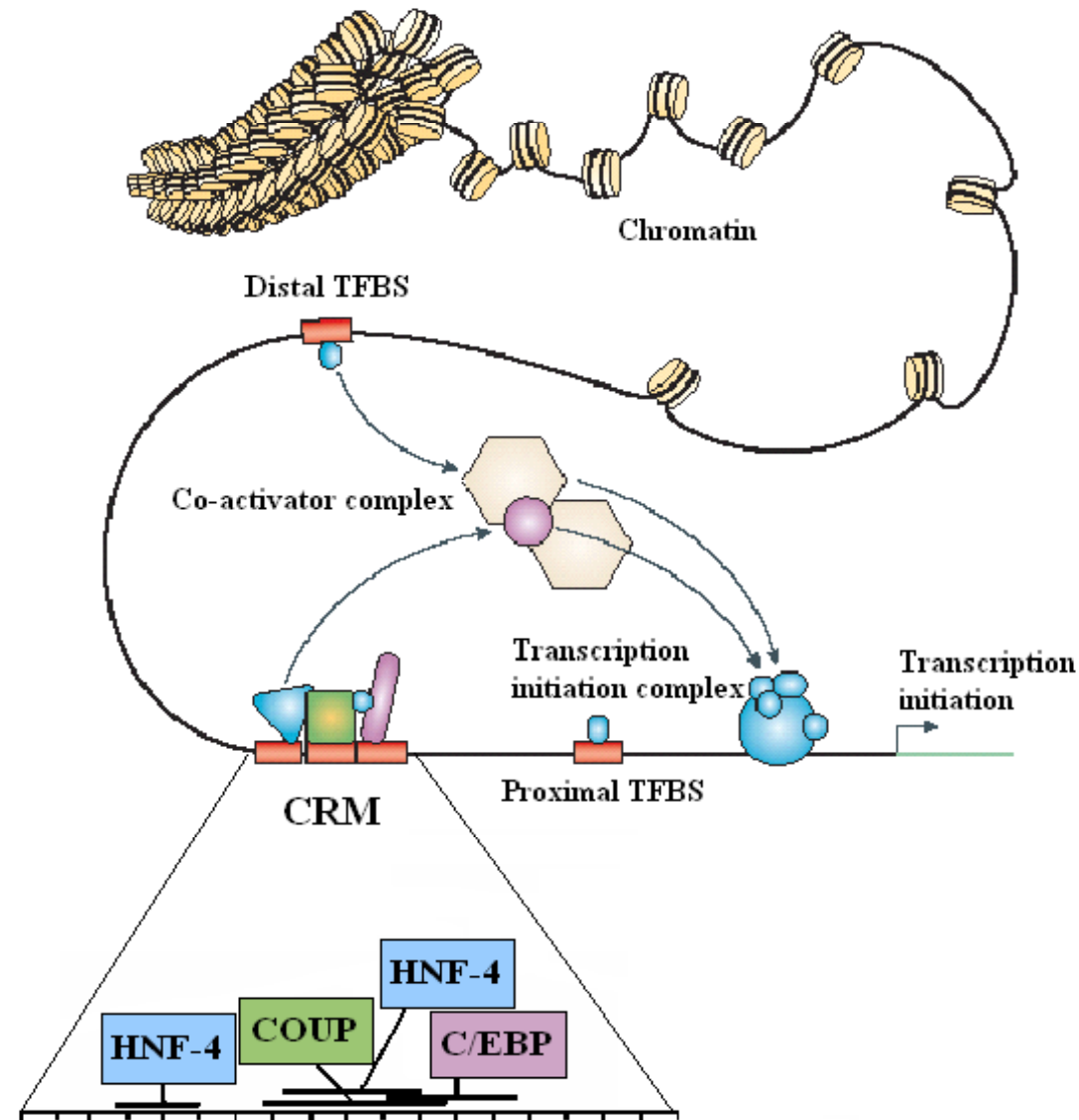


Prediction and structural analysis of the
evolutionary conserved enhancers
in *Drosophila* genomes

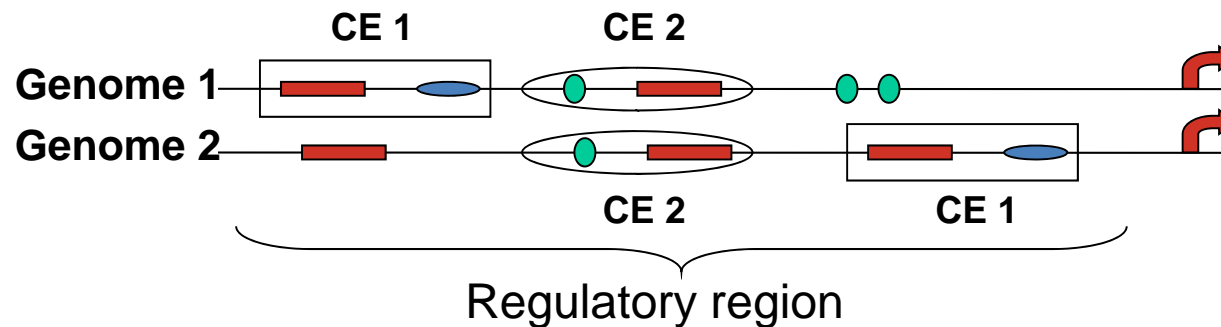
Nikulova A., Favorov A., Mironov A.

Transcriptional Regulation In Eukaryotes



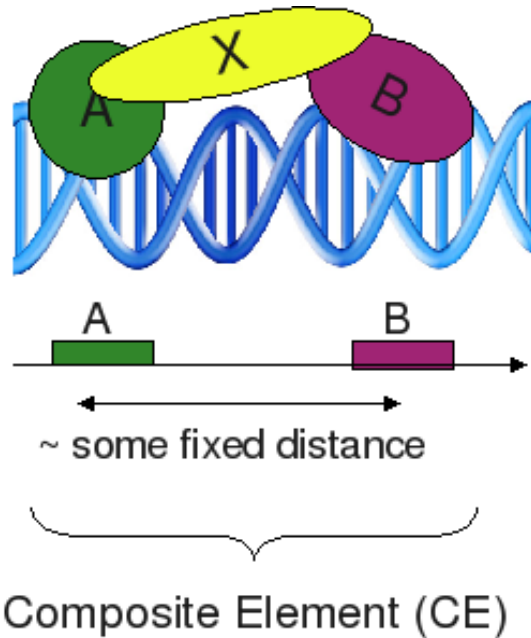
Motivation

- Transcription factor binding sites tend to cluster along the DNA strand
- In some cases sites form COMPOSITE ELEMENTS (CEs)
- Co-regulated genes have similar regulatory elements in their regulatory regions
- Functional regulatory sequences are more conserved



- During the evolution sites can undergo the significant turnover => the regulatory elements can move and reshuffle along the DNA strand

Regulatory Region Structure



- site type and their frequencies
- sites order preferences
- inter-site distance distributions



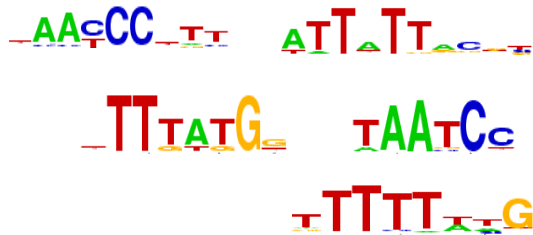
Hidden Markov Model (HMM) of a regulatory region:

- combines potential sites to form clusters
- takes into account the regulatory region structure

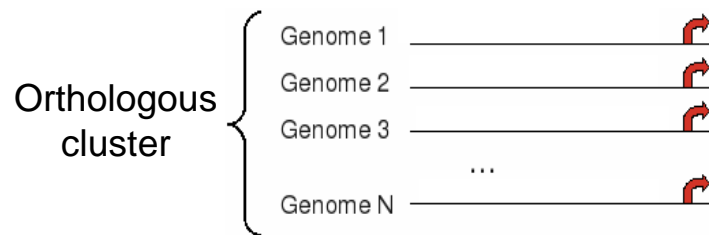
Our program (HMM-based)

Input

TF matrices



Training gene

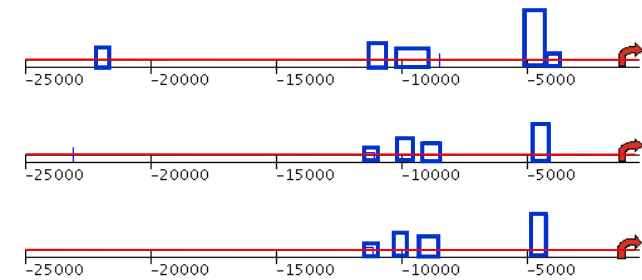


Genomes (12 flies)

```
>gi|116010444|ref|NT_033779.4| Drosophila melanogaster chromosome 2L,
complete sequence
CGACAATGCACGACAGAGGAAGCAGAACAGATATTTAGATTGCCTCTCATTCTCTCCCATATTTATAGG
GAGAAATATGATCGCGTATGCGAGAGTAGTGCCAACATATTGTGCTCTTTGATTTTTGGCAACCCAAAA
TGGTGGCGGATGAACGAGATGATAATATATTCAGTTGCCGTAATCAGAAATAAATTCATTGCAACGTT
AAATACAGACAATATATGATCGCGTATGCGAGAGTAGTGCCAACATATTGTGCTAATGAGTGCCCTCTCG
TTCTCTGCTTATATTACCGCAAACCCAAAAAGACAATACACGACAGAGAGAGAGAGCAGCGGAGATATT
TAGATTGCCATTAAATATGATCGCGTATGCGAGAGTAGTGCCAACATATTGTGCTCTCTATATAATGAC
TGCCTCTCATTCTGCTTATTTTACCGCAAACCCAAATCGACAATGCACGACAGAGGGAAGCAGAACAGAT
ATTTAGATTGCCTCTCATTCTCTCCATATATAGGGGAAATATGATCGCGTATGCGAGAGTAGTGC
CAACATATTGTGCTCTTTGATTTTTTGGCAACCCAAATGTTGGCGGATGAACGAGATGATAATATATTC
AAGTTGCCGTAATCAGAAATAAATTCATTGCAACGTTAAATACAGACAATATATGATCGCGTATGCGGA
GAGTAGTGCCAACATATTGTGCTAATGAGTGCTCTCGTTCTCTGCTTATATTACCGCAAACCCAAAA
GACAATACACGACAGAGAGAGAGAGAGAGAGAGAGAGATTTAGATTGCCTATTAATATGATCGCGTATGCG
AGAGTAGTGCCAACATATTGTGCTCTATATAATGACTGCTCTCATTCTGCTTATTTTACCGCAAAC
CCAAATCGACAATGCACGACAGAGGAAGCAGAACAGATATTTAGATTGCCTCTCATTCTCTCCCATAT
TATAGGGGAAATATGATCGCGTATGCGAGAGTAGTGCCAACATATTGTGCTCTTTGATTTTTT...
```

