

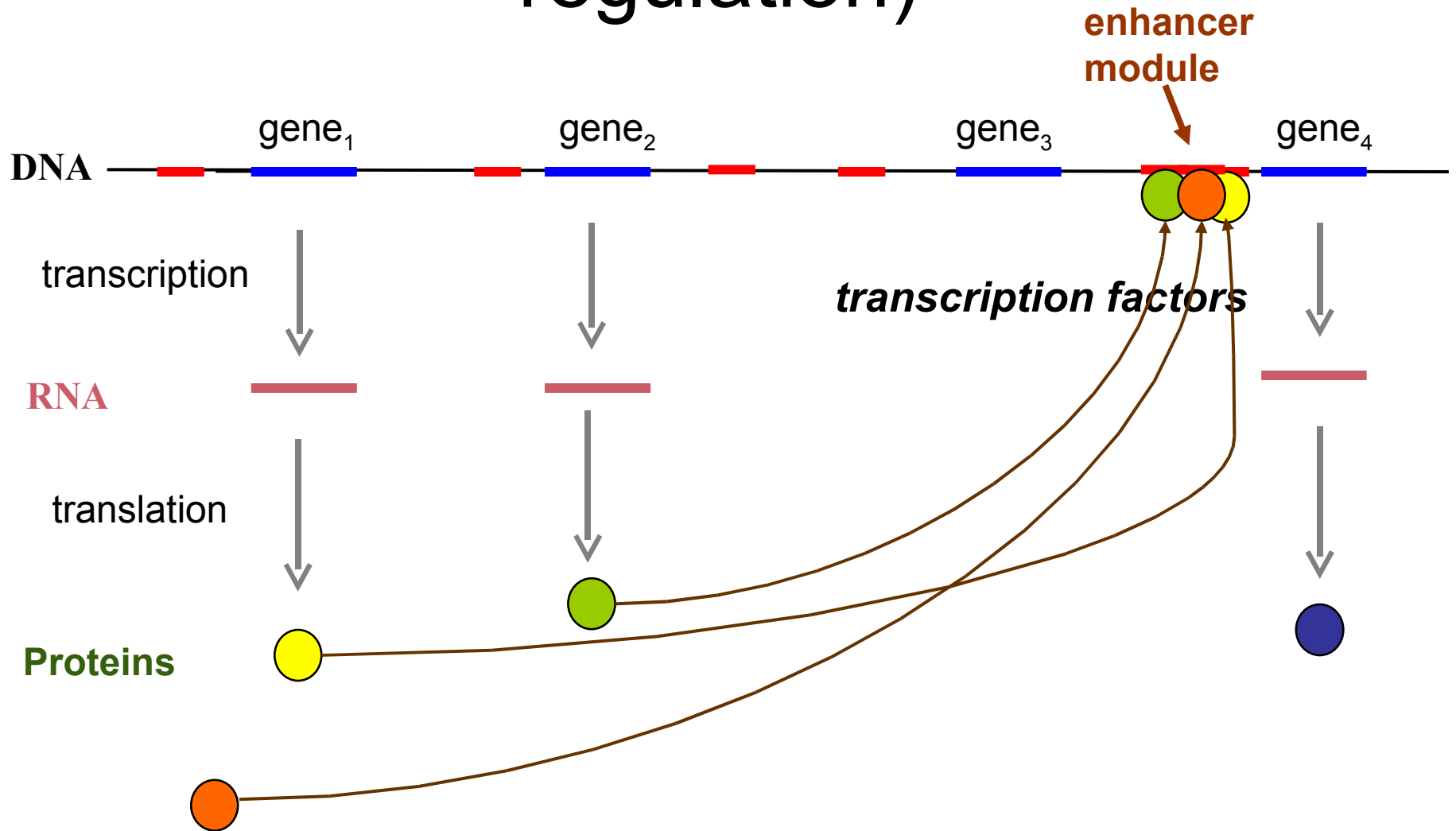
Computational Prediction of Gene Enhancer Elements by Comparative Genomics

Esko Ukkonen

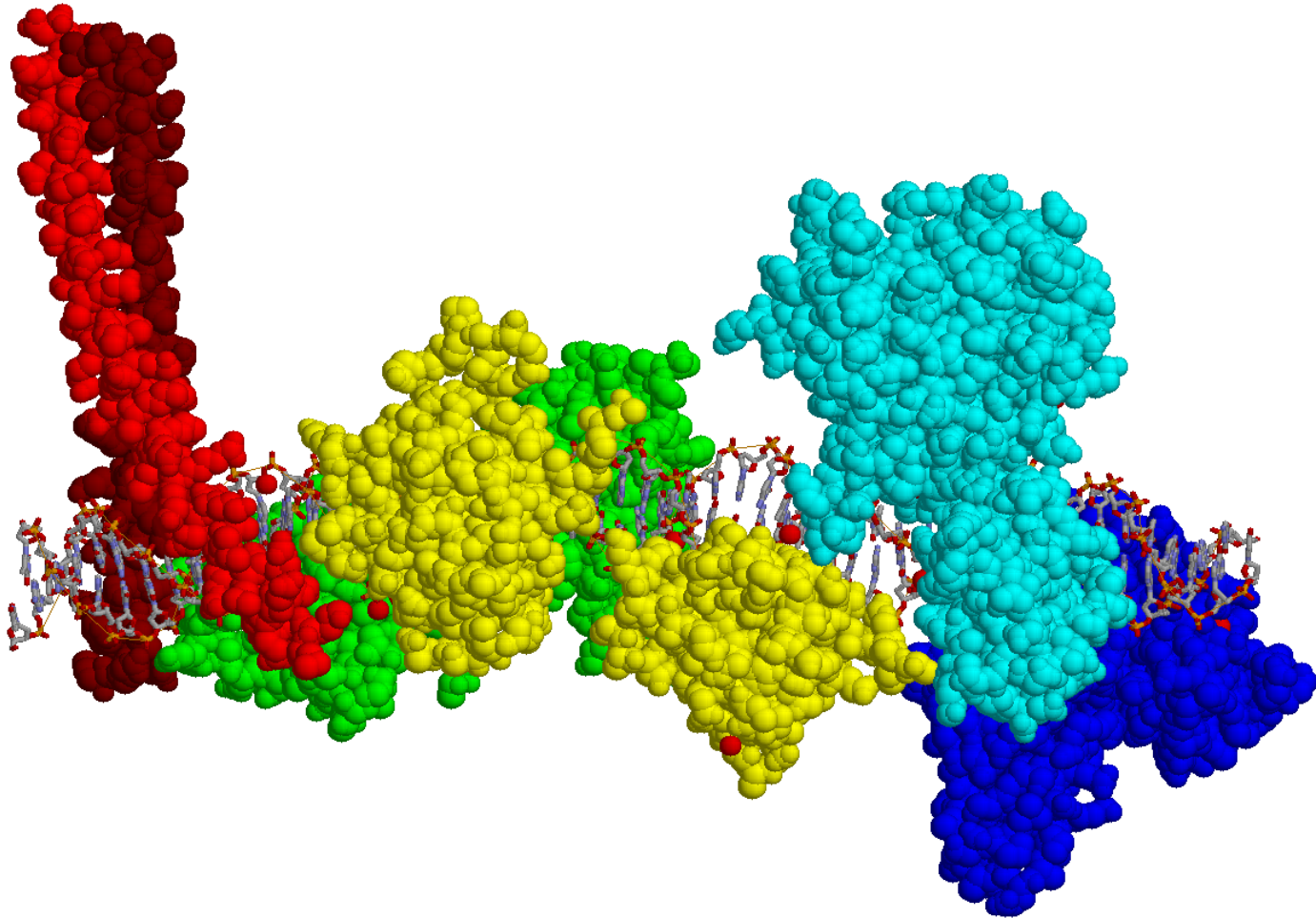
Helsinki Institute for Information Technology HIIT
University of Helsinki & Helsinki University of Technology

