BMC Bioinformatics

BioMed Central

Research article

Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo

Nikolaus Rajewsky*[1], Massimo Vergassola[1,3], Ulrike Gaul[2] and Eric D Siggia*[1]

Address: [1]Center for Studies in Physics and Biology, The Rockefeller University, 1230 York Avenue, New York, NY, USA, [2]Laboratory for Developmental Neurogenetics, The Rockefeller University, 1230 York Avenue, New York, NY, USA and [3]CNRS, Observatoire Côte d'Azur, Lab. G. D. Cassini, Nice, France

E-mail: Nikolaus Rajewsky* - nr@edsb.rockefeller.edu; Massimo Vergassola - massimo@obs-nice.fr; Ulrike Gaul - gaul@mail.rockefeller.edu; Eric D Siggia* - siggia@eds3.rockefeller.edu

*Corresponding authors

Published: 24 October 2002

BMC Bioinformatics 2002, **3**:30

Received: 29 August 2002
Accepted: 24 October 2002

This article is available from: http://www.biomedcentral.com/1471-2105/3/30

Abstract

Background: Regulation of gene transcription is crucial for the function and development of all organisms. While gene prediction programs that identify protein coding sequence are used with remarkable success in the annotation of genomes, the development of computational methods to analyze noncoding regions and to delineate transcriptional control elements is still in its infancy.

Results: Here we present novel algorithms to detect *cis*-regulatory modules through genome wide scans for clusters of transcription factor binding sites using three levels of prior information. When binding sites for the factors are known, our statistical segmentation algorithm, Ahab, yields about 150 putative gap gene regulated modules, with no adjustable parameters other than a window size. If one or more related modules are known, but no binding sites, repeated motifs can be found by a customized Gibbs sampler and input to Ahab, to predict genes with similar regulation. Finally using only the genome, we developed a third algorithm, Argos, that counts and scores clusters of overrepresented motifs in a window of sequence. Argos recovers many of the known modules, upstream of the segmentation genes, with no training data.

Conclusions: We have demonstrated, in the case of body patterning in the Drosophila embryo, that our algorithms allow the genome-wide identification of regulatory modules. We believe that Ahab overcomes many problems of recent approaches and we estimated the false positive rate to be about 50%. Argos is the first successful attempt to predict regulatory modules using only the genome without training data. Complete results and module predictions across the Drosophila genome are available at [http://uqbar.rockefeller.edu/~siggia/].

Background

In higher eukaryotes, many genes feature differential spatial-temporal expression during development and throughout the life cycle of the organism. Their complex transcription regulation is thought to be achieved by the combinatorial action of multiple transcription factors

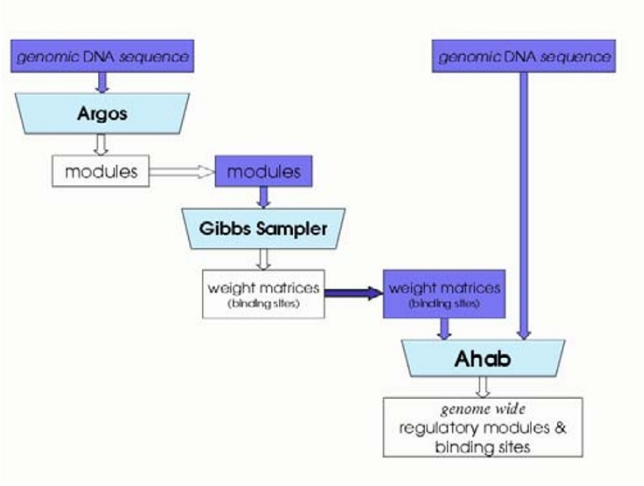

Figure 1
Summary of the input (dark blue) and output (white) of the three algorithms (light blue).

which bind to *cis*-regulatory DNA sequences. Here, transcription factors are defined as proteins which recognize and bind regulatory sites and have a potential to modulate directly or indirectly through the recruitment of cofactors the activity of the basal transcriptional apparatus of proximal genes. The number of transcription factors is a substantial part of the total number of genes in any organism, for example about 700 out of 13,500 genes in Drosophila [1].

Although combinatorial action of transcription factors has been studied throughout the life cycle of organisms [2], perhaps the most coherent picture has emerged in the context of developmental processes [3,4]. Here, a great number of experiments suggest that a major part of the gene regulatory apparatus is organized in the form of separable *cis*-regulatory modules [3]. A given module defines specific aspects of the spatio-temporal pattern of gene expression by the combinatorial action of multiple transcription factors which together define the rate of transcription. Modules thus integrate inputs from several genes and regulate another gene to form developmental networks. Modules seem to share several architectural features [5]: They are typically only hundreds of nucleotides in length and contain multiple binding sites for as many as 4–5 different transcription factors. The frequent occurence of multiple copies of the same motif as well as the enrichment of certain combinations of motifs in a module in comparison with the genome at large provide the basis for our computational strategies to predict genes which are part of the same regulatory network. Existing algorithms for discovering modules [6–8] are based on counting the number of matches of a certain minimal strength to known motifs and thus require ad-hoc parameters for each motif, resulting in parameter dependent predictions. These algorithms are also bound to miss multiple weak binding sites which are known to be present in many modules. We demonstrate a novel algorithm, Ahab, which overcomes these problems.

However, the binding sites (motifs) which reside in modules are often not known. A recent paper [9] proposed an algorithm for identfying these sites; here we show that a standard method is capable of identifying typically half of the known binding sites in a module. Its entire output can then be used as input to Ahab (Figure 1). Finally, we ask if the redundancy of sites inside modules is strong enough to predict modules using only genomic sequence. To our knowledge, our algorithm Argos is the first successful attempt to do this for a metazoan genome.

Our system of choice is the body patterning of the early Drosophila embryo which is established by a multi-tiered hierarchy of transcription factors [10]. Broadly distributed maternal factors trigger zygotic gap gene expression in discrete domains along the anterior-posterior axis of the embryo. Maternal and gap gene factors together trigger pair rule gene expression in 7 alternating stripes, which in turn regulate segment polarity and homeotic gene expression in 14 stripes. Many of these factors are known, their binding sites have been studied, and more than 20 modules have been identified. For this system, we show by comparing our predictions to literature results that we can predict regulatory modules using different levels of input – binding sites, regulatory sequence identified by 'promoter bashing'/sufficiency tests, and genomic sequence itself. Altogether, we predict roughly 200 new modules and 7 new sequence motifs and we analyze and validate the performance of each of the three algorithms.

Figure 1 summarizes the input/output levels and serves as a reference for what follows.