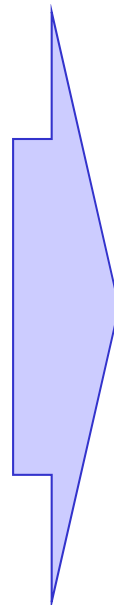
Output

Site clusters prediction

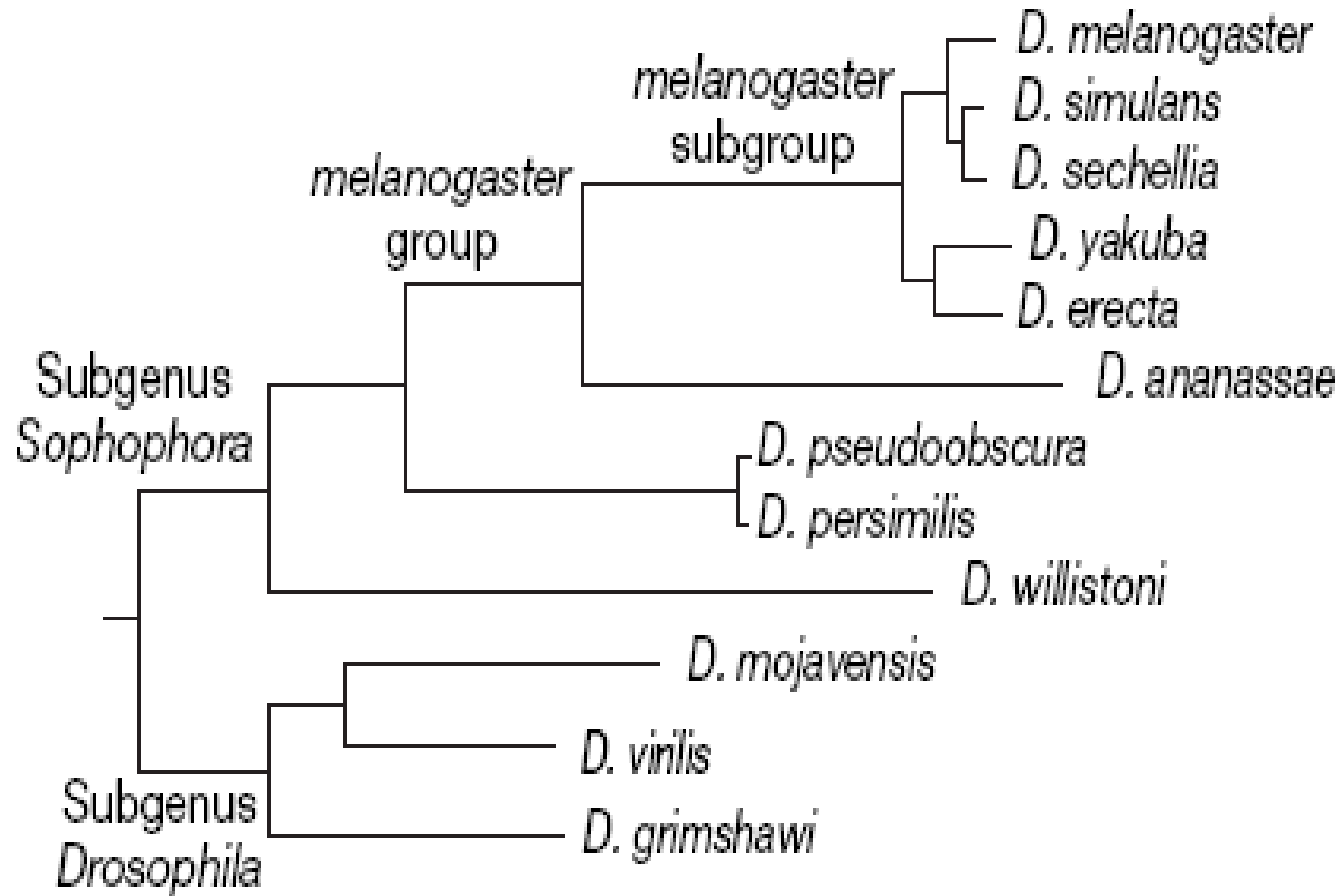


Co-regulated genes prediction

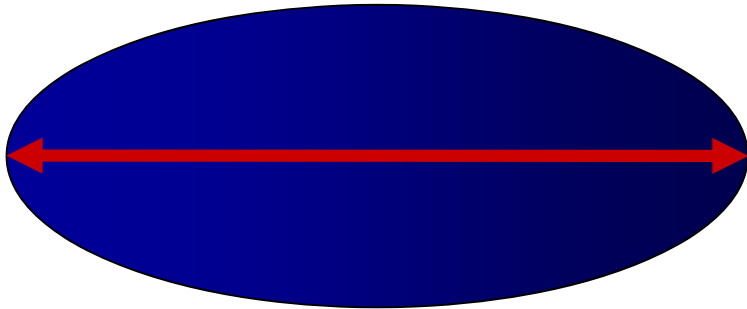
Gene	Expression Pattern	Score	Function
h		0 (61.63)	pair-rule gene, specific transcriptional repressor
tll		1 (13.48)	gap gene, transcription factor, ligand-dependent nuclear receptor
kni		2 (10.40)	gap gene, transcriptional repressor
ftz		3 (9.80)	specific transcription factor
run		4 (9.16)	pair-rule gene, transcription factor
rib		5 (9.06)	embryonic development, transcription factor
gt		6 (7.97)	gap gene, transcriptional repressor
CG9650		7 (7.45)	zinc ion binding, nucleic acid binding
Cyp6v1		8 (6.38)	cytochrome P450
slp1		9 (6.22)	pair-rule gene, transcription factor
eve		10 (6.17)	pair-rule gene, specific transcription factor
hb		11 (6.16)	gap gene, transcriptional activator
CG6486		12 (5.76)	peroxisome matrix targeting signal-2 binding
CG1958		13 (5.67)	unknown
hydra		14 (5.36)	unknown



Test Systems: *Drosophila* Developmental Genes



Anterior–posterior axis patterning system

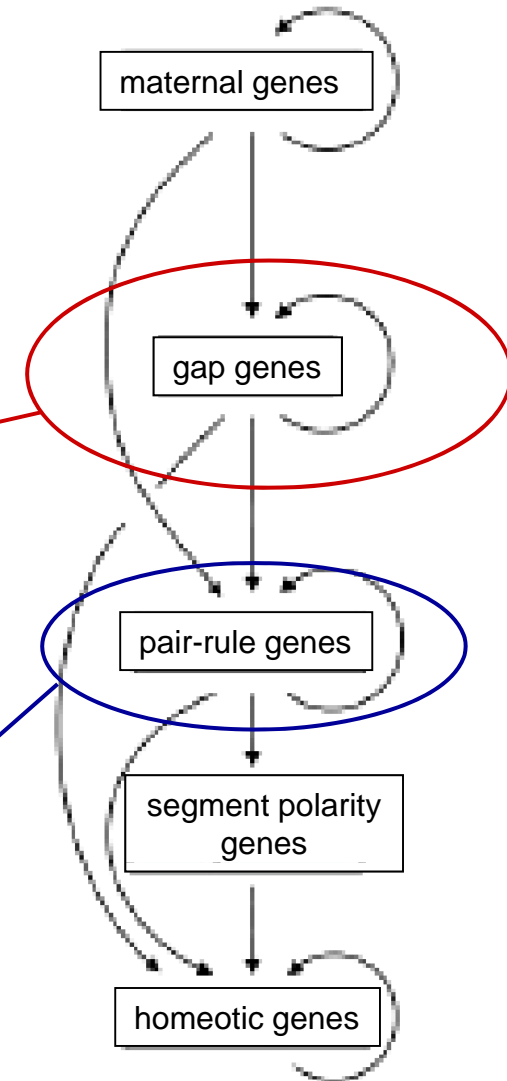


Transcription factors (gap genes):





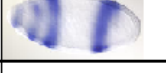


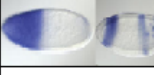


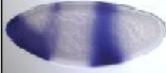
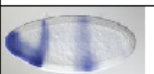


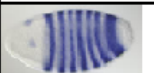

- Bicoid(bcd)
- Hunchback (hb)
- Caudal (cad)
- Giant (gt)
- Kruppel (Kr)
- Knirps (Kni)
- Tailless (Tll)

First run:

Train model on gene *hairy (h)*



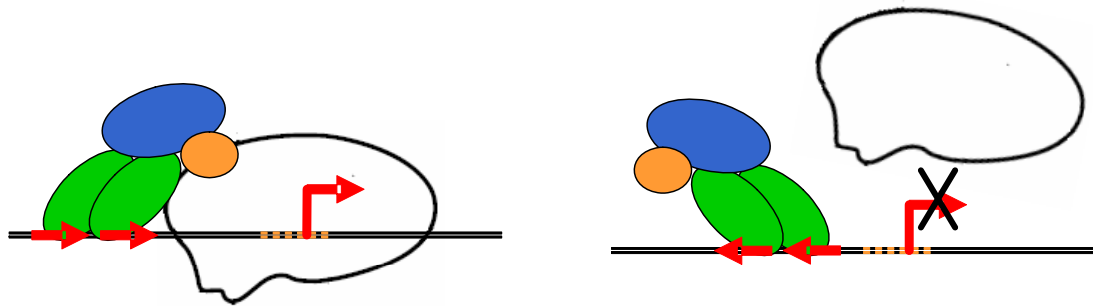
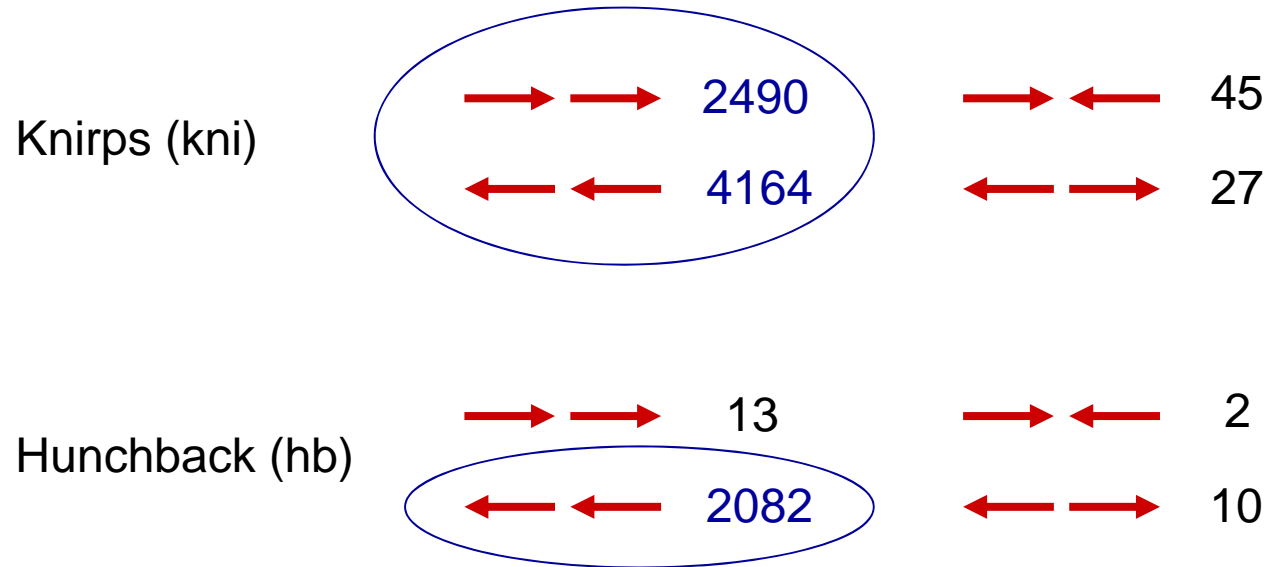
Top genes for h-trained model: expression, function

Gene	Expression Pattern	Conservation Score	Function	Gene	Expression Pattern	Conservation Score	Function
h		0 (83.33)	pair-rule gene, TF; open tracheal system development, nervous system development	CG5103		16 (12.30)	transketolase (?)
CG12496 *		1 (27.81)	unknown (next to PsGEF/CG14045 (mushroom body development))	cenG1A		17 (11.68)	small GTPase mediated signal transduction (?)
Ptx1		2 (25.60)	dendrite morphogenesis, TF	CG15485 *		18 (11.24)	endothelin-converting enzyme (?) (inside pdm2 (TF, neuron fate determination))
ftz		3 (24.49)	pair rule gene, TF; gonadal mesoderm development	gt		19 (10.40)	gap gene, TF; torso signaling pathway; terminal region determination; ring gland development
CG1958 *		4 (24.42)	unknown (next to CG9650 (also in this list))	Cyp6v1		20 (10.34)	cytochrome P450 (?)
tll		5 (20.51)	gap gene, TF; torso signaling pathway; terminal region determination, neuroblast division	ct		21 (10.10)	TF; dendrite morphogenesis, regulation of Notch signaling pathway
hb		6 (20.19)	gap gene, TF; torso signaling pathway; terminal region determination, neuroblast fate determination	CG15161		22 (9.89)	cilium assembly
CG14045		7 (18.53)	mushroom body development	CG12602 !		23 (9.82)	ATP synthesis coupled proton transport (?)
slp1		8 (16.53)	pair-rule and segment polarity gene, TF; specification of segmental identity, head	kni		24 (9.60)	gap gene, TF; dendrite morphogenesis, muscle organ development, epidermis development
CG9650 *		9 (15.44)	zinc ion binding, nucleic acid binding(?) (next to CG1958 (also in this list))	hdc		25 (9.60)	axon extension involved in development, specifically expressed in all imaginal cells
knir1		10 (15.22)	TF, knirps-related	hydra *		26 (9.36)	unknown (next to run (pair-rule gene))
eve		11 (15.07)	pair-rule gene, TF; regulation of axonogenesis; regulation of cardioblast cell fate specification	CG5397 !		27 (9.14)	sterol O-acyltransferase activity
Kr		12 (13.94)	gap gene, TF; neuroblast fate determination, axon guidance, compound eye development	robo3 !		28 (8.88)	axon guidance, mushroom body, ventral cord and central complex development
CG12163		13 (13.91)	salivary gland cell autophagic cell death	dally		29 (8.64)	sensory organ development
Nckx30C		14 (13.72)	compound eye development, sodium and calcium ion transport	CG30430		30 (8.51)	unknown
run		15 (13.31)	pair-rule gene, TF; axon guidance, dendrite morphogenesis, eye morphogenesis	prd		31 (8.50)	pair-rule gene, TF