Comparative Genomics
Univ of St Andrews, 13 June 2008

Gene enhancer modules (cis-regulation)

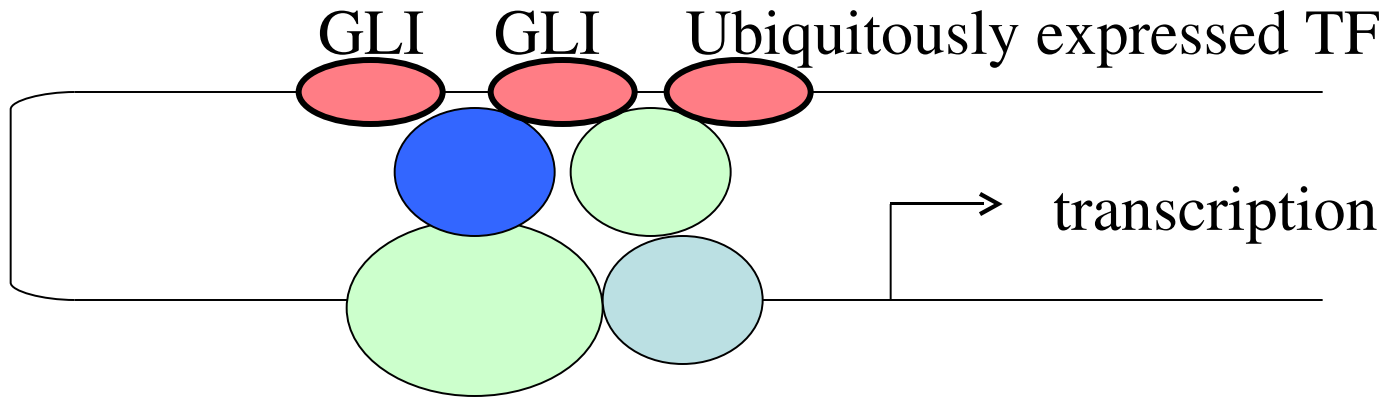


Enhancer module

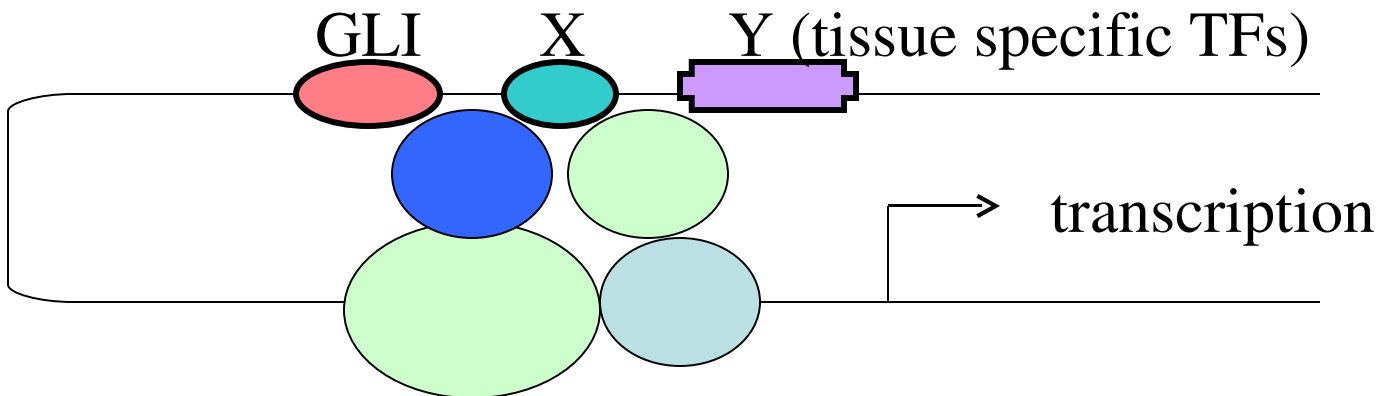


Model of cell type specific regulation of target gene expression

Common targets (e.g. Patched):



Cell type specific targets (e.g. N-myc):



Some vague remarks

- Gene expression regulation in multicellular organisms is controlled in combinatorial fashion by *transcription factors* (TFs)
- Transcription factors bind to DNA cis-elements on enhancer modules (promoters)
- Multiple factors need to bind to activate the module
- In mammals, the modules are few and far
- **The problem:** Locate functional regulatory modules, that is, find **interesting patterns**.

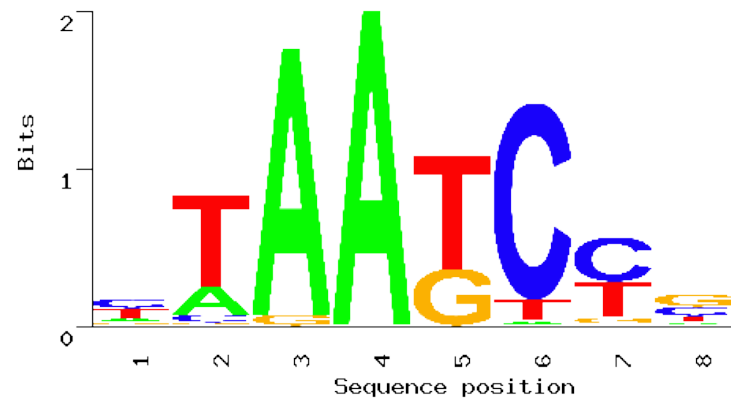
Binding affinity matrices

- The cis-elements are represented by affinity matrices.

