CORECLUST: identification of the conserved CRM grammar together with prediction of gene regulation

A. A. Nikulova, A. V. Favorov, R. A. Sutormin, V. J. Makeev and A. A. Mironov

SUPPLEMENTARY METHODS

The hidden Markov model

HMM states

The overall hidden Markov model (HMM) architecture reflects our intuition of how regulatory modules are organized (Figure 1). The HMM contains three types of generative states that represent three general types of sequence: inter-module background sequence, sites and regions between sites in modules, i.e. spacers. Each type of transcription factor binding site (TFBS) is represented by the corresponding position probability matrix (PPM) that is known *a priori*. The number of SITE states is twice the number of PPMs used for prediction, for the two DNA strands. The number of SPACER states depends quadratically on the number of SITE states, as each pair of SITE states has corresponding SPACER states.

Each generative state emits a sequence of nucleotides of varying length. This type of the HMM architecture is known as the Generalized Hidden Markov Model (GHMM) (20) or 'HMM with duration' (21, 22). This model allows for easy use of any predefined length distribution for sequences generated from a state, and not only the geometric distribution.

The emission probability distribution for every generative state can be represented as follows:

$$P_{state}(sequence)=P_{state}(sequence|L)\,P_{state}(L) \quad ,$$

where $P_{state}(sequence)$ is the probability to emit a nucleotide sequence from the state given the sequence length L , and $P_{state}(L)$ is the probability to emit any sequence of the length L from this state.

Background state. The BACKGROUND (named 'BKG' on Figure 1) state is modeled by the first order local Markov chain whose parameters are computed from the base composition within the sequence window (we use the window size 500 and recompute the transition matrix every 100 nucleotides). The length of sequences emitted in the BACKGROUND state represents our expectations about the distance between CRMs. The model assumes that it is geometrically distributed with the mean $1/p_{open}$ (p_{open} is the probability to open a module):

$$P_{BKG}(L)=(1-p_{open})^{L-1}\cdot p_{open}$$

Site states. Each SITE state (named 'S1', 'S2'... on Figure 1) emits a sequence of nucleotides according to the corresponding site model (PPM). The word length necessarily equals the PPM length:

$$P_{SITE}(L)=\begin{cases}1 if\ L=length(PPM)\\0 if\ L\neq length(PPM)\end{cases}$$

The nucleotides of the SITE state are generated independently from each other with the frequencies that are determined by the corresponding PPM:

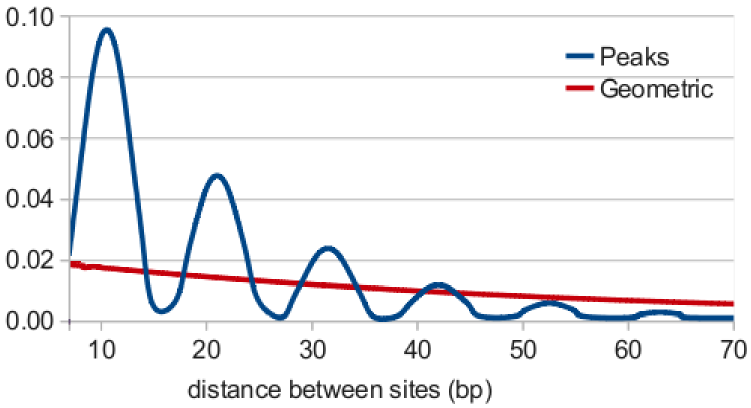
$$P_{SITE}(sequence|L)=\prod_{0\leq i<L} f_{PPM}(i,a_i) ,$$

where $f_{PPM}(i,a_i)$ is the frequency of nucleotide a_i in position i in the corresponding PPM.

Spacer states. In all SPACER states (named 'SPACER: D1' and 'SPACER: D2' on Figure 1), nucleotides are generated according to the same local Markov model as in the BACKGROUND state. The HMM has several types of the SPACER states differing by their sequence length distribution. These distributions determine the preferences in intersite distance for each motif pair (see later). Here, we use two types of the SPACER states with the following length distributions (Supplementary Figure S2): 1) the geometric distribution with the mean m , reflecting site clustering without any distance specificity, 2) the damped sinusoid with the period of 10.5 nucleotides that corresponds to the situation when interacting proteins bind to the same side of the DNA helix. The latter distribution is defined by the function:

$$f(i)=Z\cdot 2^{-round\left(\frac{i+0.5}{10.5}\right)}\cdot\left(\sin\left(\frac{2\cdot\pi}{10.5}\cdot\left(i+0.5+\frac{10.5}{4}\right)\right)+1\right) ,$$

where Z is a normalizing factor.



Supplementary Figure S2. Distributions of distances between adjacent sites in a module.

