSurveyor (Noyes *et al.*, 2008) as a predictor of known Tinman regulatory targets (Fig. 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 (BDGP; 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

Table 1. Coexpression of prediction sets with Tinman

Algorithm	Expression data	ovrlp tin expr pat (%)	*P*-value
Affinity Density	33	18 (54.5)	0.0071
TargetExplorer	6	2 (33.8)	0.3291
GenomeSurveyor	20	10 (50.0)	0.0581

Top 100 predicted Tinman targets with equivalent coexpression terms in the BDGP *in situ* expression database.

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 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). Out of top 100 candidates, 33 had associated expression terms in BDGP, and of these 18 had terms equivalent to a subset of those assigned to the Tinman expression pattern. This constituted an enrichment of 54.5% over a background of 33.8% and was statistically significant using the binomial distribution ($P = 0.0071$, Table 1).

Furthermore, we compared our binding affinity-derived target predictions with the results of our optimized conserved-motif cluster searches using TargetExplorer (Sosinsky *et al.*, 2003) and GenomeSurveyor (Noyes *et al.*, 2008). We optimized our cluster searches 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 of *tinman* coexpression in the affinity density-derived prediction list than in the conserved-motif cluster list derived from TargetExplorer output and GenomeSurveyor, which uses a different algorithm that is also based on locating clusters of statistically significant conserved binding sites, using a hidden Markov model (Sinha *et al.*, 2003) (Table 1).

Next we determined the sensitivity to known Tinman targets. We identified in the literature all published accounts where the authors present 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). We then compared how well these genes were predicted by different algorithms.

To compare the approaches, we compared the relative rankings assigned to the predictions by each algorithm. For Affinity Density, genes were ranked by decreasing score, which 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, ranking by score also reflects the relative strength of the prediction from cluster-search algorithms, because it is an expression of the likelihood of finding transcription factors associated with that gene.

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 12 published Tinman target genes, affinity density identified 7, including *zfh1*, *biniou* and *jelly belly*. The conserved motif searches from TargetExplorer (Sosinsky *et al.*, 2003) produced only four (Table 2), only one of which

Table 2. Comparison of motif-mapping methods

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

AD, Affinity Density; TE, TargetExplorer (Sosinsky *et al.*, 2003); GS, GenomeSurveyor (Noyes *et al.*, 2008); Lower rank (larger number) reflects decreasing score produced by each algorithm. Absence of a rank indicates that the algorithm did not predict the target.

(*eve*) was not predicted as a target by the affinity algorithm. The GenomeSurveyor (Noyes *et al.*, 2008) search likewise produced only one of the known targets. 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.

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 (Azpiazu and Frasch, 1993; Bodmer, 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 this dataset, 126 out of 5730 genes had a phenotype visible in the embryonic heart. 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 ($P = 0.0064$, binomial distance). In contrast, neither TargetExplorer (Sosinsky *et al.*, 2003) nor GenomeSurveyor (Noyes *et al.*, 2008)

Table 3. Predicted Tinman targets with heart phenotypes

RNAi target	Annotation	Score	Rank
zfh1	CG1322	3.86	23
pannier	CG3978	3.49	36
branchless	CG4608	3.05	52
Traf2	CG10961	2.94	65
scribbled	CG5462	2.73	76
polychaetoid	CG31349	2.75	79
Pdp1	CG17888	2.51	93

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

predicted a significant number of RNAi phenotypes. Out of 112 candidates from the TargetExplorer list, 1 had a phenotype ($P = 0.2088$). Out of 57 candidates that were tested from the GenomeSurveyor predictions, 2 had a phenotype in the RNAi screen ($P = 0.2271$). This suggests that our algorithm efficiently identified putative heart development genes with the Tinman binding motif, consistent with Tinman's requirement in heart development, whereas prediction with clustering algorithms did not.

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.

Second, 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 chooses clusters of sites using default parameters leads to an iterative and time-consuming process of optimization. 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 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 33–50%. This is the essence of the problem of aliasing (an 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.

The results from comparison of *tinman* coexpressed genes (Table 1) suggest that Affinity Density predicted a greater proportion of coexpressed genes than TargetExplorer (Sosinsky *et al.*, 2003) and GenomeSurveyor (Noyes *et al.*, 2008) in an equivalent-sized list of predictions. It is difficult to ascertain the performance of TargetExplorer from this test because there were too limited data available (six genes with *in situ* expression patterns). GenomeSurveyor performed comparably with Affinity Density by this measure, however, although the result was not statistically significant. GenomeSurveyor selects the two nearest genes to each hit region and therefore likely benefits from coregulation. This strikes us as a very reasonable assumption to make when assigning significant regions to target genes. Although our method assigns hit regions to multiple target genes, it relies upon the definition of gene region in the annotations. Future modifications could be made to include additional information about regulatory regions as it becomes available. For example, there is a well-characterized Tinman-enhancer region about 7 kb downstream of the *eve* locus which effects the transcription of the *even-skipped* gene (Knirr and Frasch, 2001). This would explain why Affinity Density failed to predict *eve* as a Tinman target in our hands (Table 2).