- A column per position
- A row per nucleotide

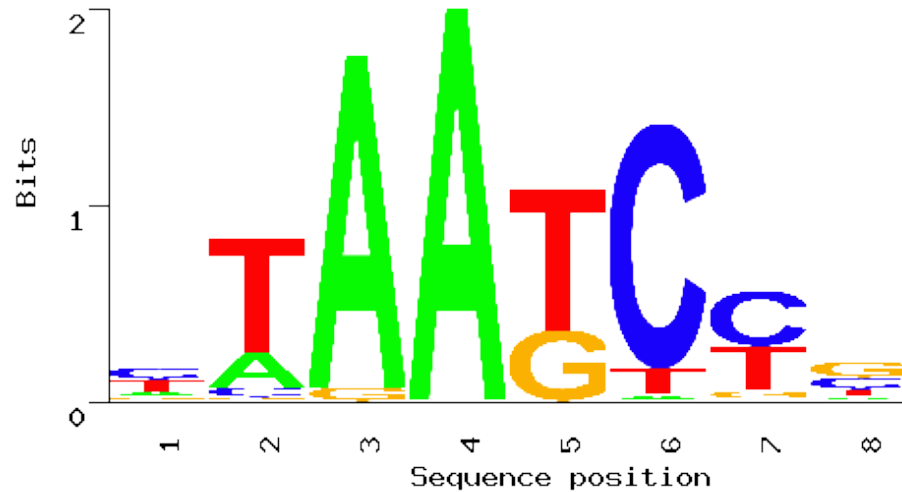
9	11	49	51	0	1	1	4
19	3	0	0	0	45	25	16
5	1	2	0	17	0	4	21
18	36	0	0	34	5	21	10

- Discovered:
 - Computationally
 - Traditional wet lab
 - Microarrays



Binding affinity matrices

9	11	49	51	0	1	1	4
19	3	0	0	0	45	25	16
5	1	2	0	17	0	4	21
18	36	0	0	34	5	21	10



Determined TF binding affinity matrices (+ JASPAR)

GLI1 GACCACCCAAG

GLI2 GACCACCCAAG

GLI3 GACCACCCAAG

Ci GACCACCCAAG

Tcf4 CCTTGAATG

c-ETS1 CCGGATGTT

Characterization of an enhancer module?

- an enhancer module (cis-regulatory module) is a collection of TF binding sites on DNA; no precise definition available
- properties of a module:
 - consists of several good binding sites of TFs
 - the sites are spatially clustered together
 - the pattern of sites is conserved

Finding conserved motifs of binding sites

- looking at one (human) genome gives too many positives
- comparative approach:
 - take the 200 kbp regions surrounding the same genes (paralogs and orthologs) of different mammals: human, mouse, chicken, ...
 - find preserved clusters (motifs) of binding sites
- **Smith-Waterman** type algorithm with a **novel scoring function**

Good binding sites

● Site of TF1 ● Site of TF2 ...

Human DNA



Clustering and conservation

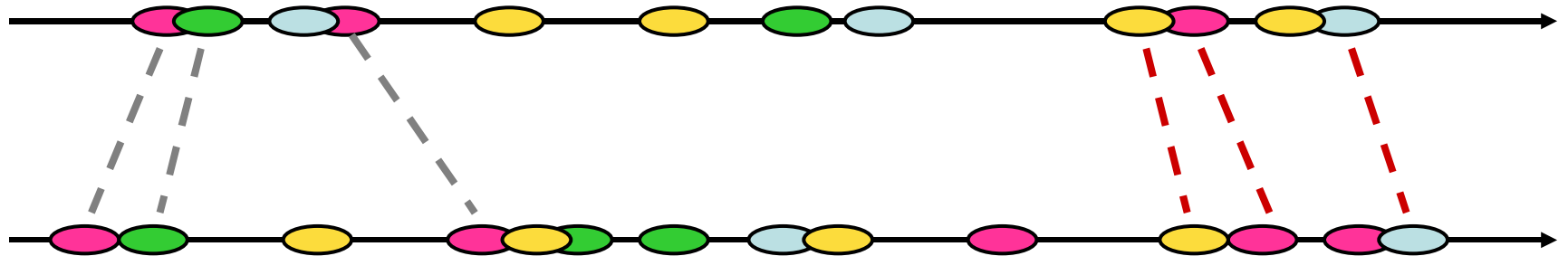
Human



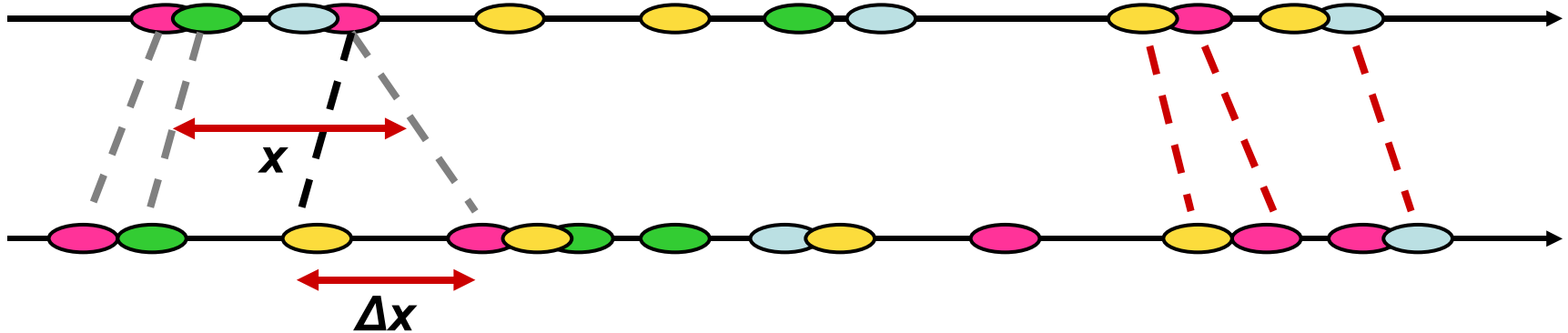
Mouse



Clustering and conservation



Alignment scoring



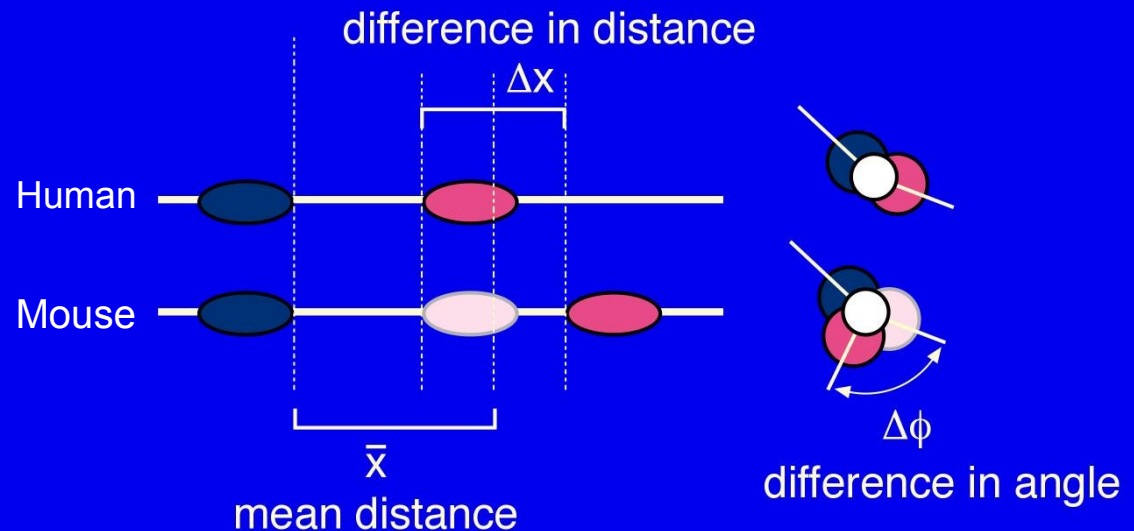
High binding affinity at the aligned sites: **bonus**

