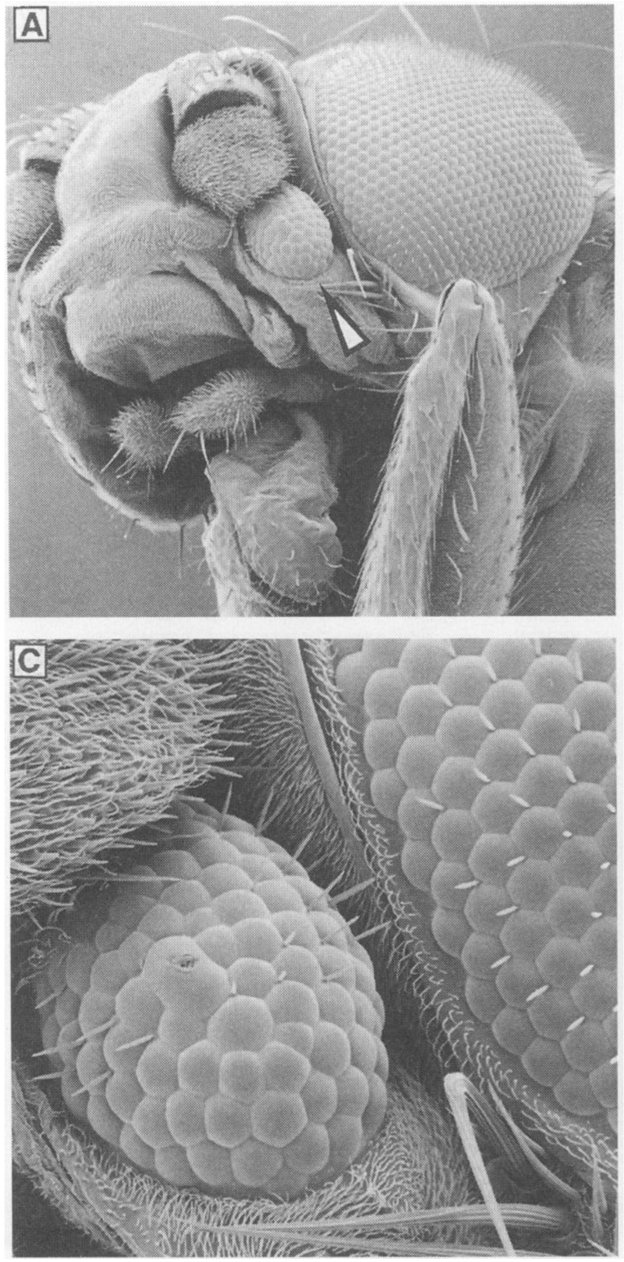
De novo identification of functionally related cis-regulatory sequences in evolutionarily distant species

Hervé Rouault

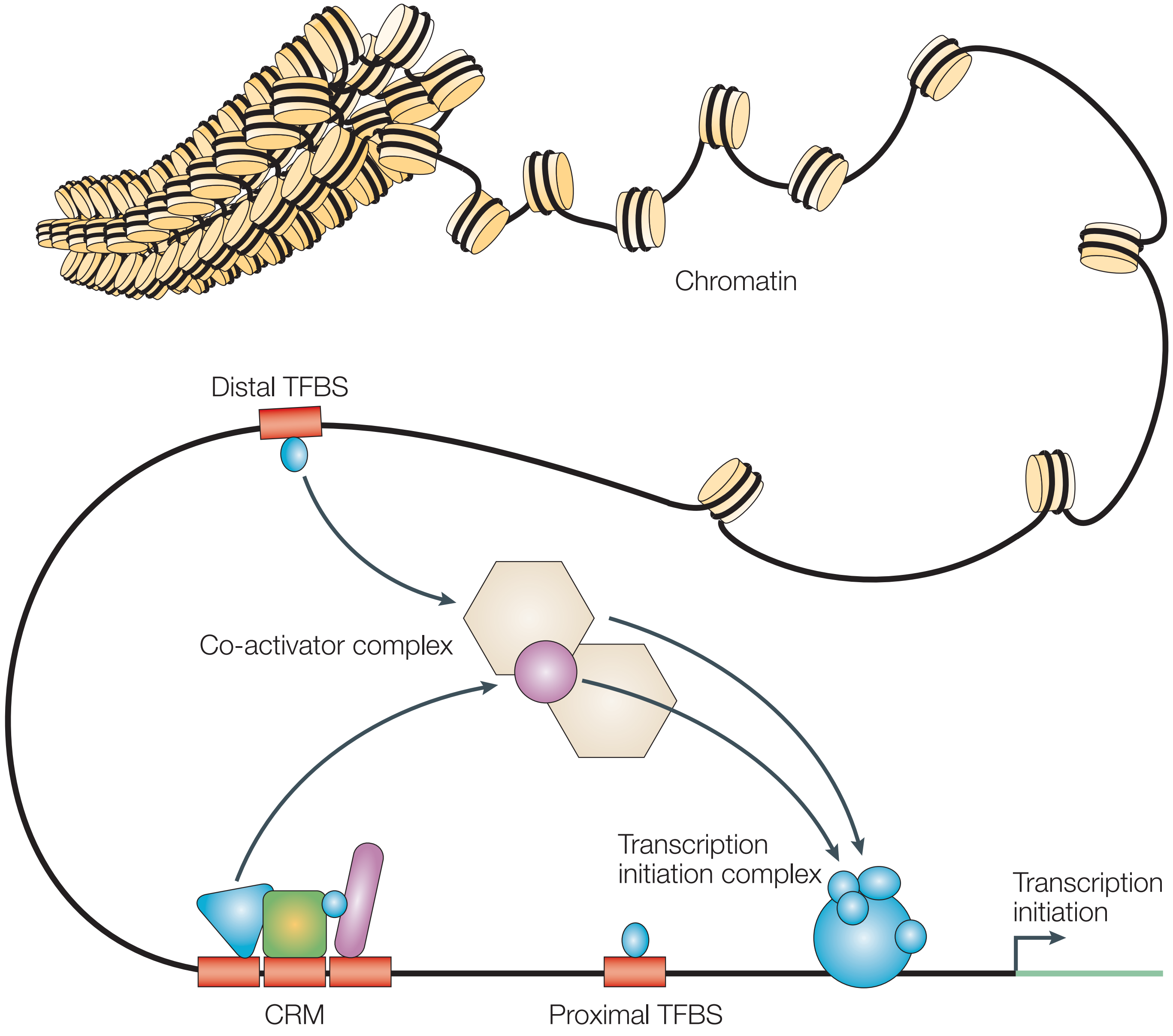
GDD, Institut Pasteur
Paris, France

Inference of the cis-regulatory code... Why?



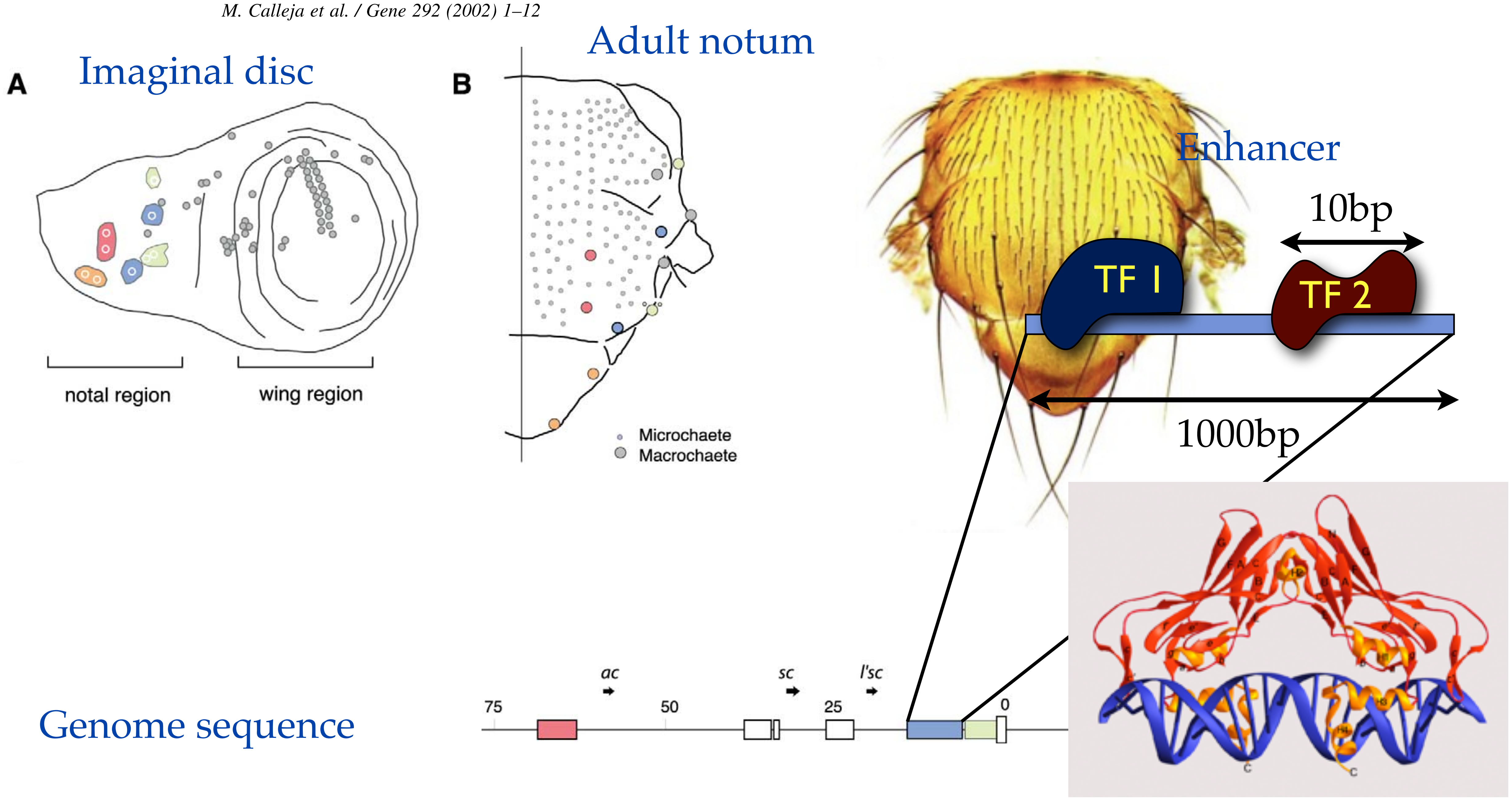
La part créative de l'évolution biochimique ne se fait pas à partir de rien. Elle consiste à faire du neuf avec du vieux. (F. Jacob)

Transcription regulation



Wasserman and Sandelin (Nat Rev Gen, 04)

Structure of the cis-regulatory code



M. Calleja et al. / Gene 292 (2002) 1–12

Imaginal disc

Adult notum

Enhancer

10bp

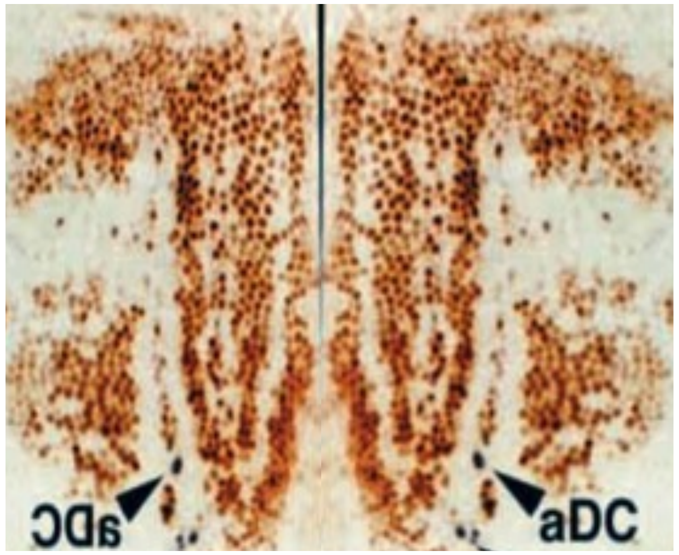
TF 1

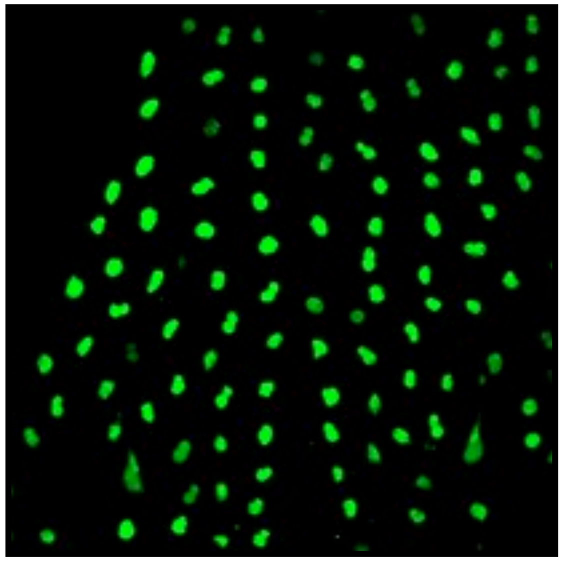
TF 2

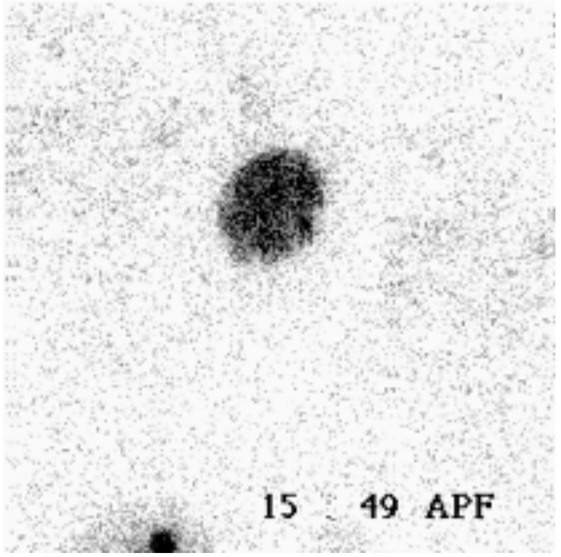
1000bp

Genome sequence

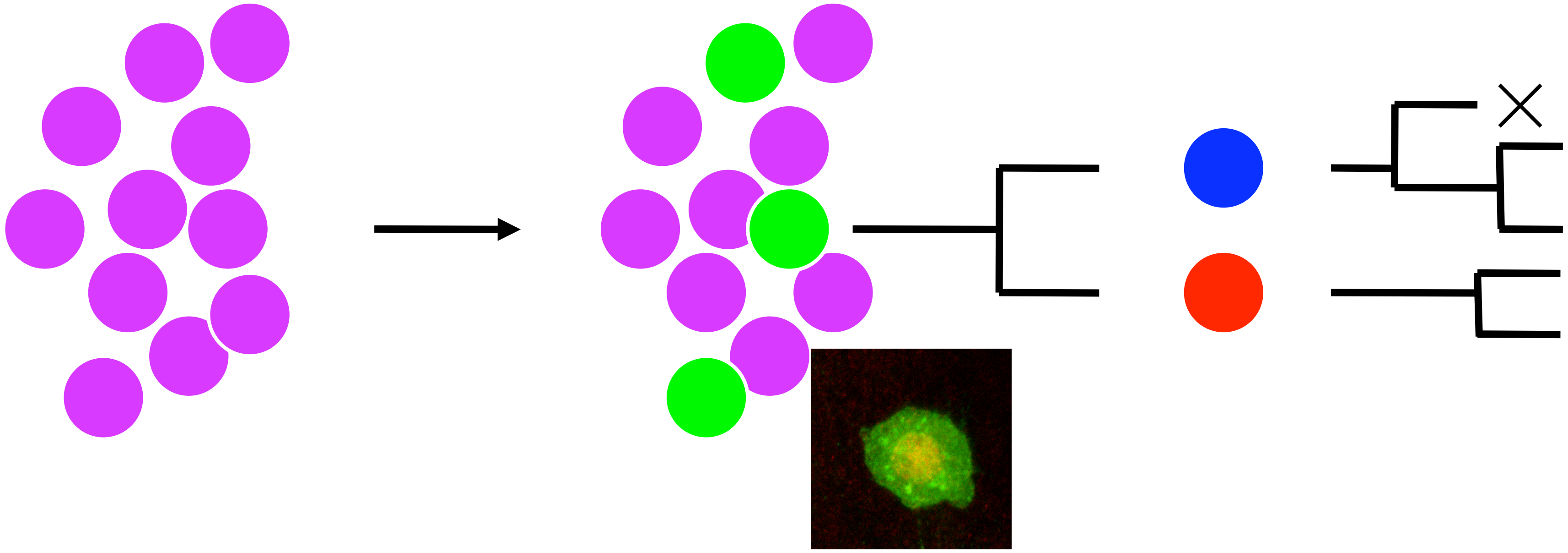
Each sensory organ develops from a single multipotent progenitor cells via a stereotyped lineage







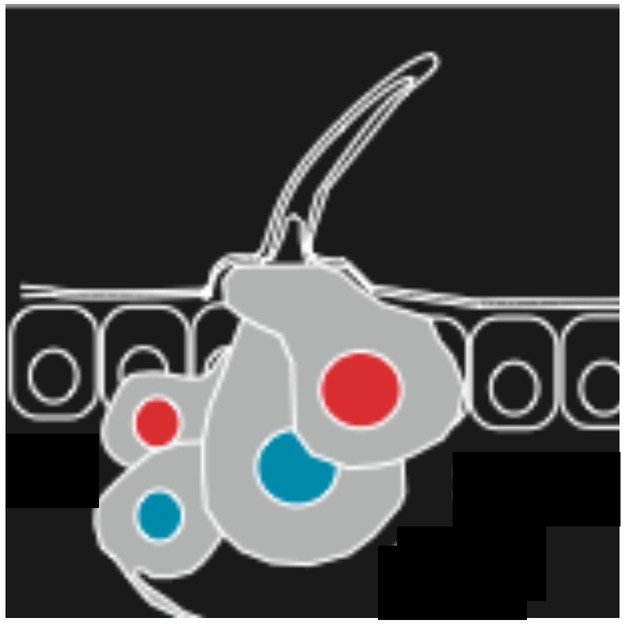
M. Gho

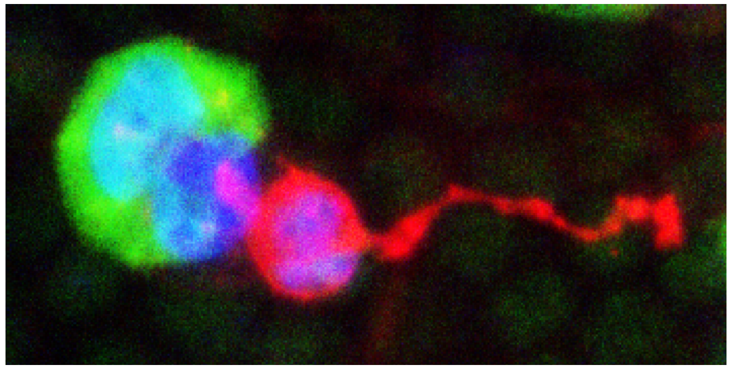


proneural
cluster cells

singling out of
sensory organ
precursor cells
(SOPs)

asymmetric
cell division





In silico determination of *a priori* unknown cis-regulatory motifs

- Why?

