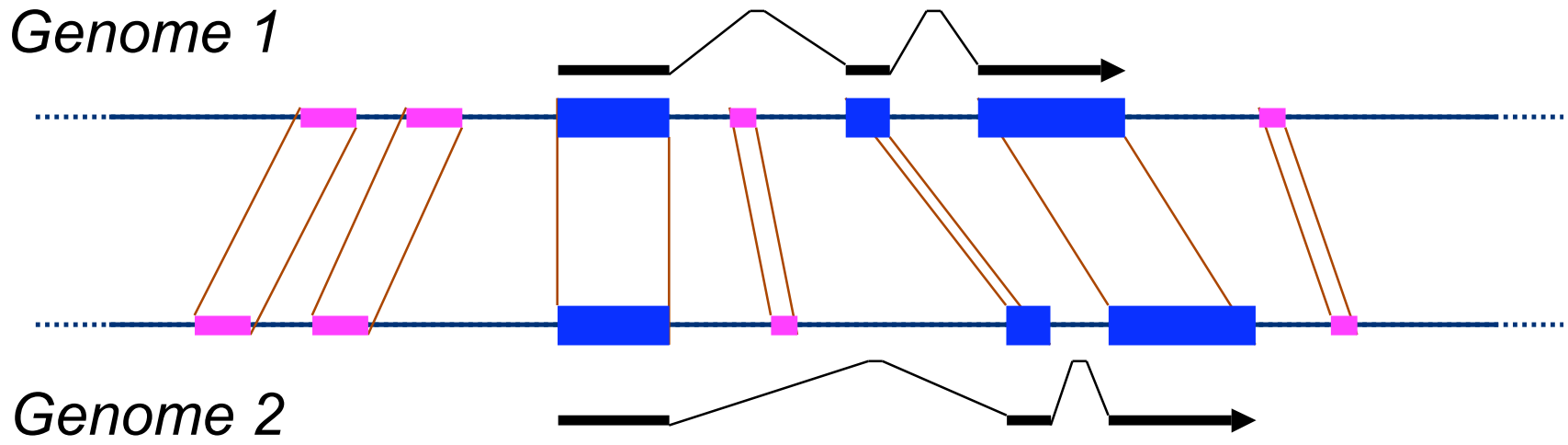
Regulatory Sequence Analysis

Applying comparative genomics to detect cis-acting elements

Jacques.van.Helden@ulb.ac.be
Université Libre de Bruxelles, Belgique
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/

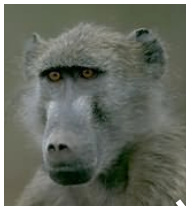
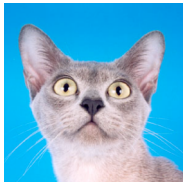
Phylogenetic footprinting to define regulatory regions