Long distance x between aligned sites: **penalty**

Non-conserved distance ($\Delta x > 0$) between aligned pairs: **penalty**

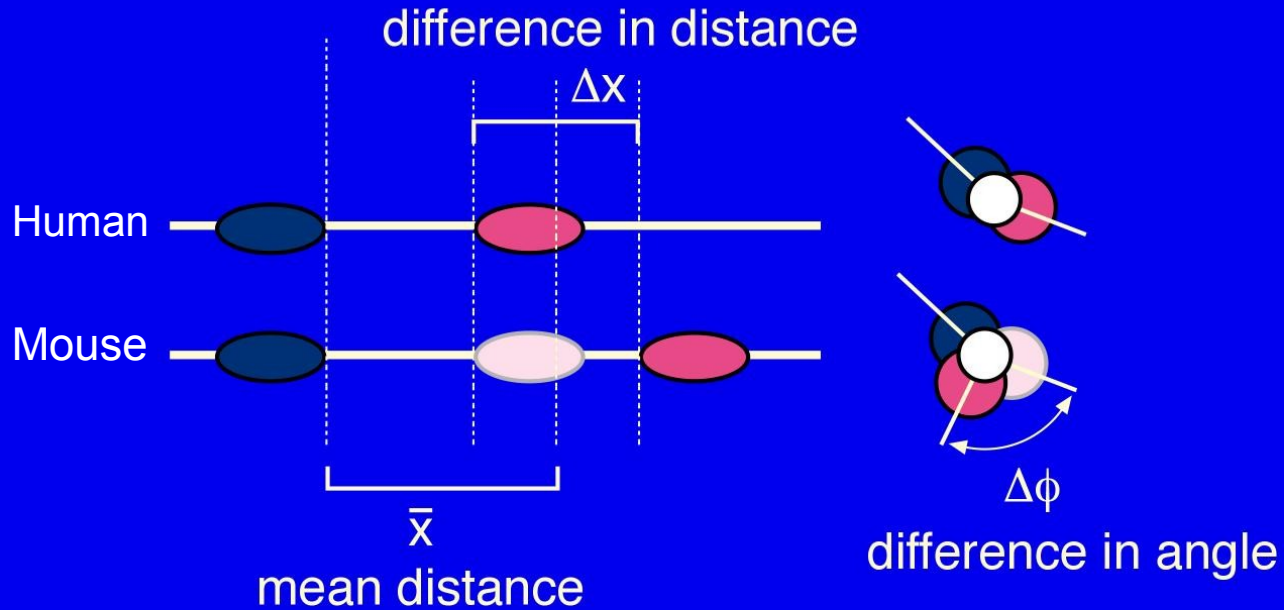
Computational identification of enhancer elements

- Preserved in evolution:
 - Affinities of functional cis-elements.
 - Spatial arrangement of elements within a module.



$$\text{Score} = \underbrace{\lambda \Delta G_T}_{\text{affinity}} - \underbrace{\mu \bar{x}}_{\text{clustering}} - \underbrace{\frac{\nu \Delta x^2 + \xi \Delta \phi^2}{2\bar{x}}}_{\text{conservation}} \quad \text{relative weights}$$

The scoring function

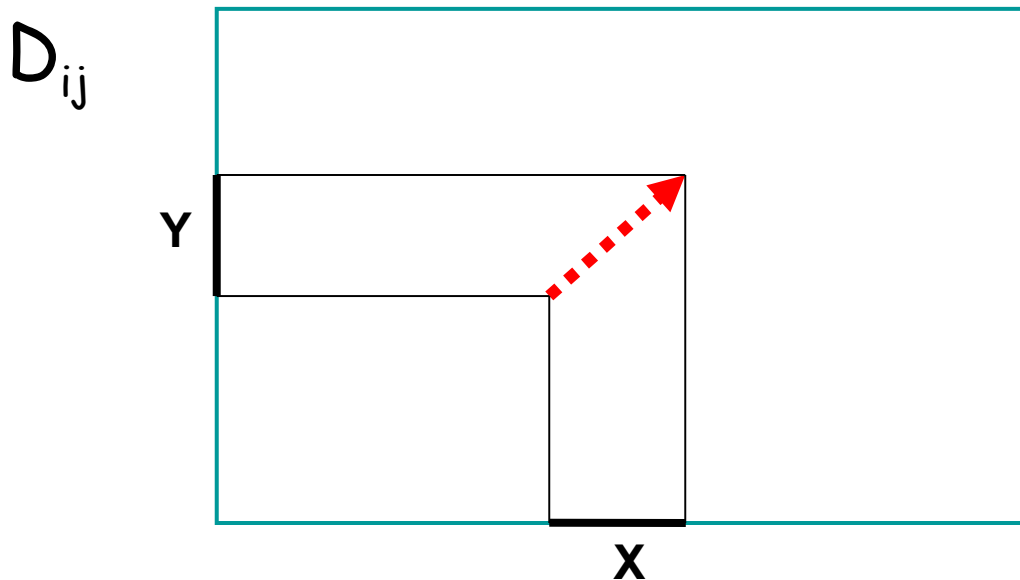


$$\text{Score} = \underbrace{\lambda \Delta G_T}_{\text{affinity}} - \underbrace{\mu \bar{x}}_{\text{clustering}} - \underbrace{\frac{\nu \Delta x^2 + \xi \Delta \phi^2}{2\bar{x}}}_{\text{conservation}}$$

relative weights

Smith-Waterman

- find the **best local alignment** of strings A and B: substring X of A and substring Y of B such that X and Y have the best scoring pairwise alignment



Dynamic programming

$$D_{ij} = \begin{cases} \max \{ \lambda w_{ij}, D_{k,l} + \lambda w_{ij} - F(p_i - q_k, p'_j - q'_l) \mid \\ 0 < p_i - q_k < 1000, 0 < p'_j - q'_l < 1000 \}, \\ \text{if } f_i = f'_j \text{ (i.e., the same TF aligned)} \\ -\infty, \text{ otherwise} \end{cases}$$