HMM transitions

In our model a CRM begins and ends with a site. So, the BACKGROUND state has transitions only to the SITE states. Each SITE state has only two transitions: back to the BACKGROUND state (which means that the module is closed) or to a silent state CLUSTER ELONGATION (named 'CE' on Figure 1), from which the BACKGROUND state can be reached only through one of the SITE states. Thus, the average number of sites in a CRM is regulated by the p_{close} parameter that is the probability of a transition from a SITE state to the BACKGROUND state. The value of p_{close} is the same for all SITE states.

The model assumes that regulatory modules may have preferred site arrangements. Firstly, sites of some types may prefer to be adjacent. This is modeled by introducing the NEXT SITE silent states (named 'NEXT: S1', 'NEXT: S2'... on Figure 1) after each CLUSTER ELONGATION state. These states determine the type of a site which follows a given site in a CRM. As all possible pairs of site types are represented, the number of the NEXT TYPE silent states next to every CLUSTER ELONGATION state equals to the number of the site types. The probability distribution of the type of the next site can vary with a given site type thus defining the sites order preferences in a module.

Secondly, we take into account the spacing between binding sites, which is controlled by the SPACER states. The NEXT SITE silent state has transitions to a set of SPACER states, differing in their distance distribution. The distance between adjacent sites of certain types is distributed according to a mixture of the SPACER states' length distributions. The weights of the distributions in the mixture differ between different motif pairs and are determined by the transition probabilities from the NEXT SITE state to the SPACER states.

The SPACER state is inevitably followed by the SITE state that has already been selected earlier (transition from the CLUSTER ELONGATION state to the defined NEXT SITE state).

Fixed parameters of the HMM

The model has fixed parameters, such as the threshold for the probability of the candidate sites (p), probabilities of CRM opening and closing (p_{open} and p_{close}), parameter for the geometric distribution of the distance between adjacent sites in CRMs (m), and the threshold for the CRM weight (w) (weight of a predicted CRM equals to the ratio of natural logarithms (base *e*) of two probabilities: the probability to obtain the nucleotide sub-sequence as generated by the CRM model and the probability to obtain it as generated by the background model). The values for these parameters were set intuitively and seemed to yield good results. For the *Drosophila* system m is set to 55, so that the average distance between adjacent sites in a module is 55 bp, p_{close} = 0.1, p_{open} = 0.001, p = 0.0045, and w = 100. For the muscle dataset, the same parameter values were used, with the exception of m set to 20, as the training sequences in this dataset contained short promoter sequences with highly clustered sites.

Time reducing

To reduce the run time, the paths in the HMM graph that have very low probability are removed. This is achieved by allowing every SITE state to begin only in those positions, in which the corresponding positional weight matrix (PWM) has a match with a relatively high score. Prior to parsing a sequence with the constructed HMM, it is searched for sites that are sufficiently strong to form a part of a regulatory module. We call them candidate sites.

In other words, the HMM does not search for the sites itself, and it just combines some of them to obtain the best CRMs. Moreover, sites are forbidden to be adjacent sites in one module if they are separated by more than 500 nucleotides.

Candidate sites searching

To search for candidate sites, all given PPMs are converted to PWMs by taking the logarithms of the PPMs' elements. Then, the distributions of the PWMs' scores on the background sequences are constructed by scanning all noncoding regions of the *D. melanogaster* genome. These distributions are used to set the PWM score cutoffs given the desired probability threshold.

Then, a candidate site is accepted in an input sequence if the probability to obtain a site with the same weight or better in a random sequence is lower than the set threshold p.

To select the probability threshold p, the following experiment was performed. We applied the algorithm to a set of AP patterning and several random genes (*btd*, *cnc*, *cad*, *gt*, *hb*, *kni*, *Kr*, *ems*, *eve*, *h*, *run*, *odd*, *ftz*, *slp1*, *slp2*, *Chd3*, *CG9855*, *CG9287*, *CG9065*, *resilin*, *rad201*, *Prosalpha6T*, *CG10345*, *capa*, *CaMKI*) with $p = 0.01$, $p = 0.02$ and $p = 0.03$. Then we derived the maximum site probability among all sites that formed the predicted modules. The maximum observed site probability was 0.0041. So, we used $p = 0.0045$. This value of p was used for all runs of CORECLUST in this study.

HMM parameters training

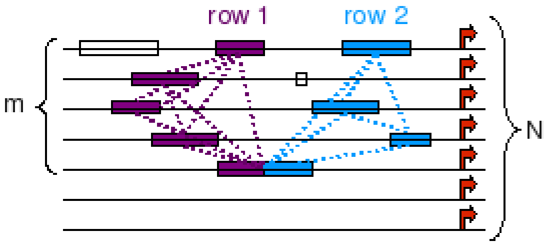
We use the Baum–Welch algorithm (26) to self-train the transition probabilities of the HMM, shown by dashed lines in Supplementary Figure S1. The prior distribution of the parameters is uniform. All probabilities are assigned according to the standard procedure (26), except the probabilities of the module-opening site types (BACKGROUND ➔ SITE transitions) that are estimated by the emission probabilities of the corresponding SITE states.