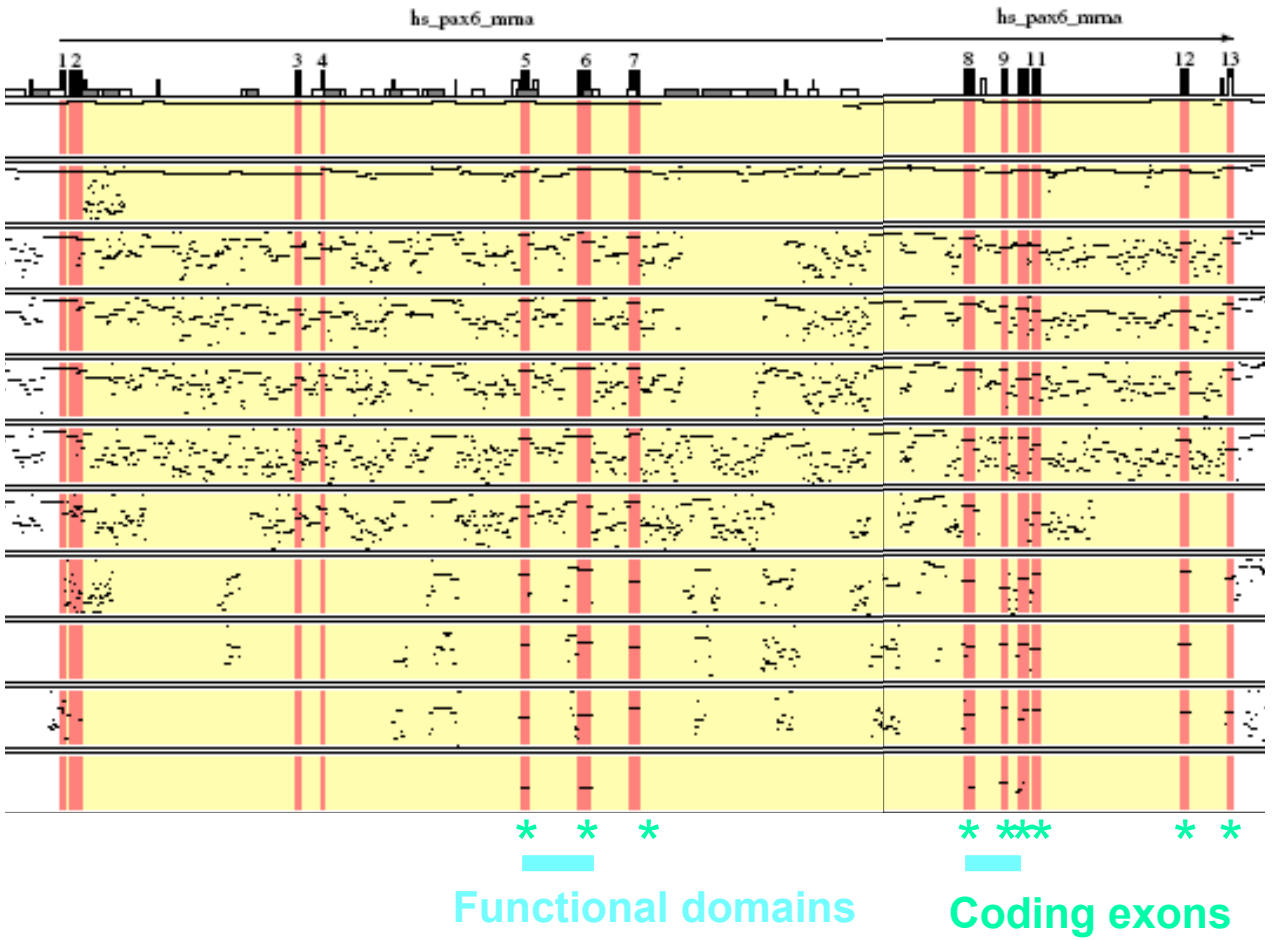


conserved exon *conserved non-coding region*

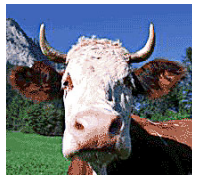
- Within non-coding sequences, regulatory elements evolve slower than their surrounding.
- Conserved non-coding sequences contain a high concentration in regulatory elements.