Results

Using known binding sites, Ahab finds 135 significant new modules in the genome

There are a sufficient number of binding sites in the literature (see additional File 1 and 2) for us to construct frequency weight matrices for the maternal transription factors Bicoid (Bcd), Caudal (Cad) and Dorsal (Dl), as well as the zygotic gap gene factors Hunchback (Hb), Kruppel (Kr), Knirps (Kni), and Tailless (Tll), and the torso response element (torRE). With the exception of giant, which is too ill-defined to model, these are all the maternal/gap genes at the top of the segmentation gene hierarchy.

Modules can be located by scanning the genome in windows, counting the number of matches to each matrix

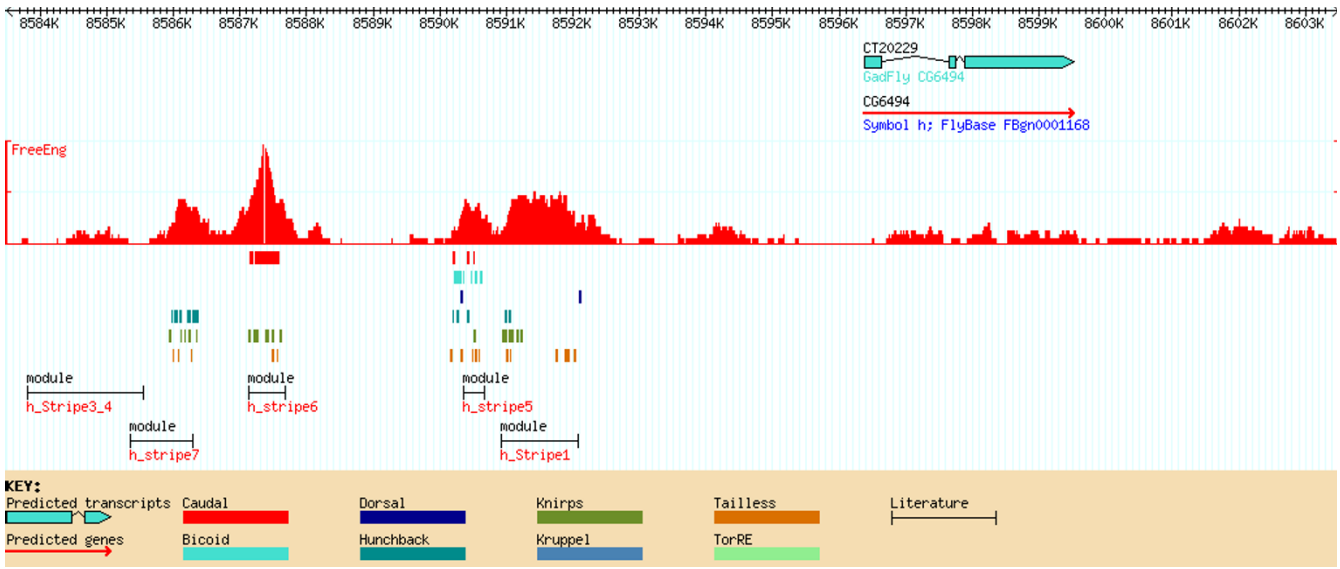

Figure 2
Ahab score for the hairy locus Module score for the hairy locus (screenshot from our interactive web browser). Plotted is the Ahab score as a function of position in the genome. Known modules are marked as "module". Four of the known modules (stripes 1 and 5–7) have high enough scores to appear among the top 146 genome wide predictions and Ahabs predicted binding sites are mapped out in these cases. The stripe3+4 module is not recovered.

with a score better than some value, and then combining scores and ranking [6,7]. We have designed an algorithm, Ahab, which eliminates all parameters other than a final rank, and, following the logic of the mobydick algorithm [11], computes via maximum likelihood the probability that the window sequence is made up by sampling from the known weight matrices or background. Mobydick was designed to find overrepresented "words" in noncoding DNA, and applied to yeast. However, in the genomes of higher eukaryotes, binding sites are much fuzzier than in yeast and simple motif models are unlikely to yield meaningful results. Therefore, Ahab generalizes the motif model to positional weightmatrices. Furthermore, we introduce a *local* background model to cut down the number of false positives arising from local variations in sequence composition and degeneracy of motifs (for example motifs that contain a poly A string). Finally, the maximum likelihood fit has to be done separately for each window and thus hundreds of millions of times for the genomes of higher eukaryotes, thus requiring an efficient implementation of the numerical procedures. Ahab tallies all possible segmentations of the sequence into binding sites (also called parsings). Thus motif overlaps are allowed and weighted according to how well they explain the data, and multiple weak copies of a factor are all counted [5,12]. Our background model is a Markov model fit to all triples of bases in the window and accounts for local compositional variation which could otherwise push up the representation of any matrix that matched one of the high copy number base triples (eg hb matches poly A tracts).

We used Ahab with a 500 bp window to scan the Drosophila genome. As a representative/typical example we show the hairy locus, Fig. 2, a pair-rule gene expressed in 7 stripes. Five modules driving individual stripes and stripe pairs have been identified experimentally. Ahab predicts four of them (stripes 1, 5, 6, 7), only the stripe 3+4 element is not recovered. Note that two modules, stripes 1 and 7, were not used as training data. The support for each local maximum corresponds nicely to the experimentally estimated sizes of the modules. Ahab also reports scores for putative binding sites in each window as defined by (4). As an example, we show the even-skipped stripe 3+7 module, one of the best studied cases in the literature (Fig. 3). Most of the known sites are recovered and some new ones predicted.