If some sequences in a training set are much closer to each other than the other ones, there is an option to decrease their total impact on the training by assigning scaling factors to the sequences according to their weights. For the *Drosophila* system, a phylogenetic tree taken from (39) is used to weight the sequences according to the Gerstein-Sonnhammer-Chothia algorithm (40).

Conservation score

To estimate the level of conservation of regulatory modules for a group of orthologous genes, the following measure was used.

The measure is based on site content of CRMs only and do not account for the order of binding sites. Only regulatory modules with weight higher than the predefined value w are considered. The module weights are not taken into account in any other way. For the sake of presentation clarity, we define the notion of a 'corresponding row' of regulatory modules. Consider a group of N orthologous genes. For m of them, regulatory modules are predicted (note that several regulatory modules may be found for one gene). Assume for the moment that we know which CRMs in different genes correspond to each other in the sense that they are similar to each other by their composition. We require that each CRM corresponds to at most one CRM in another organism. Thus, we can consider a corresponding row of CRMs represented in a subset of the given orthologous genes (Supplementary Figure S3).



Supplementary Figure S3. Corresponding rows of he predicted regulatory modules (shown by the rectangles) represented in a subset of the orthologous genes (horizontal lines). Red arrows denote gene starts. See the text for explanation.

The strength (conservation level) of a corresponding row of CRMs is computed as follows. First, we calculate the similarity score q_{ij} for every pair of regulatory modules (i and j) in the row (the module pairs are shown by dashed lines in Supplementary Figure S3). The pairwise similarity measure takes into account only site sets of the CRMs (that is, the number of sites of every type):

$$q_{ij}=n_{ij}\cdot\frac{|\cap_{ij}|}{|\cup_{ij}|} ,$$

where n_{ij} is the half-sum of the numbers of sites in the CRMs i and j , $|\cap_{ij}|$ is the size of the intersection and $|\cup_{ij}|$ is the size of the union of the site sets of the CRMs i and j . The strength of the corresponding row is calculated as the sum of q_{ij} along all pairs of CRMs in the row ($i<j$), normalized by the number of genes in the orthologous group (N). The conservation score for a given group of genes is the total strength of all corresponding rows of CRMs found for this gene group:

$$\mathit{Conservation\ score}=\sum_{rows}\frac{\sum_{i<j}q_{ij}}{N} .$$

In practice one does not know which CRMs form a corresponding row. So, for each CRM in each organism we construct its own corresponding row by selecting the most similar (in the above sense) regulatory module in every other organism. The final conservation score for a given gene group is calculated as the sum of strengths of all corresponding rows normalized by the number of orthologous genes that have predicted CRMs (m).

As an additional filter, before the conservation score is calculated, groups of orthologous genes that contain less than three genes with predicted CRMs (i.e. with $m<3$), as well as orthologous groups with less than half of genes with predicted CRMs (i.e. $m/N<0.5$), are removed from consideration.

Pseudocounts for PPMs

For every PPM, pseudocounts proportional to the square root of the number of sites, used for the matrix construction, are added (41):

$$p_{\alpha,i}=\frac{c_{\alpha,i}+0.5\cdot\sqrt{N}\cdot f_{bkg}(\alpha)}{N+0.5\cdot\sqrt{N}} ,$$

where $c_{\alpha,i}$ is the observed count of nucleotide α in position i , $f_{bkg}(\alpha)$ is the frequency of nucleotide α in the background distribution, N is the number of sites, and $p_{\alpha,i}$ is an element of the PPM.