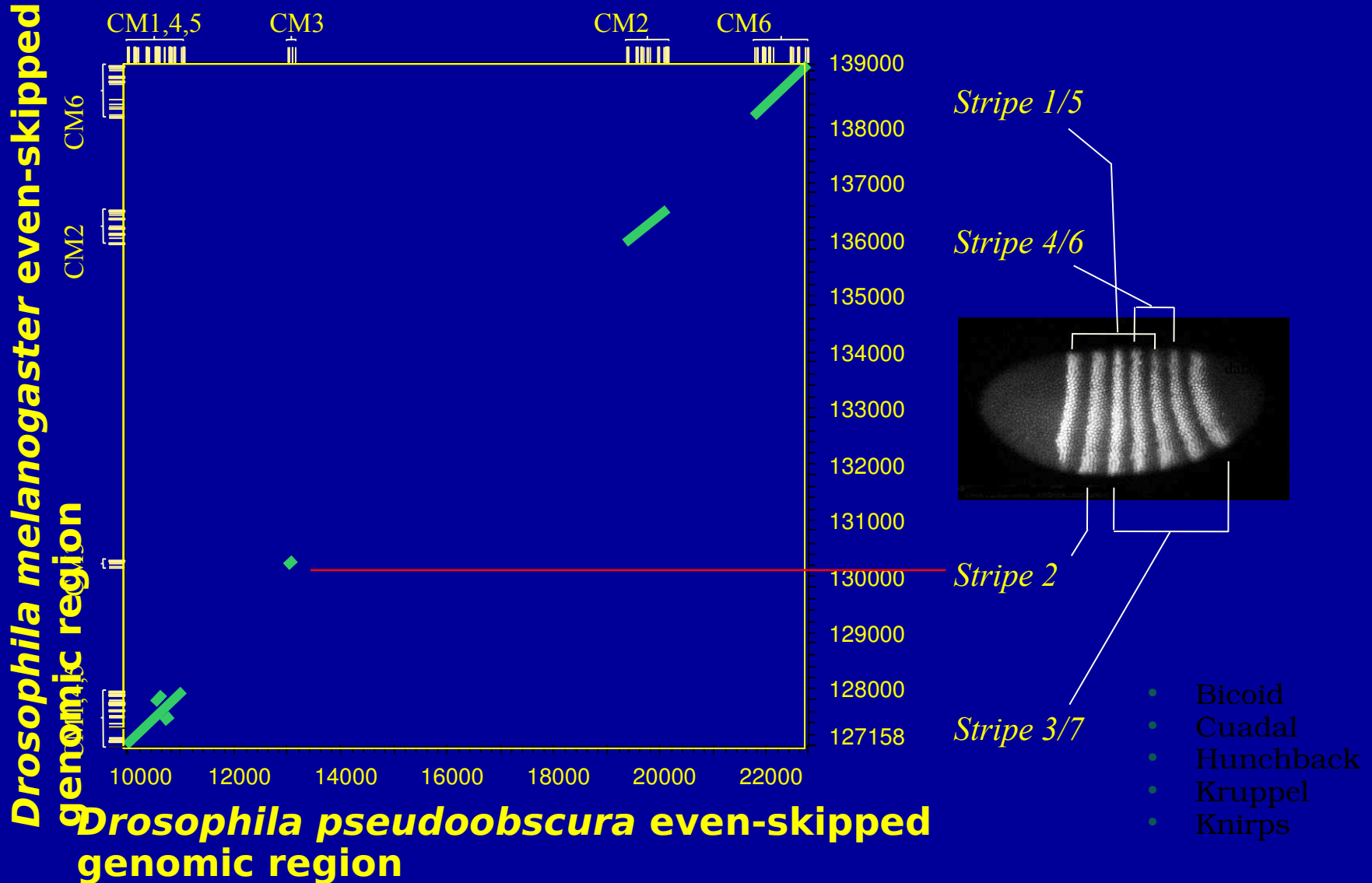
w_{ij} = sum of the binding affinities of the sites of the TF at i and j in the two sequences

$F(\Delta i, \Delta j)$ = penalty for the non-conservation and the length of the distances between adjacent sites

Parameter optimization

- scoring function has free parameters λ , μ , ν , ξ
- find good parameters by greedy hill climbing using a training data
- We used: $\lambda = 2$, $\mu = 0.12$, $\nu = 200$, $\xi = 200$

Enhancer prediction in *Drosophila*



Drosophila enhancer

- Output from EEL program
- *Drosophila* even-skipped gene stripe 2 enhancer
- Score = 487.05

	binding site		binding site ID	binding site position match		strand
	# match	score		<i>D. pseudoobscura</i>	<i>D. melanogaster</i>	
D[59]	[869]	=190.54	/TF/kni.pfm	(10508,10517)	<=> (127675,127684)	+
D[60]	[870]	=241.45	/TF/hb.pfm	(10519,10529)	<=> (127686,127696)	+
D[61]	[871]	=275.47	/TF/hb.pfm	(10531,10541)	<=> (127706,127716)	+
D[62]	[872]	=327.75	/TF/kni.pfm	(10548,10557)	<=> (127720,127729)	-
D[64]	[874]	=327.12	/TF/hb.pfm	(10663,10673)	<=> (127809,127819)	+
D[66]	[876]	=340.70	/TF/kni.pfm	(10768,10777)	<=> (127900,127909)	+
D[67]	[877]	=391.40	/TF/hb.pfm	(10780,10790)	<=> (127912,127922)	+
D[68]	[878]	=440.14	/TF/kni.pfm	(10810,10819)	<=> (127942,127951)	-
D[71]	[881]	=472.41	/TF/hb.pfm	(10850,10860)	<=> (127982,127992)	+
D[73]	[883]	=461.40	/TF/hb.pfm	(10977,10987)	<=> (128085,128095)	+
D[75]	[885]	=487.05	/TF/hb.pfm	(11031,11041)	<=> (128134,128144)	+