Ahab finds 146 highly significant modules in the genome, most located in the non-transcribed regions. 27 modules are inside introns, only 6 overlap with exons. It recovers the 11 modules used to construct our weight matrices and predicts 6 other known modules with maternal/gap gene input. Thus, 17 out of 27 known modules (see additional File 1) are found. Three of the missing 10 modules (kni64, snail, sog) are very high in score (< rank 100) but lost when filtering for the presence of at least three different factors. Three modules are lost because they contain only

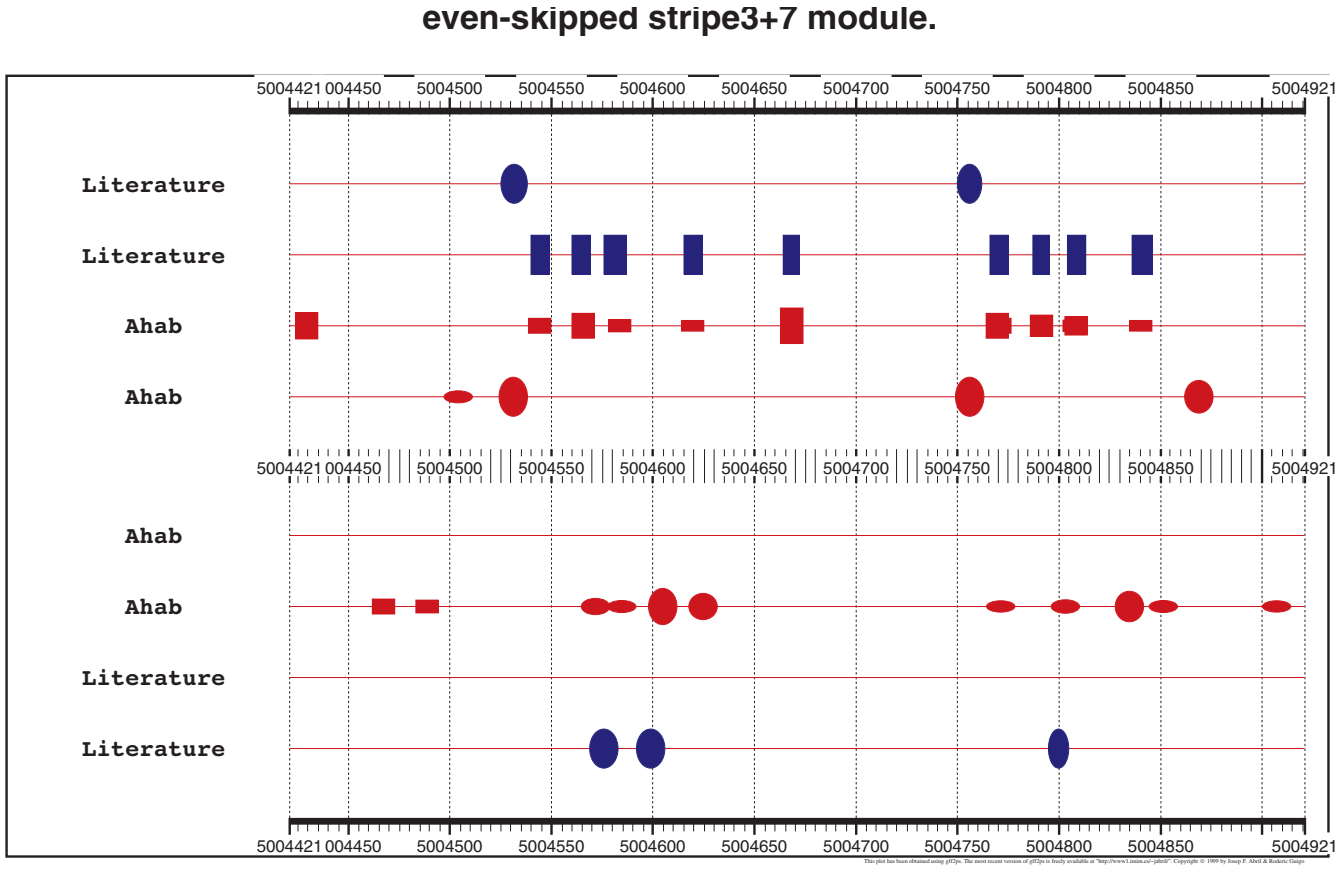

Figure 3
The even-skipped stripe 3+7 module Known binding sites (in blue) and sites predicted by Ahab (in red) for the even-skipped stripe 3+7 module. knirps sites are marked by circles, hunchback sites by boxes. The upper (lower) half depicts binding sites for the plus (minus) strand. The height of the red symbols corresponds to the score of the sites (Eq. 4).

dorsal sites (rho, twist, zen) and we do not have matrices for the other factors known to bind (such as snail and twist), however searching with only the dorsal weight matrix does recover them. The four remaining missing modules are low in score (rank > 700), but Kruppel CD2 is recovered for a window size of 700; the others are evidently low in maternal/gap gene binding sites.

To determine whether any of the 129 novel module predictions were correct, we asked whether any of the adjacent genes were patterned in the blastoderm. For 15 modules patterned expression has been reported for one of the adjacent genes; the patterns are either gap or pair rule like. An additional 11 modules are suggestive, since they are clustered with another, nonoverlapping prediction (Table 1). Interestingly, two predicted modules are in proximity to well studied segmentation genes (giant, runt) but still outside known regulatory regions.

We tested the stability of Ahab against unspecific input weight matrices by eliminating the least specific matrix in our list (Tailless). Tailless has ~600 predicted sites in the 146 modules (Table 2), twice as much as any other factor. However, we found that roughly 75 % of the predictions without using tailless (but with the same significance cutoff) were also present in the list of 146. Thus, although Tailless is rather unspecific, it makes a contribution to the predictions without dominating them.

We also varied the window size to 700 bp and masked out the repeats genome wide, before running Ahab, (see web site). Our top 150 predictions now included 11 out of the 27 known modules (see additional File 1 and 3). Lowering the cutoff value in the module score did not help, only one additional known module was recovered when including ranks 151–250. Overall, 84 modules or 58 % of the window 500 dataset are also among the top 200 of the window 700 set, and thus might be accorded more significance. Although some known modules disappeared, some very interesting new modules are predicted, in window 700 runs, including modules for key genes in the segmentation process such as caudal, grainyhead, odd

Table 1: The 32 modules and nearby genes which are known to be patterned in the early blastoderm (upper block) and 11 additional modules (lower block) for which a pair is linked to the same gene.