SUPPLEMENTARY DATA

Supplementary Tables

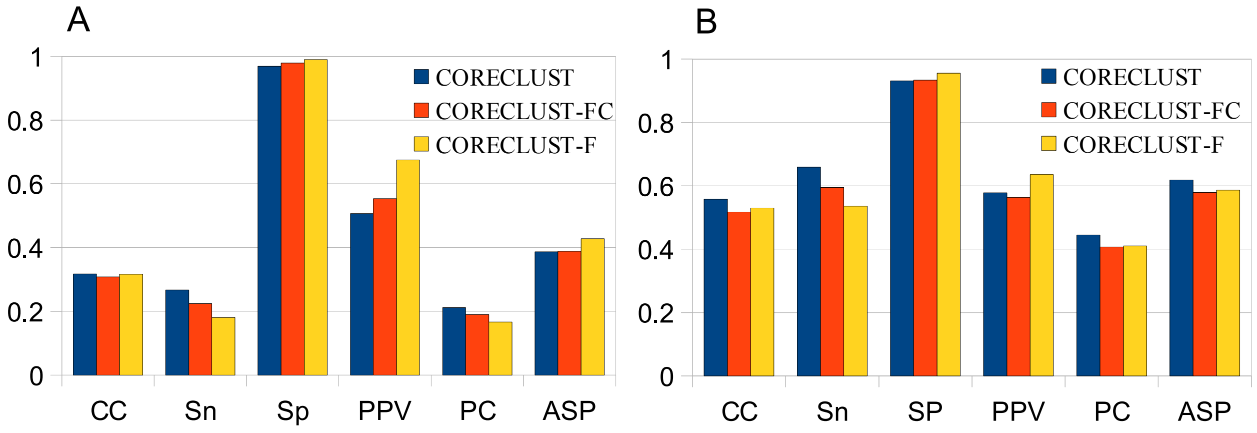
Supplementary Table S1. Genome-wide prediction of co-regulated genes for different training genes using CORECLUST and Cluster-Buster (7). *Kcor*, *Kcm*, and *Kcs* are the sizes of the intersections between the positive set and the gene lists predicted by CORECLUST and Cluster-Buster (sorted by either the maximum or the sum of the module weights) respectively; *Pcor*, *Pcm* and *Pcs* are the hypergeometric p-values of enrichment between the predicted genes and the positive gene set; *m* is the number of genes in the test list. The best p-value in each line is set in bold.

Training gene	m	CORECLUST		Cluster-Buster(max)		Cluster-Buster(sum)	
		Kcor	Pcor	Kcm	Pcm	Kcs	Pcs
h	45	16	**5.66E-22**	10	8.04E-12	8	7.18E-09
eve	40	15	**4.47E-21**	8	2.66E-09	6	1.70E-06
pdm2	43	15	**1.64E-20**	9	1.67E-10	7	1.24E-07
ftz	28	13	**6.62E-20**	7	5.11E-09	6	1.83E-07
gt	30	13	**2.08E-19**	7	8.65E-09	6	2.84E-07
run	79	17	**4.38E-19**	12	7.24E-12	8	6.68E-07
prd	21	11	**9.19E-18**	7	5.29E-10	5	1.16E-06
cad	77	16	**9.38E-18**	12	5.27E-12	8	5.48E-07
slp2	23	11	**3.47E-17**	7	1.10E-09	5	1.89E-06
tll	34	12	**1.10E-16**	8	6.59E-10	6	6.24E-07
slp1	25	11	**1.13E-16**	7	2.12E-09	6	8.79E-08
btd	108	17	**1.21E-16**	13	1.96E-11	10	5.98E-08
salm	28	11	**5.30E-16**	7	5.11E-09	6	1.83E-07
bowl	145	18	**1.12E-15**	18	**1.12E-15**	11	1.01E-07
knrl	53	13	**1.23E-15**	10	4.62E-11	8	2.78E-08
fkh	32	11	**3.09E-15**	7	1.41E-08	6	4.27E-07
ems	19	9	**3.29E-14**	7	2.33E-10	4	2.49E-05
Dfd	21	9	**1.03E-13**	7	5.29E-10	5	1.16E-06
hb	21	9	**1.03E-13**	7	5.29E-10	5	1.16E-06
kni	31	10	**1.25E-13**	7	1.11E-08	6	3.49E-07
Kr	51	11	**9.83E-13**	10	3.07E-11	8	2.03E-08
en	14	7	**1.65E-11**	6	1.62E-09	4	6.66E-06