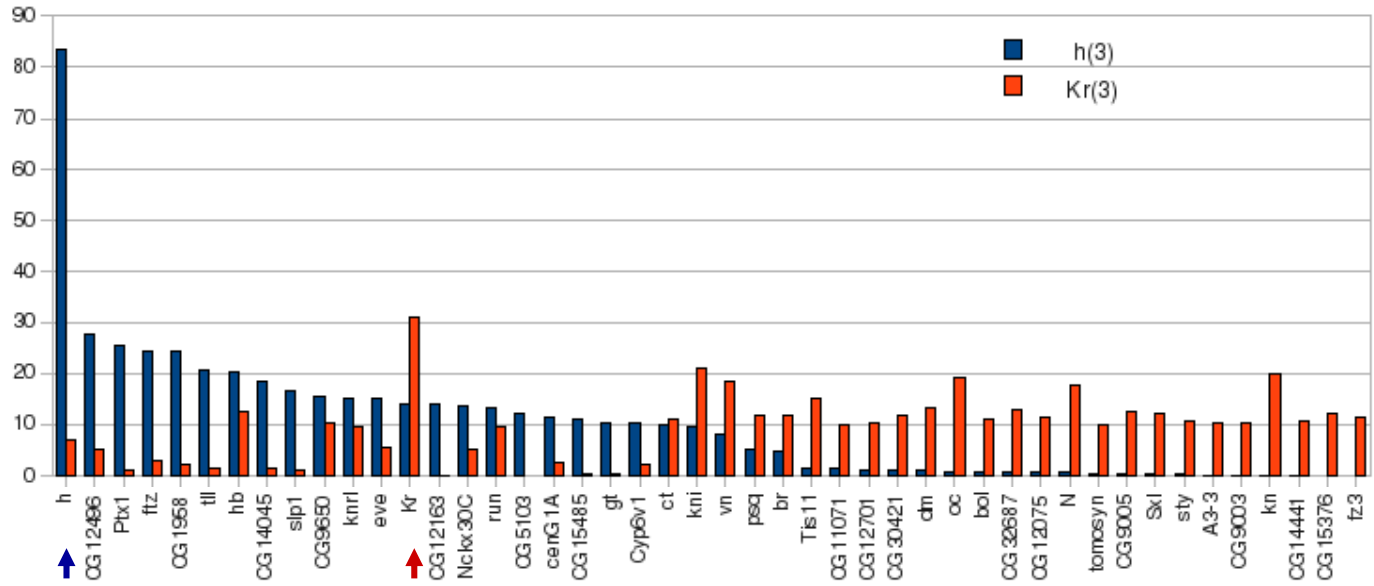
Top Genes: GO-statistics

GO Term	Top 32 genes (26)	All genes: 10558	P-Value
periodic partitioning by pair rule gene	5	6	1,64E-10
blastoderm segmentation	9	137	4,72E-09
trunk segmentation	5	15	1,81E-08
posterior head segmentation	4	15	1,60E-06
tube morphogenesis	6	102	4,49E-06

Cluster structure: sites order



Comparison of the gene lists for different training genes



h

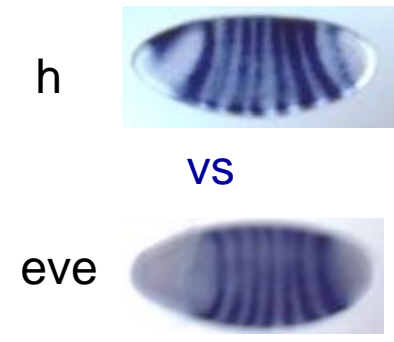
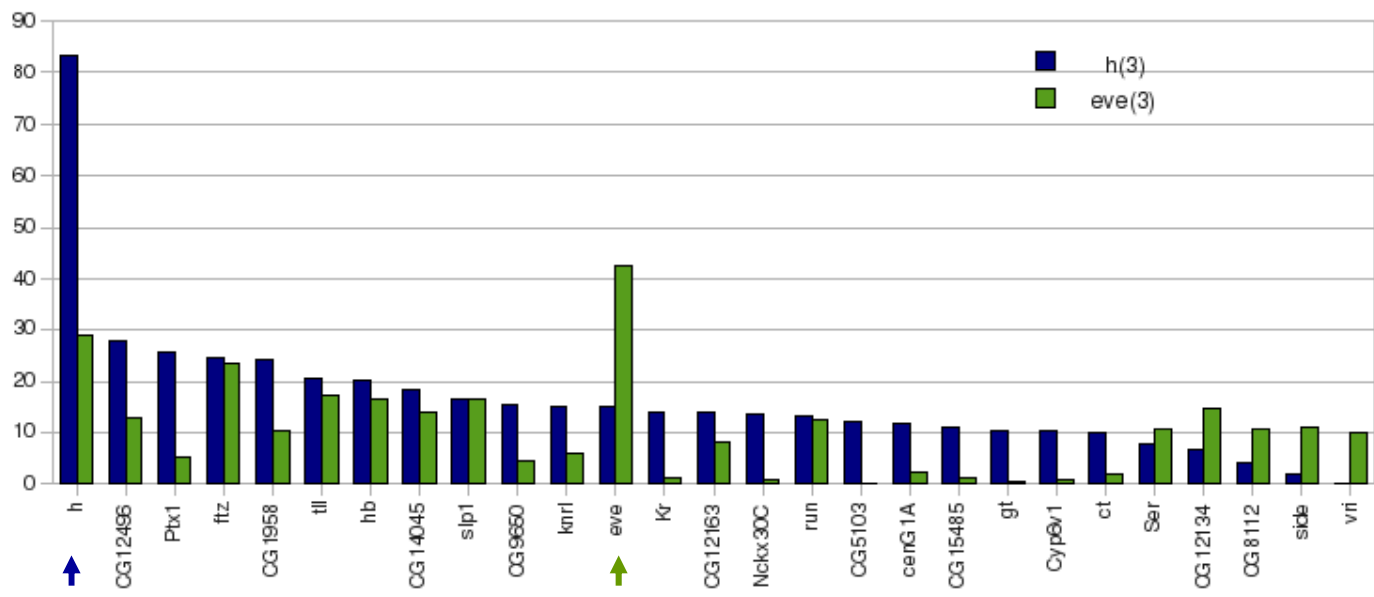
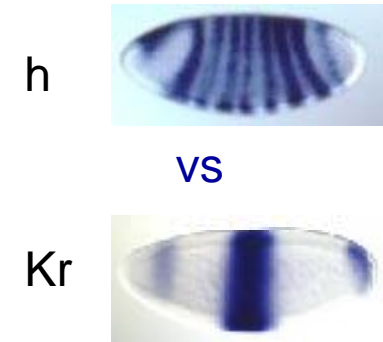
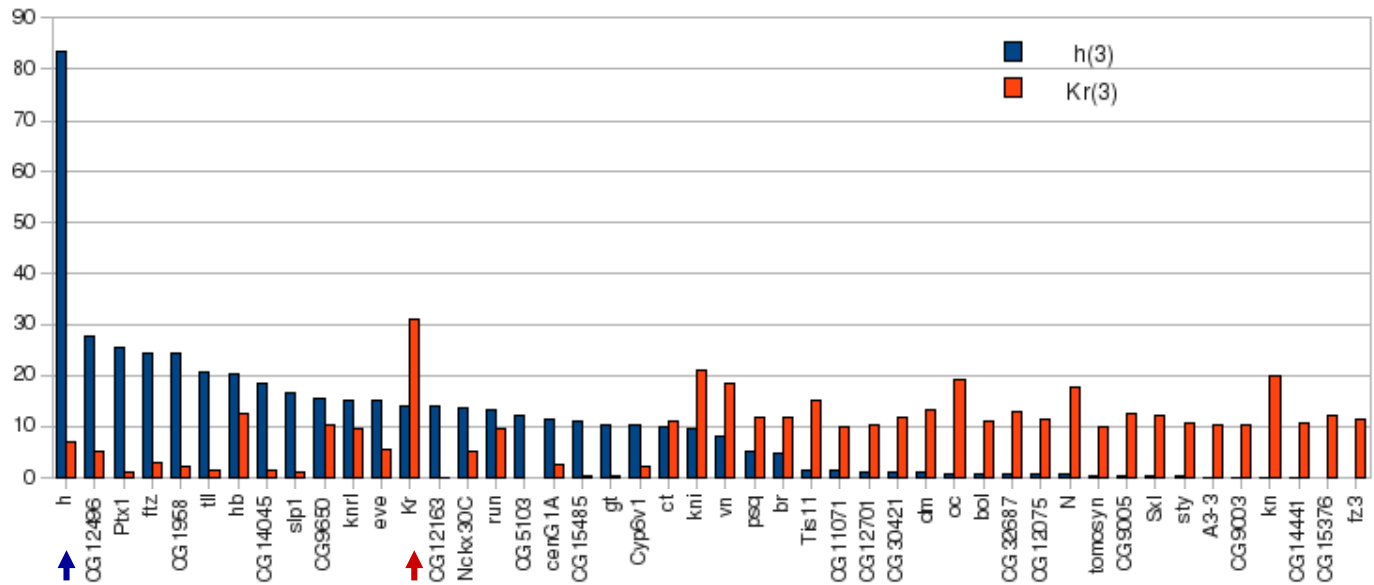