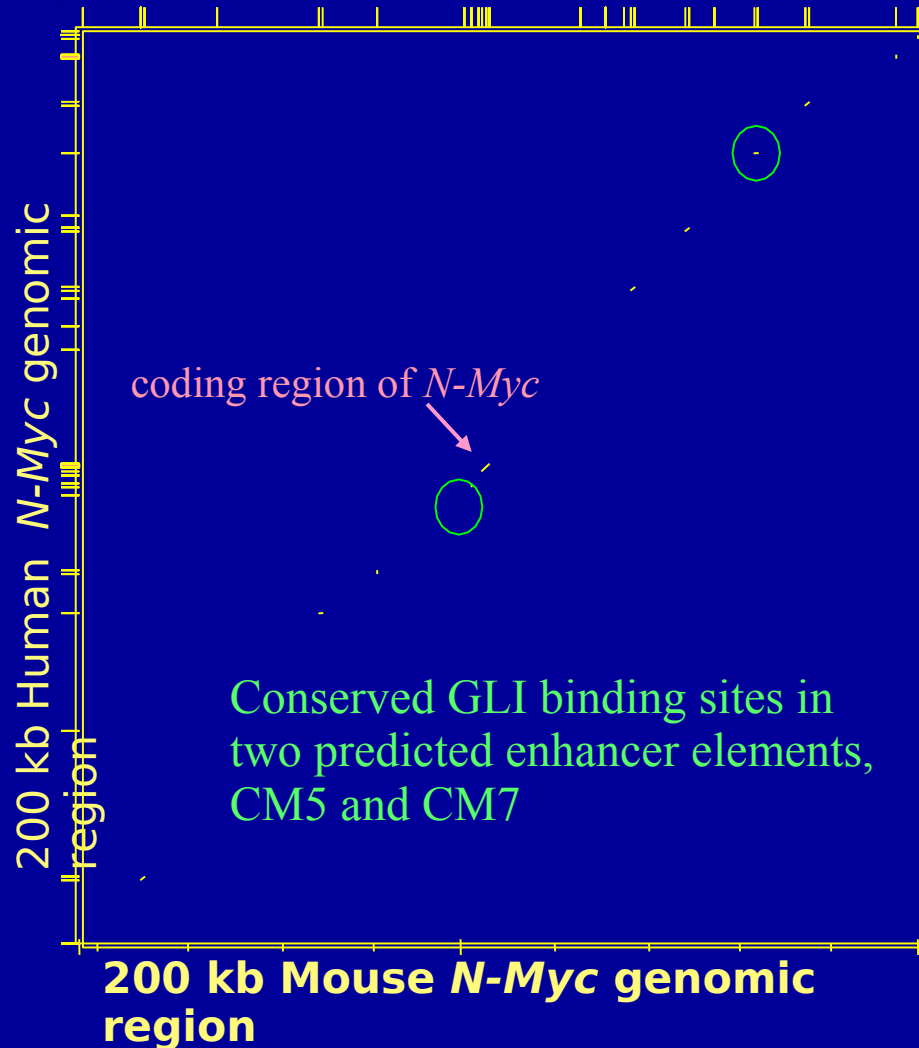

```

10460 : cccgaggatgcatcctggccctggcag-gacgacctcgctgcattagaaAACTAGATCAg
127636 : ---aggatcc-tc--gaaa-tcg-agag-cgacctcgctgcattagaaAACTAGATCAg
      Hb14                               Hb14                               Kni2
10519 : TTTTTTGTTCCT-----a---ATTTTTGTGCCctgcccTGCTCTCCTTtatgctttattggt
127686 : TTTTTTGTTCCTTggccgaccgATTTTTGTGCCcgg---TGCTCTCCTTtacg-----gtt

10571 : tatggtc-catttccatttcc-attttcatttccactcccattgtttggccgcaaaaca
127736 : tatggccgcggttcccatttcccagcttc-tttgt--tcc---g---ggct-cagaa-at
      Hb12
10629 : actccggacgggaattatggtatggtatatgcagATTTTTATGGG-cactcgggtgatct
127784 : -ct--gtatgg-aattatggtat---at--gcagATTTTTATGGGtc-c-cggc-gatcc
      Hb12
10688 : agttcgcggaatgggocgctatcctgtagcgct-gggacctcgaccggccctcggaggat
127832 : ggttcgcggaacgggagtg--cctgccgcgagaggt-cctcg-cggcgatcc-----t
      Kni4                               Hb11
10747 : atctgtatgtctatattaggaAAGTAGATCAagTTTTTTGTTCCTtttgtgcgctttttt
127883 : -t--gtc-gcccgtattaggaAAGTAGATCAcgTTTTTTGTTCCTcattgtgcgctttttt
      Kni5                               Hb10
10807 : cgcTGCGCTAGTTtttttccccgaacgcagcaaactgctctaaTTTTTTAATTCttcagc
127939 : cgcTGCGCTAGTTtttttccccgaaccagcgaactgctctaaTTTTTTAATTCttcagc

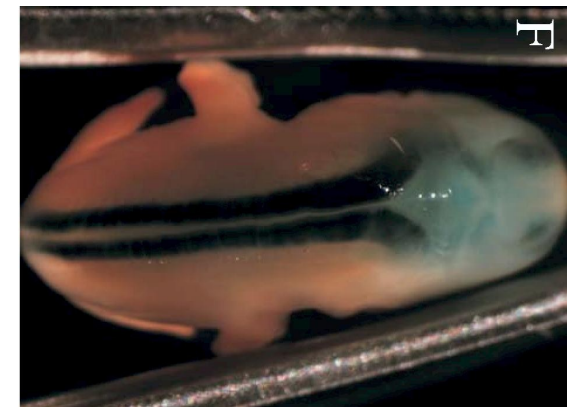
10867 : ggtttttattgtgctcctggaaaaactacggtcttccacaagggttagagcgttctagtta
127999 : gcttttcattgggctcctggaaaaacg-cgg-----acaagggttataacgctctacta
      Hb9
10927 : cctgttaattgtggcataaactcacattcacgtccgcattcagtgctctcATTTTTAAGA
128052 : cctgc-aattgtggcataaactc-----g-c---a--c--tgctctcGTTTTAAGA
      Hb8
10987 : Tatgttcttttctctgtgt-tttctgttctgttctgttcattcatATTTTTATGAGgct-
128095 : Tccgtt-tgtt-tgtgtttgttt--gtcc-gcgatgg-cattcacGTTTTACGAG-ctc
    
```

Enhancer prediction for *N-myc*

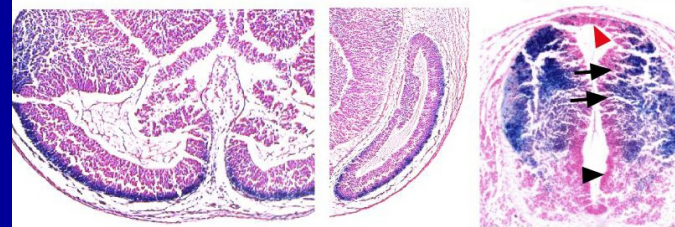
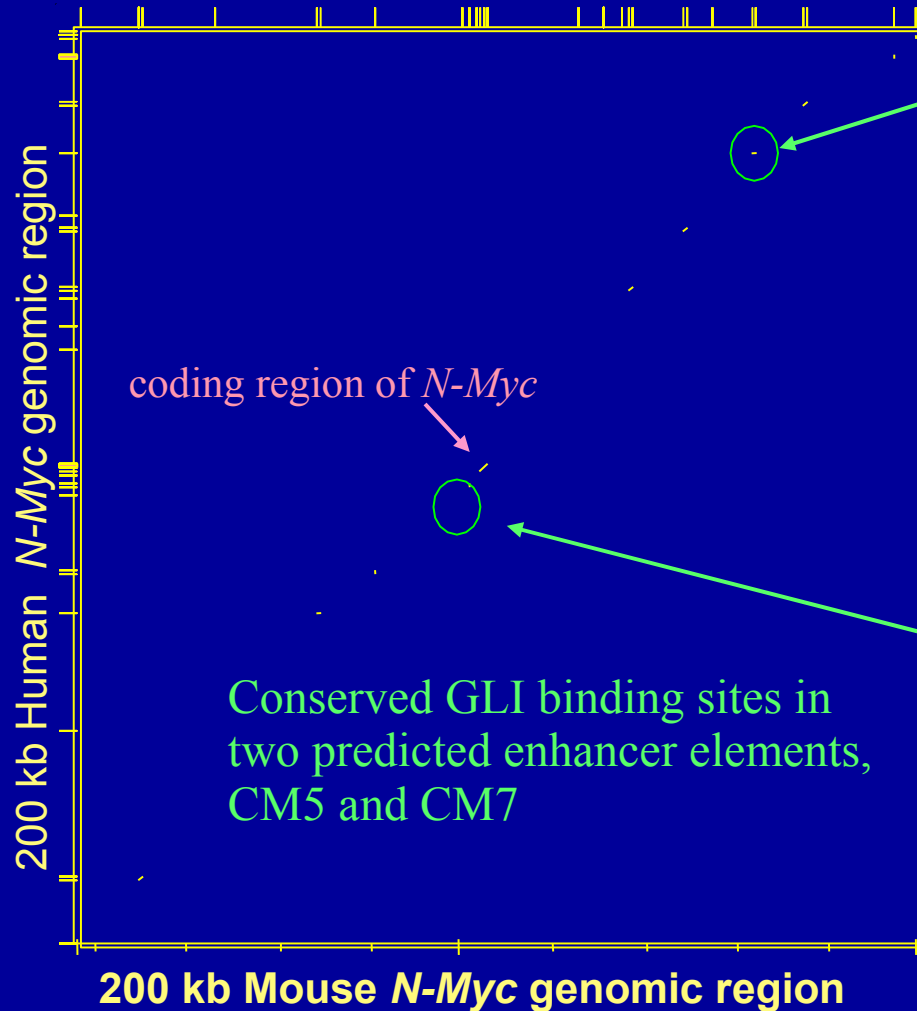