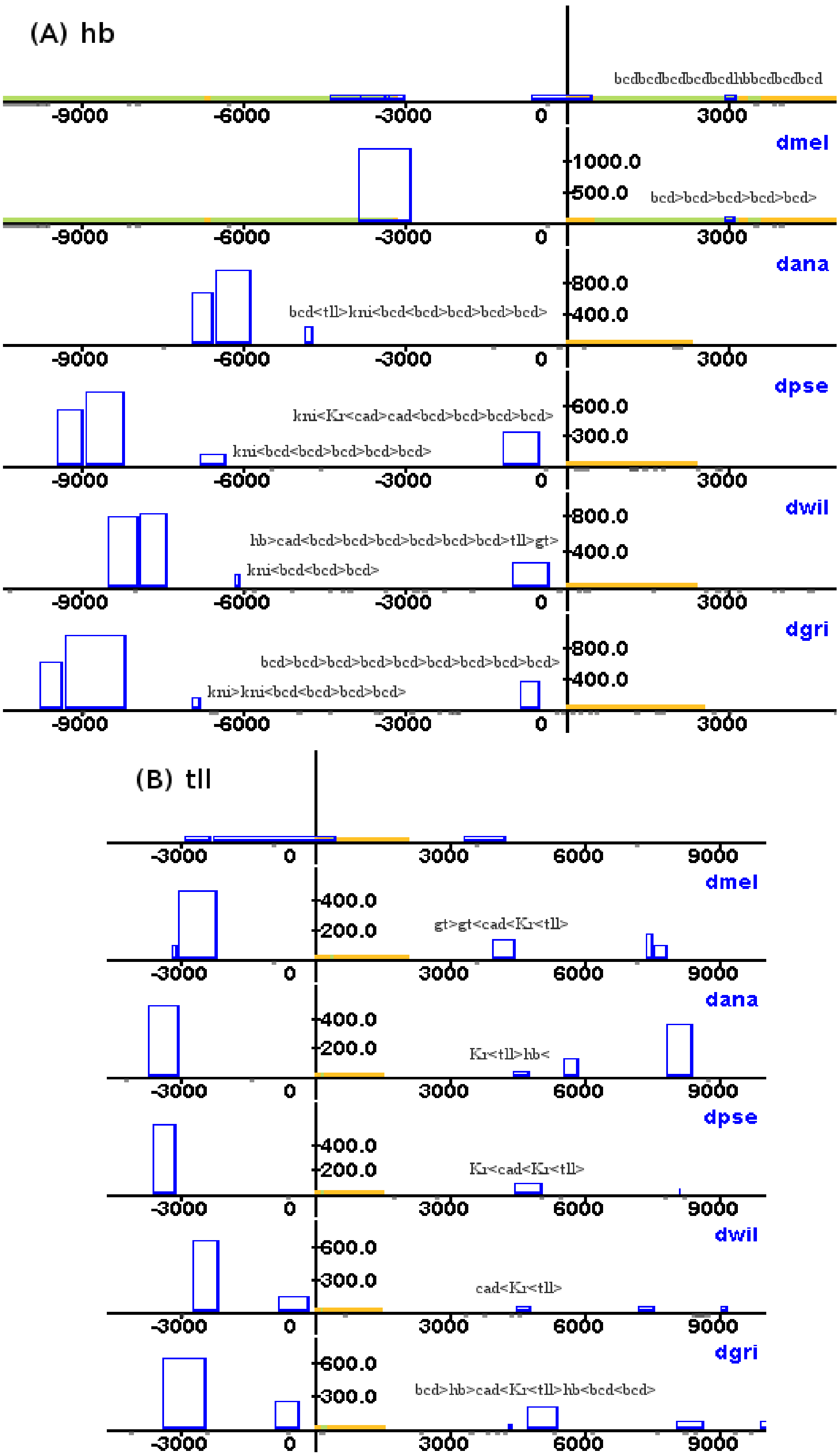
Supplementary Table S2. Correlation coefficient (CC) and positive predictive values (PPV, precision) for the predictions made by CORECLUST, CORECLUST-FC and CORECLUST-F. TOTAL row contains values calculated for the whole gene set. The maximum value in each line is set in bold.