vs

Kr

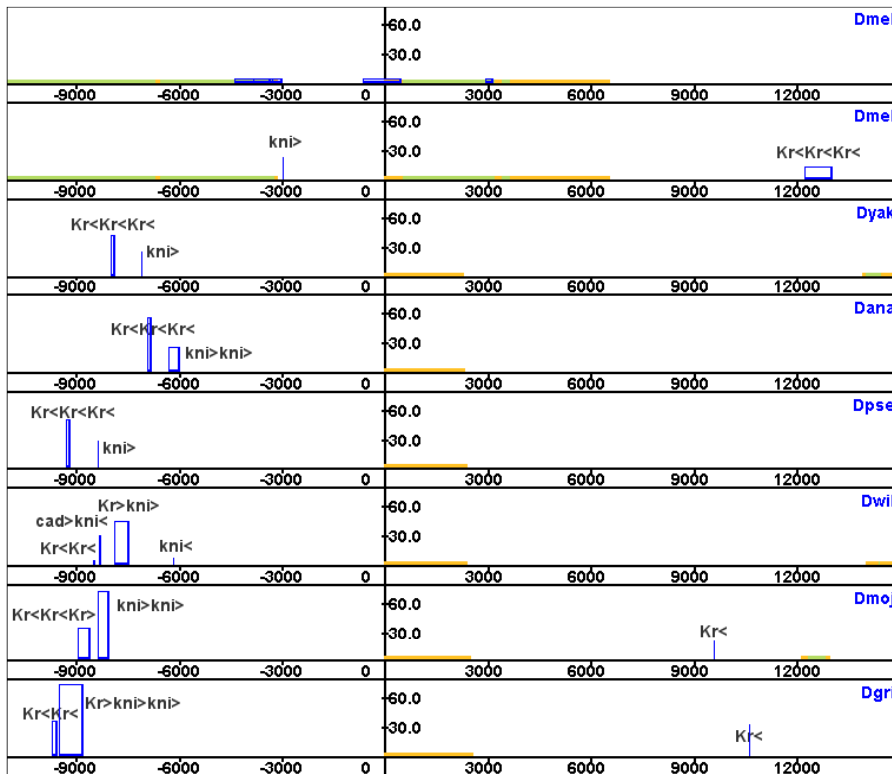


Comparison of the gene lists for different training genes



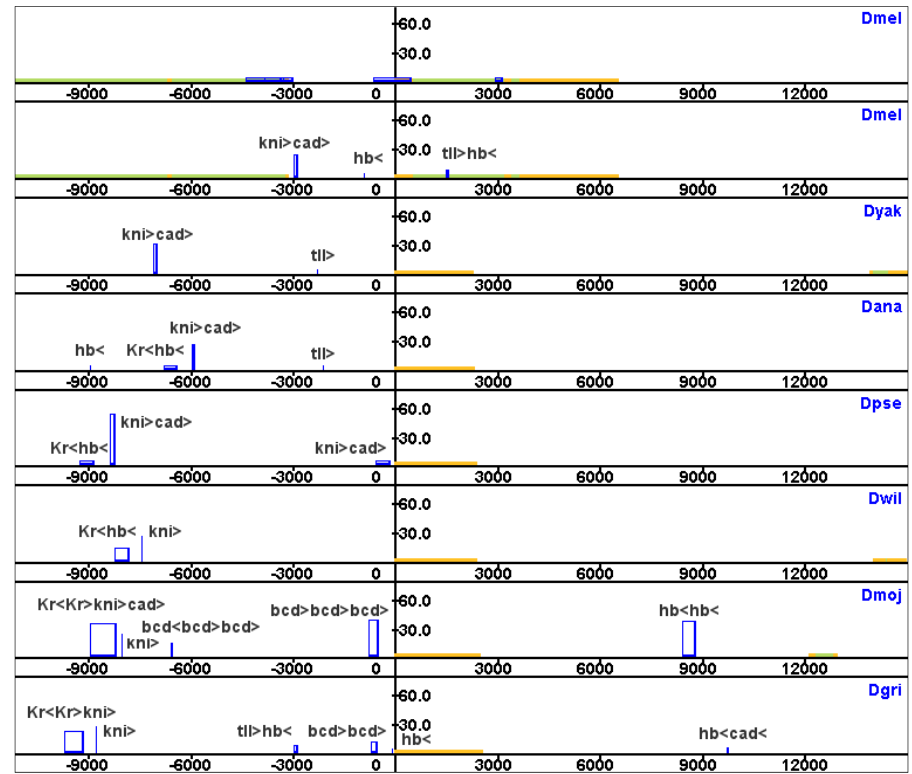
Clusters prediction: *hb*

(h – trained model)



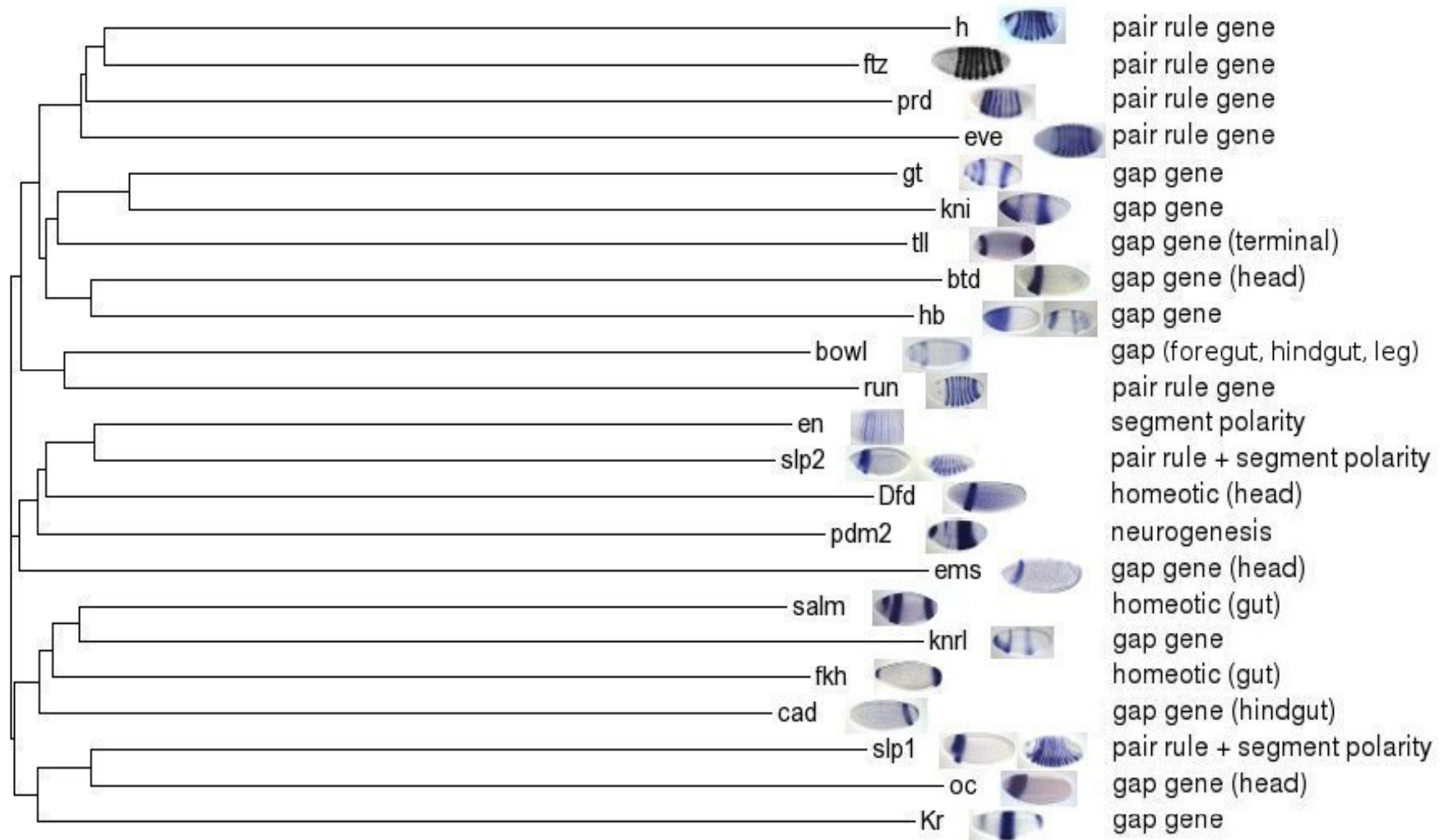
Kr<Kr<Kr<
kni>kni>

(Kr – trained model)

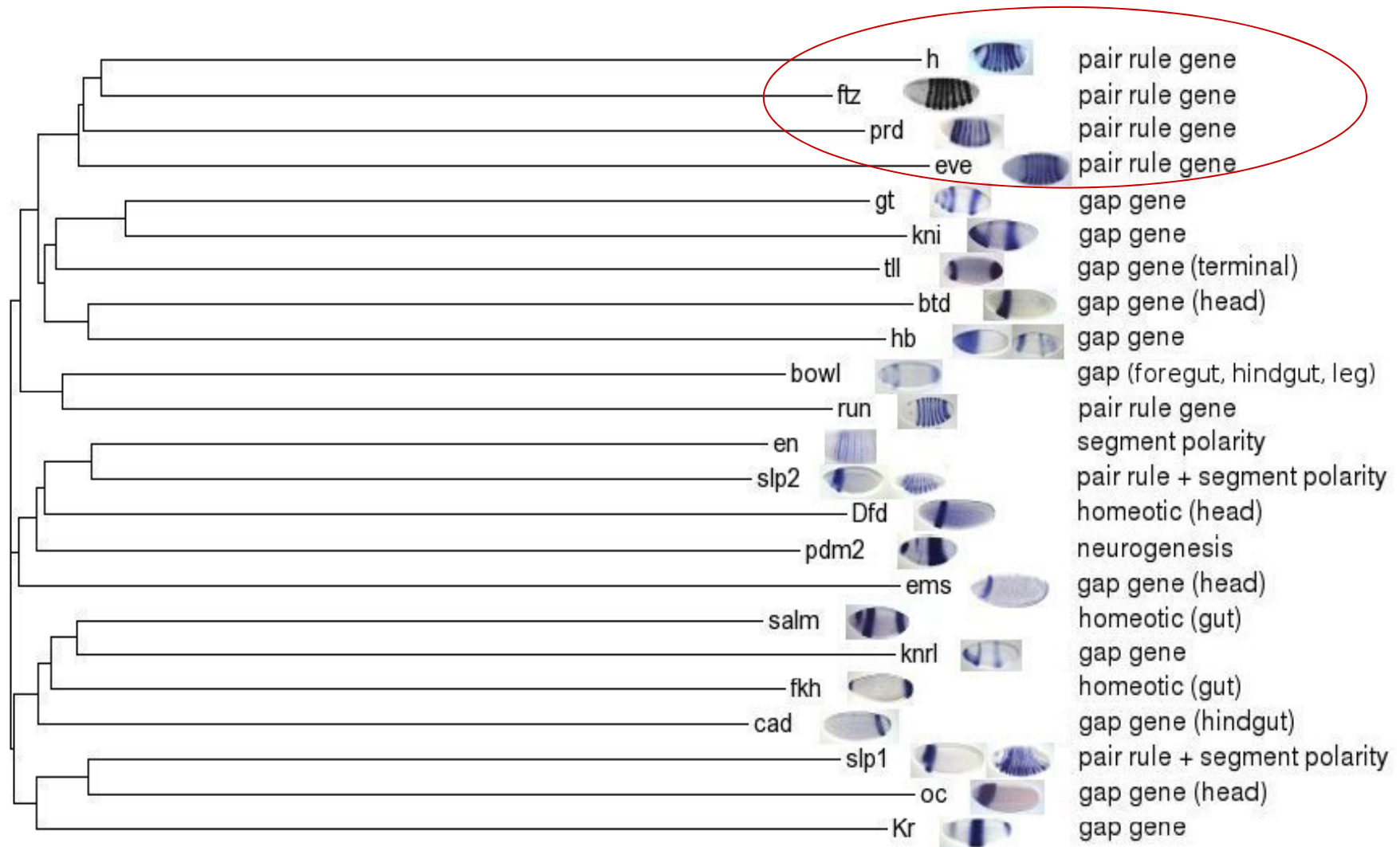


kni>cad>
Kr<hb<

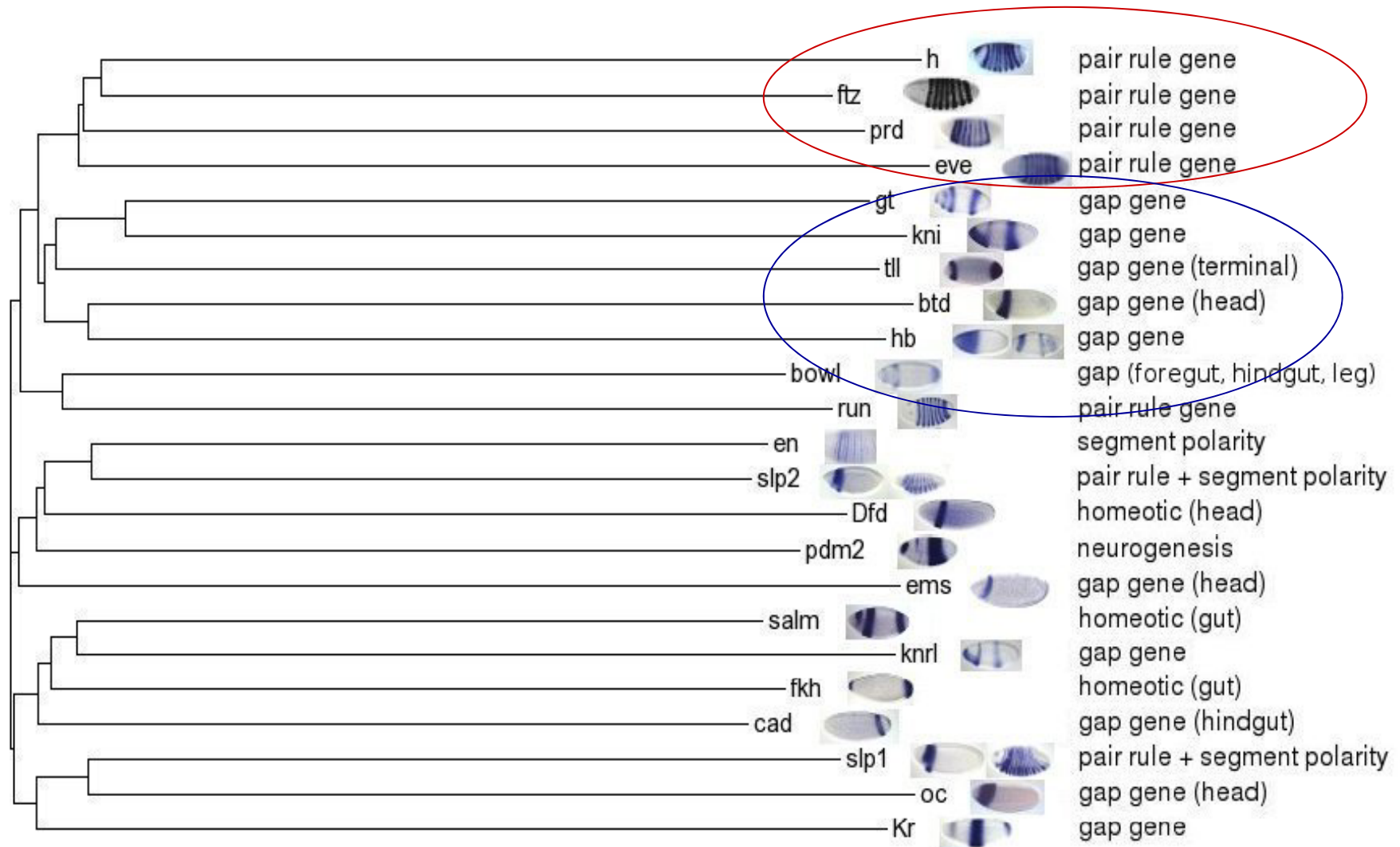
Cluster structure: comparison between genes