  General issues

  - regulation at the transcriptional level
  - small quantity of biological materials + heterogeneous

  Specific advantages

  - 12 Drosophila genomes
  - 8 SOPs enhancers have been characterized, many remain to be determined
  - 9 SOP-specific TF are known
  - experimental test *in vivo* is feasible with a reasonable investment of resources
- How?
  - Search for over-represented motifs in a set of characterized enhancers and homologous fragments from other *Drosophila* species
  - Use these motifs to look for new SOP-specific enhancers within the *D. melanogaster* genome

General idea of the approach

Training set
D. melanogaster CRMs + Homologous sequences in the other 11 *Drosophilae*

Representation of DNA binding: PWMs and motifs

a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

b

B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

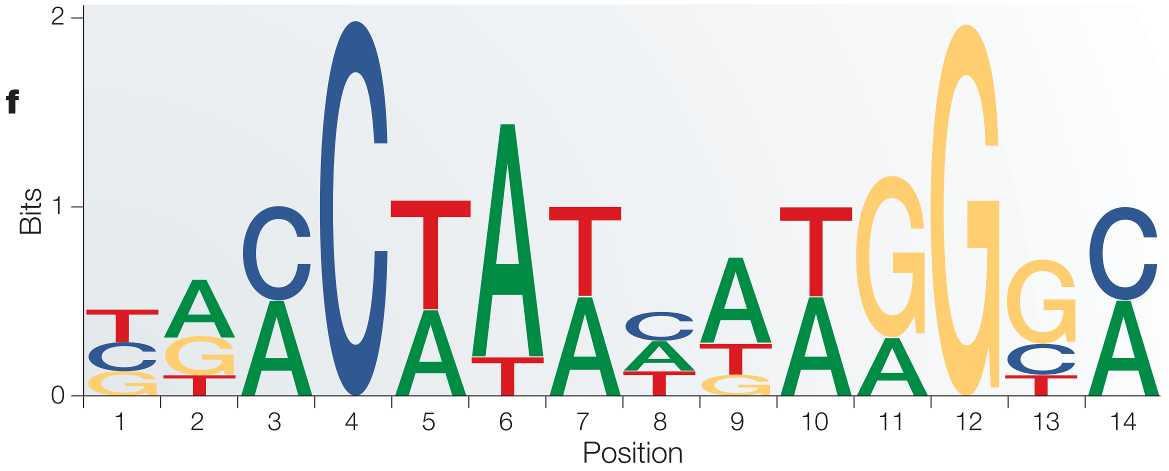
d Position weight matrix (PWM)

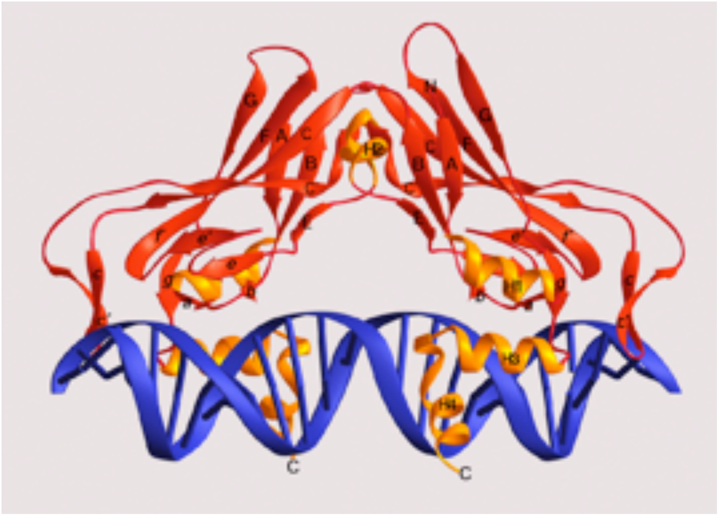
A	–1.93	0.79	0.79	–1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	–1.93	–1.93	0.79
C	0.45	–1.93	0.79	1.68	–1.93	–1.93	–1.93	0.45	–1.93	–1.93	–1.93	–1.93	0.00	0.79
G	0.00	0.45	–1.93	–1.93	–1.93	–1.93	–1.93	–1.93	0.66	–1.93	1.30	1.68	1.07	–1.93
T	0.15	0.66	–1.93	–1.93	1.07	0.66	0.79	0.00	0.00	0.79	–1.93	–1.93	–0.66	–1.93

e Site scoring

0.45	–0.66	0.79	1.68	0.45	–0.66	0.79	0.45	–0.66	0.79	0.00	1.68	–0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

$\Sigma = 5.23$, 78% of maximum





$$\epsilon_{i\alpha} = \log_2 \frac{w_{i\alpha}}{f_\alpha}$$

We define a score threshold Sth!

In silico determination of motifs and modules

i) Known PWMs, difficult task:

-binding alone does not predict functional importance (Wasserman’s « futility theorem »)

-need to take into account other informations :

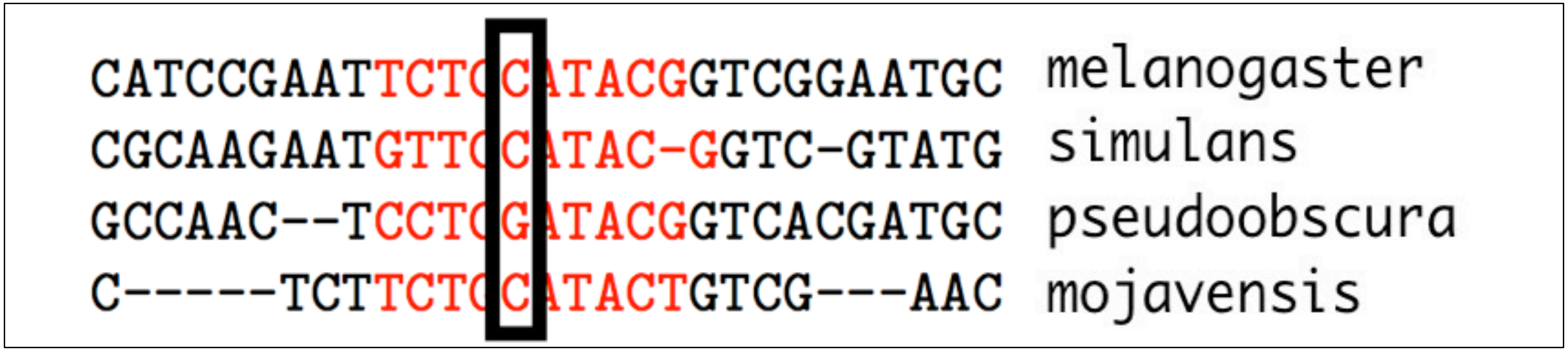
clustering of binding sites, conservation,…..

ii) PWMs unknown, even more difficult task:

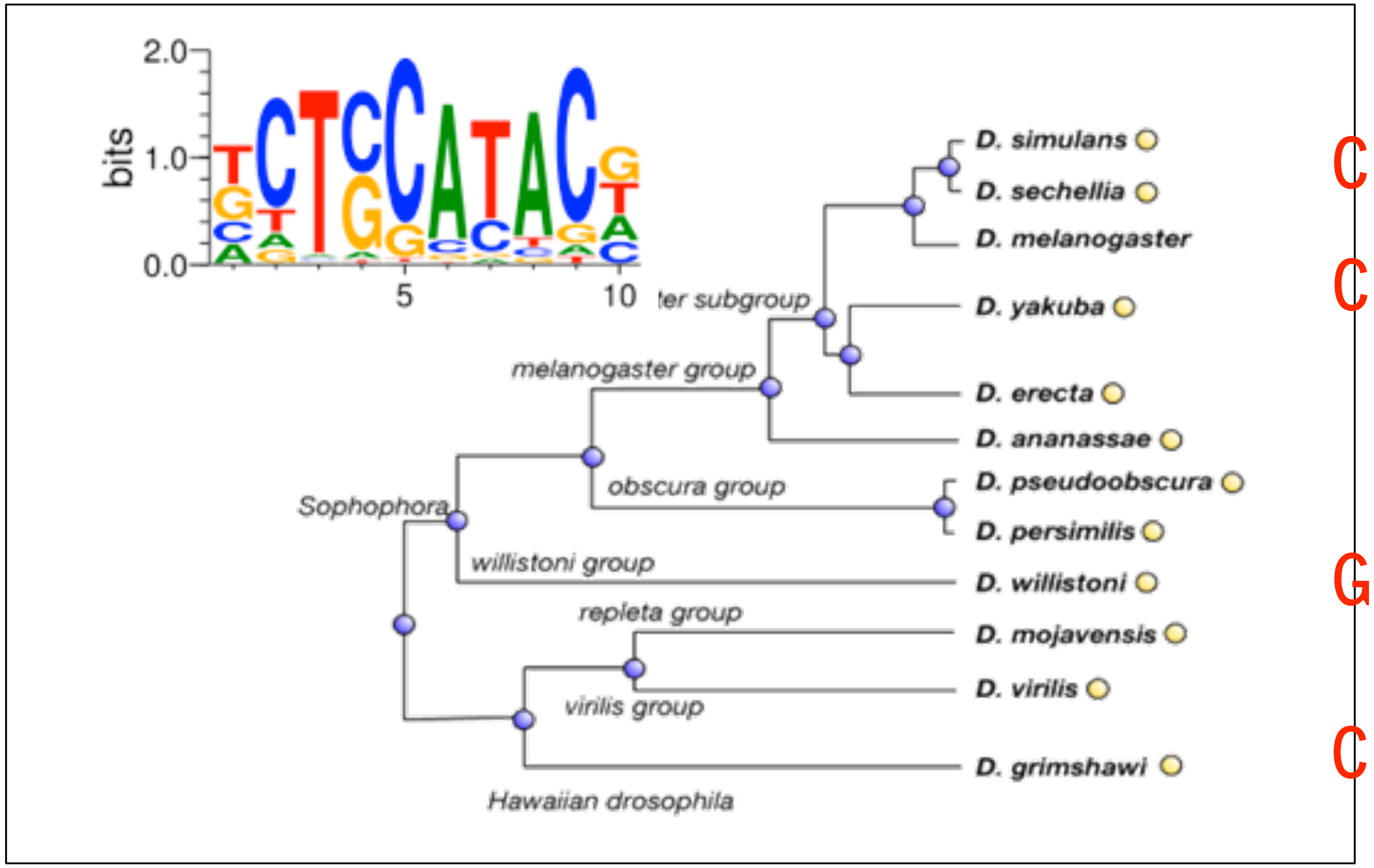
-need a training set

- use the statistics of small sequences on the training set to distinguish regulatory modules from background

Motif discovery takes into account the evolutionary distance



```
CATCCGAATTCTCCATACGGTCGGAATGC melanogaster
CGCAAGAATGTTCCATAC-GGTC-GTATG simulans
GCCAAC--TCCTCGATACGGTCACGATGC pseudoobscura
C-----TCTTCTCCATACTGTCG---AAC mojavensis
```



C

close species:
approx. 1 observation

C

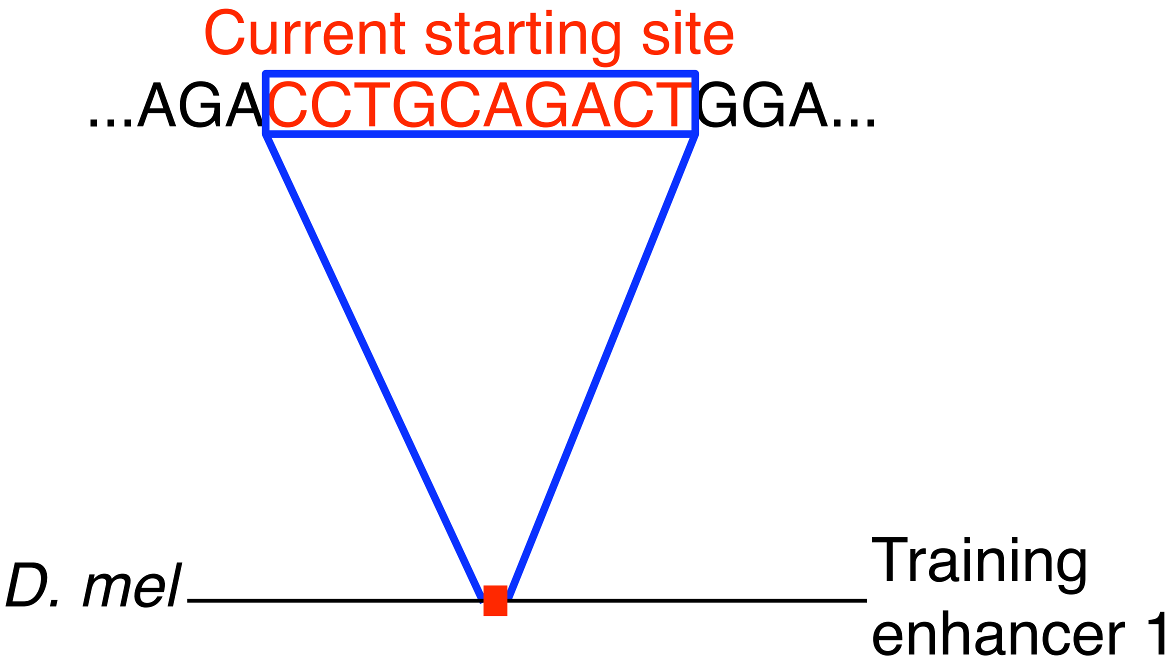
G

distant species:
approx. 2 observations

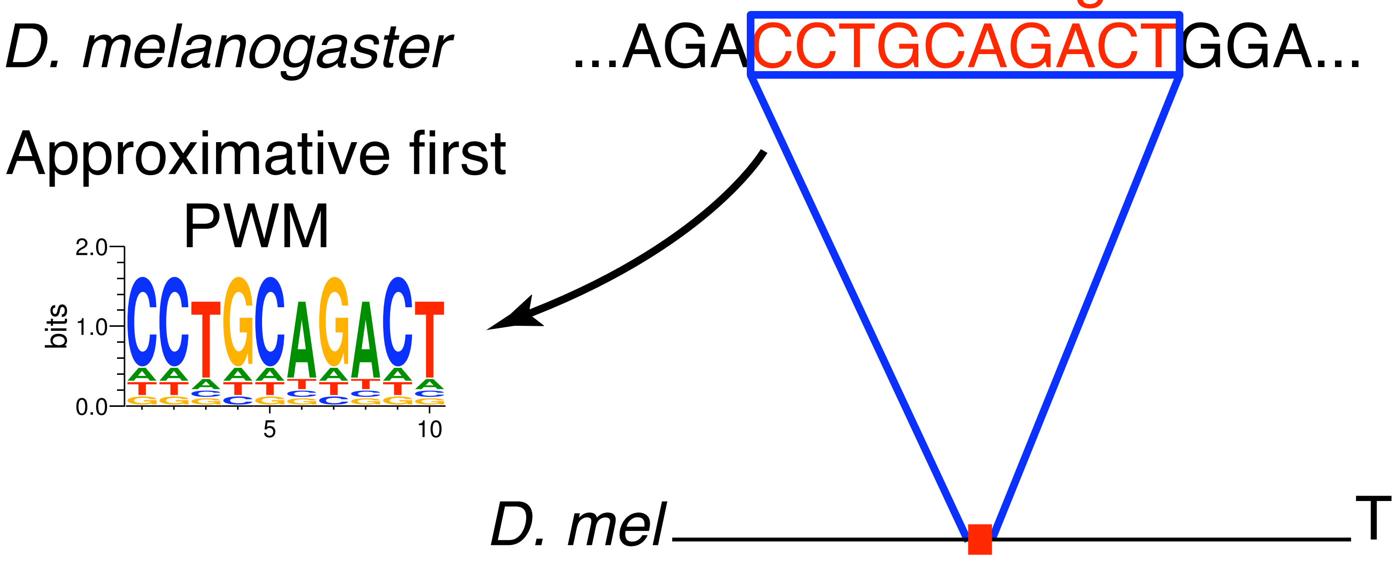
C

« Les mouches d'aujourd'hui ne sont plus les mêmes que les mouches d'autrefois… »
R Queneau

Motif discovery
step 1: creation of a list of PWMs, i.e. motifs



Motif discovery
step 1: creation of a list of PWMs, i.e. motifs



Current starting site

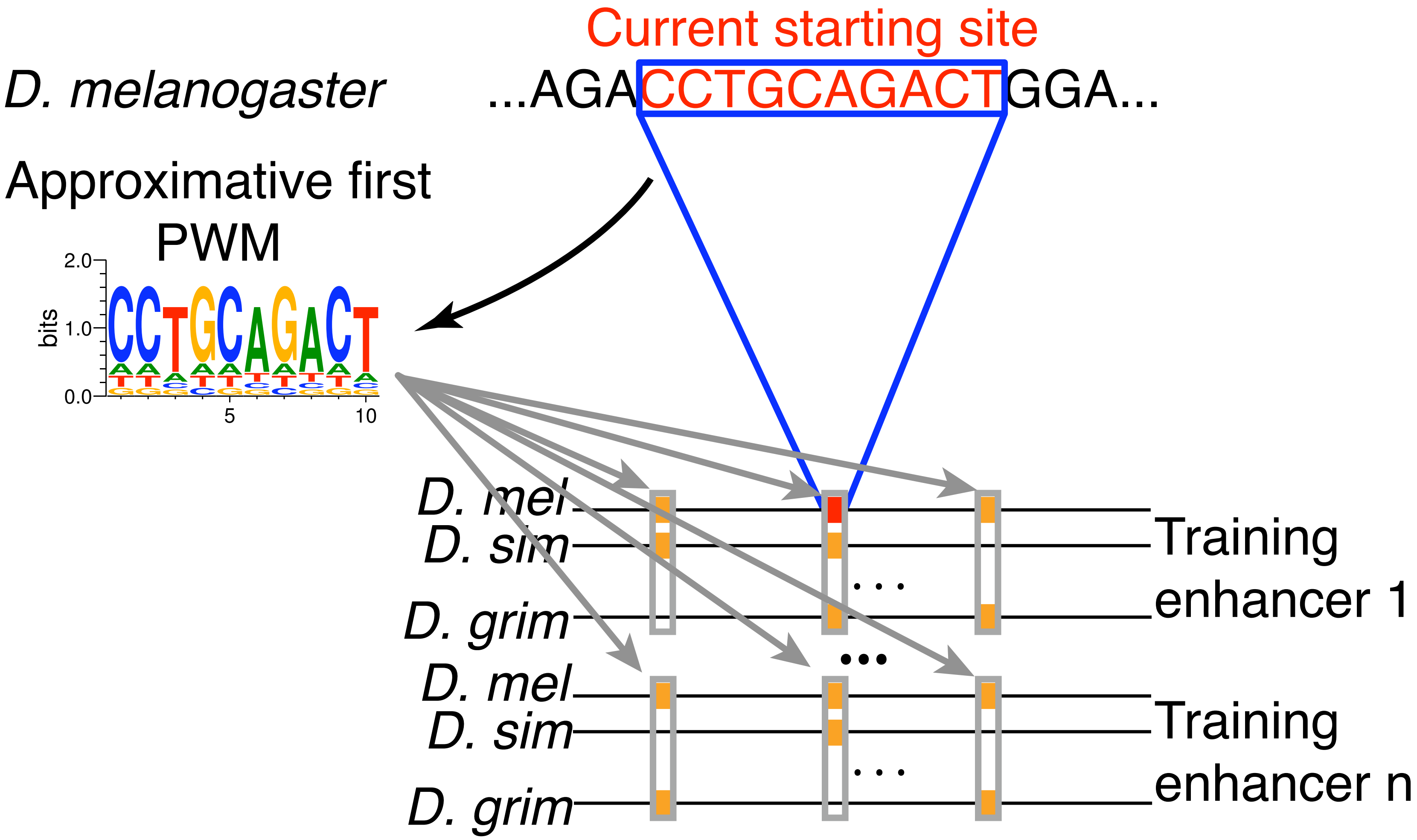
D. melanogaster ...AGACCTGCAGACTGGA...