In addition to missing certain targets due to overly stringent relative location requirements, it is also known that large genes have a bias towards being falsely predicted simply due to their greater spatial extent and therefore greater chance of being near an unrelated regulatory region (Taher and Ovcharenko, 2009). This is a factor we observed during our analysis, and our initial attempt to subtract the trend reduced sensitivity. Part of the reason for adopting the sigmoid filter is to eliminate the effect of many small signals adding up over a large region to something that compares to a strong signal over a short region.

Even without these considerations, our findings demonstrated a marked improvement of local affinity density over cluster searching as a module detection algorithm. Our method resulted in a significant increase in the representation of known targets from the literature. In addition, we predicted a significant number of genes for which RNAi yields a relevant phenotype in embryos. In contrast, neither TargetExplorer (Sosinsky *et al.*, 2003) nor GenomeSurveyor (Noyes *et al.*, 2008) predicted a significant number of RNAi genes. These data give us increased confidence that our prediction set includes a large number of novel true targets of Tinman (Table 4). Many of these genes are coexpressed with Tinman or in tissues derived from Tinman-expressing precursors.

Table 4. Tinman regulatory target predictions

I	II	III	IV
fas^{fmh}	CG31708	CG14250	scrib^{h}
tin^{fmvh}	CG12772	bnl^{mvh}	tutl
mXr	msi	LpR2	CG31647
Bap^{mvh}	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^{mv}	CG10301	ptc^{mh}	Spn
CG18262^{m}	CG10300	Btk29A^{h}	CG6271
sm	**pnr**^{mvh}	knrl	UGP
rols^{mv}	Sox21b^{f}	CG13862	CG32040
ed^{fm}	mspo^{h}	CG5391	Irk2
CG33100	cnc^{m}	CG14559	nuf^{mvh}
Lmpt^{v}	Sulf1^{m}	Traf2	CG3599
beat-IIa^{v}	aPKC^{fm}	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^{mvh}	Doc3^{mh}	CG10882	CG31221
baz	Argk^{mv}	unc-5	Ggamma30A
Sema-1a	CheB93a	lqf^{mh}	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.

Ultimately, the goal is to apply this approach to the targets of other transcription factors or groups of factors. As a first-order attempt to determine whether this method is likely to be generally applicable, we conducted a cursory survey with simple regular expressions to represent various other transcription factors with well-characterized binding sites, and without any sigmoidal filtering. Out of nine transcription factors, we were able to enrich coexpressed genes as in Table 1 for five factors including Krüppel, gooseberry, paired, snail and Ultrabithorax, with the remaining four factors, serpent, twin-of-eyeless, twist and HLHm5, showing enrichment but not statistical significance. This suggests that the method is broadly applicable even with a very crude motif-recognition algorithm and no background filter. Ward and Bussemaker (2008) successfully used a similar affinity based approach, and compared their affinity score across yeast genomes. Our method differs in that we measure affinity across entire gene regions instead of narrowly defined promoter sequences, and take advantage of an anti-aliasing kernel procedure to reduce artifacts. Together these results support the use of affinity-density based calculations for the identification of regulatory targets.

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 3D object, the genome of interest, as 2D. 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 kilo base 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' (Gehring, 1987; Zaffran and Frasch, 2005) with high affinity. Since Bagpipe acts in concert with Tinman to induce visceral mesoderm (Azpiazu and Frasch, 1993), we assume that its targets are a subset of Tinman's. However, Vnd is required for cell-fate specification in the developing CNS (Skeath *et al.*, 1994; White *et al.*, 1983). 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 the 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 cannot 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 these computational predictions at all.

Studies of well-characterized pathways in which the authors analyzed several transcription factors simultaneously (Berman *et al.*, 2002; Stathopoulos *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 5′ or 3′ regions 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 Filipski, 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).

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.

ACKNOWLEDGEMENTS

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

Funding: National Institutes of Health (1T32HD049309 to D.J.H. for Developmental Biology Training Program at OHSU); RO1-HL075498-01A1 (J.B.W.).