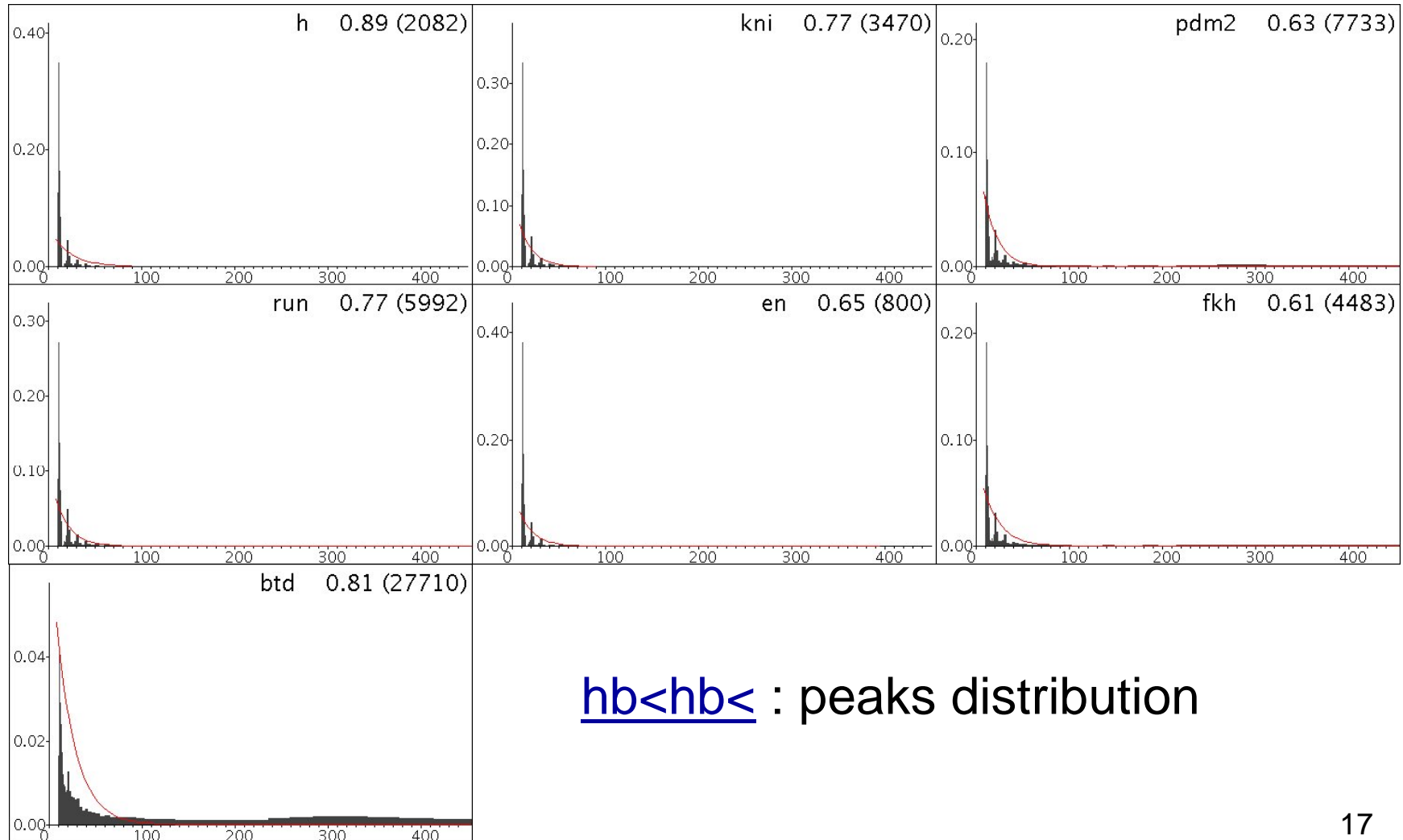
Cluster structure: comparison between genes



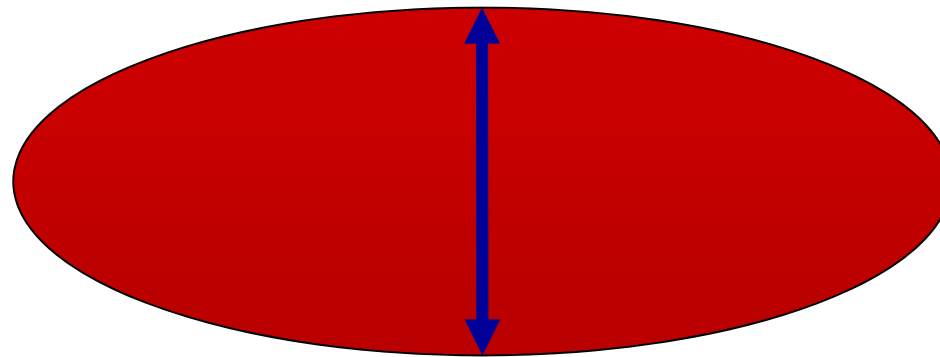
Cluster structure: comparison between genes



Cluster structure: site-to-site distances





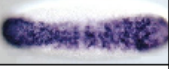


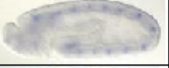


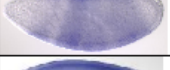





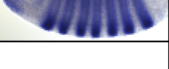
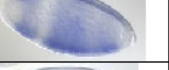


Dorsal–ventral axis patterning system



Transcription factors:

- Dorsal (dl)
- Snail (sna)
- Twist (twi)
- Brinker (brk)
- Su(H)
- Mad

Top genes: expression, function

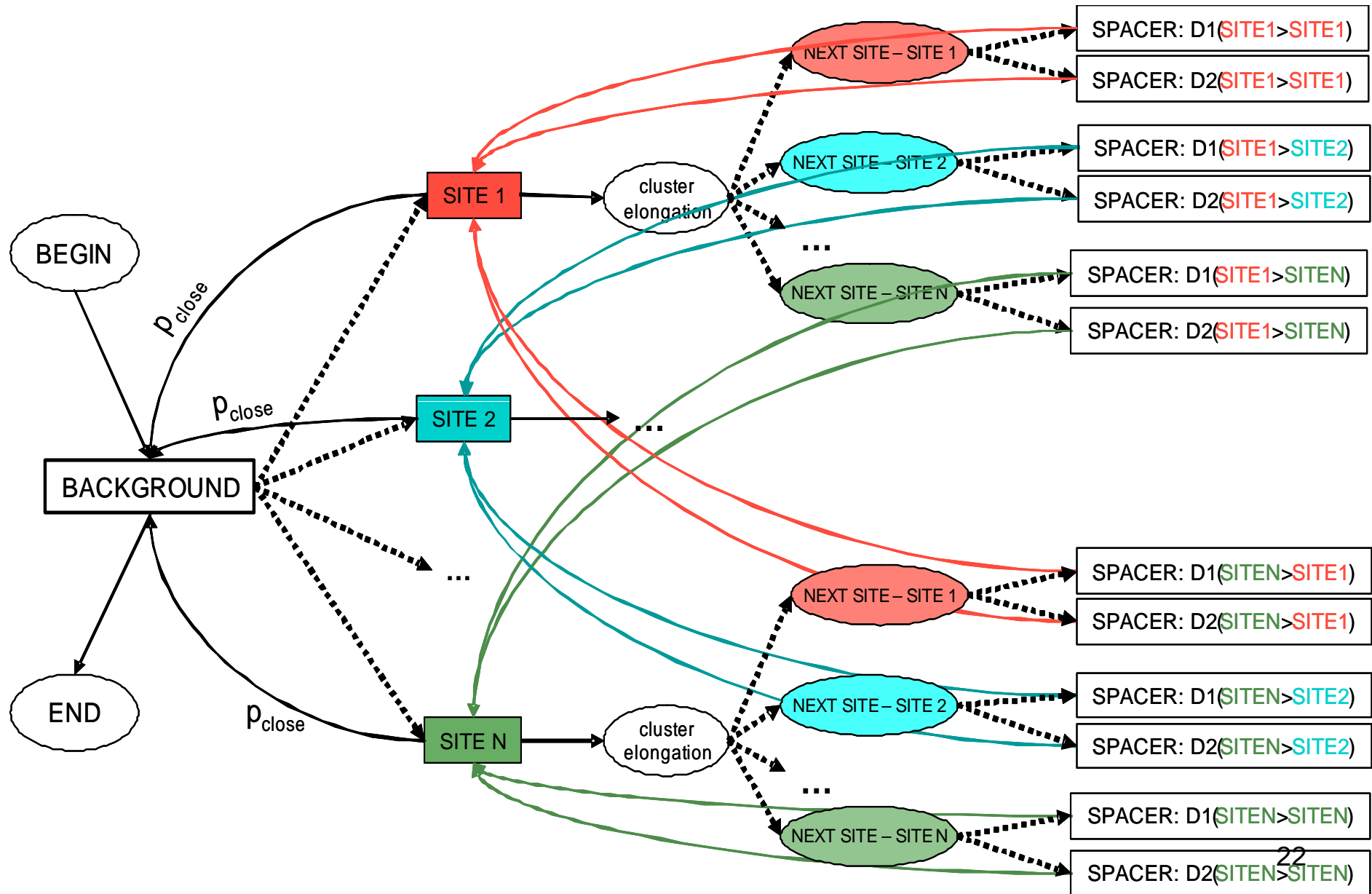
Gene	Expression Pattern	Conservation Score	Function	Gene	Expression Pattern	Conservation Score	Function
vn		0 (23,41)	epidermal growth factor (Egf) receptor signaling pathway	ths		16 (11,56)	glial cell differentiation, hindgut, heart and somatic muscle development, mesoderm development
cact		1 (20,47)	dorsal/ventral axis specification	aret		17 (11,39)	oogenesis, negative regulation of oskar mRNA translation, germ cell development
Tim17b2 *		2 (19,11)	protein targeting to mitochondrion	sesB		18 (11,31)	ATP:ADP antiporter; neuron and muscle homeostasis (next to Imp (nervous system development))
DI		3 (16,94)	Notch binding, neurogenesis	lh *		19 (11,10)	transmembrane transport (?) (next to conv (open tracheal system development))
nvy		4 (16,80)	transcription factor; axon guidance, dendrite morphogenesis; muscle organ development	cv-2		20 (11,05)	Dpp pathway, imaginal disc-derived wing vein specification
shn		5 (16,11)	transcription factor, ectoderm development	hbs		21 (10,88)	regulation of striated muscle tissue development
numb		6 (14,97)	protein localization; Notch pathway, muscle cell fate specification, regulation of neurogenesis	CG32306		22 (10,58)	unknown
Atg5 *		7 (14,46)	salivary gland cell autophagic cell death (between Dok (dorsal closure) and brk)	chn		23 (10,55)	transcription factor; peripheral nervous system development
CG31190 !		8 (13,10)	cell adhesion(?)	CG13833 *		24 (10,55)	oxidation reduction (?) (next to lmd (muscle organ development))
lbe		9 (12,93)	transcription factor; segment polarity, muscle organ development, dendrite morphogenesis	psq		25 (10,45)	DNA binding; imaginal disc-derived wing morphogenesis
vnd		10 (12,32)	neuroblast development	emc		26 (10,23)	transcription repressor; sensory organ development, neuromuscular process
CG3394 *		11 (12,32)	long-chain fatty acid transporter activity (?) (between nvy and betaTub60D (axonogenesis))	CG33900		27 (10,08)	transcription factor, development of central neuroendocrine system
pxb		12 (12,22)	mushroom body development	CG32060 !		28 (10,05)	unknown
CG4998 !		13 (11,85)	proteolysis (?)	Ptp99A		29 (9,94)	motor axon guidance; defasciculation of motor neuron axon
E2f		14 (11,78)	transcription factor; neuron development, muscle tissue development	Gadd45		30 (9,77)	JNK cascade (next to Ady43A*)
tup		15 (11,57)	transcription factor; torso signaling pathway, motor axon and dendrite guidance, dorsal closure	sog		31 (9,71)	torso signaling pathway; dorsal/ventral axis specification