Rank	Score	Gene	Location
1 *	37.82	hairy (stripe 6)	up/9.2 kb
2 *	28.29	knirps	up/1.6 kb
3 *	27.24	tailless	up/2.6 kb
5 *	25.32	knirps	up/1.1 kb
8	20.80	runt (stripe 7)	up/3.2 kb
9	19.89	optix = six3	down/11 kb
10	19.75	Dichaete	down/2.3 kb
17	18.88	Tenascin-m	up/110 kb
18	18. 87	giant	down/14.5 kb
20 *	18.76	Kruppel	up/4.1 kb
23	18.30	ken	intra
24	18.29	giant (posterior)	up/2.1 kb
25	18.29	hairy (stripe 1)	up/4.7 kb
27	18.06	hairy (stripe 1 or 5)	up/5.4 kb
34	17.36	hairy (stripe 7)	up/10.4 kb
36 *	17.25	even skipped (stripe 3+7)	up/3.5 kb
37	17.15	knirps-like	intra
41 *	17.02	hairy (stripe 5)	up/6.2 kb
43	16.94	brinker	up/10.9 kb
45	16.83	pipsqueak	intra
46	16.82	teashirt	intra
48 *	16.80	short gastrulation	intra
51	16.76	abd-A	up/17 kb
54	16.65	abd-B	up/15.3 kb
61	16.20	vnd	intra
75	15.90	cap n' collar	up/4.3 kb
76 *	15.89	even skipped (stripe 2)	up/1.7 kb
91	15.69	runt (stripe 3)	up/9.67
120 *	15.34	tailless (proximal torso)	up/0.64 kb
124	15.32	proboscopedia	intra
126	15.29	runt	up/17.2 kb
129 *	15.24	hunchback (central stripe)	up/3.34 kb
6	23.97	Cyp6V1	down/1.6 kb
11	19.66	CG13595?	down/4.5 kb
14	19.32	Cyp6V1	up/6 kb
55	16.62	echinoid	up/58 kb
58	16.34	CG2118/Acf1/faf	intra
69	16.01	faf/Acf1/CG2118	intra
93	15.65	echinoid	intra
105	15.51	bruno 3	up/95 kb
117	15.37	CG5060	up/31.1 kb
130	15.22	bruno 3	up/25.1 kb
132	15.18	CG5060	up/30.1 kb

Modules which were used to construct weight matrices are marked with stars. The columns give the rank of each module, the score, the gene, information about the location of the module in respect to the gene (up/downstream or intragenic). For References and additional material see additional File 3.

skipped (odd), and sloppy paired 1,2 (slp1,2) (see additional File 3). Thus Ahab could be improved by allowing for variable window lengths.

We estimated the false positive rate by scrambling the columns (positions) in the input frequency matrices. The new matrices are thus unlikely to be functional, but retain

Table 2: Statistics of factor binding sites for the set of 146 modules predicted by Ahab.

	bcd	cd	dl	hb	kni	Kr	tll	torRE
sites	179	99	143	213	302	203	597	24
specific	3.3	3.3	3.6	3.4	2.6	3.6	2.8	4.3
modules	65	35	62	72	84	76	119	4

The specificity is defined as a standard error, eg 3 (4) implies a 0.14% (0.003%) probability of getting a match as good as the median data match from random sequence.

the same specificity. Ahab found roughly half as many "modules" for the same score cutoff as for the original set ie a 50% false positive rate. Note that this is a conservative estimate because part of the consensus motif recognized by hunchback and caudal is largely a poly T motif, and thus preserved by scrambling. Moreover, only very few known patterned gene are predicted by the scrambled matrices.

Another perhaps more straightforward estimate of statistical significance and our true positive rate is to use the experimental result (see footnote in [6]) that less than 2% of the genome or ~300 genes are patterned during the blastoderm. Since we can not tell which of two neighboring genes is regulated by each of the 102 intergenic modules we predict, we are obliged to label 237 genes (adding the 33 genes with intragenic modules) as potentially patterned. For 237 random predictions one expects 2% or five genes to be patterned, and the probability to get 21 or more genes (see additional File 3) by chance (p-value) is $\sim 10^{-10}$. It should be stressed that the true success rate of Ahab will be much higher since the number of genes for which blastoderm expression is demonstrated is (< 100) or 1/3 of the total. Thus we expect an additional 50 genes in our set to be patterned, so a 50% overall positive rate for our module predictions.

The Gibbs sampler finds binding sites within experimentally characterized modules

Binding site information for most of the transcription factors relevant to any developmental process is only rarely available. More common are modules obtained by 'promoter bashing' from several genes with similar expression. Thus it is natural to ask, in view of the site repetition within modules, whether standard motif finders are able to recover good weight matrices from modules and if so can these be used as input to Ahab to find genes with similar regulatory inputs.

We have tailored the Gibbs algorithm (see methods) to this problem by searching for only one site at a time, and then masking only the central 1–2 bases of each sequence motif found before iterating. The results were thereby much more reproducible between runs. Most importantly, motifs were allowed to overlap, a very common occurrance in modules and arguably important for their function [3].

To gain confidence in the capabilities of the Gibbs algorithm, we prepared two synthetic data sets representative of the data we wanted to examine (eg several kb long, 30–50% of the sequence covered by motifs, the remaining sequence random and 60% A/T). Data set 1 was made by equally sampling our Hb, Cad, Kr and Tll matrices. Data set 2 was generated from four synthetic frequency matrices of specificity equal to the known ones. The customized Gibbs virtually perfectly recovers all the synthetic weight matrices from dataset 2. By contrast, only half of the natural sites were recovered from dataset 1. The sites overlapped and their delineation was imperfect. Thus Gibbs detects sequence correlations among the factors we chose, but probably exaggerates them, because it computes significance based on single base frequencies.

We then ran the Gibbs algorithm on several modules with extensive binding site data, Table 3. In accord with experiment, Gibbs predicts that about 30–50% of the sequence is covered by motifs. The specificity of the Gibbs motifs is typically higher than for the experimental ones, presumably because a smaller number of sites is sampled. Even when the majority of the sequences composing the Gibbs motif match a single factor, there are a few other strong factor matches generally in different positions and perhaps the other orientation. Nevertheless the Gibbs motifs are largely reproducible between runs. Generally, we recover about half the factors known to influence the module, and interestingly predict several new motifs. The lack of a 1:1 correspondence between the experimental motifs (generally composed from a wider range of data then presented to Gibbs) and those we find, points to a real ambiguity, we believe, in how to parse a sequence into binding sites.

Table 3: Motifs derived by Gibbs sampling the indicated modules (see additional File **1**)

Module	recovered factors (copies)	novel motifs
eve stripes 2, 3+7	kni(15), bcd(9), hb/cad(9)	
eve stripes 5, 4+6	hb/cad(6),	RTTNSRCGSAAT(9),
h stripes 5,6,7	kni/hb/cad(12)	ATYCYGCARY/ *bcd (6)*
	Kr(12), hb/cad(14) *kni (7)*	GRCNWG[T/G]TSNSA (9)
hb (both mods)	*Kr/tll(6)*, hb/cad(8)	ATTTTCCNSC (9)
kni (1.1 k)	Kr(7), tll(5), hb/cad(9)	GWGWG [A/C] GWGYG(7)
Kr (700 bp)	bcd(5), hb/cad(7)	TWNTGATCCWS (6)
tll (3 mods)	kni(9), cad(9), *Kr(7)*, hb/cad(8)	TCRAWSAAT/*torRE (8)*

The criterion for a motif to match a known factor is given in Methods; copies refers to all sequences in the Gibbs derived motif. Only the consensus sequences of the most prominent unclassified motifs are shown with the abbreviations (R = A/G, W = A/T, S = G/C). Matches in italics are marginal and names linked by / co-occur within the same Gibbs motif, possibly on opposite strands.