Approximative first PWM

D. mel

Training enhancer 1

Motif discovery
step 1: creation of a list of PWMs, i.e. motifs



Current starting site

D. melanogaster ...AGACCTGCAGACTGGA...

Approximative first PWM

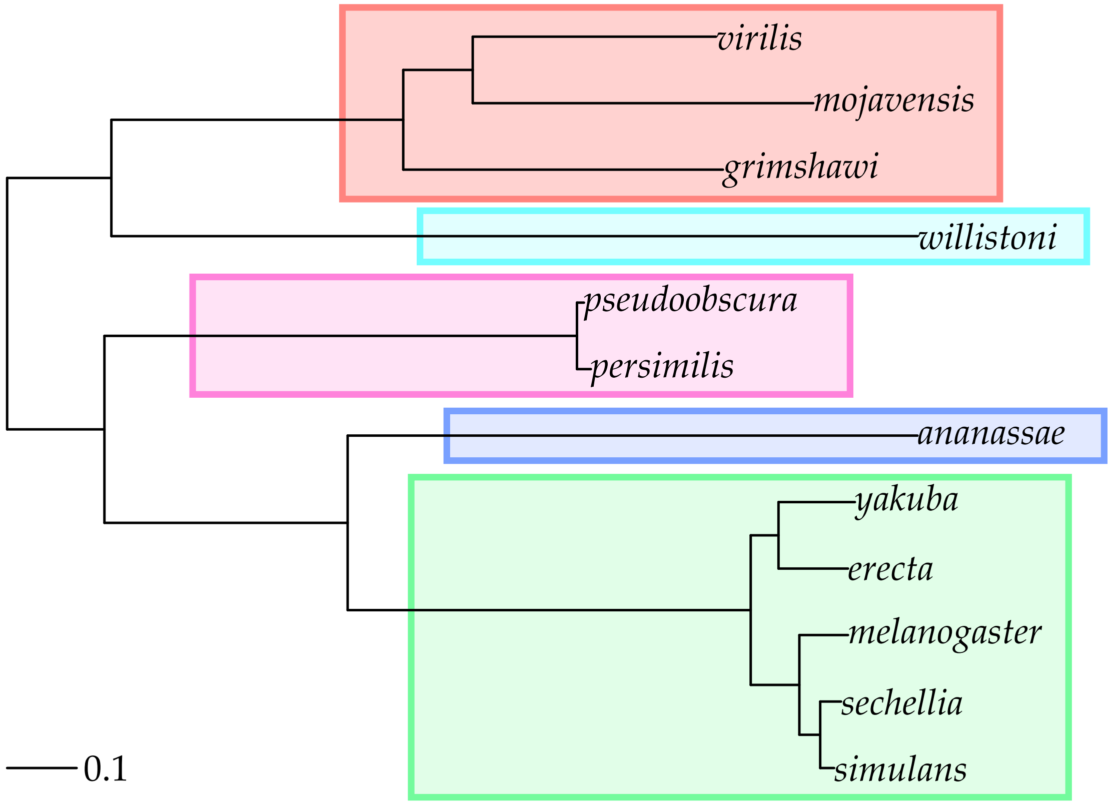
D. mel
D. sim
D. grim
Training enhancer 1

D. mel
D. sim
D. grim
Training enhancer n

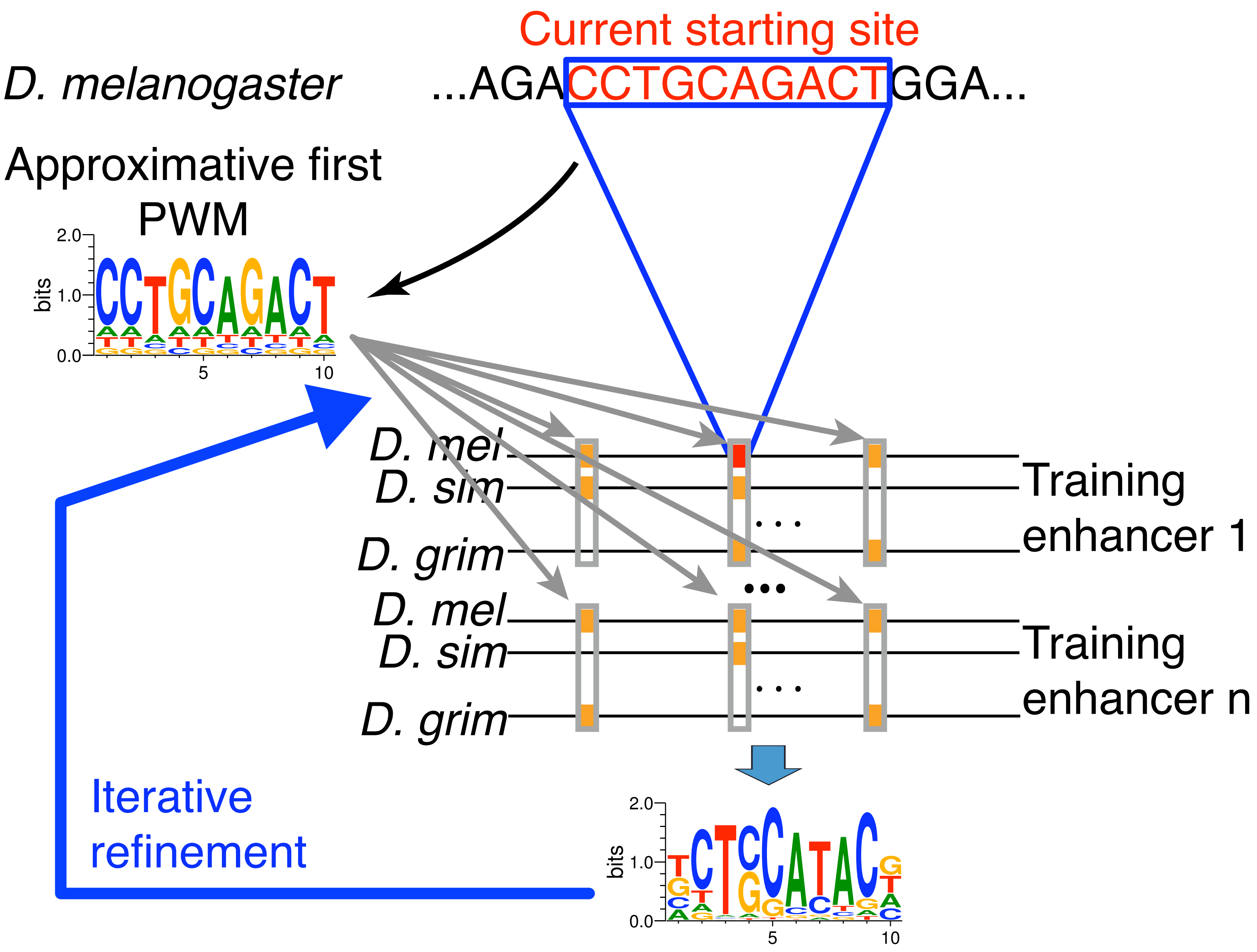
High selectivity imposed by a high score threshold, i.e. set to recognize 0.1-0.5 sites per 10 kb depending on sequence composition

Scanning is done on both strands of the chromosomes.

Defining evolutionarily related groups



Motif discovery
step 1: creation of a list of PWMs, i.e. motifs



Current starting site

D. melanogaster ...AGACCTGCAGACTGGA...

Approximative first PWM

D. mel

D. sim

D. grim

Training enhancer 1

D. mel

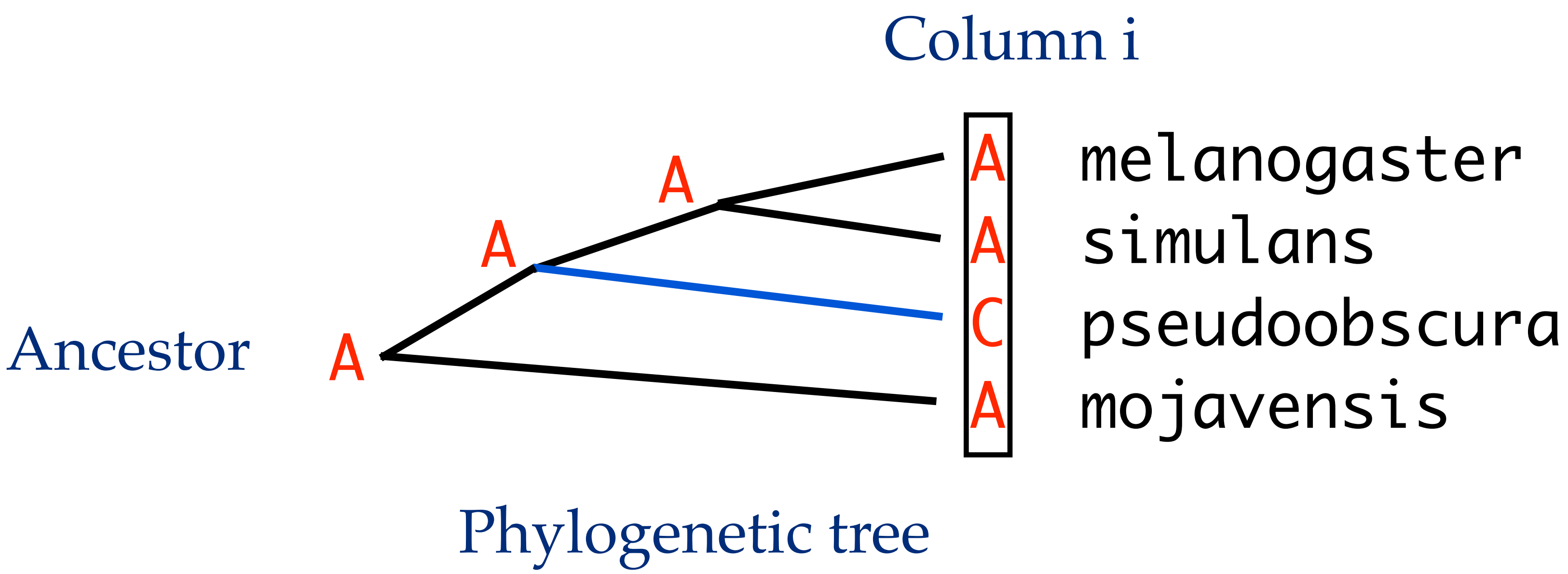
D. sim

D. grim

Training enhancer n

Iterative refinement

Matrix inference



Column i

Ancestor A

A

A

A melanogaster

A simulans

C pseudoobscura

A mojavensis

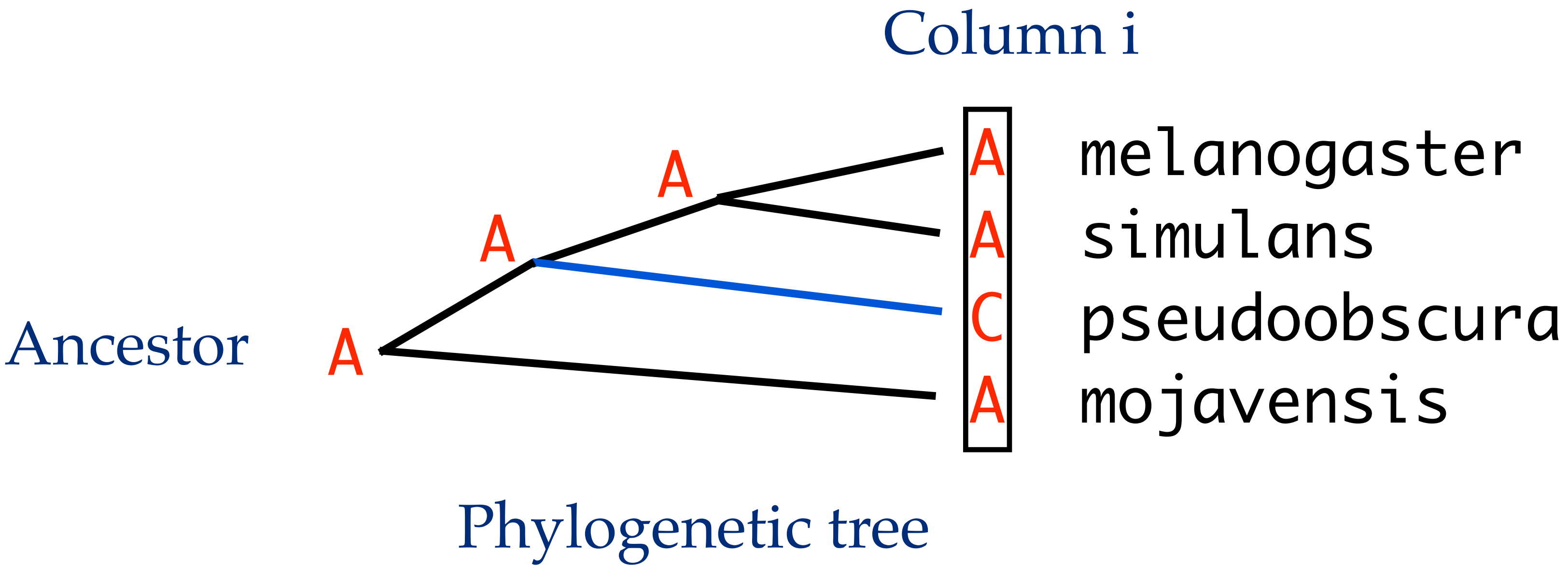
Phylogenetic tree

$p(w|\text{set of sites})$ is obtained by Bayes' theorem

$$p(w|\text{set of sites}) \propto p(\text{set of sites}|w)p(w)$$

$$p(\text{set of sites}|w) = \prod_{\text{sites } S_i} p(S_i|w)$$

Matrix inference

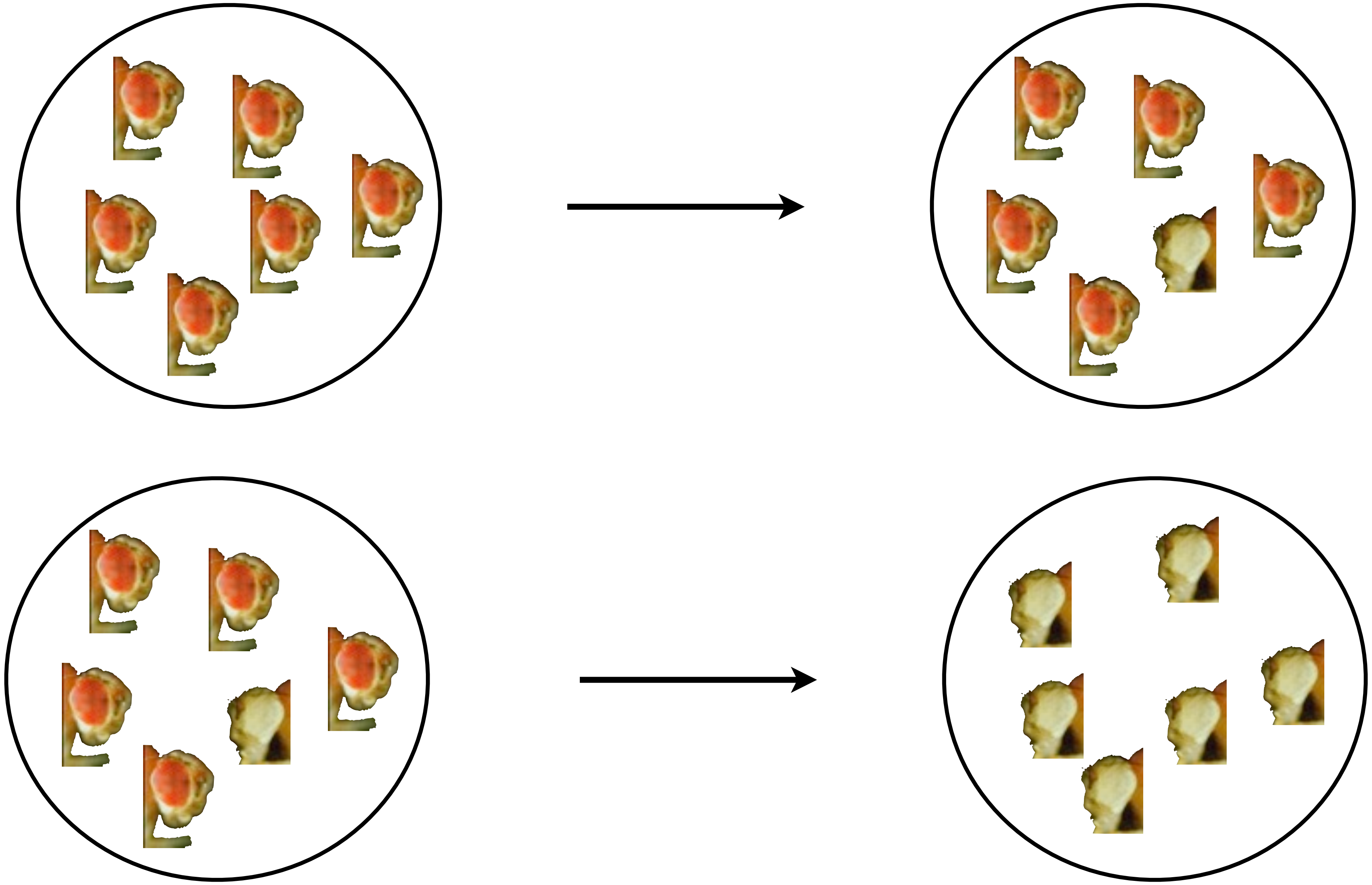


$$p(S_i|w)\ ?$$

Felsenstein '81 $$p(B \to B') = q\delta_{B,B'} + (1-q)w_{B'}$$

$$q = \exp\left(-\frac{d}{1/2 + 4\pi_{A,T}\pi_{C,G}}\right)$$

A refined model of evolution : Halpern and Bruno '98



$$p(A \longrightarrow C) = p^{\text{apparition}}(A \longrightarrow C) \quad \text{x} \quad p^{\text{fixation}}(A \longrightarrow C)$$