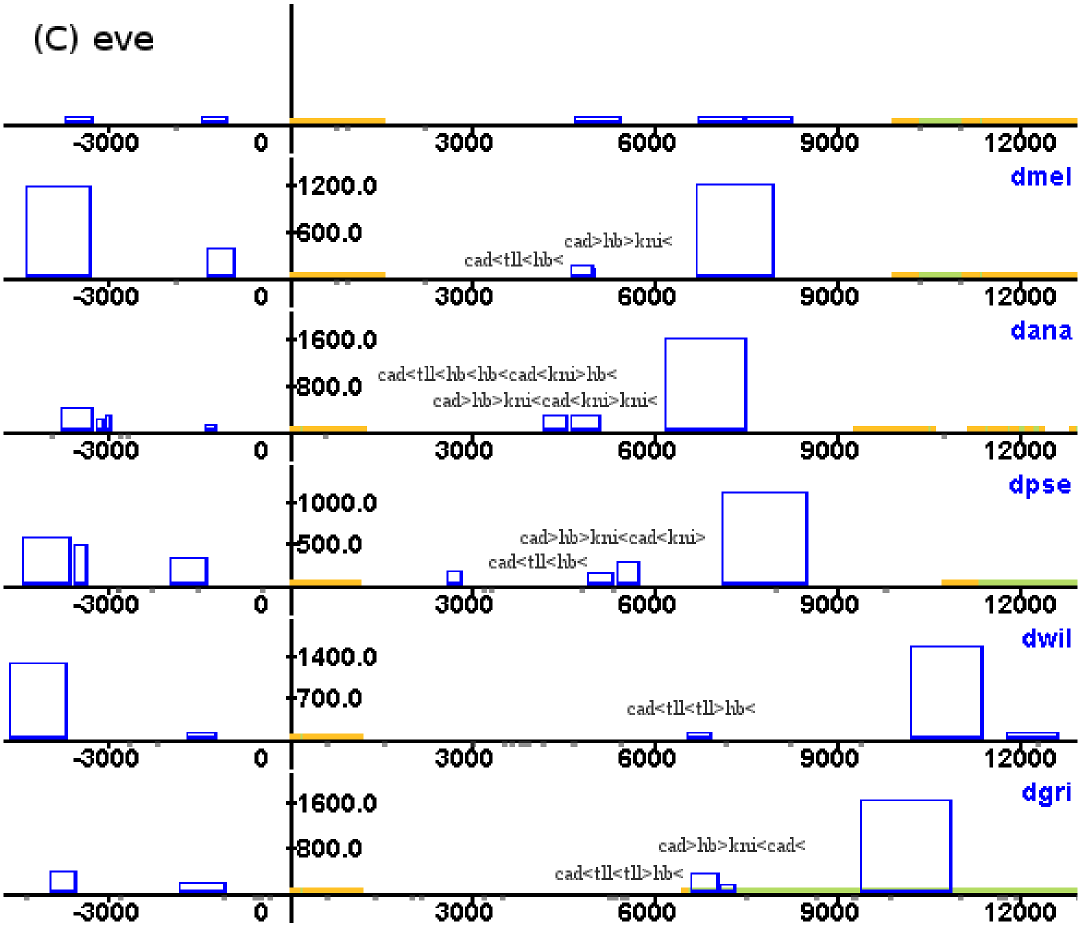
gene	CC			PPV		
	CORECLUST	CORECLUST-FC	CORECLUST-F	CORECLUST	CORECLUST-FC	CORECLUST-F
h	**0.69**	0.62	0.59	**0.83**	0.82	0.82
kni	0.43	0.43	**0.53**	0.62	0.62	**0.84**
hb	0.32	0.28	**0.35**	0.35	0.34	**0.48**
ftz	**0.31**	**0.31**	0.11	**0.51**	**0.51**	0.31
eve	**0.73**	0.65	0.62	**0.76**	0.70	0.70
run	0.08	**0.15**	0.10	0.59	**0.81**	0.72
tll	0.26	0.23	**0.29**	0.42	0.47	**1.00**
gt	0.41	**0.42**	0.35	0.78	0.78	**1.00**
Kr	0.45	0.53	**0.61**	0.58	0.65	**1.00**
cad	-0.03	-0.03	-0.03	0.00	0.00	0.00
prd	**0.26**	NaN	0.17	**1.00**	NaN	**1.00**
ems	-0.02	**-0.01**	NaN	0.00	0.00	NaN
btd	0.45	**0.71**	0.54	0.26	**0.59**	0.36
slp1	0.35	0.35	**0.43**	0.68	0.70	**1.00**
bowl	**0.20**	**0.20**	NaN	**0.15**	**0.15**	NaN
salm	0.23	**0.27**	NaN	0.16	**0.23**	NaN
fkh	**0.31**	-0.04	NaN	**0.28**	0.00	NaN
TOTAL	**0.32**	0.31	0.32	0.51	0.55	**0.68**