Using only three modules as input, Ahab finds 63 significant modules in the genome

Next we tested whether Gibbs derived matrices can be fed to Ahab for a genome-wide search for modules. As samples, we used three known modules that drive expression of the pair rule gene hairy in stripes 5, 6, and 7, which are known to receive input from Bcd, Cad, Hb, Kni, Kr, and Tll (see additional File 1).

Customized Gibbs finds 6 highly significant weight matrices within the 2 kb composite module (see additional File 4). One matches the Kr weight matrix with high quality, another Kni, and a third represents a mixture of Hb and Cad, whose matrices are indeed quite similar due to a poly T motif. The other three motifs are new. Using these 6 weight matrices as input, Ahab finds 63 highly significant modules genome-wide (Table 4, additional File 5).

The top four modules overlap with those used in the Gibbs sampling. In addition, 13 new modules were contiguous to genes that are known to be patterned in the blastoderm, and two fall close to a single gene of unknown function. One of the top scoring modules is the hunchback central stripe module, other particularly interesting hits are in the intron of knirps and 18 kb upstream of hairy outside the known regulatory region, and proximal to emc, a known transcriptional co-repressor. Compared to the Ahab predictions, we find more modules near homeotic genes (Abd-A, abd-B, Ubx, hth), due to the presence of the novel Gibbs motifs. The statistical significance of our predictions and the inferred true positive rate are comparable to the segmentation gene results.

Argos: prediction of regulatory modules from raw genomic sequence

As a final generalization, we ask whether there is enough repetition of sequence motifs within a module for its discovery using no information other than the noncoding sequence in the genome. To determine whether a given motif is locally overrepresented, its frequency of occurrence has to be scored against some statistical model. Both Ahab and Gibbs, use counts of short strings or single bases, within the window of interest to compute the significance of longer motifs. We attempted running Gibbs on successive windows of sequence and scoring the resulting motifs with Ahab, but were not able to discriminate the known modules from the remainder of the upstream region.

We therefore devised an alternative strategy that uses the information available in all the noncoding sequence and thus extrinsic to the window of interest. We enumerate all motifs in a class (a consensus of length 8 and 2 mutations worked best), and use their frequency to assign a probability for observing an overrepresentation of any one of them in the window of interest. The actual binding motifs may be longer, we only need to capture the most significant region. Typically several hundred motifs are significant for each window. They heavily overlap so the individual motif scores can not be simply added for an overall score. Although we tried using Ahab to eliminate redundant motifs, we got better results with a greedy algorithm that looks only at the motifs without placing them on the sequence. The greedy algorithm winnows the list of motifs down by starting with the highest scoring one and eliminates any motif related to it under shifts and a limited number of mutations. The next remaining motif is retained and overlaps with it are eliminated, until ~5 quasi independent motifs are obtained whose scores can be added (details in additional File 7).

We evaluated the results of Argos for the the modules in additional File 1. Log probability scores > 70 were taken as significant, since they are found very rarely using rand-

Table 4: Modules predicted by Ahab from the hairy derived Gibbs motifs that are patterned in the early blastoderm.

Rank	Score	Gene	Location
1 *	42.16	hairy (stripe 7 element)	up/10.9 kb
2 *	36.43	hairy (stripe 6 element)	up/9.2 kb
3 *	31.43	hairy (stripe 5 element)	up/6.2 kb
4 *	25.08	hairy (stripe 7 element)	up/10.4 kb
5	22.58	hunchback (hb central stripe)	up/3.3 kb
7	21.91	abd-A (in iab-7 region)	up/83.2 kb
8	20.02	homothorax	intra
9	19.41	bxd (in bxd reg region)	up/18.8 kb
10	18.55	frizzled 2	up/37.9 kb
16	17.76	hairy (not in known modules)	up/18.9 kb
17	17.73	abd-A	up/0.2 kb
21	17.43	abd-A	up/35.8 kb
24	16.69	fd64A	up/1.4 kb
25	16.55	nubbin	up/2.6 kb
31	16.10	extra macrochaetae	down/2.2 kb
37	15.90	knirps	intra
46	15.90	Btk29A	intra
11	18.55	CG6559	up/30 kb
39	15.72	CG6559	up/45 kb

The last two modules are distinguished by proximity to a common gene. Modules which were used to construct weight matrices are marked with stars. The format follows Table 1. For references and additional material see additional File 5.

omized noncoding sequence [1]. With this threshold and at least l00 bp overlap, half of the modules were recovered, indicating a 50% false negative rate. The number of false positives is difficult to assess, because the code looks for *any* regulatory module. However for several well studied segmentation genes (specifically even-skipped, giant, hairy, hunchback, knirps, Kruppel, and tailless), with 15 known modules, we squarely hit 7 when looking over the entire 10 kb upstream of translation start (gt, kni, Kr-730, Kr-CD2, eve3-7, eve autoregulatory element, h-7), Fig. 4, but only three predictions are outside of known modules (one for tailless and two for hunchback). This indicates a low false positive rate. Genome wide we are predicting about one module per 5 kb of noncoding sequence averaged over the genome (with a strong bias for noncoding vs coding), which corresponds to roughly one module per gene.

Discussion

We have demonstrated algorithms that exploit three very different levels of prior information and lead to statistically highly significant predictions for early developmental modules in the fly. The Ahab algorithm is perhaps closest to the 'calculation' actually performed by the cell. The weight matrix match is a surrogate for the energetic preference of a transcription factor for a particular sequence, and Ahab models the competition of several factors and their binding energies for a stretch of DNA (a module). Ahab ignores distances between binding sites and the actual factor concentrations. Thus, the success of Ahab suggests that just modeling the binding energies is already predictive. It will be interesting to see how well Ahab performs in situations where the concerted binding of cofactors constrains the spacing of binding sites [13,14].

Finding overrepresented weight matrices is a well studied problem for which Gibbs sampling constitutes a reasonable solution if the data consists of distinct motifs separated by random bases. The difficulty we have encountered with this algorithm in dissecting regulatory modules for binding sites is not rare or diffuse motifs but rather too much signal, namely the overlay of motifs of different sizes and specificities. The Gibbs statistical model is not strictly correct for our data. A more adequate algorithm would allow competition among motifs of different lengths [15]. Irrespective of technical problems, the discovery of binding motifs by site repetition is qualitatively a more difficult task than their recognition by transcription factors [16]. Thus our ability to recover plausible motifs for about half the known factors was not obvious in advance and is another manifestation of the redundancy in module design.