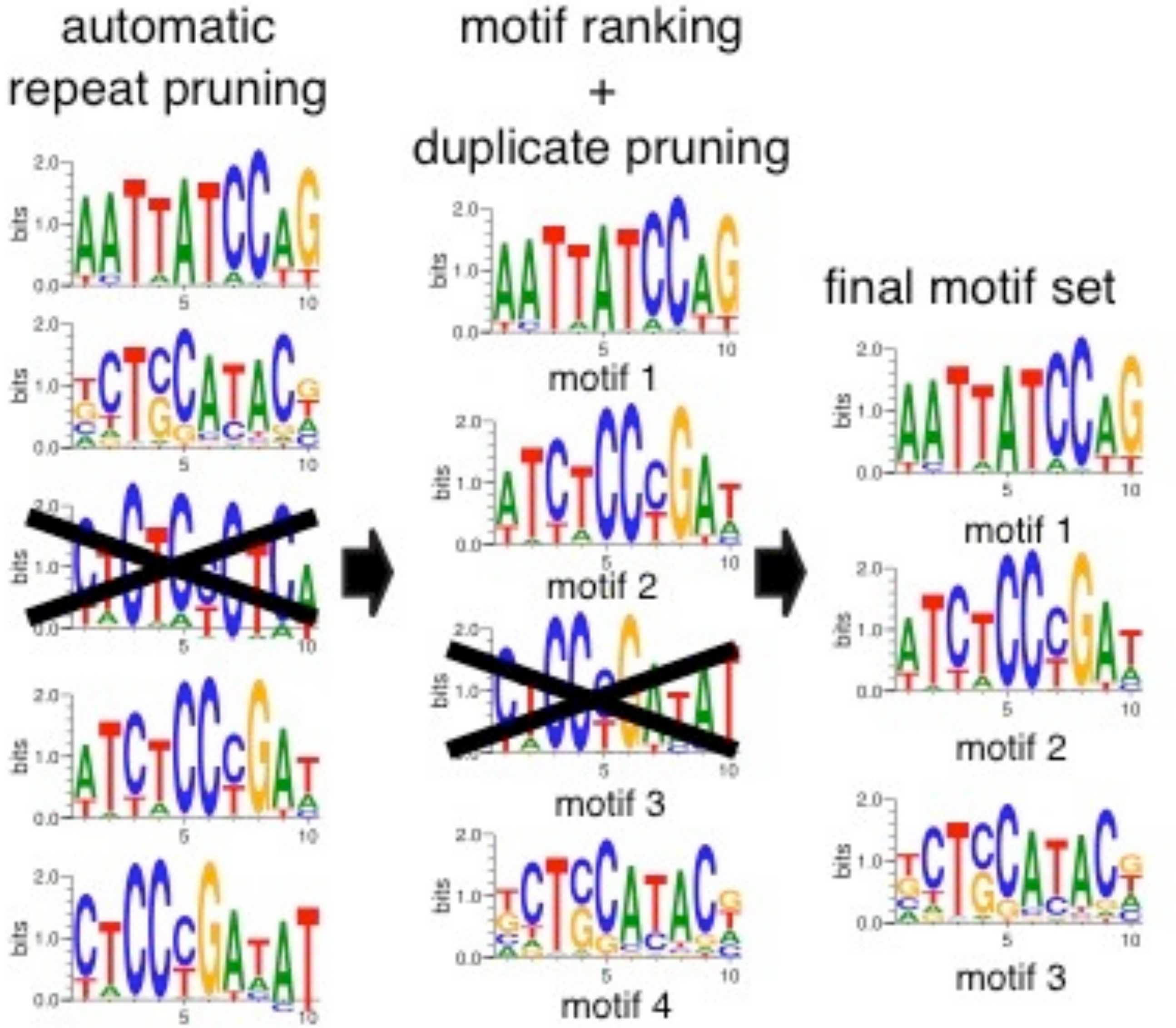
A refined model of evolution : Halpern and Bruno ’98

$$p(A \rightarrow C) = \; p^{\text{apparition}}(A \rightarrow C) \quad \text{x} \quad p^{\text{fixation}}(A \rightarrow C)$$

Kimura ’69 $$p^{\text{fixation}}(A \rightarrow C) = \frac{1 - e^{-4s}}{1 - e^{-4Ns}} \approx \frac{4s}{1 - e^{-4Ns}}$$

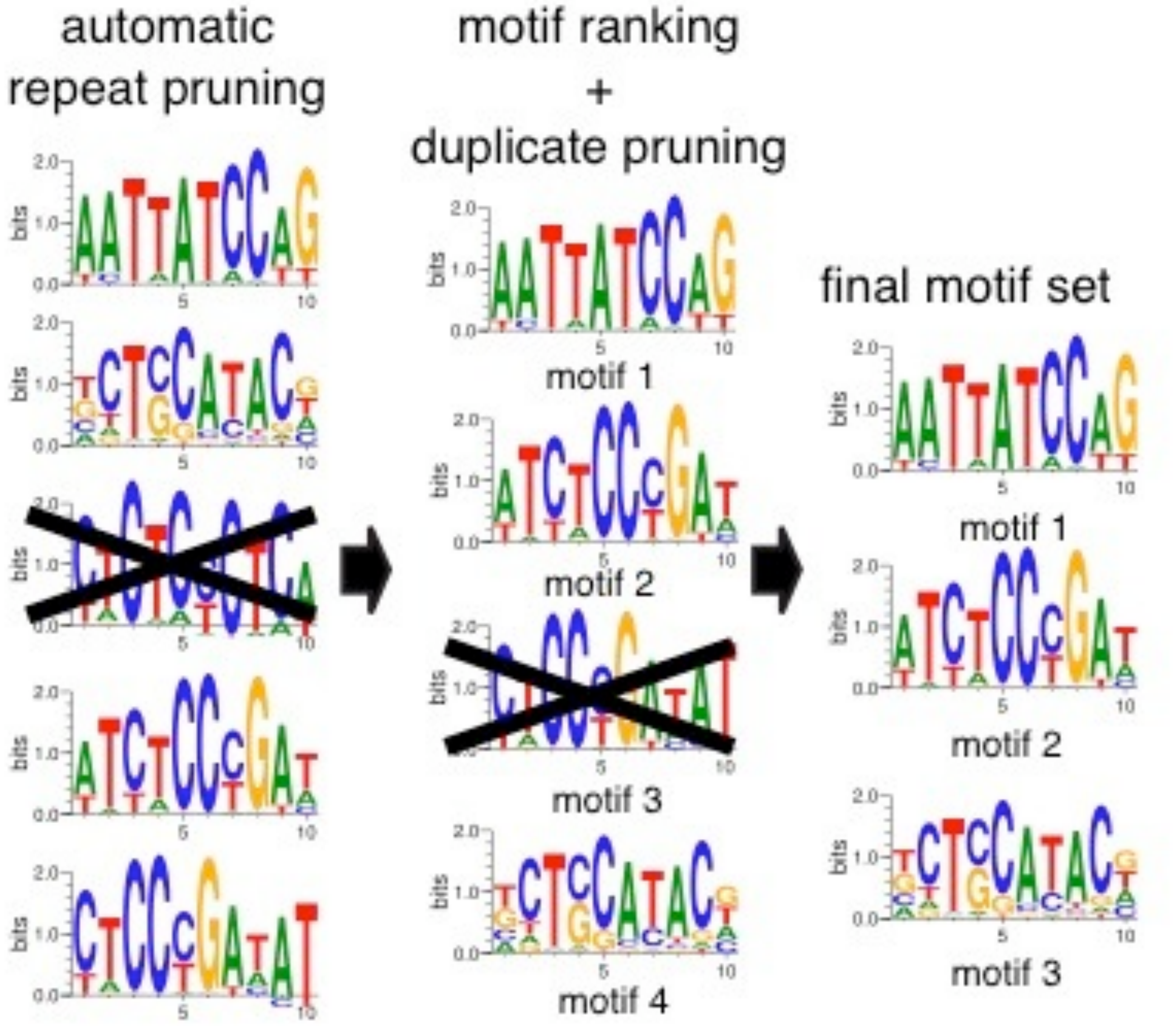
$$\frac{f_A w_C}{f_C w_A} = \frac{p^{\text{fixation}}(A \rightarrow C)}{p^{\text{fixation}}(C \rightarrow A)} = e^{4Ns}$$

Motif discovery
step 2: filtering and ranking



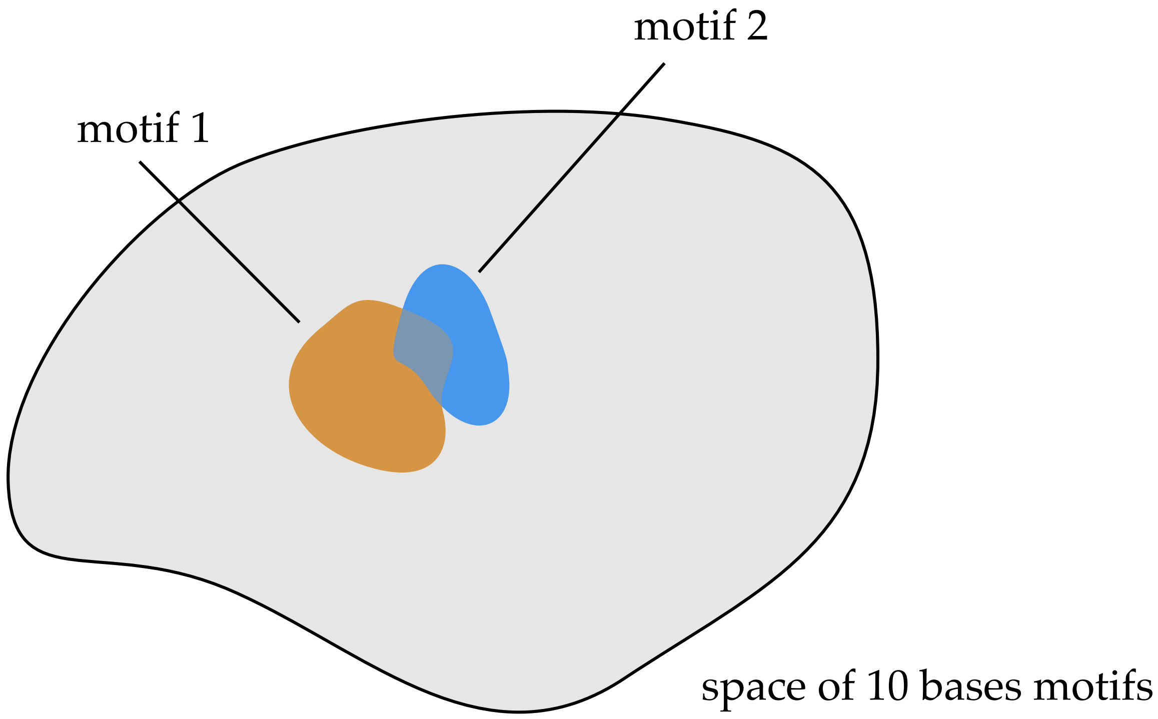
1. elimination of repeated motifs, i.e. that are distributed in a very non-poissonian manner in the background set (10 000 intergenic sequences of 2 kb)

Motif discovery
step 2: filtering and ordering the list of motifs



2. elimination of duplicated motifs, i.e. that recognize largely overlapping sets of sequences

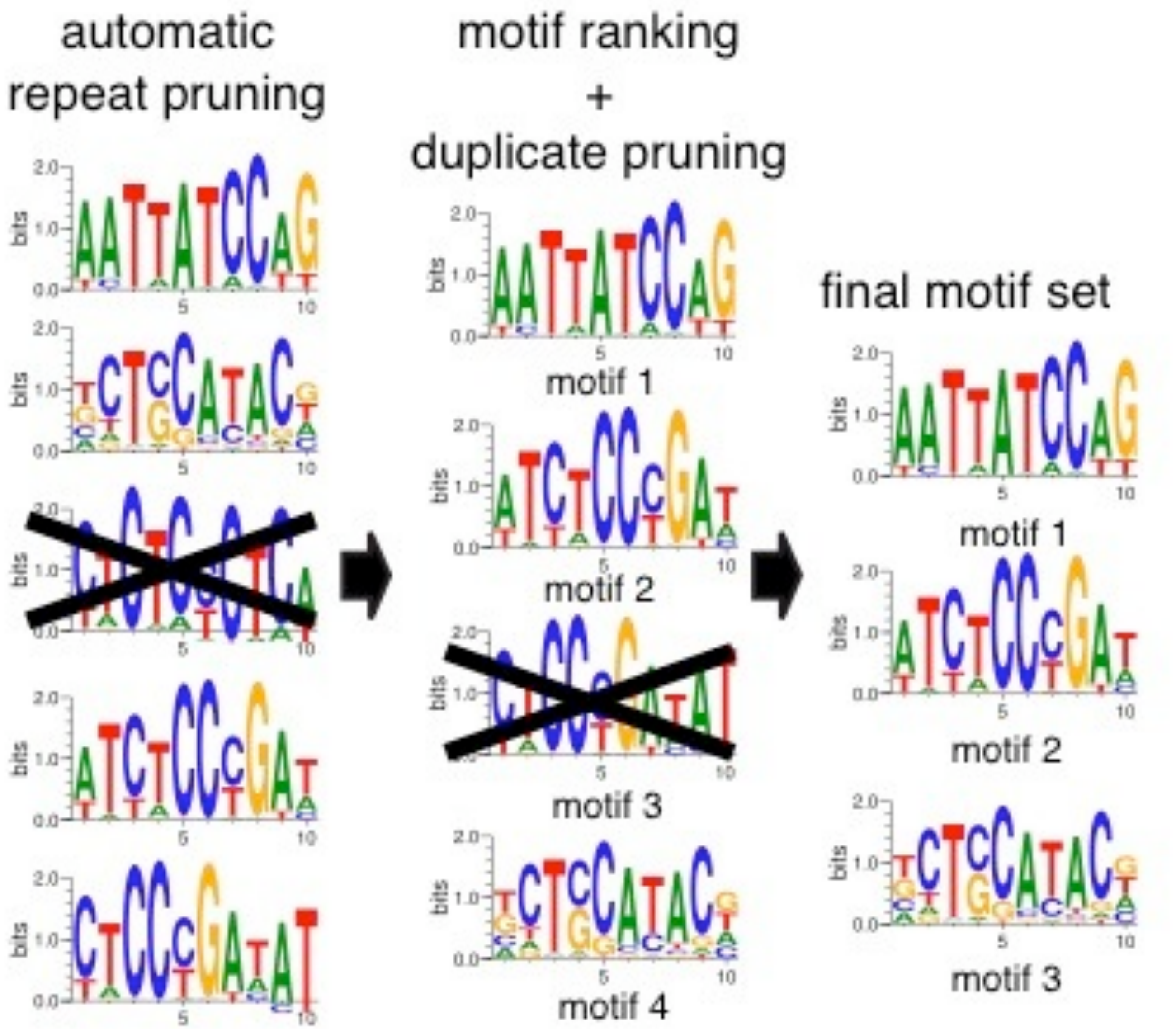
Proximity between PWMs



$$\mathrm{Prox}(\mathbf{w}^{(1)},\mathbf{w}^{(2)}) = 2\frac{\mathcal{P}\left\{\left[S\left(\mathbf{s},\mathbf{w}^{(1)}\right) > S_{th}\right] \text{ and } \left[S\left(\mathbf{s},\mathbf{w}^{(2)}\right) > S_{th}\right]\right\}}{\mathcal{P}\{S(\mathbf{s},\mathbf{w}^{(1)}) > S_{th}\} + \mathcal{P}\{S(\mathbf{s},\mathbf{w}^{(2)}) > S_{th}\}}$$

$$\mathcal{P}\{S(\mathbf{w},\mathbf{s}) > S_{th}\} = \sum_{\mathbf{s}} p(\mathbf{s})\Theta(S(\mathbf{w},\mathbf{s}) - S_{th})$$

Motif discovery
step 2: filtering and ordering the list of motifs



3. Ranking based on the departure of its distribution in the training set from Poisson distribution at the density measured in the background set
Both density and clustering contribute to the motif score

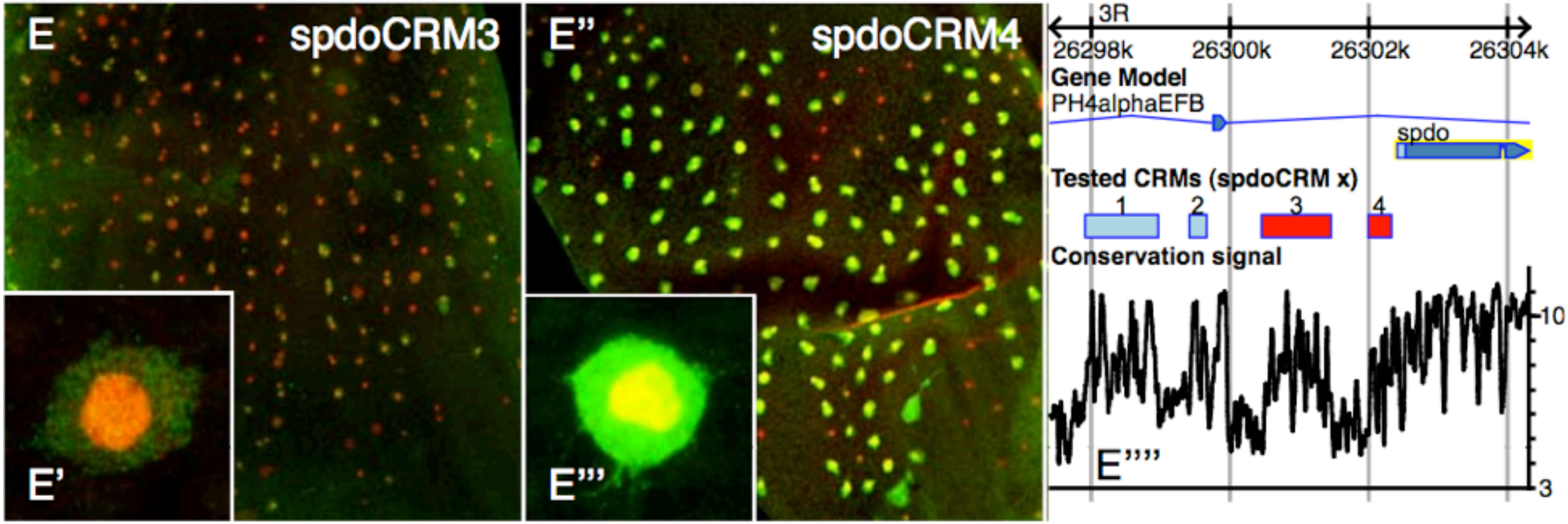
Training set

- 14 CRMs (144-2398 nt)

  8 known CRMs, previously validated in vivo using reporter assays)

  6 new CRMs identified based on their :