Top Genes: GO-statistics

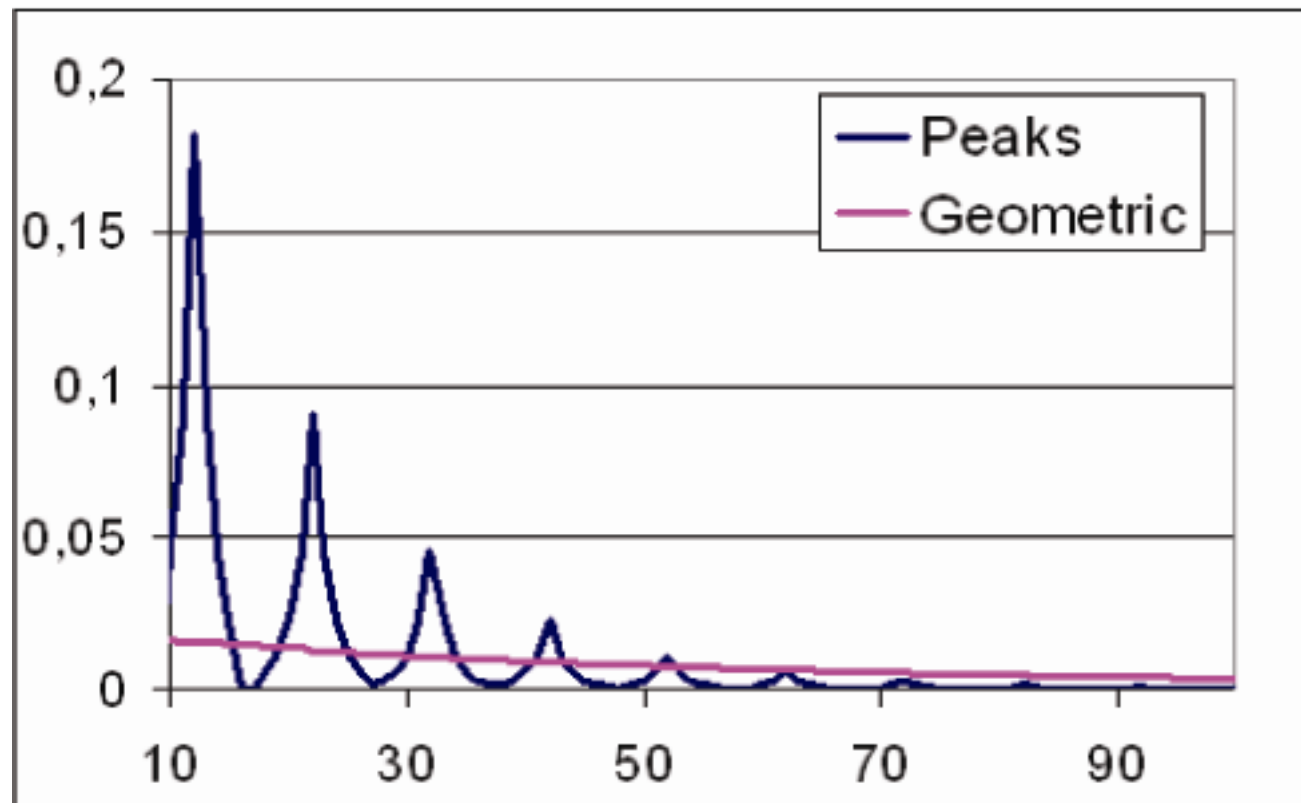
GO Term	Top 32 genes (30)	All genes: 10558	P-Value
organ development	18	1101	2,94E-08
regionalization	10	407	5,49E-06
patten specification process	10	426	7,56E-06
regulation of transcription, DNA-dependent	9	390	3,45E-05
nervous system development	10	565	7,34E-05
imaginal disc-derived wing vein morphogenesis	4	36	9,53E-05
imaginal disc-derived appendage morphogenesis	7	237	9,64E-05

THANK YOU ! 😊

HMM Scheme

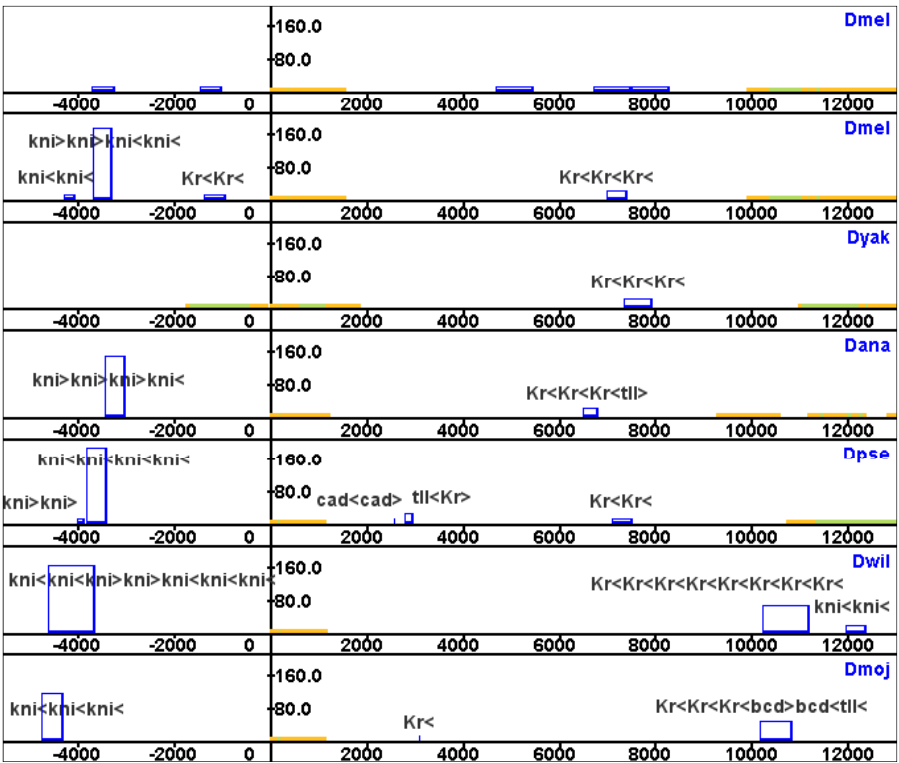


Inter-site Model Distance Distributions

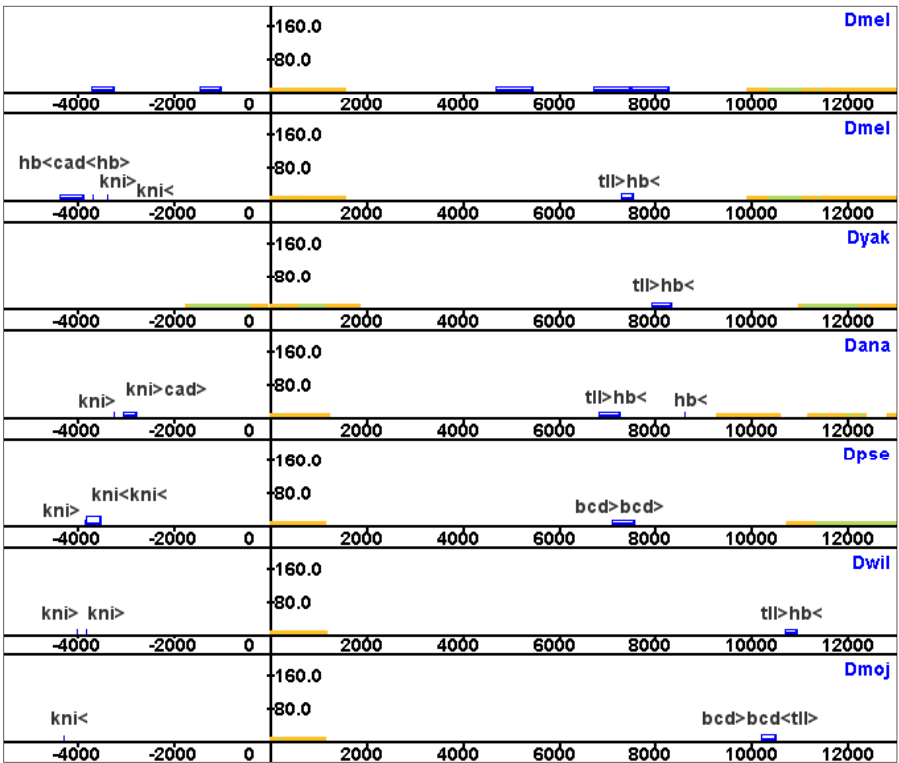


Clusters prediction: *eve*

(h – trained model)



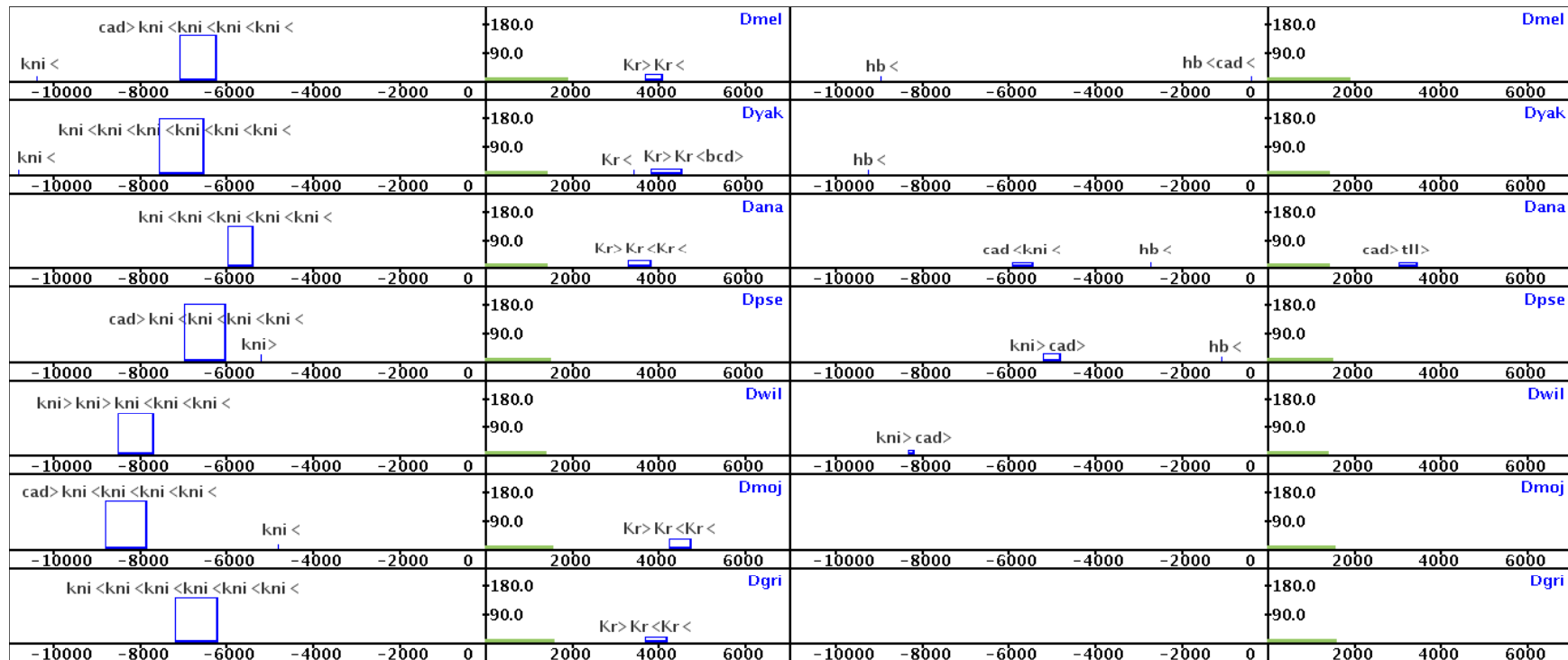
(Kr – trained model)