Reference [6] describes another approach to locating modules from clusters of known weight matrices. They count the number of matches of each weight matrix in an interval with a score above some empirically defined cutoff, and then score a 700 bp window as significant when the total number of matches for all factors is large enough. Information about the background is implicitly encoded in their choice of threshold. We do not have factor specific cutoffs, and use a locally defined background model, which renders our algorithm more automatic and less sensitive to local variation in sequence composition, eg poly A runs.

Although we are predicting many more modules than in [6], the positive hit rates are comparable between the two methods (50% vs 10 positives out of 28 predictions [6]). A more detailed comparison of both data sets reveals, however, that the 28 modules predicted by [6], with the exception of the giant one, do not overlap with any of the top 137 modules predicted by Ahab, although there are 4 genes in common to our sets. More strikingly, the 10 modules for which experimental results in [6] suggest functionality based on blastoderm expression of a neighboring gene fall below 500 in our ranking with exception of giant and one of the hairy derived modules for nub, Table 4. Presumably due to the difference in background

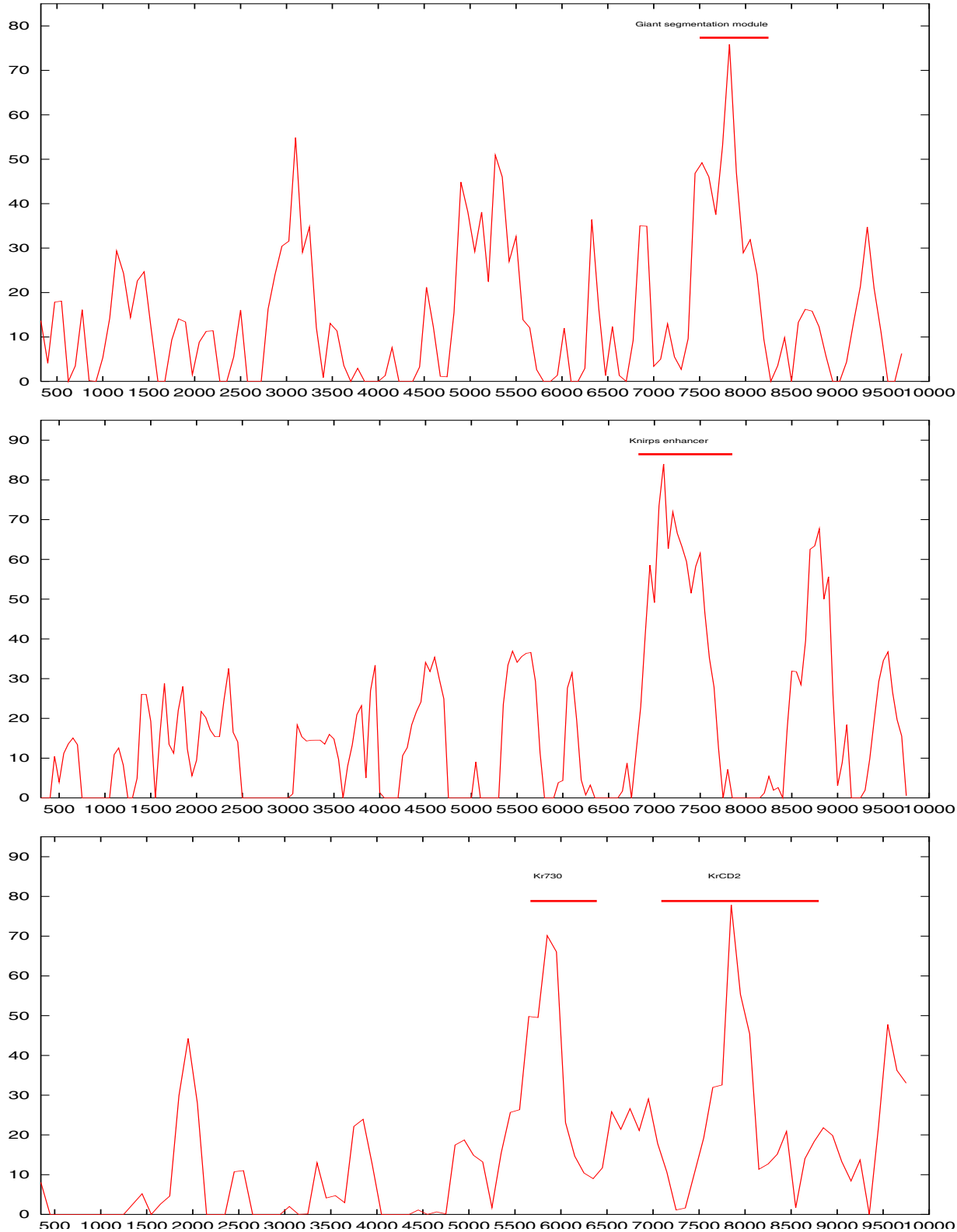

Figure 4
Argos score for the upstream regions of giant, knirps and Kruppel Argos score to observe a 500 bp module upstream of giant, knirps and Kruppel. The bars mark known modules and translation start is at the right most base.

model, their modules are dominated by Hb sites, while ours are not, which contributes considerably to the divergence of the predictions. Clearly, only direct experimental validation of predicted modules through reporter gene fusions will help to compare the different methods. In this fashion, we plan to test a number of the new modules predicted for key genes in the segmentation system such as h, run, gt, odd, prd, slpl/2 and cad.

In order to understand regulatory networks of genes, it is useful to generalize from a few genes or modules with common functions to new candidates. When control is combinatoric, a purely experimental approach tends to be more tedious than screening a modest list of candidates. Thus a potentially important aspect of our work is the combination of motif discovery from modules via Gibbs sampling and generalization to the entire genome with Ahab. We have demonstrated the feasibility of this procedure when we worked from the hairy stripe 5–7 modules. Interestingly, the candidate list of similar modules genome wide was quite small, but had little overlap with the top scoring modules predicted from the full set of gap gene weight matrices. Hopefully some of the new motifs discovered by Gibbs sampling are real; perhaps they are binding sites for corepressors. Clearly the first step is to confirm a striped expression for some of the genes in Table 4.