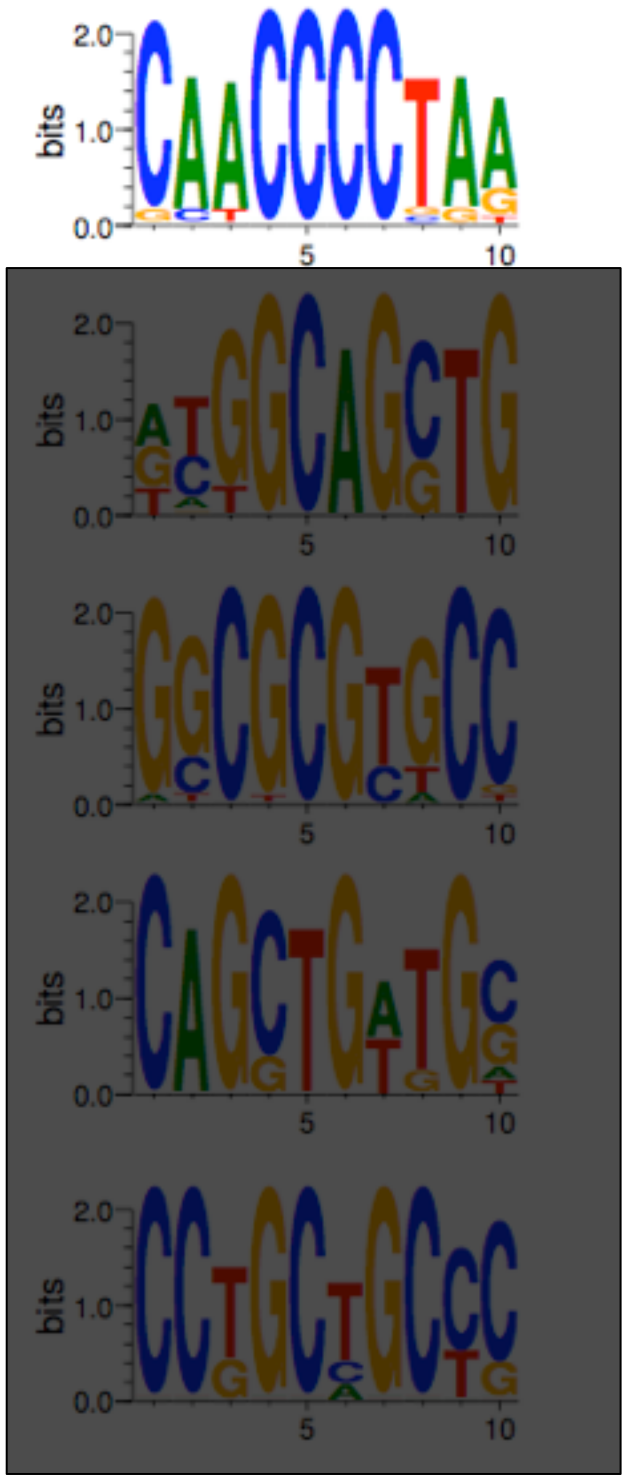
  - proximity to SOP-specific genes
  - sequence conservation within the 12 *Drosophila* species

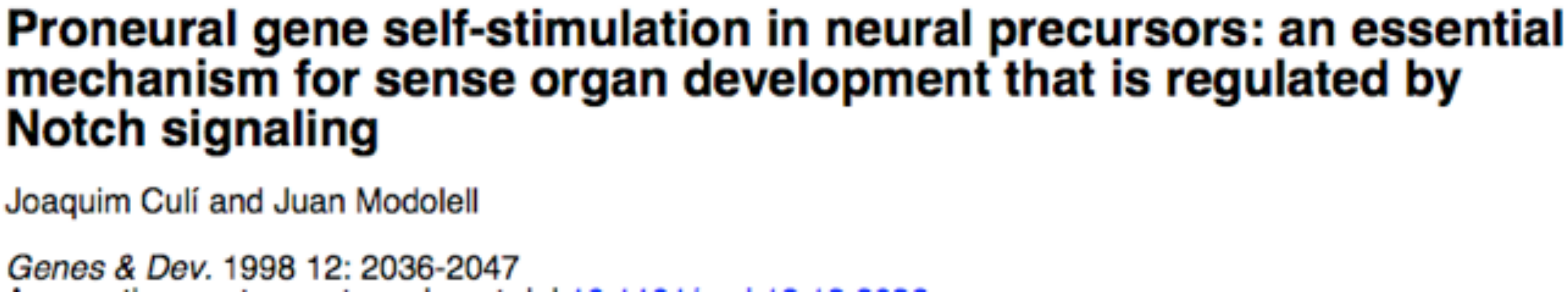


- 31 conserved genomic fragments (250-1320 nt)

total length: 34,703 nt (0.04% of the repeat-masked non coding DNA)

motif 1 corresponds to the α2 box

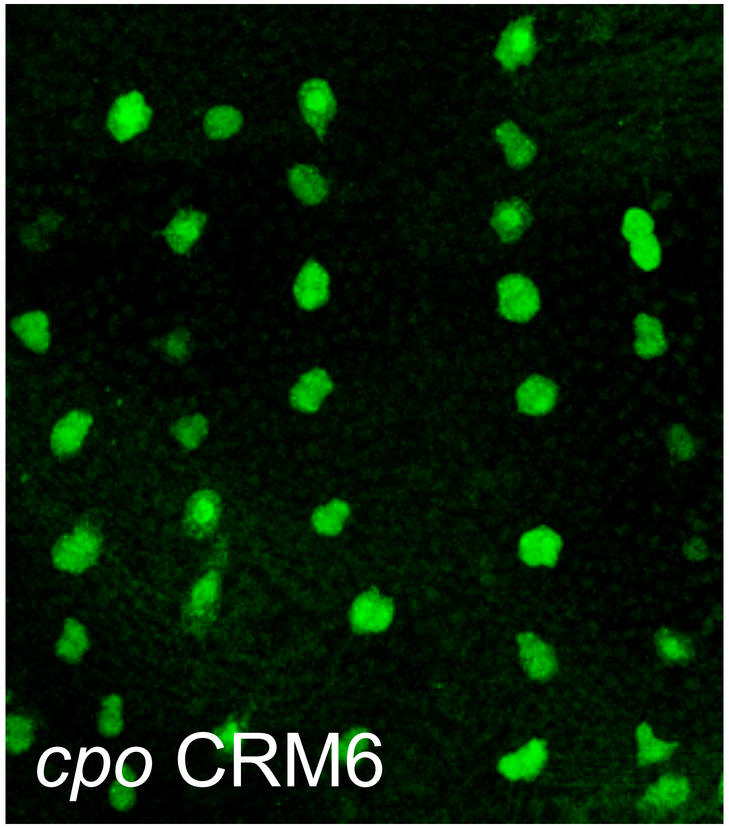


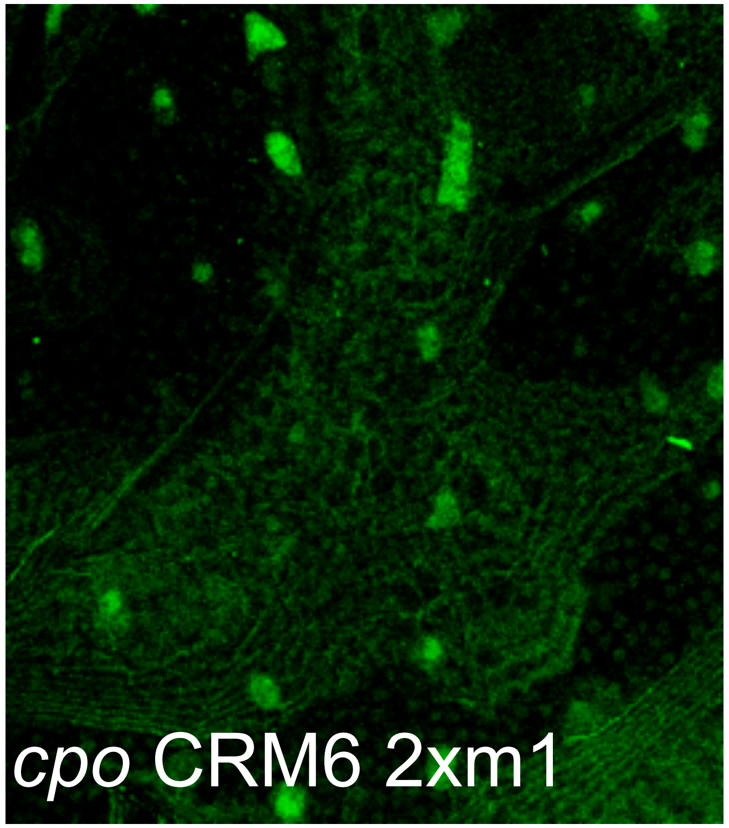


Proneural gene self-stimulation in neural precursors: an essential mechanism for sense organ development that is regulated by Notch signaling

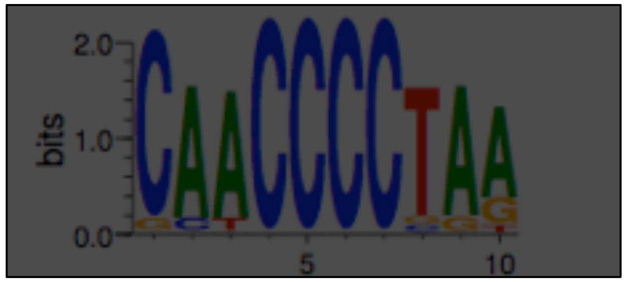
Joaquim Culí and Juan Modolell

Genes & Dev. 1998 12: 2036-2047

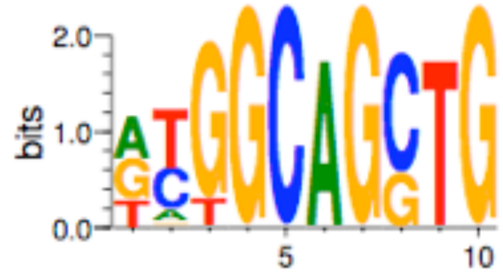

cpo CRM6

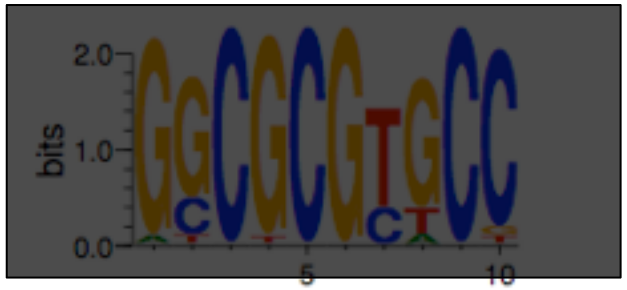

cpo CRM6 2xm1

motifs 2 and 4 predict binding sites for proneural bHLH factors

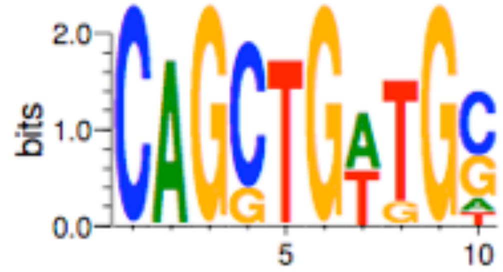


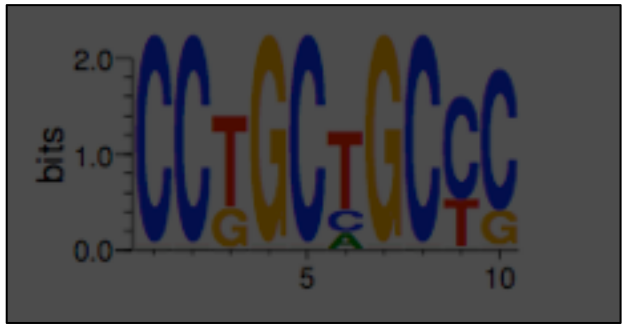
motif 2
E-box



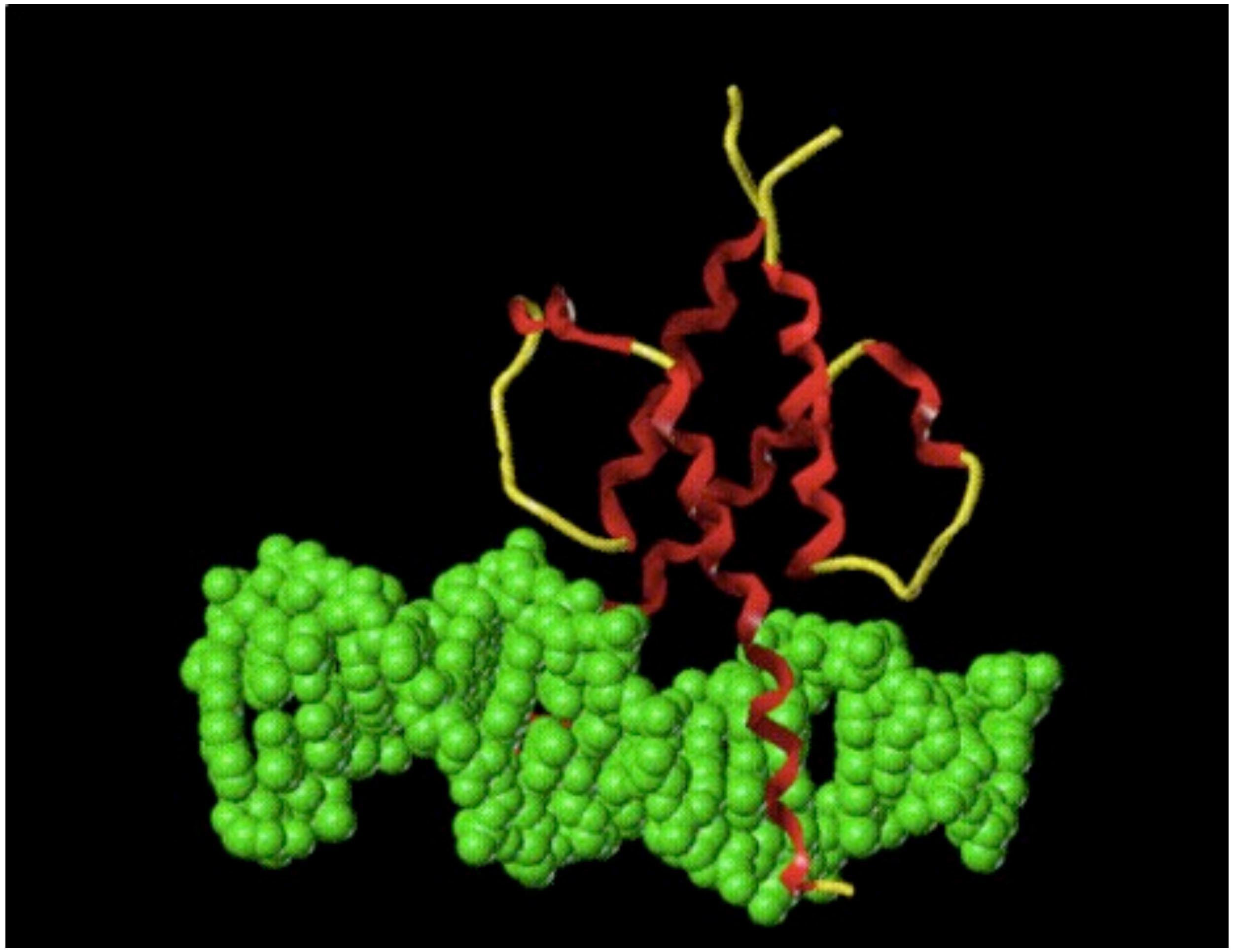


motif 4
E-box



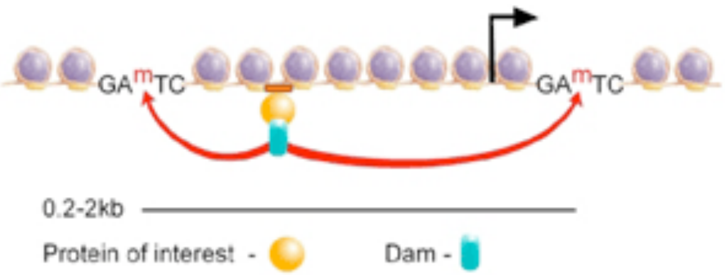


bHLH heterodimers

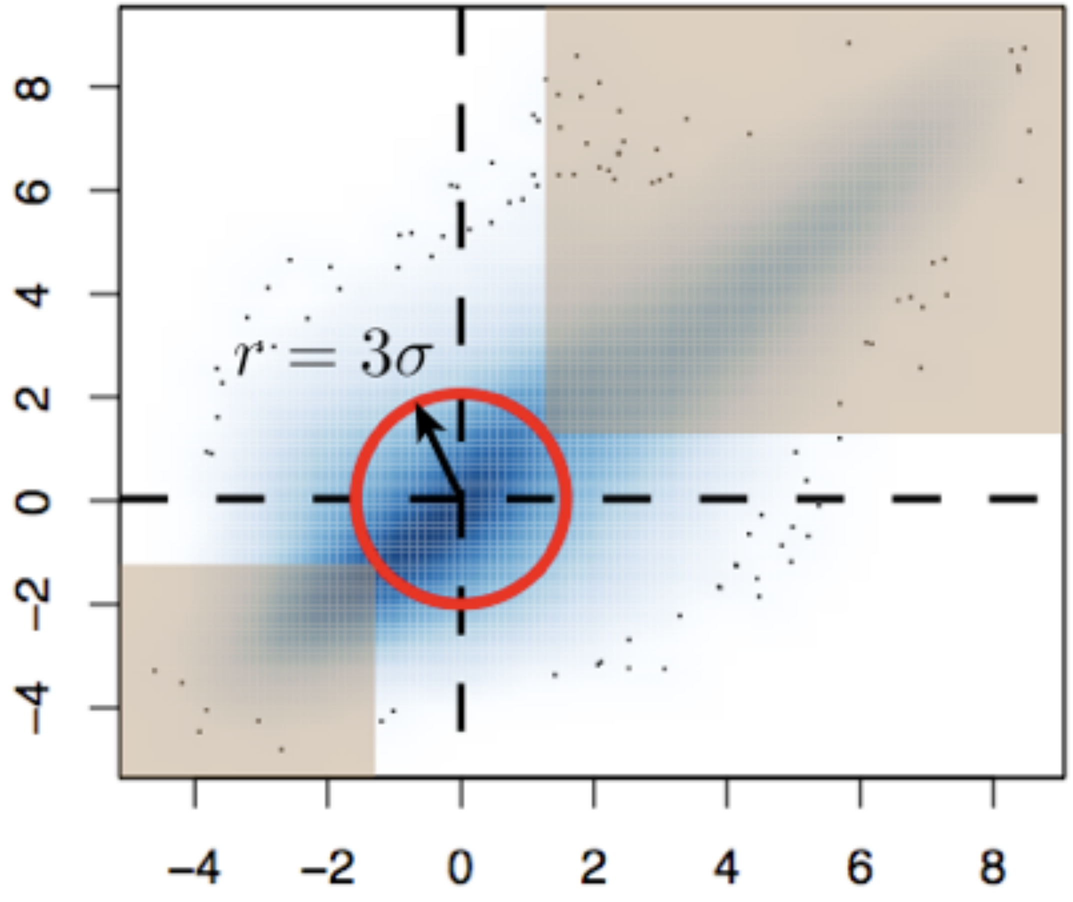


E-box: CAnnTG

Achaete binding sites cross-correlate with predicted CRMs

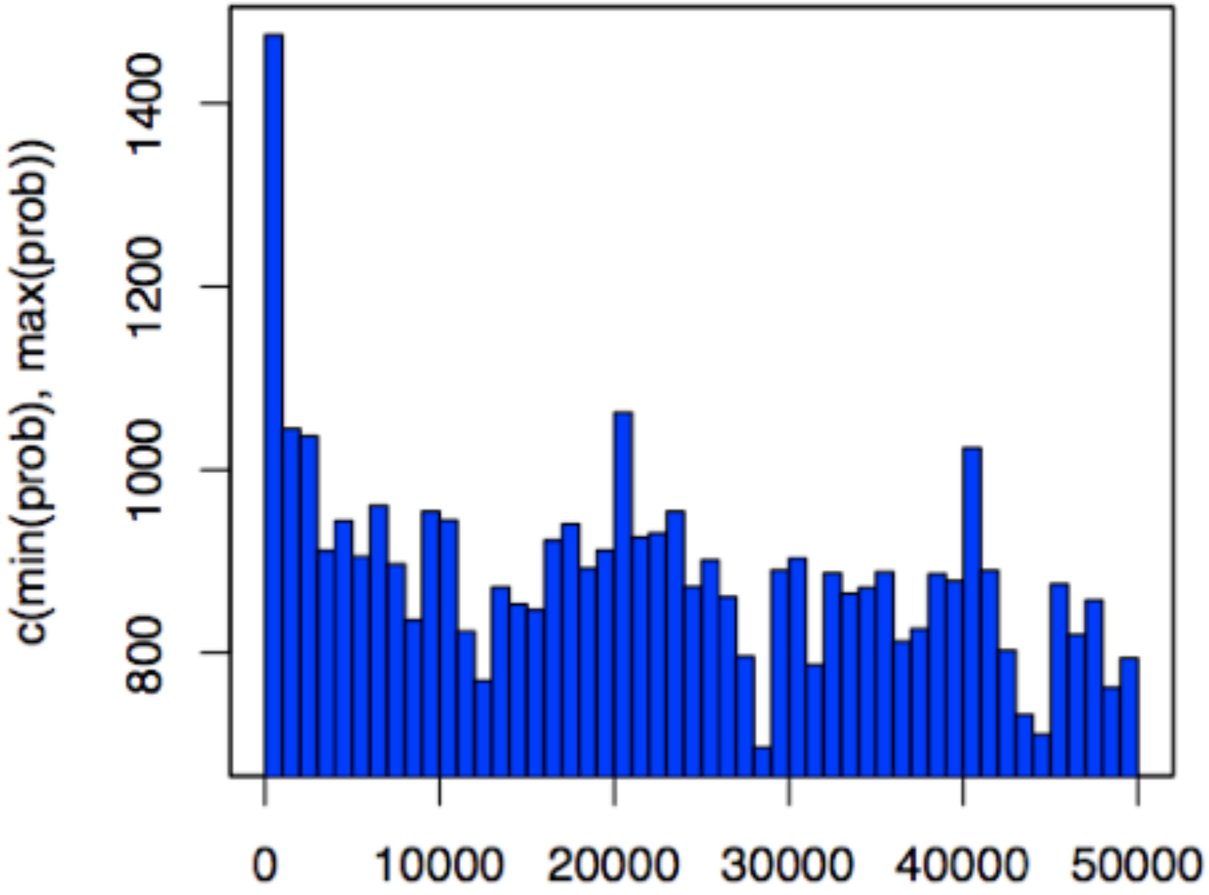


Dam-Achaete *vs* Dam alone expressed in proneural clusters (scaGAL4 Gal80ts driver)