Supplementary Figures

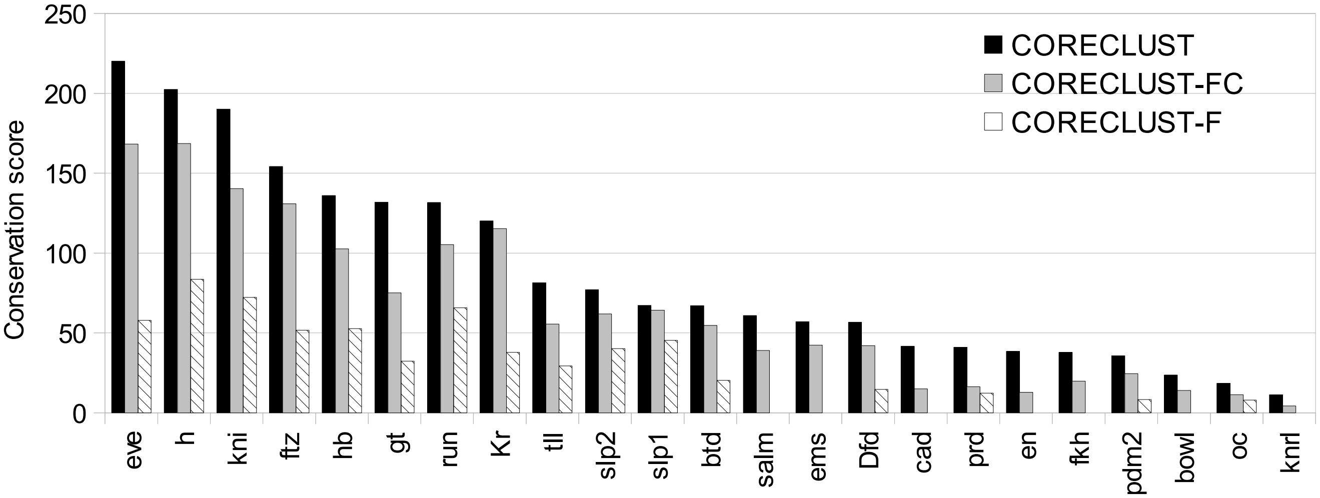


Supplementary Figure S4. An effect of inclusion of structural aspects on the prediction accuracy for *Drosophila* (**A**) and vertebrate (**B**) datasets. The measures are correlation coefficient (CC), sensitivity (Sn), specificity (Sp), positive predictive value (PPV), performance coefficient (PC), and average site performance (ASP).


(A) hb
bcdbcdbcdbcdbcdhbbcdbcdbcd
-9000
-6000
-3000
0
3000
dmel
1000.0
500.0
bcd>bcd>bcd>bcd>bcd>
-9000
-6000
-3000
0
3000
dana
800.0
400.0
bcd<tll>kni<bcd<bcd>bcd>bcd>bcd>
-9000
-6000
-3000
0
3000
dpse
600.0
300.0
kni<Kr<cad>cad<bcd>bcd>bcd>bcd>
kni<bcd<bcd>bcd>bcd>bcd>
-9000
-6000
-3000
0
3000
dwil
800.0
400.0
hb>cad<bcd>bcd>bcd>bcd>bcd>bcd>tll>gt>
kni<bcd<bcd>bcd>
-9000
-6000
-3000
0
3000
dgri
800.0
400.0
bcd>bcd>bcd>bcd>bcd>bcd>bcd>bcd>bcd>
kni>kni<bcd<bcd>bcd>bcd>
-9000
-6000
-3000
0
3000
(B) tll
-3000
0
3000
6000
9000
dmel
400.0
200.0
gt>gt<cad<Kr<tll>
-3000
0
3000
6000
9000
dana
400.0
200.0
Kr<tll>hb<
-3000
0
3000
6000
9000
dpse
400.0
200.0
Kr<cad<Kr<tll>
-3000
0
3000
6000
9000
dwil
600.0
300.0
cad<Kr<tll>
-3000
0
3000
6000
9000
dgri
600.0
300.0
bcd>hb>cad<Kr<tll>hb<bcd<bcd>
-3000
0
3000
6000
9000



Supplementary Figures S5. Examples of CRM predictions made by CORECLUST, demonstrating advantage of accounting for regulatory structure. In all cases, CORECLUST managed to found CRMs corresponding to the known modules, while CORECLUST-FC and CORECLUST-F failed to identify these CRMs at all, or scored them with weights much lower than a threshold. Regulatory modules are shown by blue rectangles. The heights of the predicted CRMs correspond to their weights. At the top line the known regulatory regions from REDFly database (34) are presented. Other lines correspond to orthologous genes from different genomes: dmel - *D. melanogaster,* dana - *D. ananassae,* dpse - *D. pseudoobscura,* dwil - *D. willistoni,* and dgri - *D. grimshawi.* Genes are shown in green, exons are colored by orange. Gray rectangles below the lines denote repeat regions. The site compositions of CRMs of interest are shown next to them. The site strand is shown by symbols '>' (positive strand) and '<' (negative strand). **(A)** Predictions made for gene *hb*. The CRM of interest is a poly-Bcd module at ~3Kbp downstream the gene start in *D.melanogaster* genome. **(B)** Predictions made for gene *tll*. The CRM of interest is located downstream the gene and contains Cad, Kr and Tll binding sites. **(C)** Predictions made for gene *eve*. The CRMs revealed by CORECLUST are two closely spaced CRMs at ~5Kbp downstream the gene start in *D.melanogaster* genome.



Supplementary Figure S6. Comparison of the conservation level of the CRMs predicted by CORECLUST, CORECLUST-FC and CORECLUST-F. The model training and CRM search were done separately for every gene.