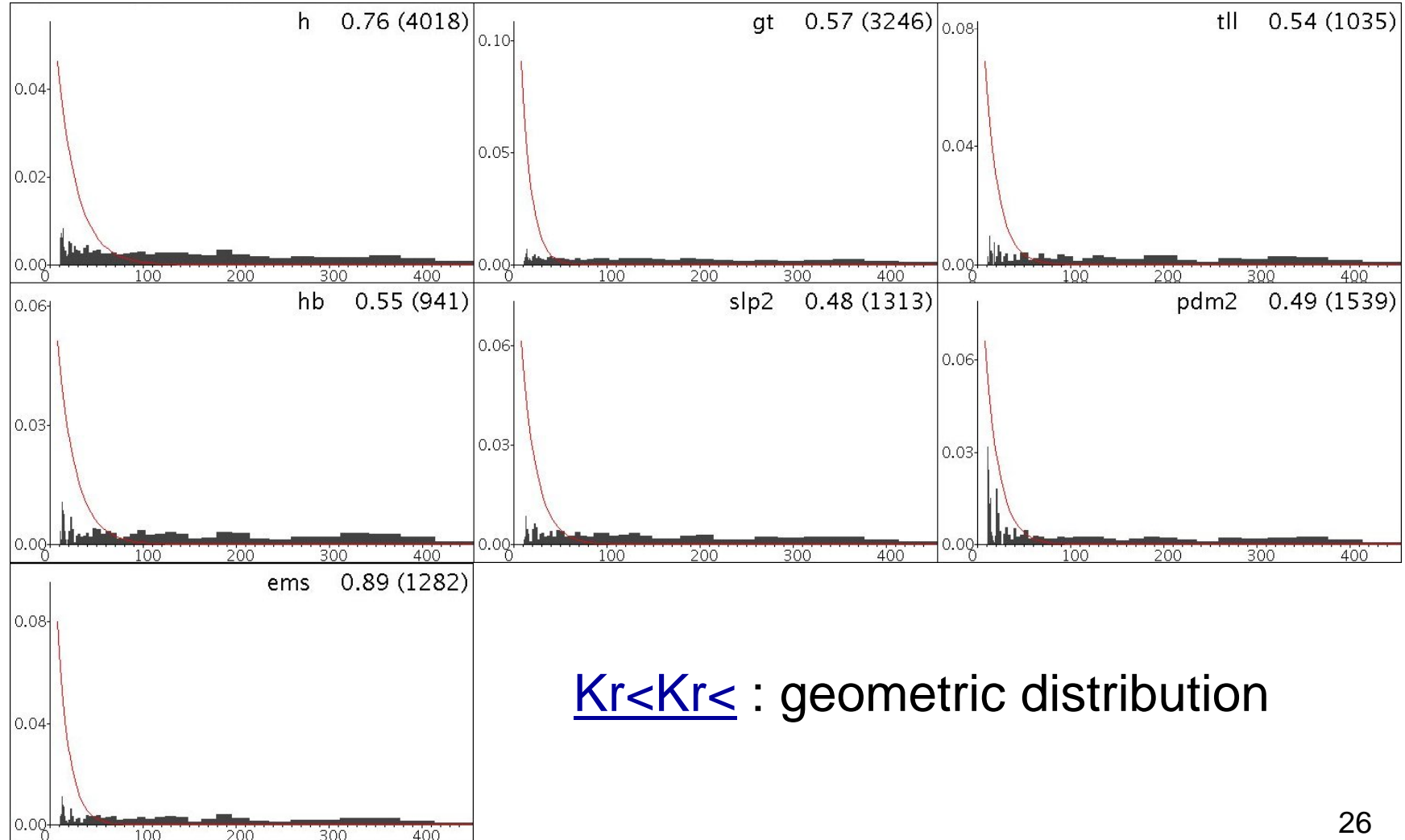
Clusters prediction: *ftz*

(h – trained model)

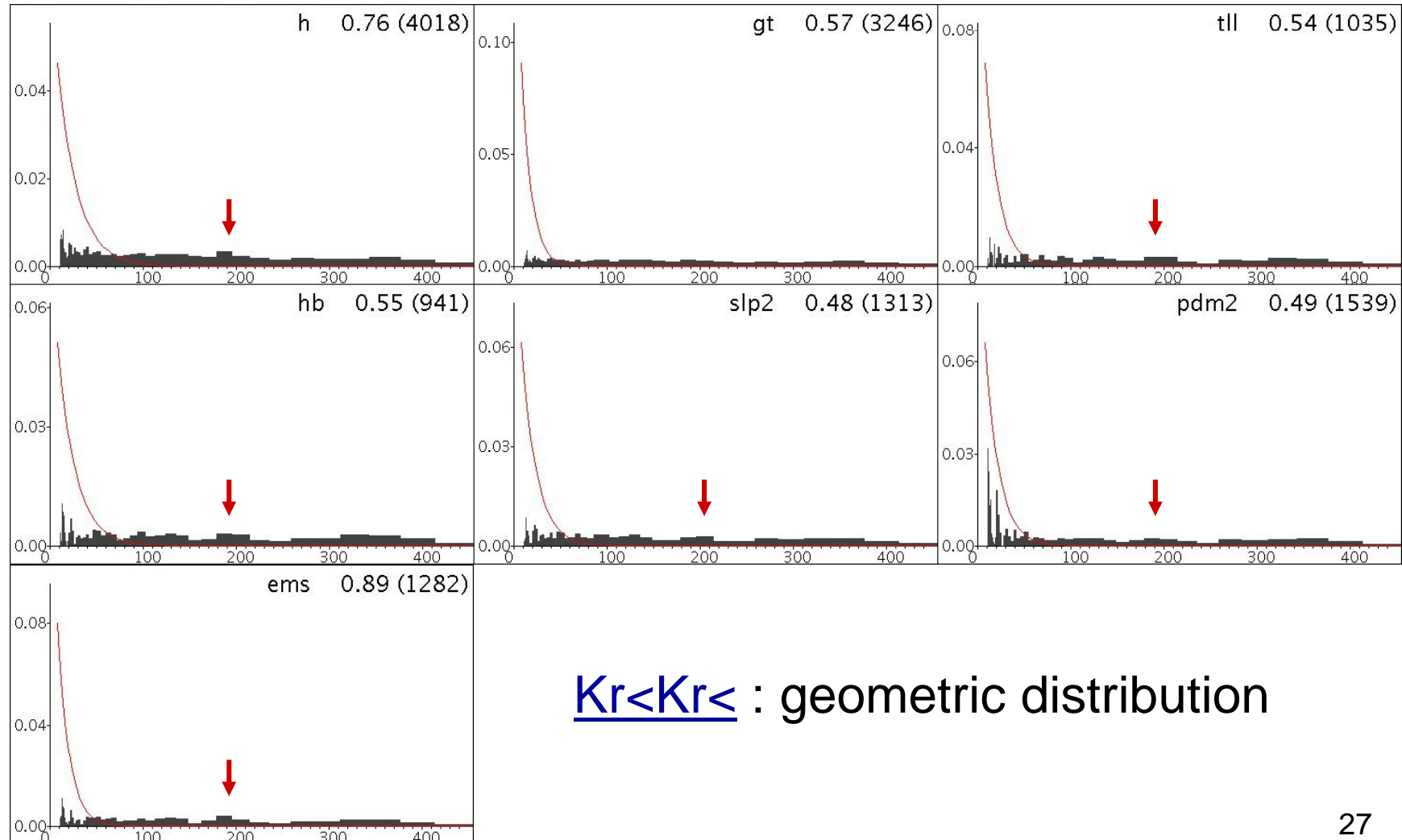
(Kr – trained model)