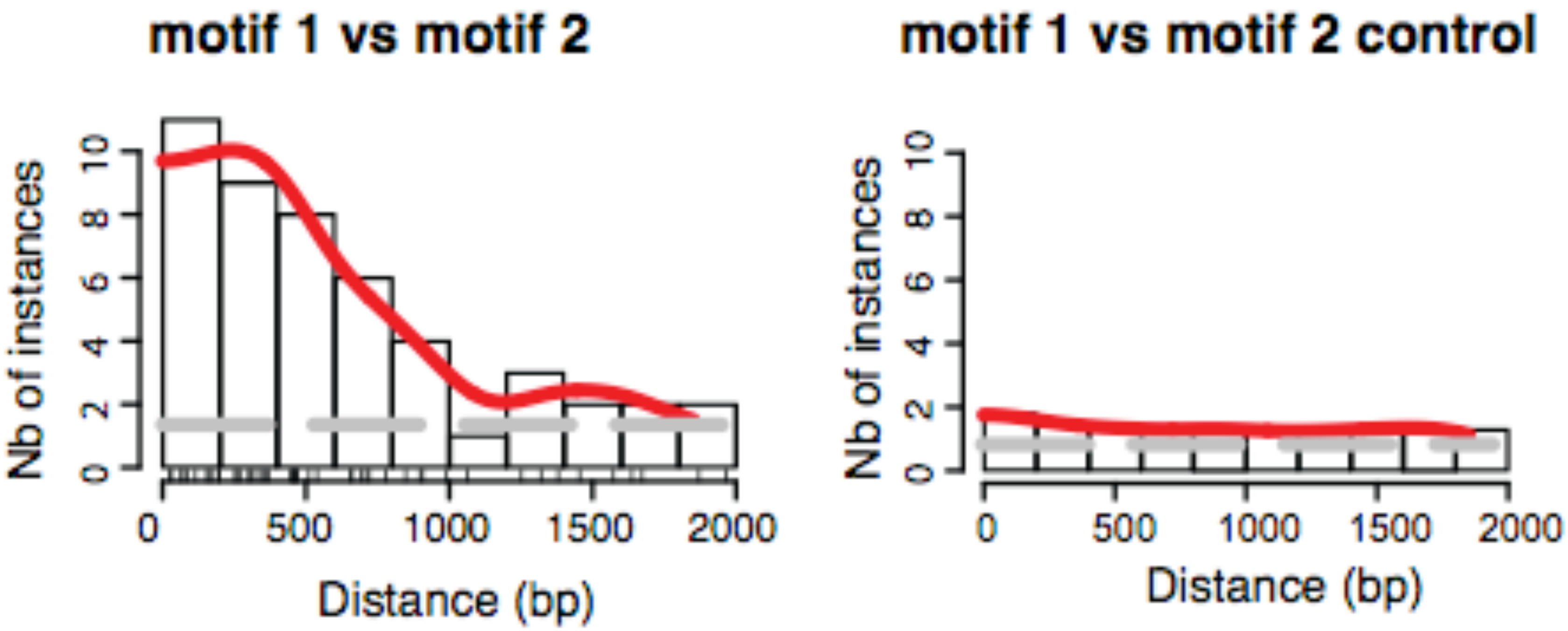
K Mazouni *et al* (unpublished)

Achaete DamID fragments vs predicted CRMs

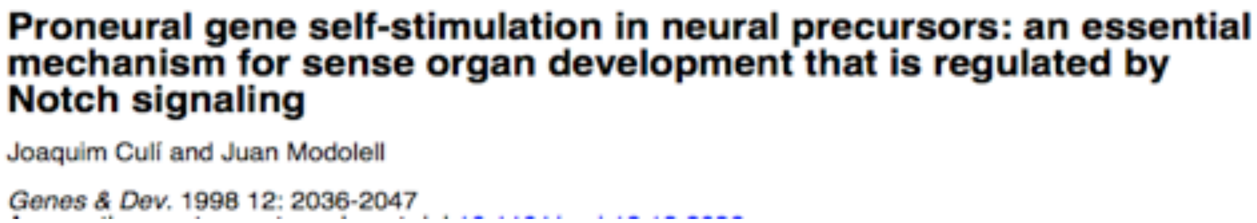


bin = 1 kb

The α2 box / E box combination

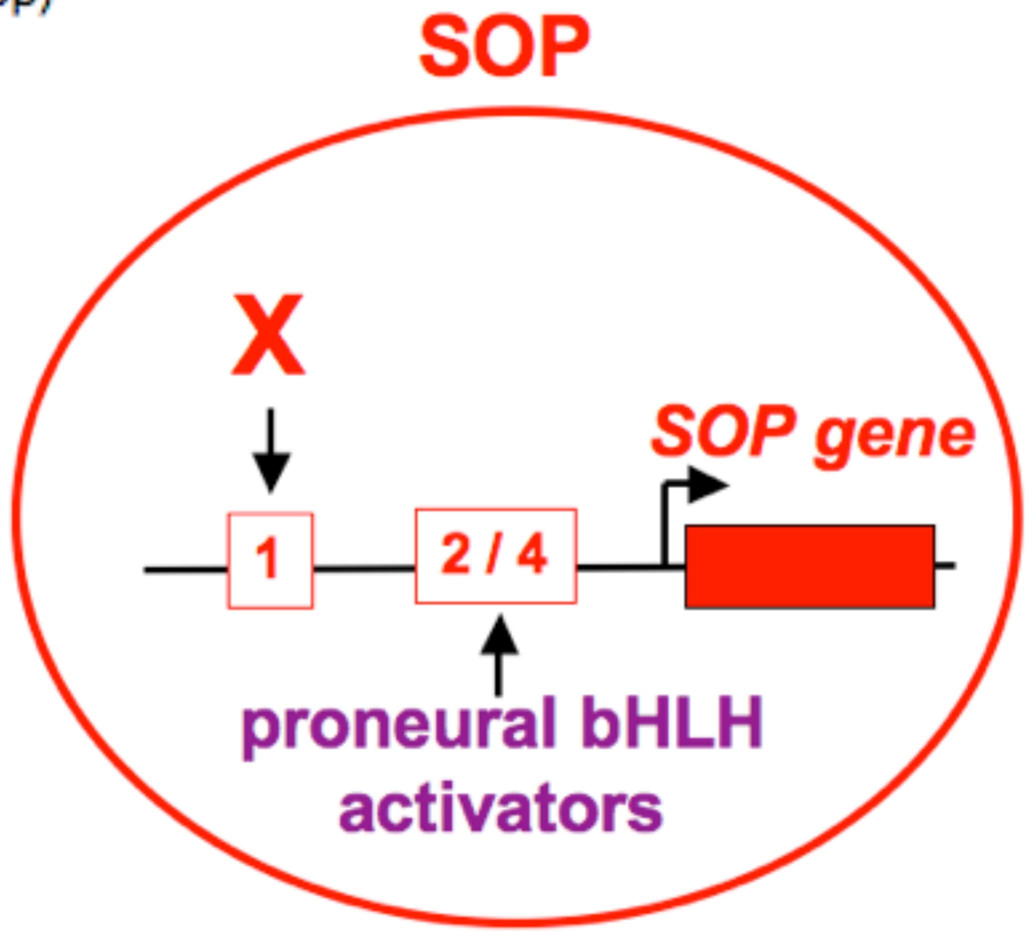


Motifs 1 and 2 cross-correlate
in the *D. melanogaster* genome


Proneural gene self-stimulation in neural precursors: an essential mechanism for sense organ development that is regulated by Notch signaling

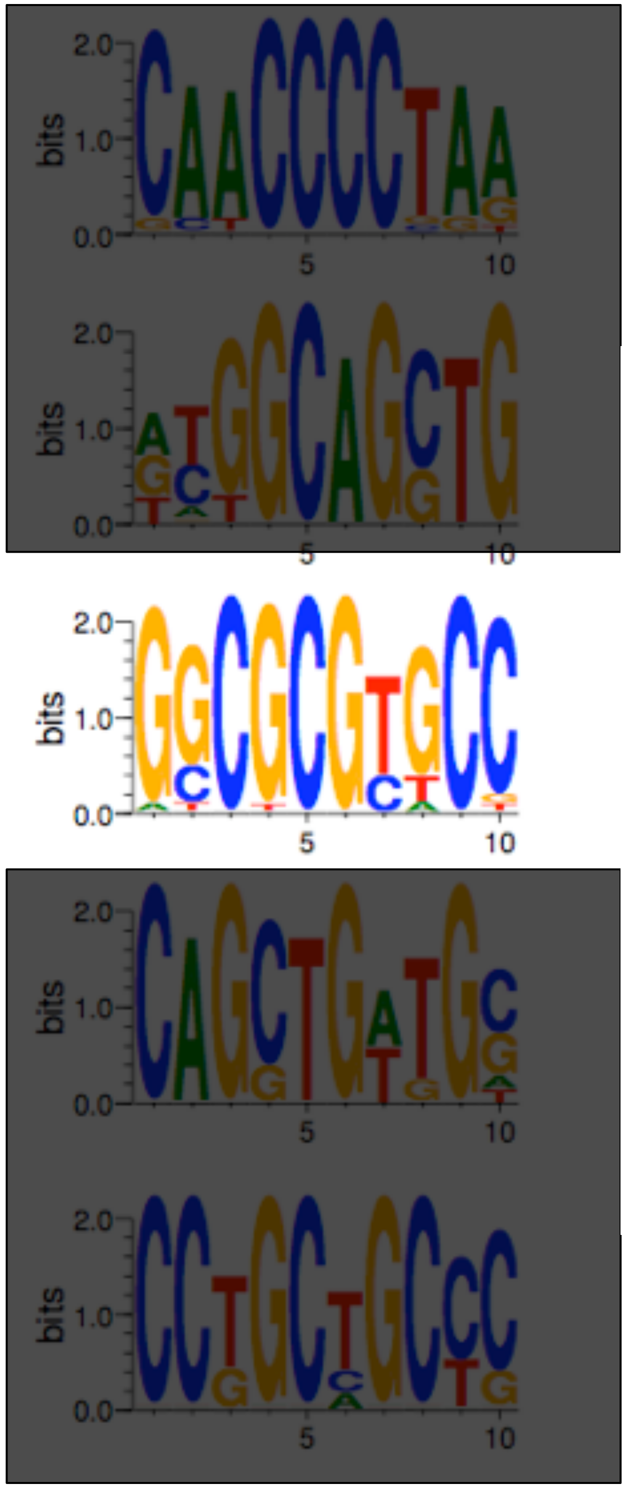
Joaquim Culí and Juan Modolell

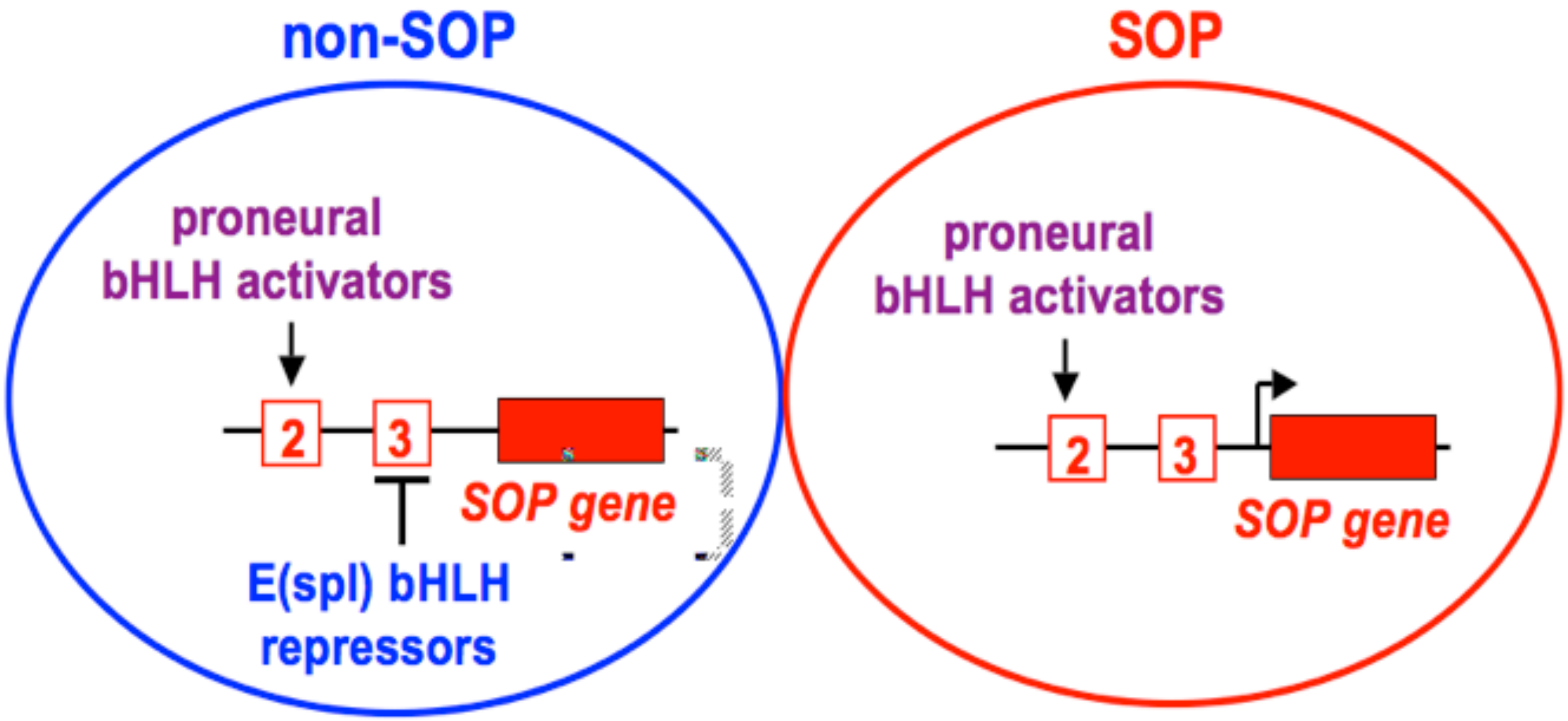
Genes & Dev. 1998 12: 2036-2047



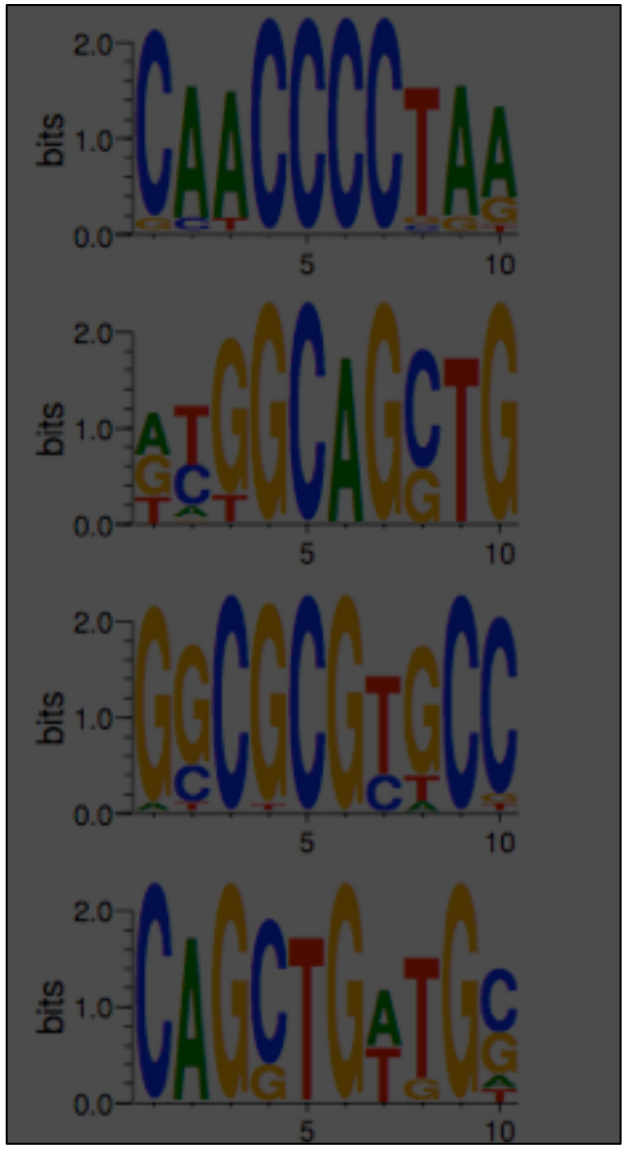
Motif 3 may predict binding sites for E(spl) bHLH repressors