Wet-lab verification

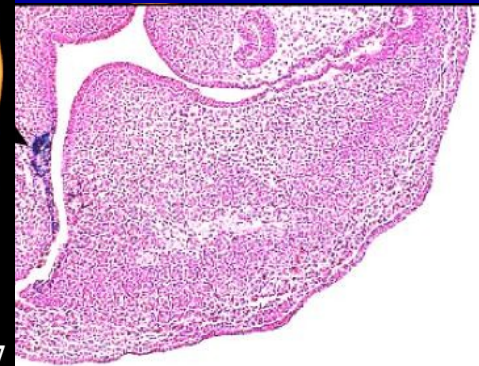
- Selected predicted cis-modules for wet-lab verification
- Fused 1kb DNA segment containing the predicted enhancer to a marker gene (LacZ) with a minimal promoter and generated transgenic embryos.



Enhancer prediction for *N-myc*



Tissue-specific findings



Genome scale comparisons: EEL

- Whole genomes can be analyzed with our implementation EEL (**Enhancer Element Locator**)
- We have compared: human genome vs mouse, rat, chicken, fugu, tetraodon and zebrafish
 - 100 kb flanking regions on both sides of the gene
 - Coding regions masked out
 - About 20 000 comparisons for each pair of species
 - About 2 min computing time for each pair

Summary of the protocol

- input: +/- 100 kb sequences of orthologous pairs of genes from human and mouse; TF affinity matrices
- find all good enough TF binding sites from the sequences
- find the best local alignments of the binding sites using the EEL scoring function
- output: the sequences in good local alignments; these are the putative enhancers
- Post-processing: an expert biologist selects most promising predictions for wet lab verification; hopefully he/she has good luck!

Availability of EEL

- EEL is available at <http://www.cs.helsinki.fi/u/palin/EEL/>
 - Paper: Nature Protocols (2006)
- A database of EEL predictions: sysdb.cs.helsinki.fi/~tkt_bsap
- paper: Hallikas, Palin *et al*, Genome-wide prediction ..., *Cell* 124,1 (Jan 13, 2006), 47-59.

Improving EEL: Scoring parameter optimization

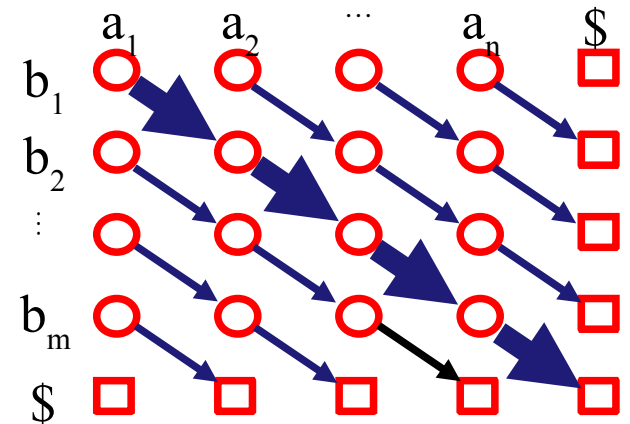
- The scoring function has 3 free parameters.
- Find good parameters by greedy hill climbing using a training data
- Better training data?

Significance of Alignment Scores of Evolutionarily Related Sequences

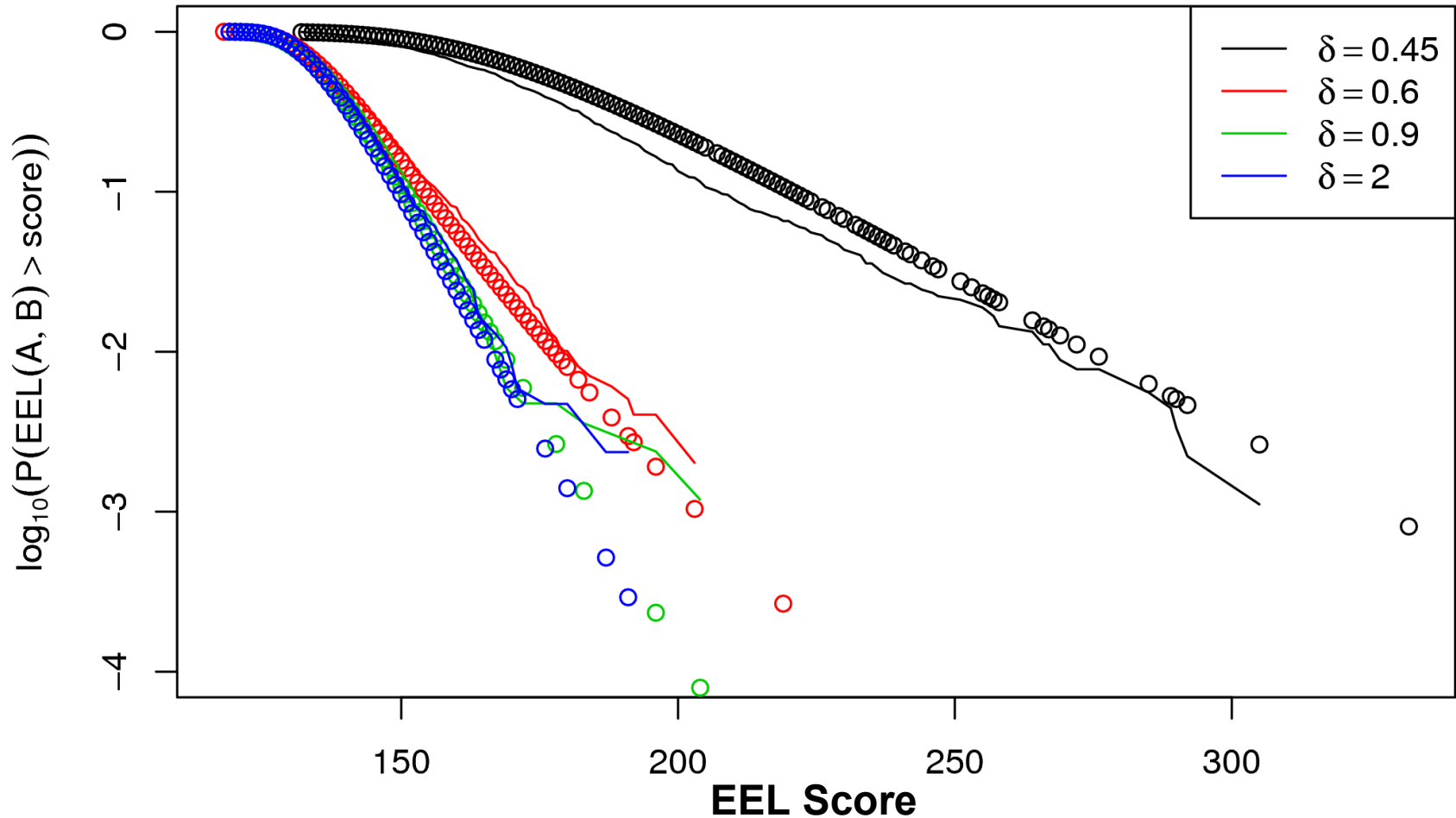
- The classic Karlin-Altschul statistics for independent sequences (a model for maximal segment sums) does not apply directly
- Two types of high scoring local alignments:
 - Independent characters off-diagonal
 - Dependent characters on-diagonal
- Two Component Extreme Value (TCEV) Distribution

$$P(S(A,B) < t) \approx$$

$$\text{Exp}(-K(n-1)me^{-\lambda t} - K'me^{-\lambda't})$$



TCEV distribution of EEL Scores



$n = m = 256\text{kbp}$ nucleotides from uniform distribution as ancestor. Evolutionary distance δ . Lines empirical from 10^3 samples. Points from fitted distribution.