Phylogenetic footprints for the pax6 gene






Slide from Philippe Gautier

Pourcentages de positions identiques (PIP) dans la région chromosomique de Pax6

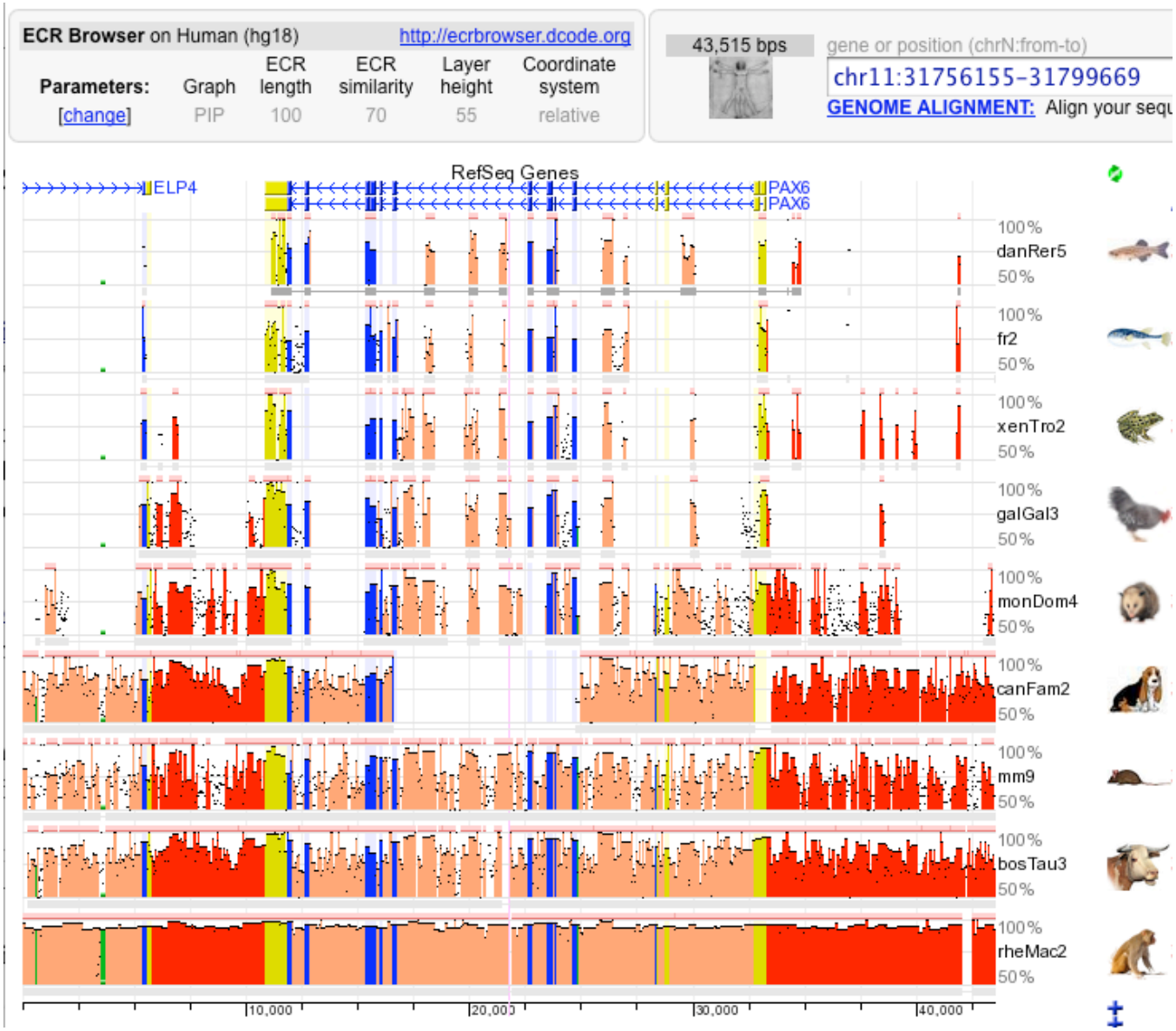
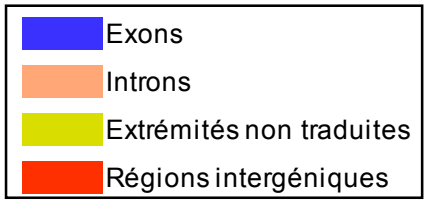
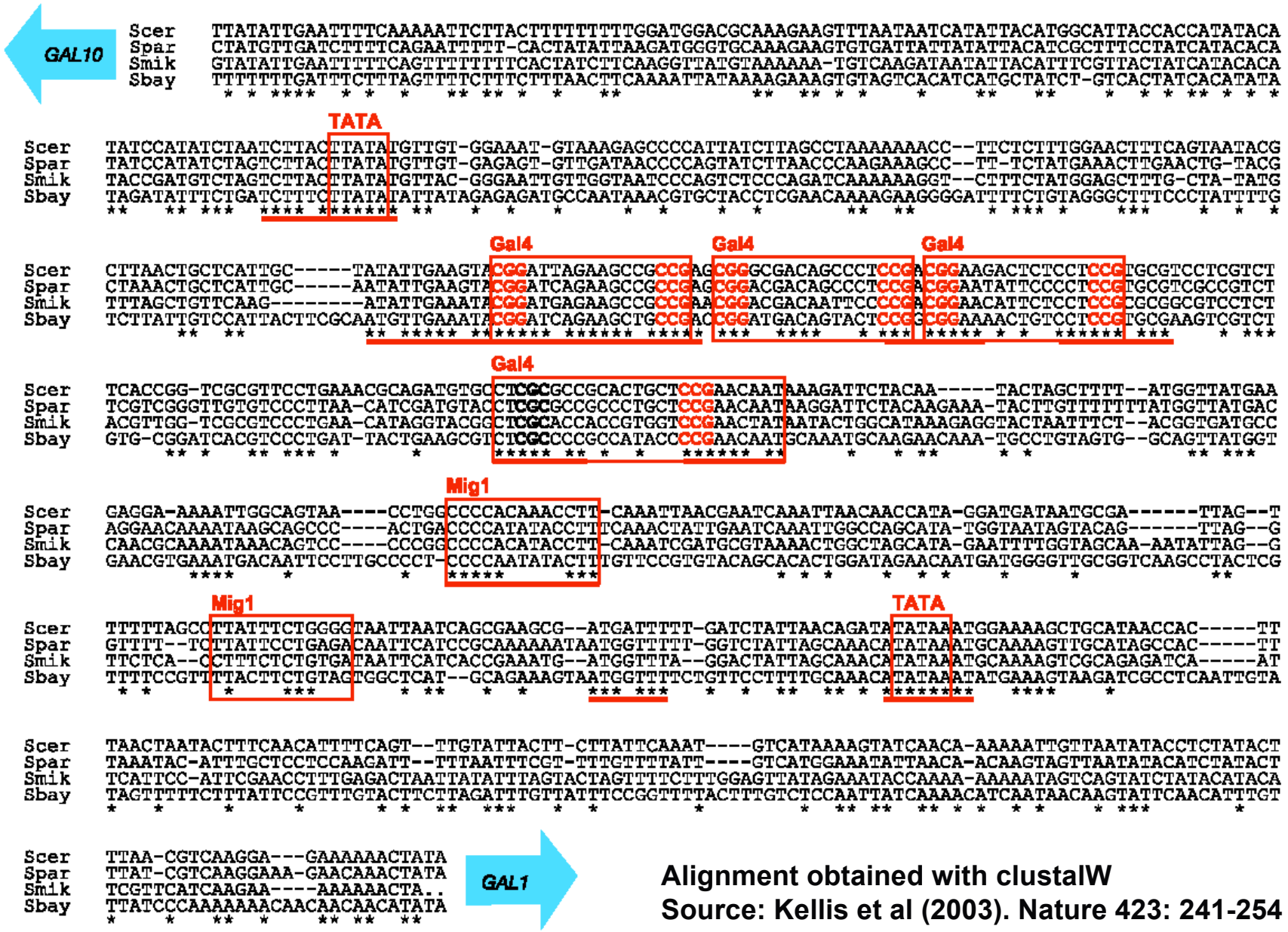


Image générée sur l'ECR browser (http://ecrbrowser.dcode.org/)

- La génomique comparative permet d'améliorer la localisation des gènes.
- Alignement de la région génomique contenant le gène Pax6, entre le génome humain et une série d'organismes de plus en plus distants évolutivement (de bas en haut).
- Les blocs de séquences conservées reflètent souvent la présence de fragments codants.
- Cependant, il existe également des segments conservés dans les régions non-codantes.