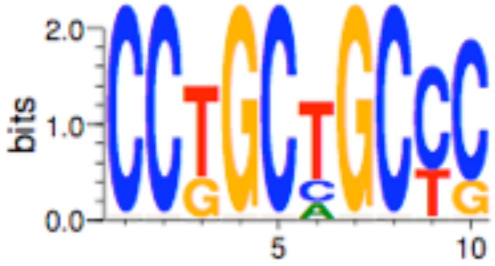
motif 3
N-box



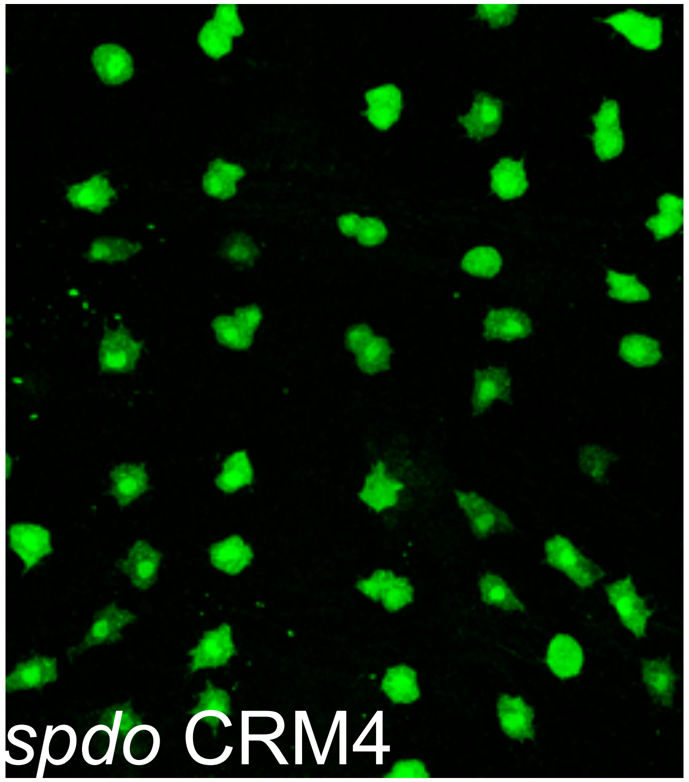


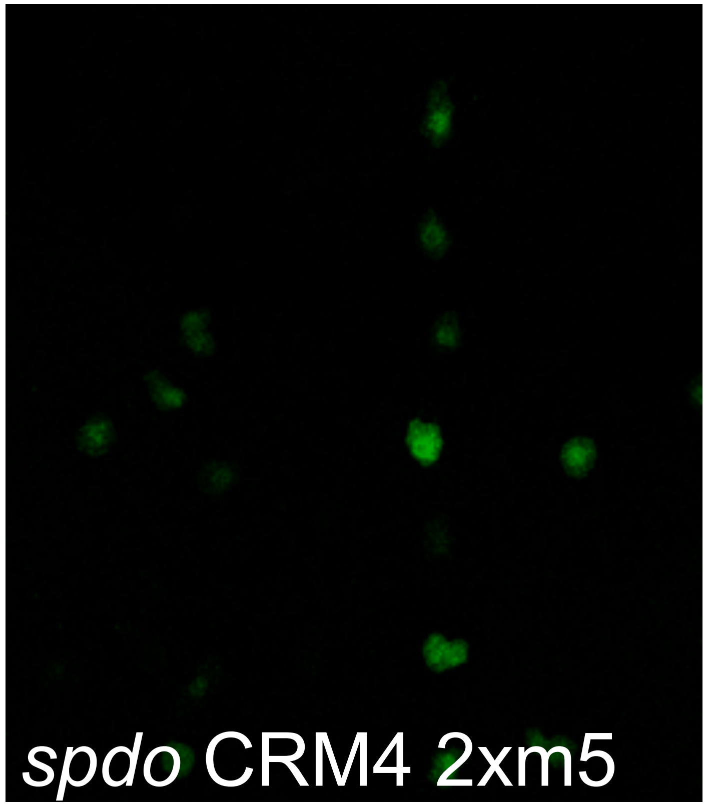
Motif 5 is novel





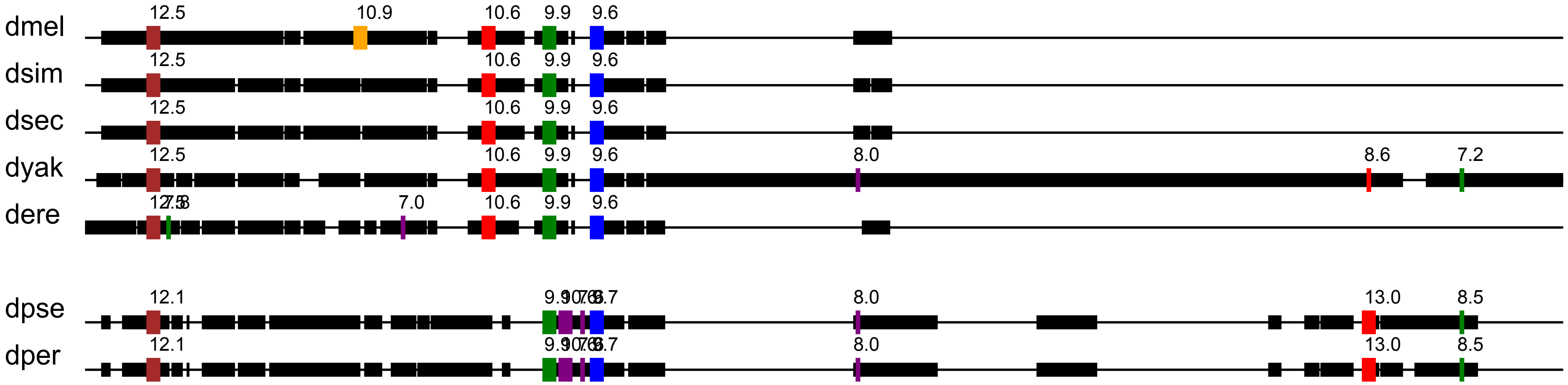
motif 5



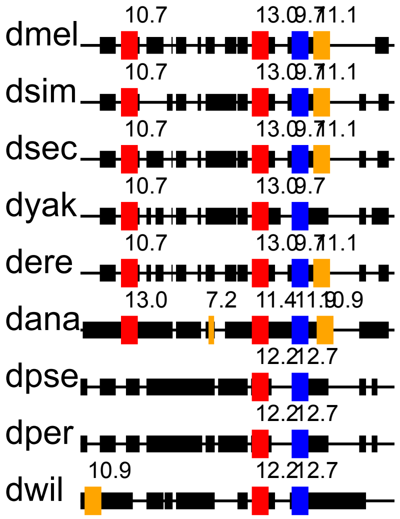


Conservation of motifs on training set

CG32150 enhancer



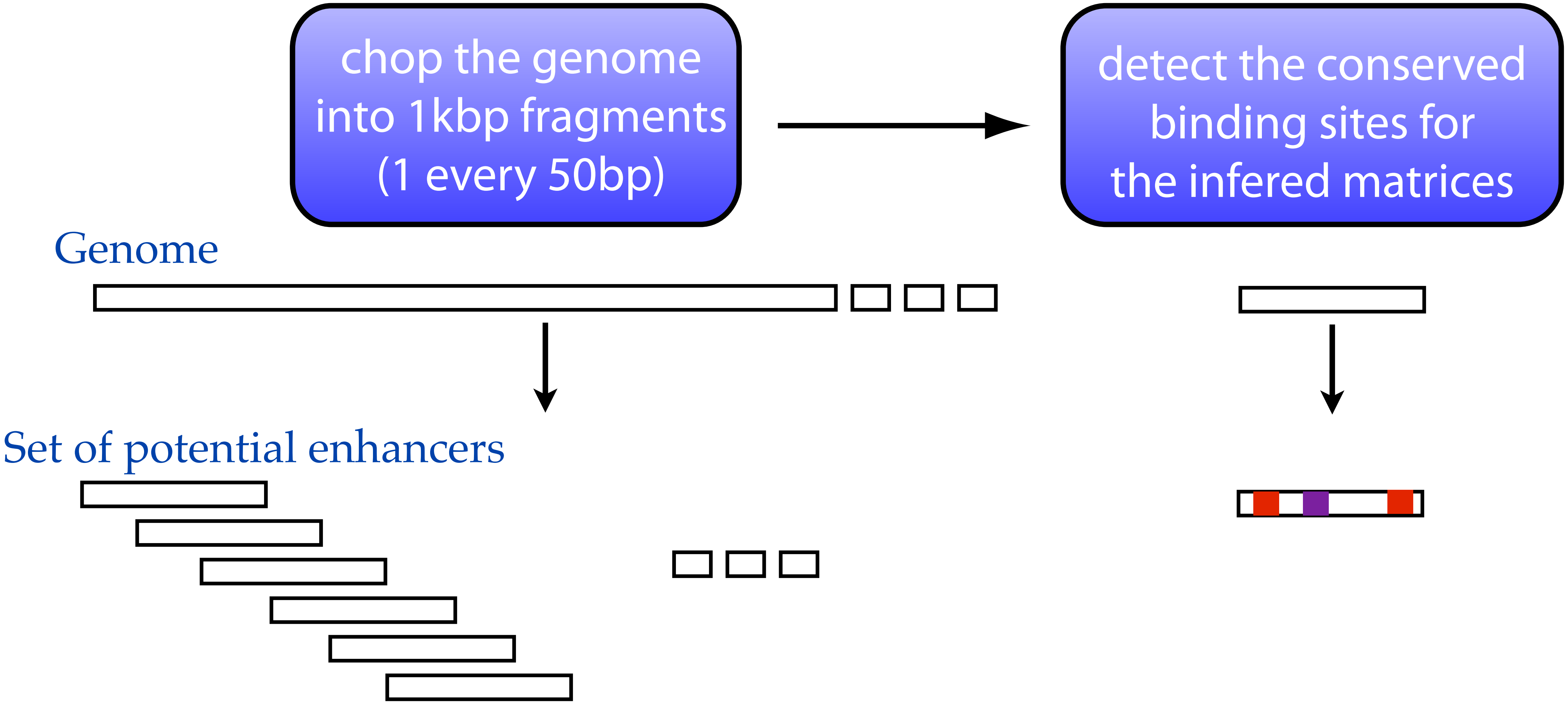
Neur enhancer



Predicted motifs move within the enhancers through evolution (in few cases)

lundi 17 janvier 2011

Enhancer prediction



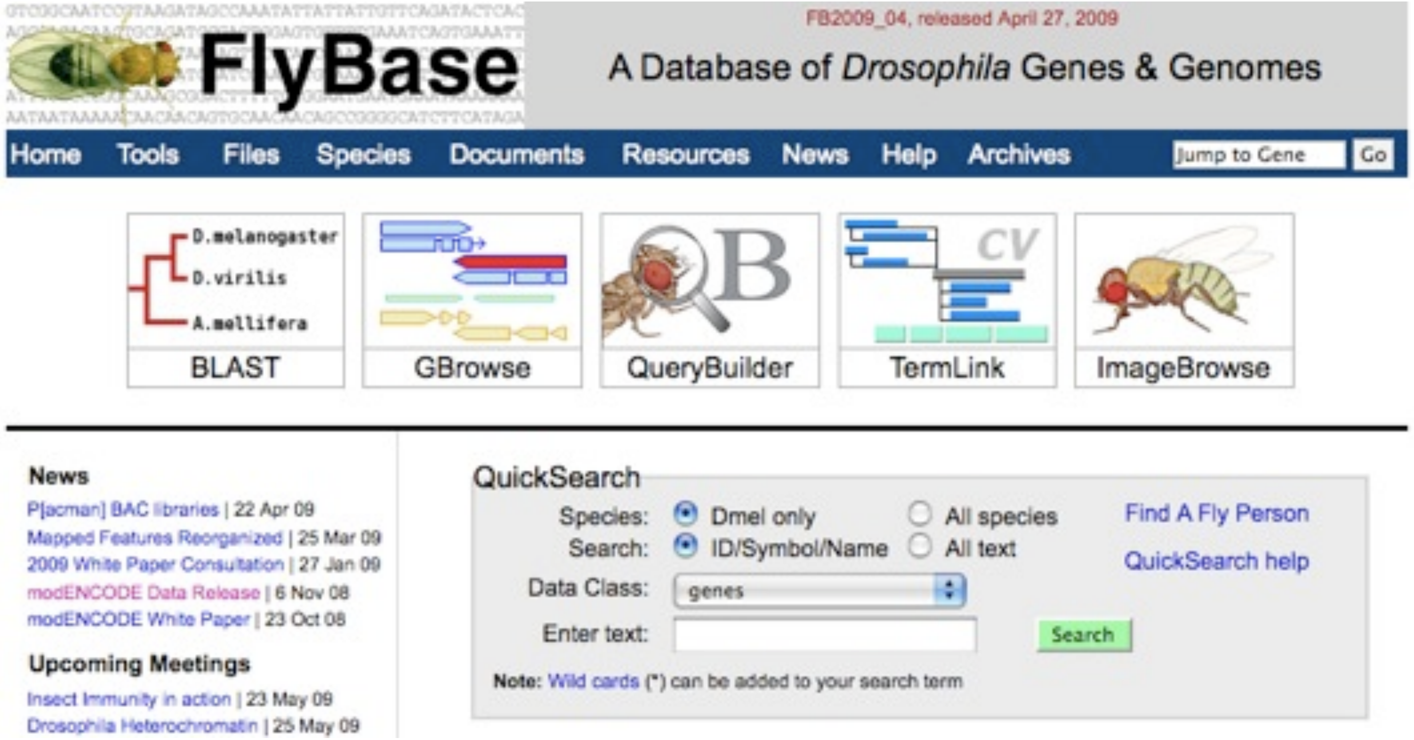
chop the genome into 1kbp fragments (1 every 50bp)

detect the conserved binding sites for the infered matrices

- each fragment is given a score according to the motif over-representation

$$S(E) = \sum_{\text{PWM w}} n_w(E) \ln \left[\frac{\lambda_w^{(tr)}}{\lambda_w^{(bg)}} \right]$$

Associating enhancers with GO categories



FB2009_04, released April 27, 2009

A Database of *Drosophila* Genes & Genomes

News

P[acman] BAC libraries | 22 Apr 09
Mapped Features Reorganized | 25 Mar 09
2009 White Paper Consultation | 27 Jan 09
modENCODE Data Release | 6 Nov 08
modENCODE White Paper | 23 Oct 08

Upcoming Meetings

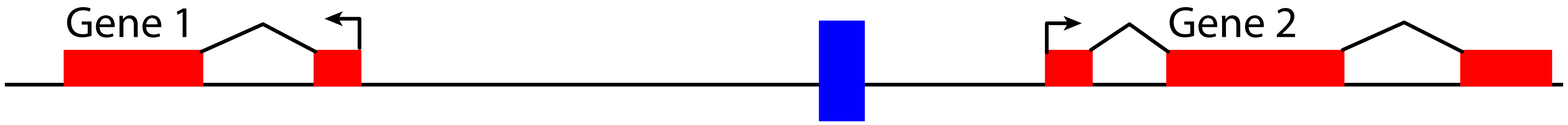
Insect Immunity in action | 23 May 09
Drosophila Heterochromatin | 25 May 09

QuickSearch

Species: ● Dmel only ○ All species
Search: ● ID/Symbol/Name ○ All text
Data Class: genes
Enter text:
Search

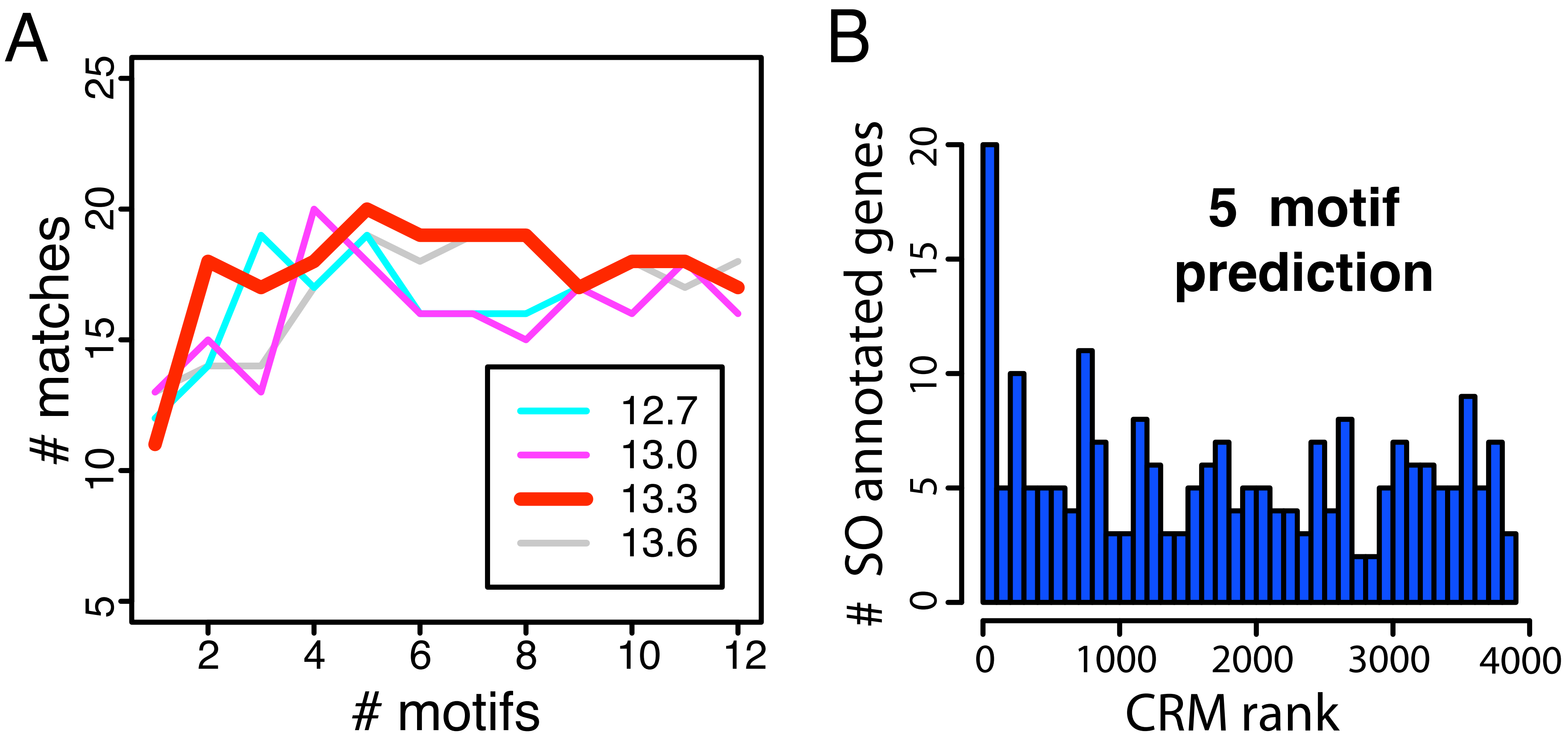
Note: Wild cards (*) can be added to your search term

Find A Fly Person
QuickSearch help



Predicted enhancer

Choosing parameters and first «test».