Our algorithm Argos for predicting enhancers from raw genomic sequence works astonishingly well. It will be most interesting to use this approach together in conjunction with the customized Gibbs sampler and Ahab in situations where nothing is known experimentally about the transcriptional regulation of genes of interest to identify co-regulated genes. Namely, following the hierarchical structure in Figure 1, Argos could be used to predict modules, then the customized Gibbs sampler to predict binding sites (weight matrices) and finally Ahab to predict, genome wide, genes in the same regulatory network.

Several recent papers [6,7,9] as well as ours have taken only the very first steps in applying computational approaches to the the elucidation of *cis*-regulatory modules. For body patterning in the fly, it is very encouraging that such limited information as we have used works so well. It remains to be seen if the same approaches work on systems where a single master regulatory gene initiates a developmental cascade, or where integration of developmental cues occurs partly at the level of signal transduction.

Conclusions

Predicting and understanding transcriptional regulation is a fundamental problem in biology. We have designed new algorithms for the detection of *cis*-regulatory modules in the genomes of higher eukaryotes which is a first step in unraveling transcriptional regulatory networks. We have demonstrated, in the case of body patterning in the *Drosophila* embryo, that our algorithms allow the genome-wide identification of regulatory modules when the motifs for the transcription factors are known (algorithm Ahab), or when only related modules are known (customized Gibbs sampler in conjunction with Ahab), or when only genomic sequence is analyzed with Argos. We believe that Ahab overcomes many problems of recent studies and we estimated the false positive rate to be about 50%. Argos is the first successful attempt to predict regulatory modules using only the genome without training data. All our results and module predictions across the *Drosophila* genome are available at [http://uqbar.rockefeller.edu/~siggia/]. The Ahab code is available upon request from the authors.

Methods

Genomic data

We downloaded the Release 2 genomic sequence and annotation for *Drosophila melanogaster* from "Gadfly" (Genome Annotation Database of *Drosophila,* [http://www.fruitfly.org/] (Oct 2000)). Using a map provided by Chris Mungall (private communication) we mapped the annotation, which was done on separate contigs, to chromosomal coordinates. "Flybase" [http://flybase.bio.Indiana.edu/] provides a curated assembly of genetic and molecular data from the existing literature. Our web sites links this database to the Gadfly annotation using a map provided by David Emmert (private communication). Since our algorithms are based on searching for clusters of common sites, microsatellites can score high, but tend not to be functional (for an exception see [9]). We used the Tandem Repeat Remover [17] to mask microsatellites, with scoring parameters (2 5 5 75 20 20 500) (respectively; match, mismatch, indel scores; percentage priors for mismatch and indels; minimum score, and maximum length) which are as promiscuous as possible yet did not detect appreciable microsatellites in random sequences. With these settings, ~5.7% of all non coding sequences are masked.

We collected from the experimental literature modules that drive blastoderm specific expression of a reporter gene in response to several of the factors in our list. In many cases the module was shown to be the minimal element. The modules mapped to chromosomal coordinates are reported in the additional File 1.

We collected a total of 199 experimentally characterized sites for the factors bicoid (30 sites), caudal (21), dorsal (32), hunchback (43), knirps (27), kruppel (20), tailless (20), and torRE (6). Giant sites are too few in number and ill defined to be useable. Each binding site was mapped to

the genome and padded with six bases on both sides. The multiple alignment program WCONSENSUS [18] (options, -f -d -sl -a and background frequencies representative of noncoding sequence (60% A/T)) was used to align and orient the sites and create a weight matrix for each factor. Results and references are in the additonal File 2.

Algorithm Ahab: fitting multiple weightmatrices to sequence

Ahab computes an optimal probabilistic segmentation of a sequence S into binding sites and background for a fixed set W of sequence motifs modelled by weight matrices. Ahab is related to the mobydick algorithm [11], but has several novel key features described in detail in the results section and the additional File 6. It fits the probabilities p_w of the (fixed) matrices w and background p_B, so as to maximize the likelihood of generating S under a certain explicit model. Namely, select a weight matrix or background according to its probability p_w, p_B, sample according to the predetermined frequencies, and add the resulting bases to the sequence under construction. The fit of the model to S is accomplished by defining the probability of a particular segmentation T of S as

$$P(T) = \prod_{k=1}^{N(T)} p_{w_k} m(s \mid w_k), \qquad (1)$$

where $k = 1, 2, ..., N(T)$ labels the weight matrices (or background) which were used in segmentation T. The quality of the match between the weight matrix w_k and the subsequence $s = (n_1,...n_l)$ is incorporated in $m(s|w_k) = \prod_{j=1}^{l} f_j(n \mid w_k)$, where $f_j(n|w_k)$ are the normalized frequencies of nucleotide n at position j for weight matrix w_k.

An important consequence of equation (1) is that multiple binding sites with weak matches to the weight matrix for the same factor (p_{w_k} large, $m(s|w_k)$ small) may make an important contribution to $P(T)$. In many cases these redundant sites with low weight matrix scores have been observed in experimentally known modules. Any algorithm that would just count matches of sequences to a weight matrix above a certain threshold would have to use ad-hoc measures to incorporate these sites into the score. Note also that the weight matrix only captures the sequence dependent part of the binding energy, so 'weak' binding could equally well be termed 'nonspecific'. We know too little about the physical binding energies of transcription factors, and their cofactors and protein concentrations in vivo, to calculate whether any modules are actually occupied by factors.

Ahab uses a *local* Markov model of order q for background sequence, that is, a single base n at site j is segmented with probability $p_B f_j(n|B)$, where $\sum_n f_j(n|B) = 1$, and $f_j(n|B)$ is contingent on the q preceding bases following the usual Markov model definitions. The f_j are computed by enumerating all $q + 1$ tuples of bases in S, which has the effect of suppressing the number of copies of any w which match frequent triples of bases, eg poly A tracts (we typically use $q = 2$).

The likelihood Z to observe S is then

$$Z(S) = \sum_T P(T). \qquad (2)$$

Dynamic programming allows the calculation of $Z(S)$ in a time proportional to the sequence length and the number of weight matrices. The maximization of $Z(S)$ then determines $p_{w,B}$ (see additional File 6). The likelihood Z_B that S comes from background only (ie $p_B == 1$) is trivially computed from the Markov model. The score R that S is a regulatory module is then in log-odds units

$$R = \log\left(\frac{Z_{\max}}{Z_B}\right). \qquad (3)$$

At each position $i = 1, ..., L$ in S the probability $P_i(w|S)$ to observe the start of weight matrix w of length l_w is computed by standard posterior decoding. Let $Z(i, j)$ denote the likelihood for the sequence S from position i up to j. (The symbol $Z(S)$ used above, is just $Z(1,L)$.)

$$P_i(w \mid S) = \frac{Z(1, i-1)\, p_w m(s \mid w)\, Z(i + l_w, L)}{Z(1, L)}, \qquad (4)$$

using the optimized p_w's. The sum of equation (4) over all positions is the average copy number of w in the data. Summing over all segmentations (1) naturally allows for overlapping sites and the 'profile' (4) quantifies the competition between different factors for the same bases.