Global alignment of intergenic regions

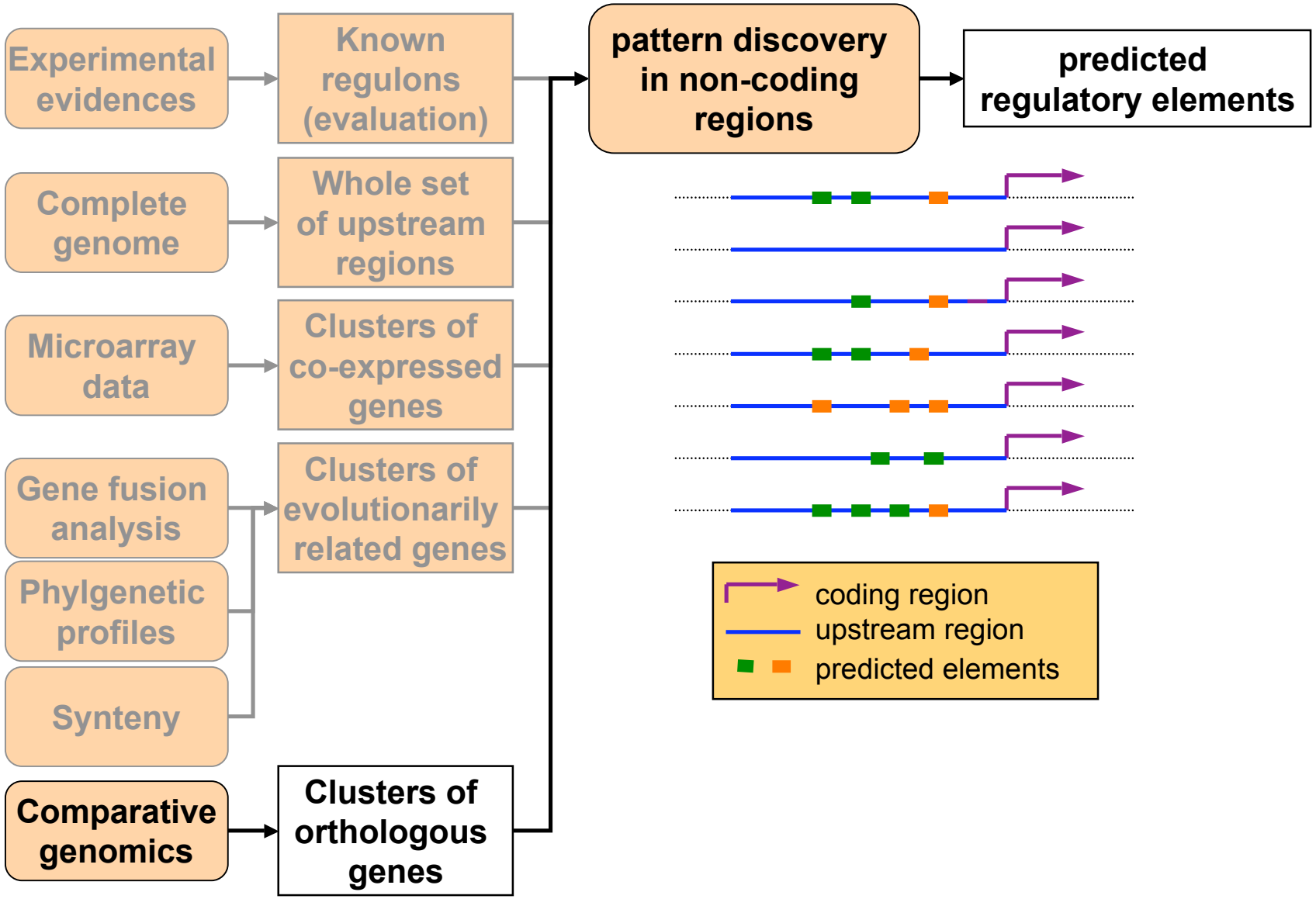


Alignment obtained with clustalW
Source: Kellis et al (2003). Nature 423: 241-254