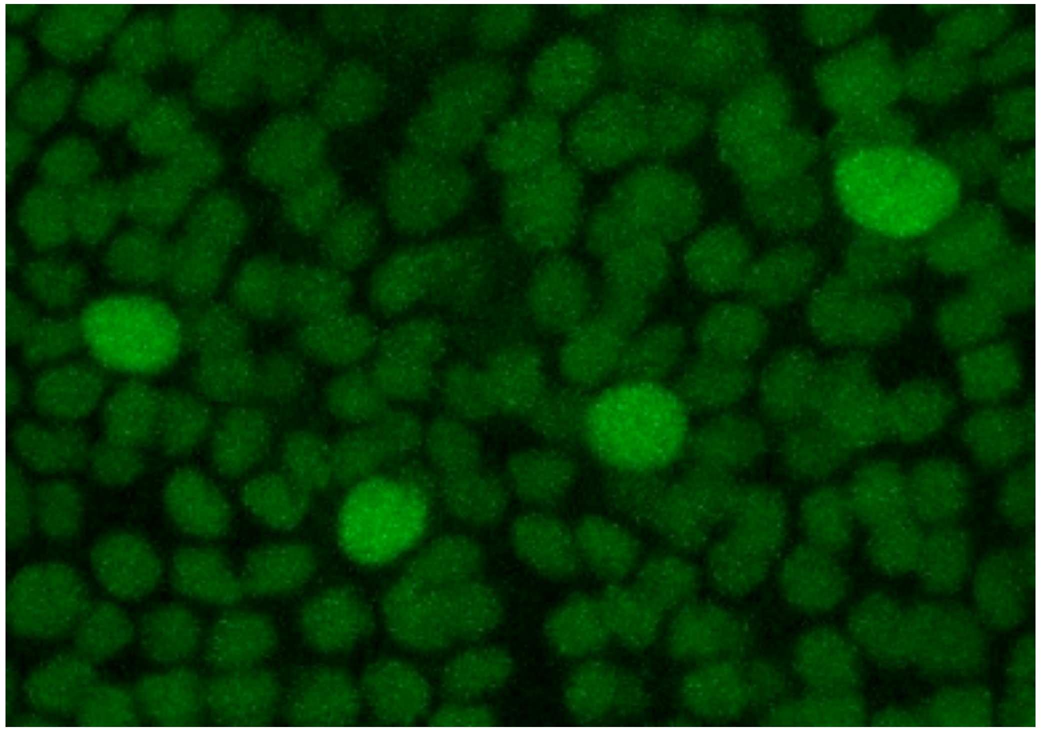
Rouault et al, PNAS 2010

Results obtained with the Felsenstein model

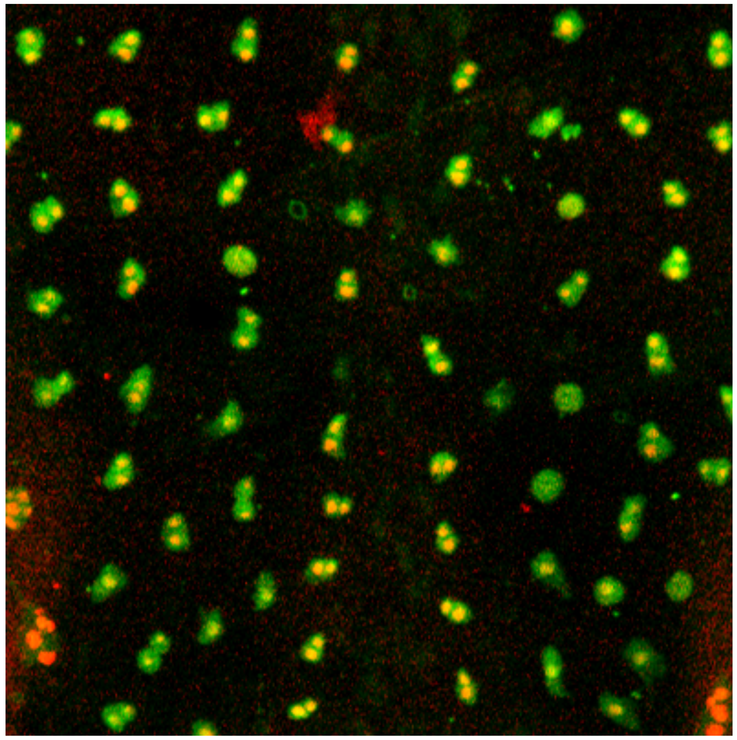
Parameters of the algorithm have be optimized on this criterion

Lola

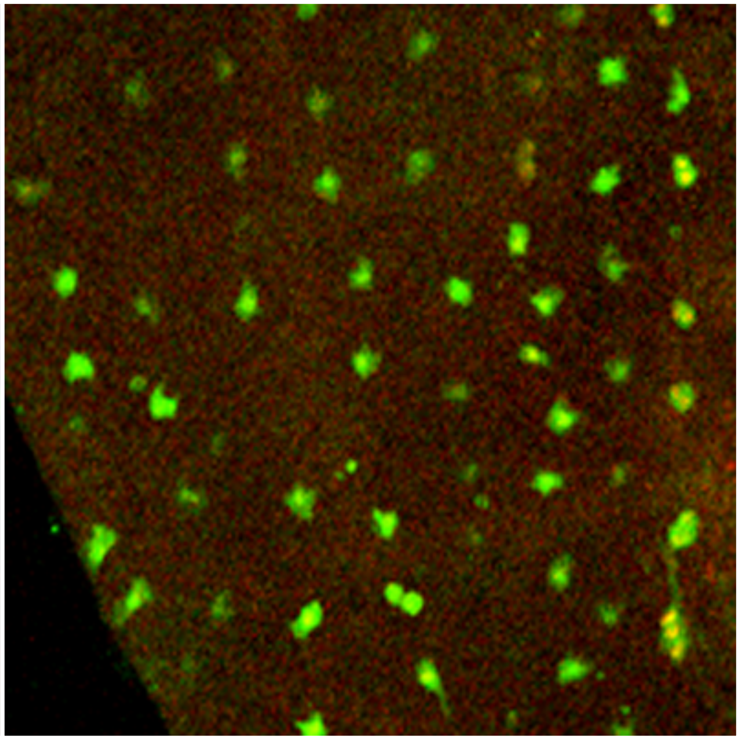
Lola protein presence



CRM20



CRM40

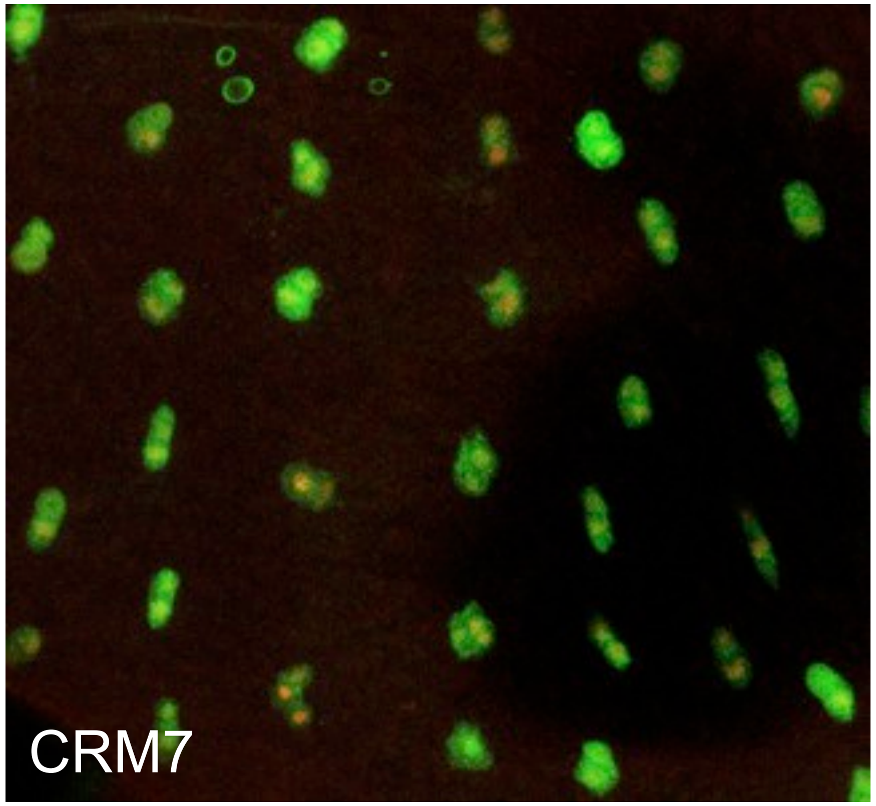


transcriptional repressor
antagonizes Notch in the R3/R4 decision in the eye

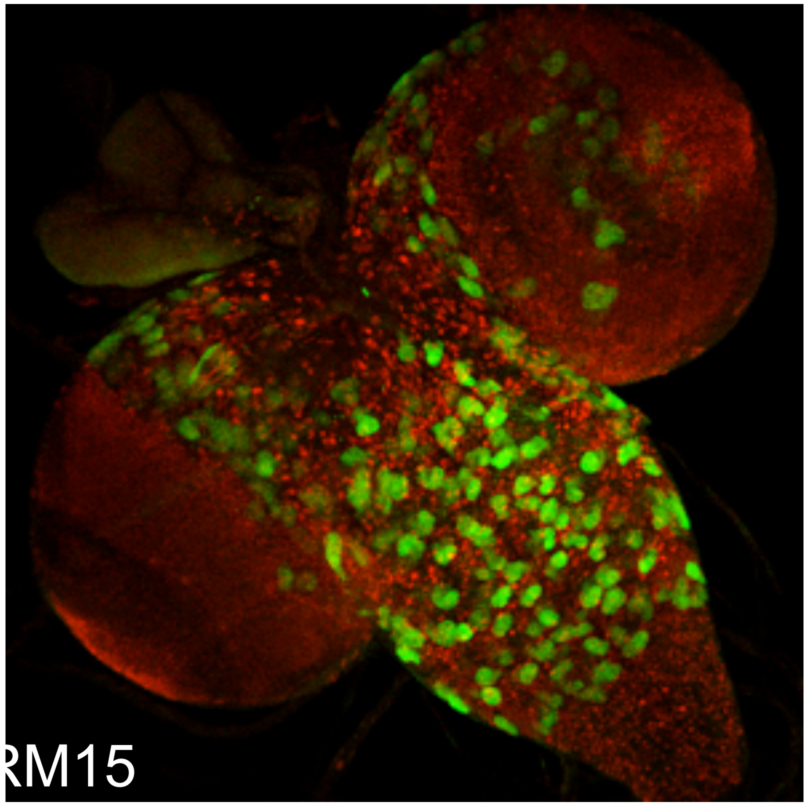
Hypothesis: represses Notch target genes in SOPs

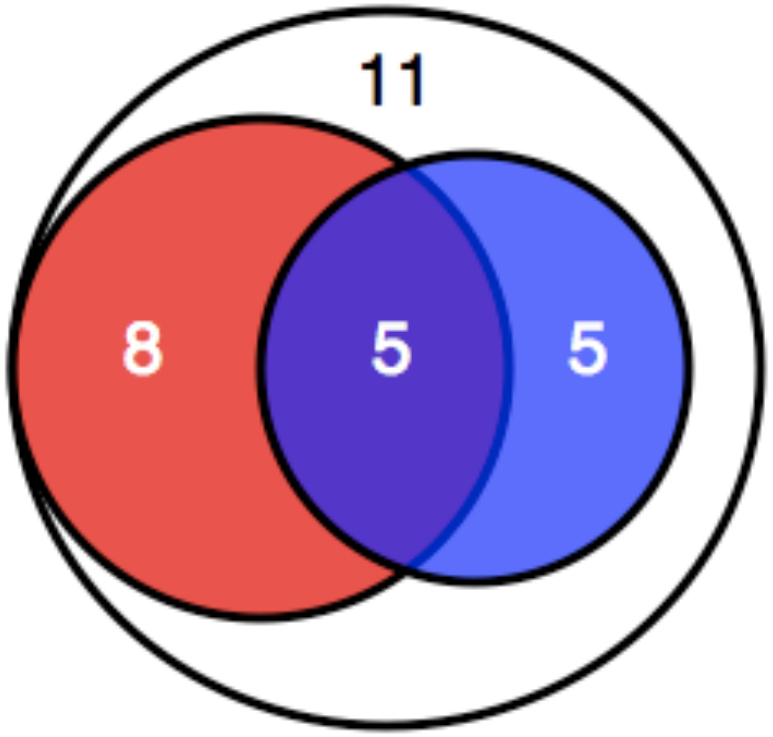
In vivo activity of predicted CRMs

pupal notum (SOPs)



larval brain (NBs)





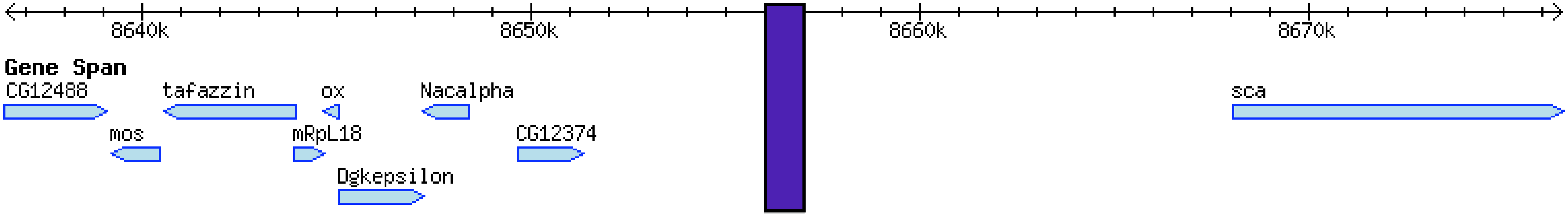
CRM prediction using the 5 top-ranked motifs:

10/29 are expressed only or predominantly in SOPs (3 also in PNCs)

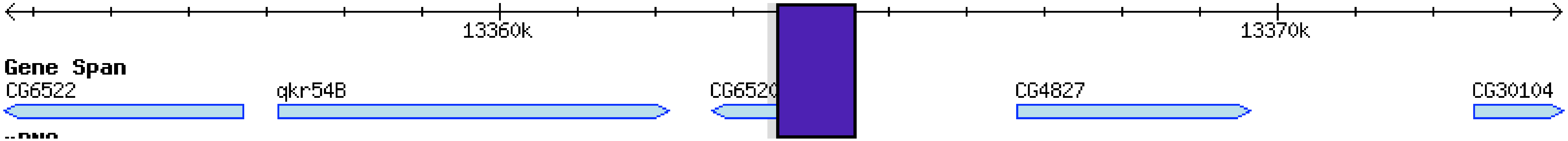
13/29 are expressed in neuroblasts of the larval brain

Identifying genes up-regulated in SOPs

known genes

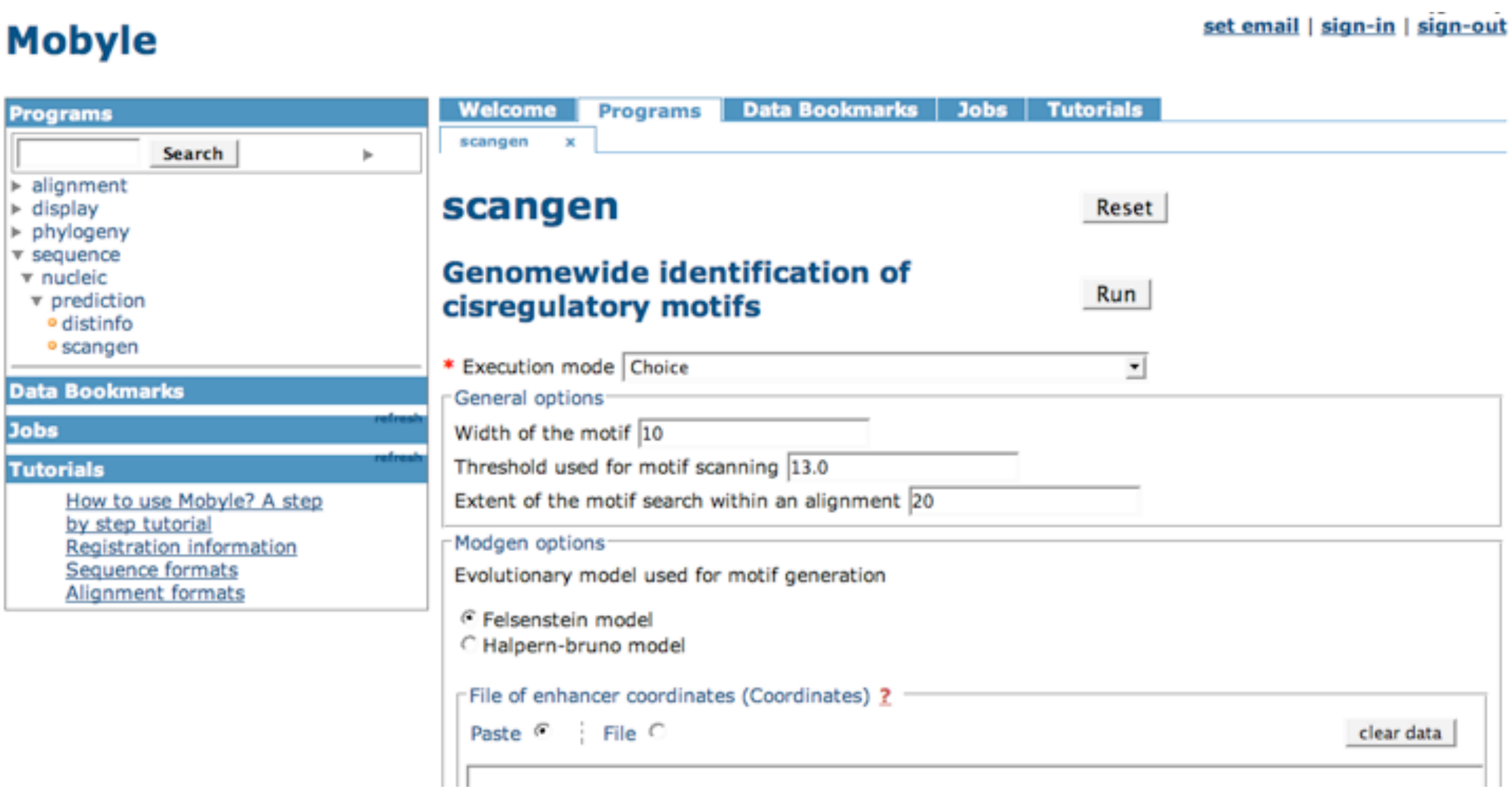


new genes



- analysis of expression patterns by in situ hybridization in larval discs

Available on the web very soon



Tutorials:

- How to use Mobyle? A step by step tutorial
- Registration information
- Sequence formats
- Alignment formats

scangen

Genomewide identification of cisregulatory motifs

* Execution mode: Choice

General options

Width of the motif: 10

Threshold used for motif scanning: 13.0

Extent of the motif search within an alignment: 20

Modgen options

Evolutionary model used for motif generation

● Felsenstein model
○ Halpern-bruno model

File of enhancer coordinates (Coordinates) ?

Paste ● File ○

Outlooks (ongoing)

- Improve predictions :

  Combine in silico data with high-throughput experiments (DamID, ChIP on chip)

- Role of the identified CRMs/genes in patterning :

  Dynamics of the CRM expression

- Extend this in silico approach to

  other stages and/or tissue specific enhancers

  other organisms, metazoans (vertebrates)

Acknowledgments

ENS, Laboratoire de physique statistique

V. Hakim

Institut Pasteur, Drosophila Developmental Genetics

K. Mazouni, L. Couturier, F. Schweisguth