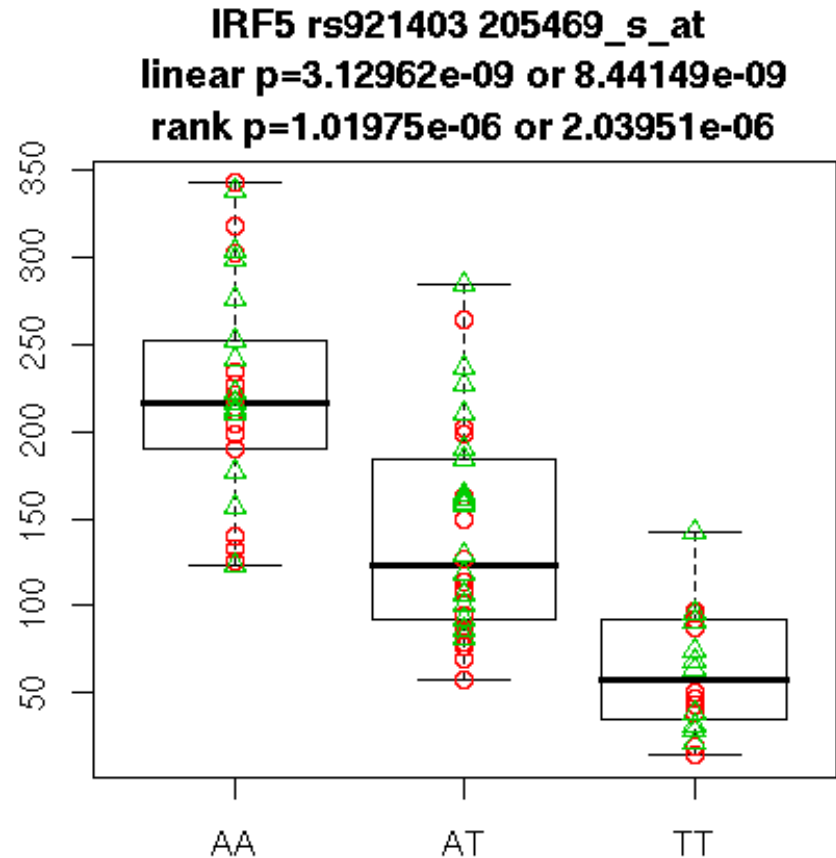
Improving EEL: Multiple alignment



- Star-alignment improves specificity
- Sum-of-pairs multiple alignment has been implemented but very slow (aligning three 10 kb long sequences takes a few minutes)

Predicting regulatory SNPs: Example rSNP from a training data

**Regulatory SNP:
(SNP, gene) pairs
with strong
correlation between
allele and
expression**



Morley et.al Nature 2004, HapMap

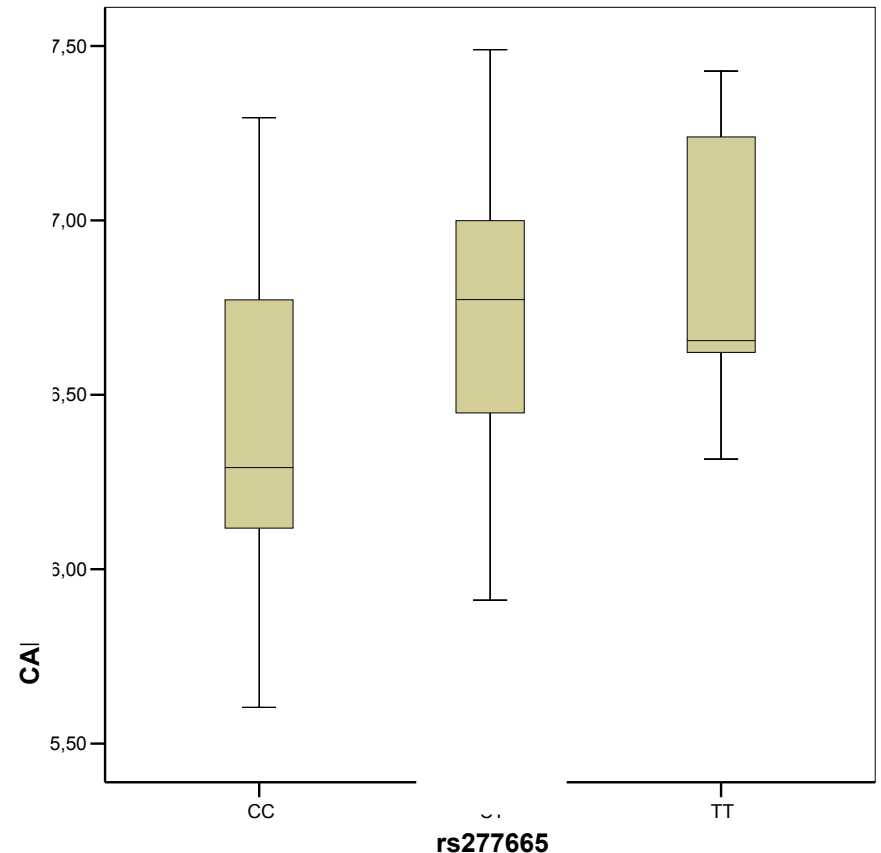
Predicting new rSNPs

- 18 000 genes, about 4 000 000 SNPs (ENSEMBLE)
- 4 relevant parameters
 - EEL score
 - expression level of the gene
 - the effect of SNP allele on the TFBS affinity
 - TFBS' affinity / maximal affinity of the TF
- Find parameter space regions where the training data is relatively frequent => the (SNP, gene) pairs in such regions are predicted to have regulatory SNP ('data mining') => genotyping ...

Testing of predictions

- Quite weak signal so far
- Improvements:
 - More careful modeling
 - better measurement of expression
 - allele-specific expression

CAMTA1__225693_s_at__ENSG00000171735



Disease-associated rSNP: colorectal cancer


CAGATAAGATAATGTAGTCAAAG**G/T**gcaggttaa

G seems to mean higher
cancer risk, by factor 1.5

Europe: 46% of people have **G**

China: 39% of people have **G**

SNP rsX occurs in an
EEL module



Acknowledgements

- Kimmo Palin (UH Computer Science, now at Sanger Instit.)
- Outi Hallikas (UH Biomedicum)
- Jussi Taipale (UH Biomedicum)