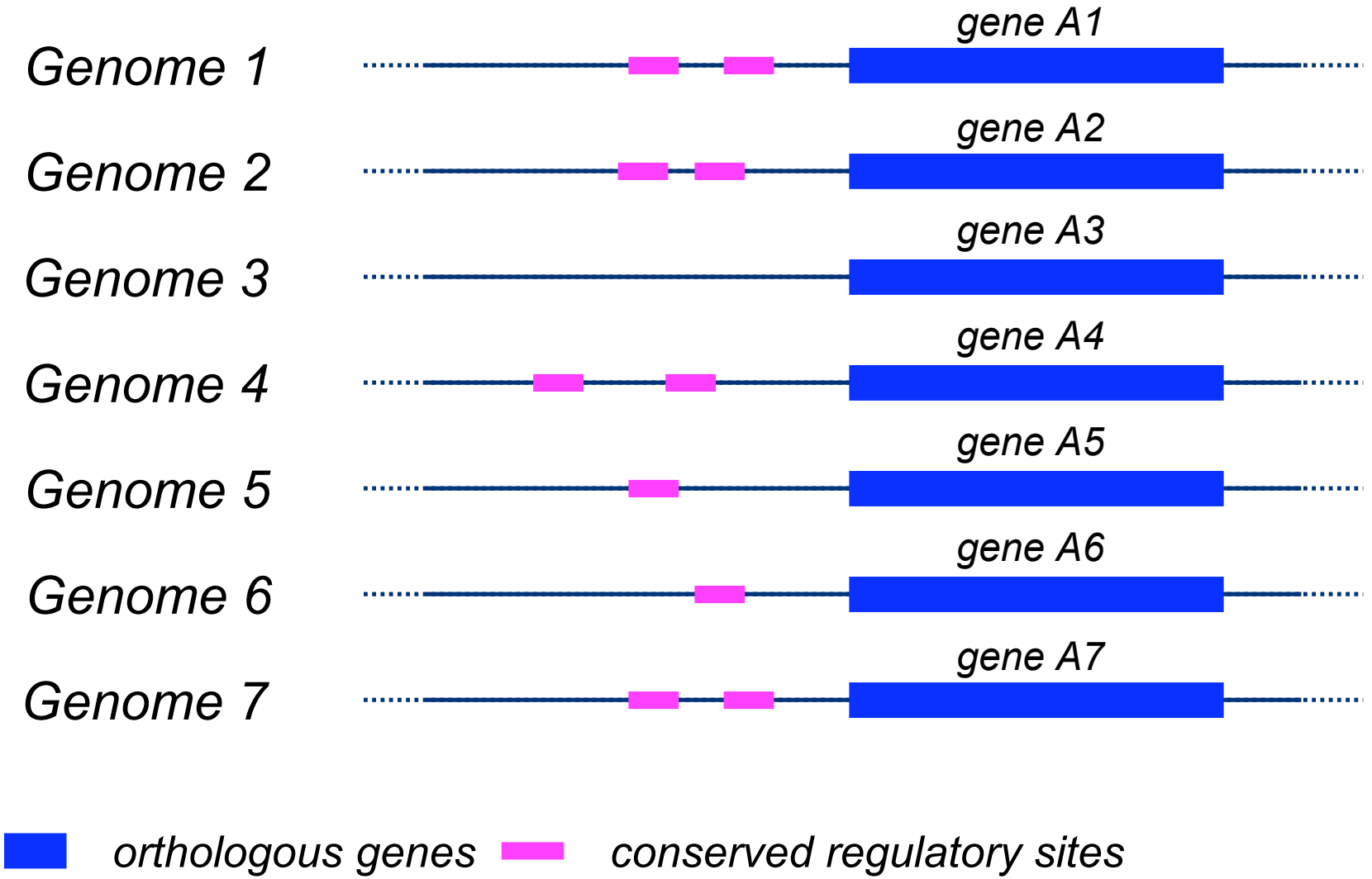
Another alignment in the same genomes

```
GAL80 (YML051W) upstream regions
Scer            ATGGCGCAAGTTTTCCGCTTTGTAATATATATTTATACCCCTTTCTTCTCTCCCCTGCAA
Spar            AGGGGCCAAAGCTCCCGCTCTGTAAAATATATTTATATCCCTTCCTTCTCTCCCCTGCAA
Smik            TAGGGACAAAGCCCGCCTTTTGTAATATATACTTATACCCTCTCCTTCTCTCCCCTGCAA
Sbay            ............................................................
                  **  ***      *  * ***** ***** ***** **  * ****************

Scer            TATAATAGTTTAATTCTAATATTAATAATA---TCCTATATTTTCTTCATTTACCGGCGC
Spar            TATAATAGTTTAATTCTAATATTAATAATA---TCCTATATTTTCCTTACC-ACCGGCGC
Smik            CATAATAGTTAACTCCTAATATTAATAATAATATCCTACAATTTCCTTAGC-ACCGGGGC
Sbay            ............................................................
                 ********* * * ***************   ***** * **** * *   ***** **

Scer            ACTCTCGCCCGAACGACCTCAAAATGTCTGCTACATTCATAATAACCAAAAGCTCATAAC
Spar            ACTCTCGCCCGAACGACCTCAAAATGCTTGCTACATTCATAATAATCAAAAGCTTATAAC
Smik            ACTCTCGCCCGAACGACCTCAAAACGCTTGCTACATCCATAATATTCAGAACTACATCAC
Sbay            ............................................................
                ************************ *  ******** *******  ** **    ** **

Scer            TTTTTTTTT----TGAACCTGAATATATATACATCACATATCACTGCTGGTCCTTGCCGA
Spar            TTTTTTTTTCCTTTGTACCTGAATATATATACATCTCATGTCACTGCTGGTCCTTGCCGG
Smik            TTTTTTTTT-----GTACATAAAAATATATAC--CACATGTCACTGCTGATCCTTGCTGA
Sbay            ............................................................
                *********     * ** * ** ********  * *** ********* ******* *

Scer            CCAGCGTATACAATCTCGATAGTTGGTTT-C-CCGTTCTTTCCACTCCCGTCATGGACTA
Spar            CCAGCGTATACAACCTCGATAGCTGGTTTTC-CCGTTCTTCCCACTCCTGTCATGGACTA
Smik            CGAGCGTATACAAGCTCGATAGCTGGTCTTTACCGTGCCATTCCCTGCCGTCATGGACTA
Sbay            ............................................................
                * ********** ******** **** *   **** *    * ** * **********
```