Cluster structure: site-to-site distances

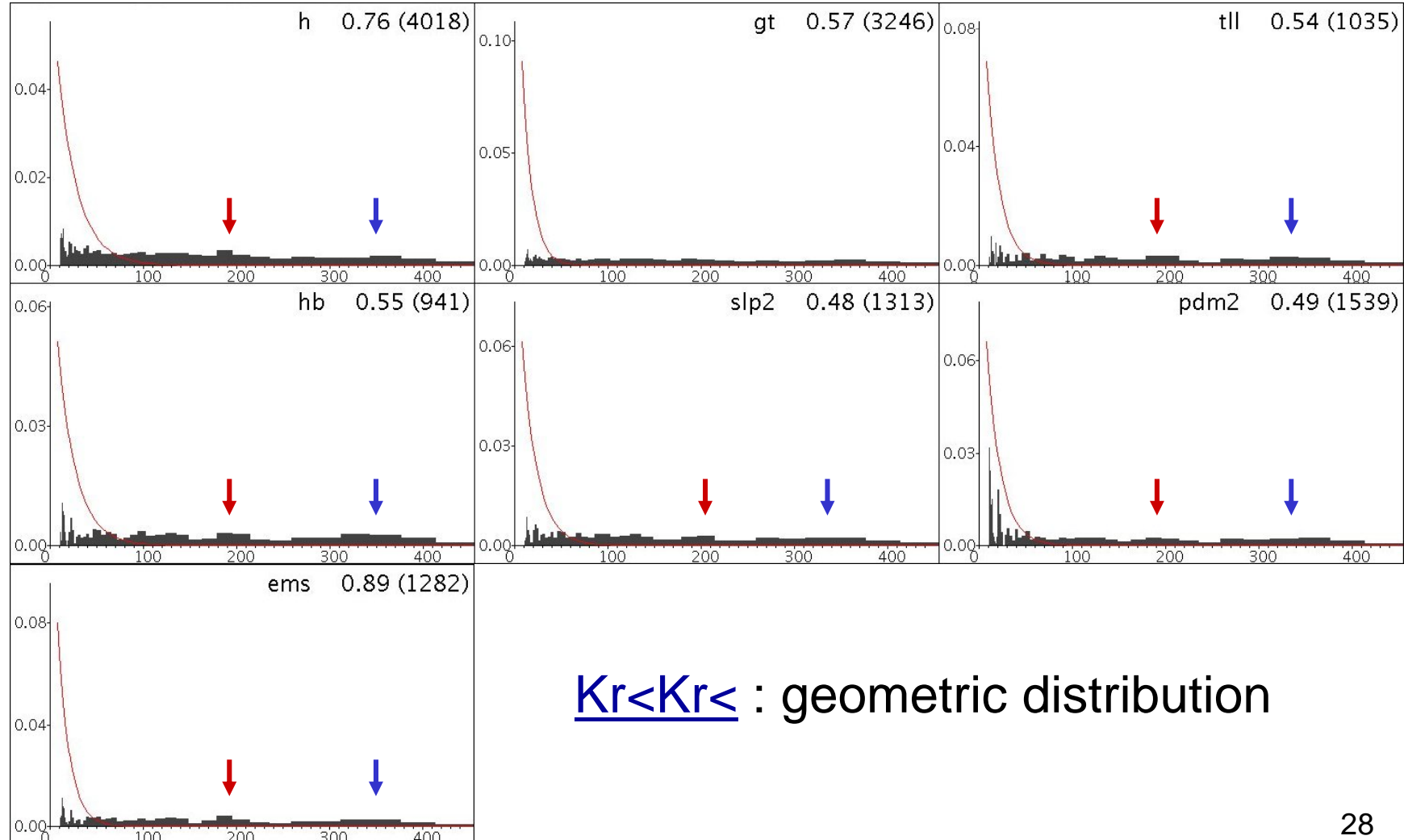


Cluster structure: site-to-site distances



$K_{r < K_r <}$: geometric distribution

Cluster structure: site-to-site distances

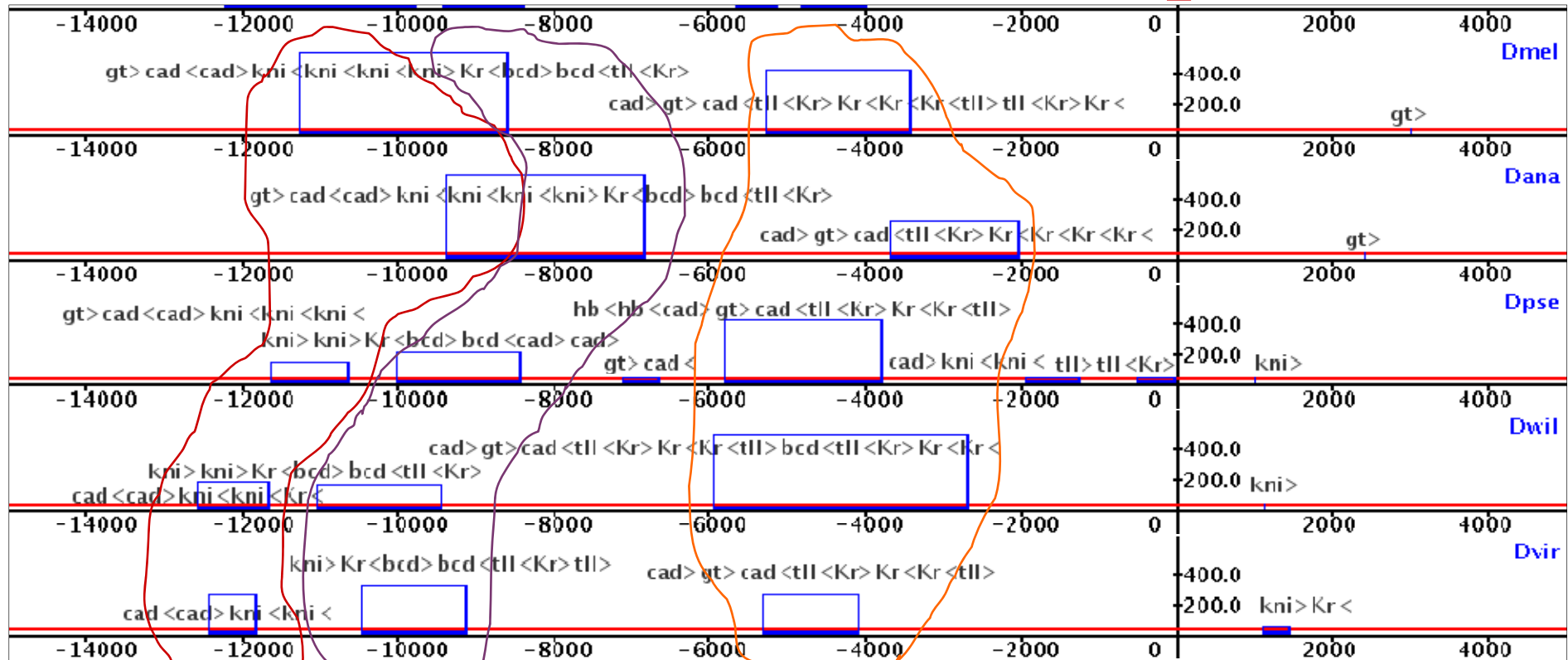


$K_{r < K_{r <}}$: geometric distribution

Example: pair-rule gene *hairy* (*h*)



stripe 3+7 stripe 6 stripe 1+5



kni+cad

kni+cad+bcd+Kr+tll

Kr+(tll+cad+gt)

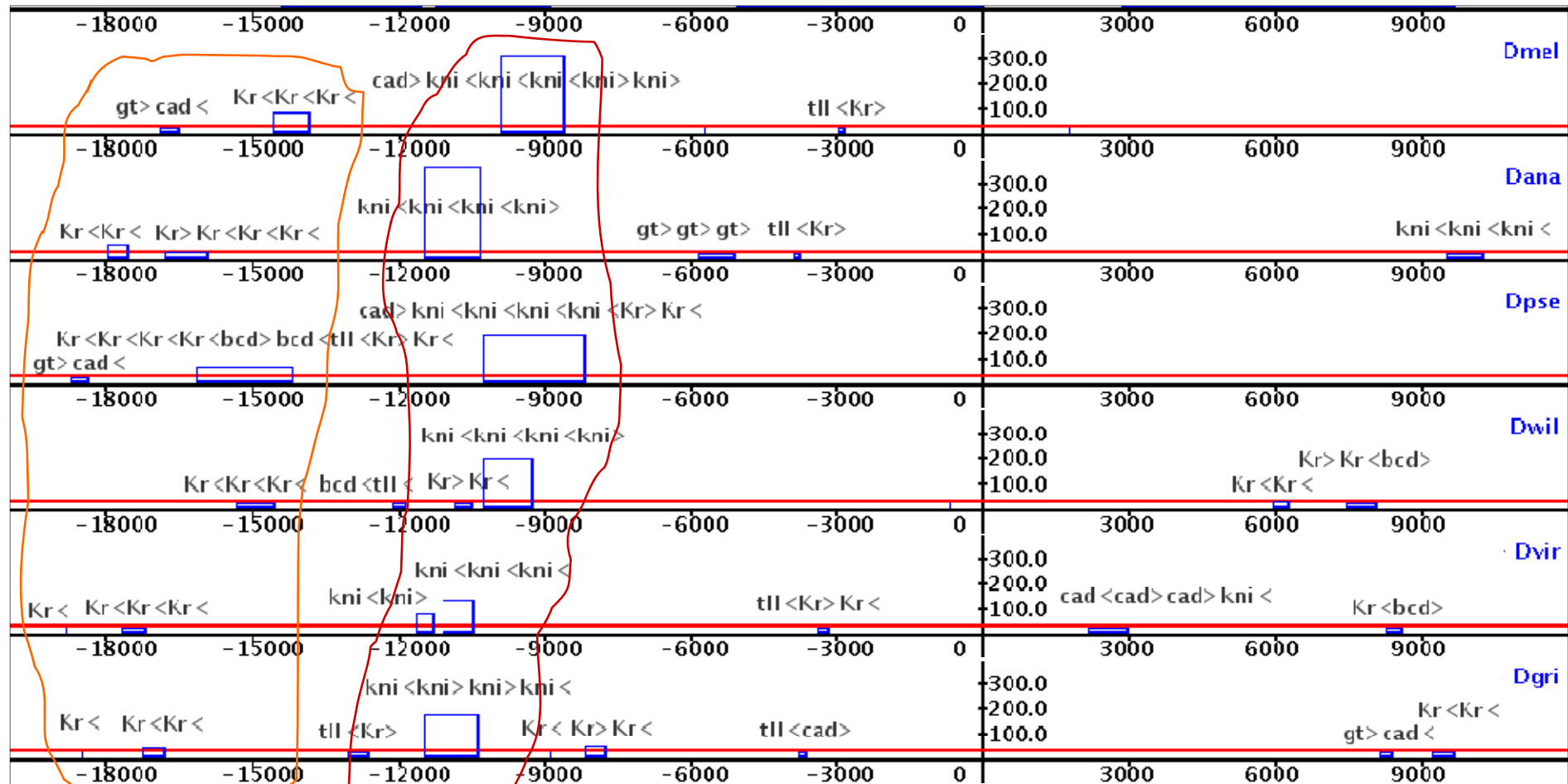
Example: pair-rule gene *runt* (*run*)



stripe 1+5

stripe 3+7

stripe 6

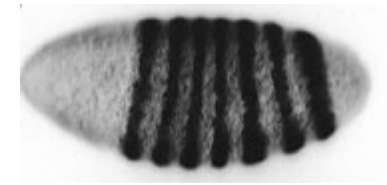


Kr+(tll+cad+gt)

kni+(cad)

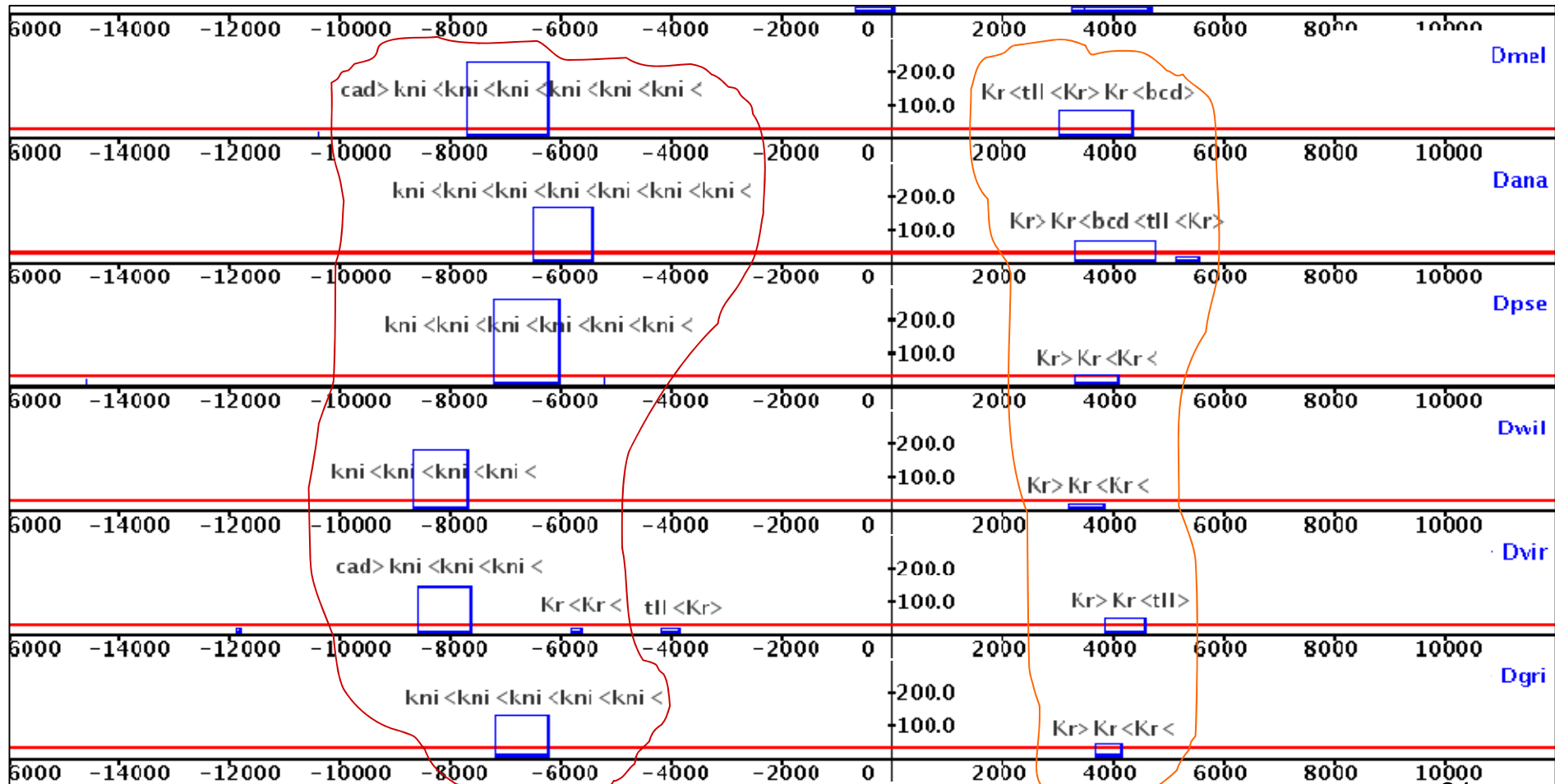
kni+cad+bcd+Kr+tll

Example: pair-rule gene *fushi tarazu* (*ftz*)



stripe 3+7 ???

stripe 1+5



kni+(cad)

Kr+(tll+bcd)