It takes a modern LINUX workstation about 2 days to run Ahab over the entire genome with a 500 bp window moved in 20 bp steps, fitting the gap gene weight matrices we collected. We enumerated all local maxima in the score R larger than 15 and eliminated those within 500 bp of a higher scoring peak and obtained 216 disjoint regions. If we insist that at least 3 different factors contribute to the module with individual average copy numbers of at least 1 the number of modules reduces to 169, and eliminating all candidates with 80 or more basepairs masked by the Tandem Repeat Remover gave a final list of 146 modules.

Predictions with more than 200 bases overlap with a known module or with an overlap of at least 50% of the length of the known module were considered to be a recovered known module.

Determination of motifs from modules

We ran locally the Gibbs sampler algorithm provided by C. Lawrence's group at [www.wadsworth.org/resnres/bioinfo/], as described in Results. We generally used a motif length of 11 bp and allowed the algorithm to vary this by +-2 if the data warranted, and took a prior copy number of 7 when fitting 1–2 kb of data (again the algorithm will adjust this number). Other parameters were taken as default, and under these conditions typically 5 distinct motifs were fit.

To decide whether Gibbs derived motifs matched known ones we ran the known weight matrices in both orientations over the individual sequences composing the motif plus five flanking bases and computed the position that maximized the information score. The Gibbs motif was deemed to match a known factor if:

1) a single known matrix was the top match to a majority of the sequences,

2) the optimal match occurred at the same position (with some variability allowed for factors such as hb, with poly-A regions),

3) the information score of the match was comparable to the score of the sequences which define the matrix to the matrix itself and not dominated by the flanking bases.

4) the preceding conditions were met in two independent runs (with typically 2–4 runs done for each data set).

Algorithm Argos

Argos is described in detail in the results section and additional File 7.

Genome-wide display of our results

Our webpage [http://uqbar.rockefeller.edu/~siggia/] contains all our predictions. For each module, all nearby genes were extracted from the annotation, their position relative to the module (ie up/down stream, intronic), and Flybase links for gene function were collected into a table. The number of binding sites for each factor in the module is listed and their position and score along with known binding sites can be viewed in graphs which were produced with "gff2ps" [19].

To view our results interactively on a larger scale together with the current fly annotation we installed the Gbrowse software from L. Stein's Generic Model Organism Systems Database Project [http://www.gmod.org]. Our modules (predicted and experimental), binding sites, and restriction sites are included in the display. A function was added that allows to plot the Ahab score (equation 3) along the genome. Thus, the user can explore where additional putative modules fall relative to any gene of interest.

Contributions

Authors 1, 2 and 4 carried out the computional part of this study, author 3 annotated the Ahab results. All authors read and approved the final manuscript.

Note

[1]Randomized sequence was produced by randomly pooling and concatenating 100 basepair chunks from genomic noncoding sequence

Additional material

Additional file 1

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-3-30-S1.pdf]

Additional file 2

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-3-30-S2.pdf]

Additional file 3

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-3-30-S3.pdf]

Additional file 4

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-3-30-S4.pdf]

Additional file 5

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-3-30-S5.pdf]

Additional file 6

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-3-30-S6.pdf]

Additional file 7

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-3-30-S7.pdf]

Acknowledgements

We thank Nicholas D. Socci for customizing the Generic Genome Browser such that it allows the genome wide display of the Ahab score (see Fig. 2). Ulrich Unnerstall helped typesetting the supplementary material. Chris Mungall provided help with the fly annotation. The Ahab code is available from the authors. Support was provided by the NSF grant DMR 0129848, and the NIH grant GM-66434.

References

1. Rubin GM, Yandell MD, *et al*: **Comparative Genomics of the Eukaryotes.** *Science* 2000, **287:**2204-15
2. Brivanlou AH, Darnell JE Jr: **Signal transduction and the control of gene expression.** *Science* 2002, **295:**813-8
3. Davidson EH: **Genomic regulatory systems.** *Academic Press, San Diego* 2001
4. Davidson EH, Rast JP, *et al*: **A genomic regulatory network for development.** *Science* 2002, **295:**1669-78
5. Arnone MI, Davidson EH: **The hardwiring of development: organization and function of genomic regulatory systems.** *Development* 1997, **124:**1851-64
6. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the Drosophila genome.** *Proc Natl Acad Sci USA* 2002, **99:**757-762
7. Markstein M, Markstein P, Markstein V, Levine MS: **Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo.** *Proc Natl Acad Sci USA* 2002, **99:**763-768
8. Halfon MS, Grad Y, Church GM, Michelson AM: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.** *Genome Res* 2002, **12:**1019-28
9. Papatsenko DA, Makeev VJ, Lifanov AP, Regnier M, Nazina AG, Desplans C: **Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers.** *Genome Res* 2002, **12:**470-81
10. Wieschaus E: **Embryonic transcription and the control of developmental pathways.** *Genetics* 1996, **142:**5-10
11. Bussemaker HJ, Li H, Siggia ED: **Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis.** *Proc Natl Acad Sci USA* 2000, **97:**10096-100
12. Jiang J, Levine M: **Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen.** *Cell* 1993, **72:**741-52
13. Ryoo HD, Marty T, Casares F, Affolter M, Mann RS: **Regulation of Hox target genes by a DNA bound Homothorax/Hox/Extradenticle complex.** *Development* 1999, **126:**5137-48
14. Courey AJ, Jia S: **Genes and Development Transcriptional repression: the long and the short of it.** *Genes and Development* 2001, **15:**2786-96
15. Zhang M, Siggia ED, Li H: **private communication.**
16. van Nimwegen E, Zavolan M, Rajewsky N, Siggia ED: **Probabilistic clustering of sequences: inferring new bacterial regulons by comparative genomics.** *Proc Natl Acad Sci USA* 2002, **99:**7323-7328
17. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-80
18. Hertz GZ, Hartzell III GW, Stormo GD: **Identification of consensus patterns in unaligned DNA sequences known to be functionally related.** *Comput Appl Biosci* 1990, **6:**81-92
19. Abril JF, Guigo R: **gff2ps: visualizing genomic annotations.** *Bioinformatics* 2000, **16:**743-744