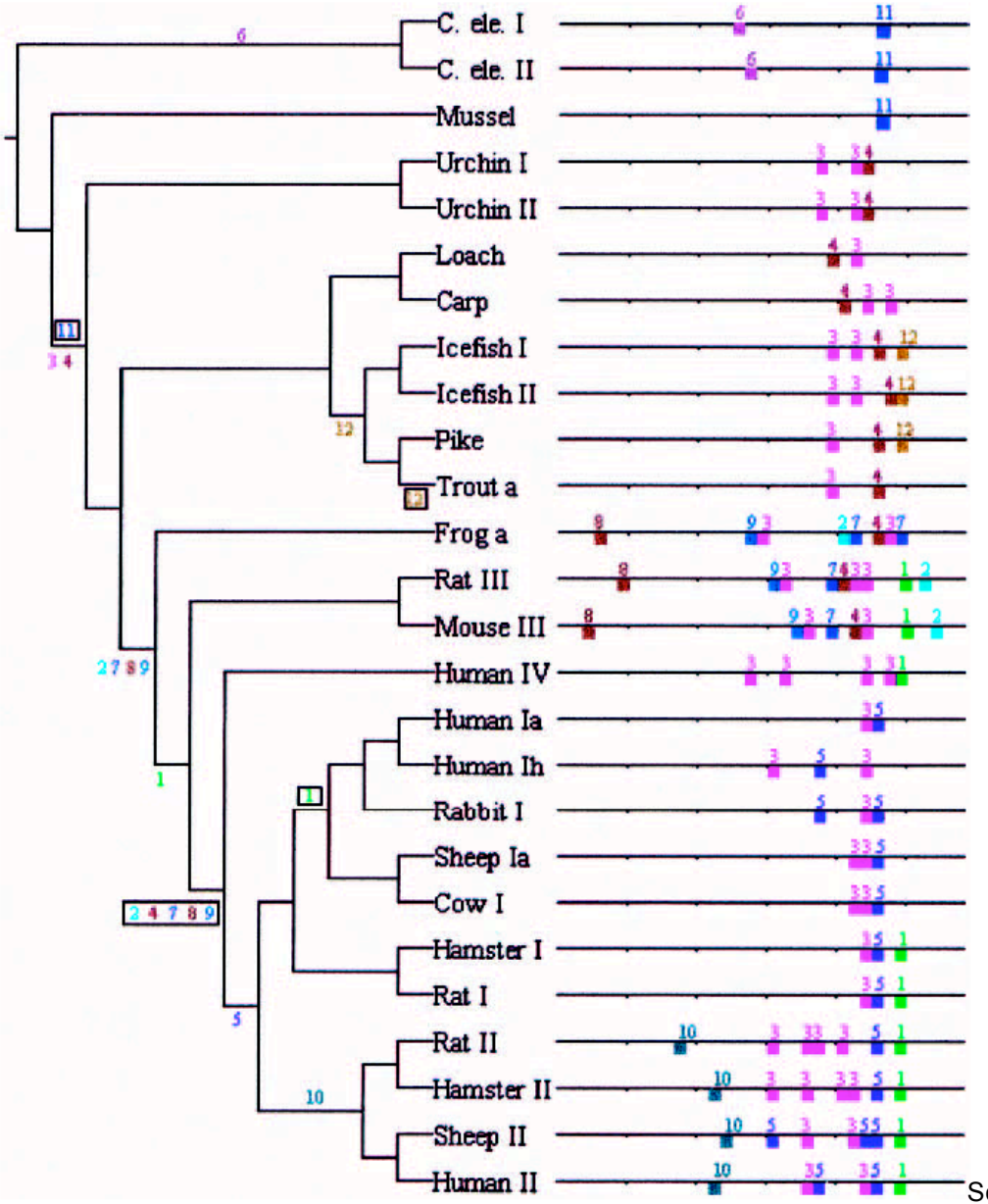
Motifs in clusters of orthologous genes (COGs)



Phylogenetic footprinting to predict regulatory sites



Footprinter example metallothionein

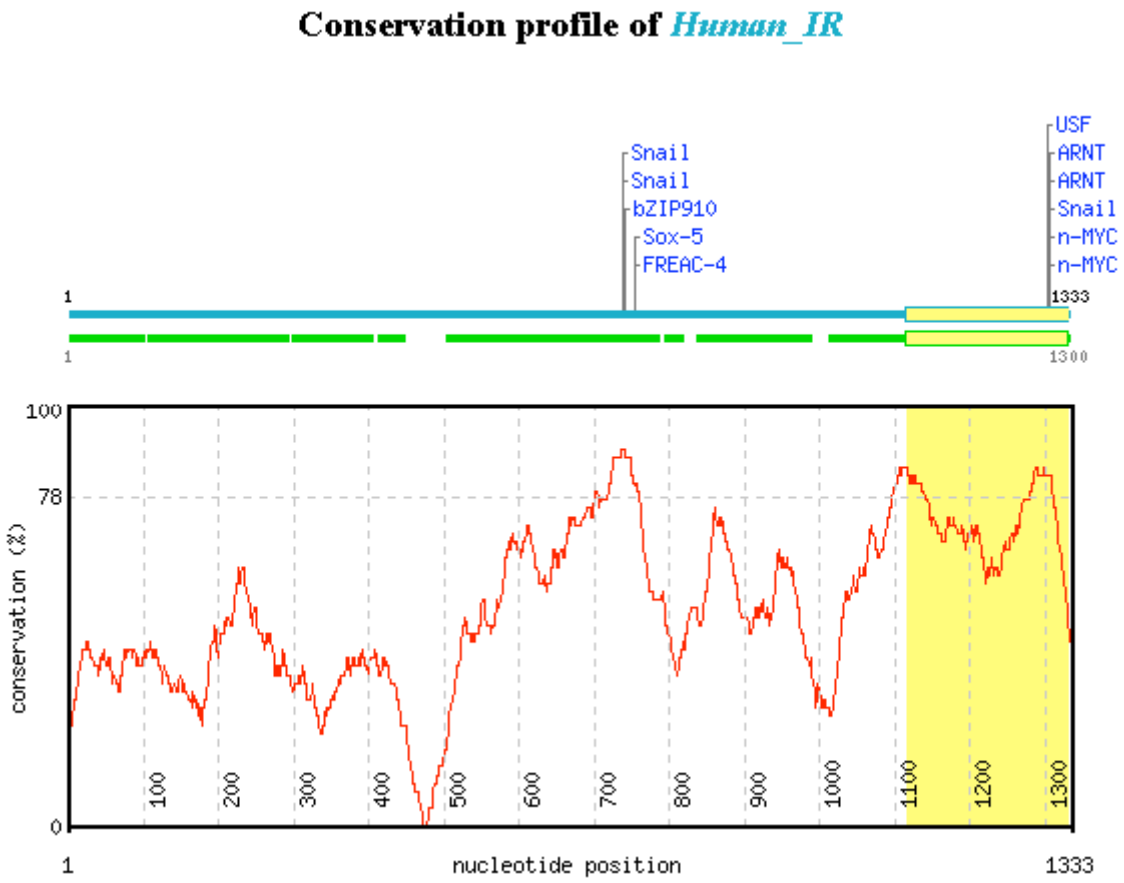


- 590 bp upstream of the same gene (methallothionein) in different species.
- 12 highly conserved motifs are detected.
- Each motif can be associated to a given internal node of the phylogenetic tree.