Conflict of Interest: none declared.

REFERENCES

Azpiazu,N. and Frasch,M. (1993) *tinman* and *bagpipe*: two homeobox genes that determine cell fate in the dorsal mesoderm of *Drosophila*. *Genes Dev.*, **7**, 1325–1340.

Berman,B.P. *et al.* (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.

Berman,B.P. *et al.* (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol.*, **5**, 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. *et al.* (2006) Computational identification of ftz/ftz-f1 downstream target genes. *Dev. Biol.*, **299**, 78–90.

De Graeve,F. *et al.* (2004) The *ladybird* homeobox genes are essential for the specification of a subpopulation of neural cells. *Dev. Biol.*, **270**, 122–134.

Frith,M.C. *et al.* (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.*, **30**, 3214–3224.

Gehring,W. (1987) Homeoboxes in the study of development. *Science*, **246**, 1245–1252.

Halfon,M.S. *et al.* (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.*, **12**, 1019–1028.

Hamming,R.W. (1998) *Digital filters*, 3rd edn. General Publishing Company, Ltd., Toronto, Ontario.

Jagla,K. *et al.* (1997) *ladybird*, a new component of the cardiogenic pathway in *Drosophila* required for diversification of heart precursors. *Development*, **124**, 3471–3479.

Jagla,T. *et al.* (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**, 645–663.

Kasahara,H. *et al.* (2001) Characterization of homo- and heterodimerization of cardiac CsxNkx2.5 homeoprotein. *J. Biol. Chem.*, **276**, 4570–4580.

Kim,Y.-O. *et al.* (2004) A functional genomic screen for cardiogenic genes by RNA interference in developing Drosophila embryos. *Proc. Natl Acad. Sci. USA*, **101**, 159–164.

Knirr,S. and Frasch,M. (2001) Molecular integration of inductive and mesoderm-intrinsic inputs governs *even-skipped* enhancer activity in a subset of pericardial and dorsal muscle progenitors. *Dev. Biol.*, **238**, 13–26.

Kumar,S. and Filipski,A. (2007) Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome Res.*, **17**, 127–135.

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. *Proc. Natl Acad. Sci. USA*, **96**, 1451–1456.

Noyes,M.B. (2008) A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. *Nucleic Acids Res.*, **36**, 2547–2560.

Segal,E. *et al.* (2008) Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature*, **451**, 535–541.

Simpson,J.H. *et al.* (2000) Short-range and long-range guidance by slit and its robo receptors: a combinatorial code of robo receptors controls lateral position. *Cell*, **103**, 1019–1032.

Sinha,S. *et al.* (2003) A probabalistic method to detect regulatory modules. *Bioinformatics*, **19** (Suppl. 1), i292–i301.

Skeath,J.B. *et al.* (1994) The ventral nervous system defective gene controls proneural gene expression at two distinct steps during neuroblast formation in *Drosophila*. *Development*, **120**, 1517–1524.

Sosinsky,A. *et al.* (2003) Target Explorer: an automated tool for the identification of new target genes for a specified set of transcription factors. *Nucleic Acids Res.*, **31**, 3589–3592.

Stathopoulos,A. and Levine,M. (2004) Whole-genome analysis of *Drosophila* gastrulation. *Curr. Opin. Genet. Dev.*, **14**, 477–484.

Stathopoulos,A. *et al.* (2002) Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo. *Cell*, **111**, 687–701.

Taher,L. and Ovcharenko,I. (2009) Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements. *Bioinformatics*, **25**, 578–584.

Ward,L.D. and Bussemaker,H.J. (2008) Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics*, **24**, i165–i171.

Watada,H. *et al.* (2000) Intramolecular control of transcriptional activity by the NK2-specific domain in NK-2 homeodomain proteins. *Proc. Natl Acad. Sci. USA*, **97**, 9443–9448.

White,K. *et al.* (1983) Genetic and developmental analysis of the locus *vnd* in *Drosophila melanogaster*. *Genetics*, **104**, 433–448.

Yu,H.-H. *et al.* (1998) The transmembrane semaphorin Sema I is required in *Drosophila* for embryonic motor and CNS axon guidance. *Neuron*, **20**, 207–220.

Zaffran,S. and Frasch,M. (2005) The homeodomain of Tinman mediates homo- and heterodimerization of NK proteins. *Biochem. Biophys. Res. Commun.*, **334**, 361–369.

Zikova,M. *et al.* (2003) Patterning of the cardiac outflow region in *Drosophila*. *Proc. Natl Acad. Sci. USA*, **100**, 12189–12194.