	mef2>	mef2<	myf>	myf<	sp1>	sp1<	srf>	srf<	tef>	tef<
mef2>	0,00 / 0,49	0,00 / 0,50	0,00 / 0,53	0,96 / 1,00	0,00 / 0,63	0,00 / 0,60	0,00 / 0,50	0,00 / 0,50	0,00 / 0,50	0,01 / 0,77
mef2<	0,00 / 0,50	0,00 / 0,81	0,17 / 0,00	0,01 / 0,89	0,00 / 0,33	0,41 / 0,01	0,00 / 0,45	0,18 / 0,01	0,00 / 0,39	0,22 / 0,51
myf>	0,25 / 0,83	0,10 / 0,00	0,00 / 0,22	0,00 / 0,79	0,16 / 1,00	0,00 / 0,51	0,11 / 0,79	0,00 / 0,42	0,37 / 1,00	0,00 / 0,56
myf<	0,00 / 0,50	0,00 / 0,61	0,22 / 0,99	0,00 / 0,52	0,45 / 0,98	0,22 / 0,10	0,00 / 0,53	0,09 / 0,77	0,00 / 0,50	0,01 / 0,82
sp1>	0,00 / 0,50	0,34 / 0,01	0,00 / 0,52	0,01 / 0,84	0,01 / 0,39	0,59 / 0,99	0,01 / 0,72	0,02 / 0,77	0,02 / 0,95	0,00 / 0,47
sp1<	0,16 / 0,01	0,00 / 0,64	0,34 / 0,43	0,00 / 0,84	0,00 / 0,51	0,06 / 0,94	0,39 / 0,68	0,00 / 0,08	0,00 / 0,63	0,04 / 0,99
srf>	0,11 / 1,00	0,00 / 0,52	0,00 / 0,52	0,00 / 0,54	0,00 / 0,48	0,11 / 0,00	0,07 / 0,99	0,38 / 0,21	0,32 / 0,00	0,00 / 0,50
srf<	0,00 / 0,51	0,00 / 0,51	0,00 / 0,51	0,15 / 0,00	0,00 / 0,42	0,65 / 0,00	0,19 / 0,99	0,00 / 0,17	0,00 / 0,52	0,00 / 0,52
tef>	0,01 / 0,06	0,02 / 0,01	0,00 / 0,45	0,01 / 0,95	0,28 / 0,00	0,52 / 0,00	0,02 / 0,15	0,00 / 0,49	0,14 / 1,00	0,00 / 0,50
tef<	0,00 / 0,50	0,50 / 1,00	0,00 / 0,48	0,33 / 0,81	0,13 / 0,01	0,01 / 0,61	0,01 / 0,46	0,00 / 0,50	0,01 / 0,18	0,00 / 0,51

Supplementary Figure S7. Parameters of the model, trained on the muscle dataset. Each cell (*i*, *j*) of the table contains the conditional probability to observe a site of type *j* next to a site of type *i* (first number), and the probability that the distance between these sites is distributed according to the helical phasing distribution (second number). The color of a cell corresponds to the conditional probability value. The site strands are shown by symbols '>' (positive strand) and '<' (negative strand).

Captions for Supplementary Figures S8 and S9 (in separate files)

Supplementary Figure S8. Parameters of the model trained on eleven *Drosophila* developmental genes. See Supplementary Figure S7 for notations.

Supplementary Figure S9. Distributions of intersite distance for the overrepresented site pairs predicted for the eleven *Drosophila* developmental genes. The distribution plots are organized as a table where rows correspond to the first site in a pair and columns correspond to the second site; the site strands are shown by symbols '>' (positive strand) and '<' (negative strand). Each cell of the table corresponds to one site pair. The correlation coefficient for the sites in a pair and the number of observed site pairs are shown at each plot. The site pair is shown if it is overrepresented, which means that it was observed at least 70 times and the absolute value of the correlation coefficient for its sites is more than 0.2 (both are presented on the plot). As an exception, the distributions for the site pairs Gt>Gt> and Gt<Bcd> are shown, in spite of a very low number of the observations. The width of the border is proportional to the absolute value of the correlation coefficient for the pair. The distributions were build based on the regulatory modules with weight ≥ 100. Predictions were made for the well-known developmental genes in all analyzed *Drosophila* genomes. The distance was measured between the sites' starts. The random distance distribution is shown by the red line.

References

7. Frith,M.C., Li,M.C., and Weng,Z. (2003) Cluster-Buster: Finding Dense Clusterss of Motifs in DNA Sequences. *Nucleic Acids Res.*, 31, 3666-3668.

20. Kulp,D., Haussler,D., Reese,M.G. and Eeckman,F.H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **4**, 134–142.

21. Rabiner,L.R. (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257—286.

22. Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.

26. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press.

34. Halfon,M.S., Gallo,S.M. and Bergman,C.M. (2008) REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. *Nucleic Acids Res.*, **36**, D594–598.

44. Stark,A., Lin,M.F., Kheradpour,P., Pedersen,J.S., Parts,L., Carlson,J.W., *et al.* (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, **450**, 219-232.

45. Gerstein,M., Sonnhammer,E.L. and Chothia,C. (1994) Volume changes in protein evolution.

J. Mol. Biol. , **236**, 1067-1078.

46. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.