Source: Blanchette and Tompa (2002). Genome Research. 12, 739–748

Cross-matches in promoters of orthologous genes

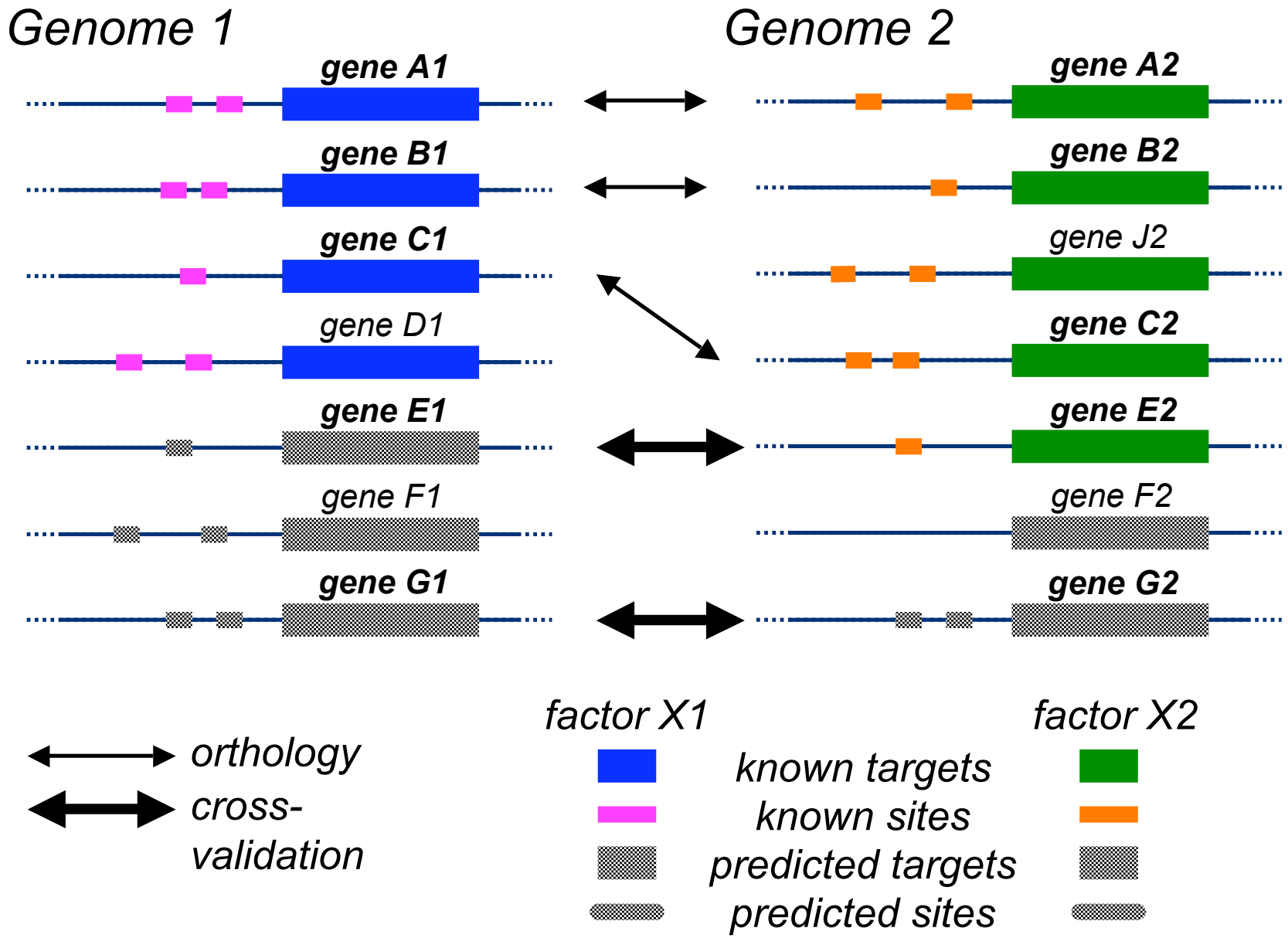
- Lenhard et al. (2003). J.Biology 2:13.
- 100 PSSM for known mammal transcription factors
- Searching for conserved matches in Human and mouse increases the selectivity by 85%.
- **Consite**: http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite/



Cross-validation of genome-scale pattern matching

- Genome-scale pattern matching raises many false positive
- Cross-validation :
  - gene *A* from genome *X* has a good match in its upstream sequence
  - ortholog *A'* from genome *Y* has a good match in its upstream sequence

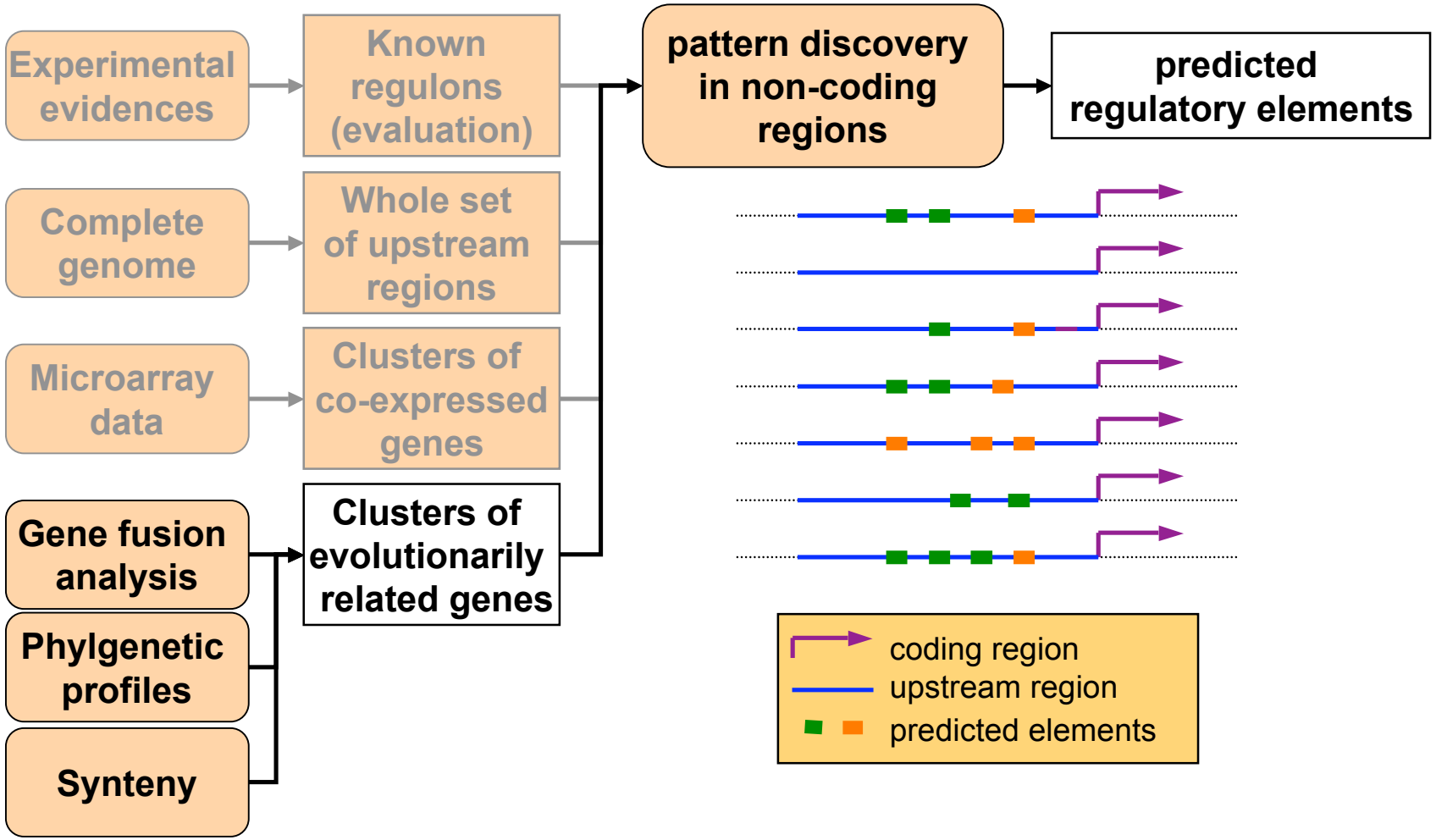
Cross-validation of pattern matching



Detection of functional clusters of genes

- Various methods of comparative genomics allows to detect clusters of functionally related genes
  - Operon conservation
  - Gene fusion analysis
  - Phylogenetic profiles (synteny)
- These functional clusters can be used to discover regulatory motifs in their upstream regions.

Clusters predicted from comparative genomics



Pattern discovery in predicted regulons

Organism	Cluster	pattern	reverse_complement	score
Escherichia_coli_K12	EC21	ccccctcaccctctt	aagagggtgaggggg	13.01
Escherichia_coli_K12	EC21	cccccctcaccccctt	aaggggtgaggggg	13.01
Escherichia_coli_K12	EC21	gccctcaccccctc	gaggggtgagggc	13.01
Escherichia_coli_K12	EC21	ggggagagggtgagggga	tcccctcaccctctcccc	13.01
Escherichia_coli_K12	EC21	cctcaccctcaccctctcccctc	gaggggagagggtgagggtgagg	13.01
Escherichia_coli_K12	EC3	ccccctcgcccctt	aaggggcgaggggg	12.73
Escherichia_coli_K12	EC3	aagggcgaggggg	ccccctcgccctt	12.73
Escherichia_coli_K12	EC3	gccctcgcccctc	gaggggcgagggc	12.73
Escherichia_coli_K12	EC3	ccccctcaccccctt	aaggggtgaggggg	12.73
Escherichia_coli_K12	EC3	ccccctctcccctt	aaggggagaggggg	12.73
Mycoplasma_pneumoniae	MP1	tataatact	agtattata	11.75
Mycoplasma_pneumoniae	MP1	cttaatactaat	attagtattaag	11.75
Escherichia_coli_K12	EC17	cccctctccctt	aagggagagggg	10.63
Escherichia_coli_K12	EC17	cccctctcccctt	aaggggagagggg	10.63
Escherichia_coli_K12	EC17	cccctcgccctt	aagggcgagggg	10.63
Mycoplasma_pneumoniae	MP1	aataataag	cttattatt	10.4
Mycoplasma_pneumoniae	MP1	aataatattatt	aataatattatt	10.4
Mycoplasma_pneumoniae	MP1	taataataagnnnnnaataa	ttattnnnnncttattatta	10.4
Mycoplasma_pneumoniae	MP1	cttagtattatt	aataatactaag	10.4
Mycoplasma_pneumoniae	MP1	taataataagnnnnnaataa	ttattnnnnncttattatta	10.4
Mycoplasma_pneumoniae	MP1	aataatattaaga	tcttaatattatt	10.4
Mycoplasma_pneumoniae	MP1	cttagtatataataatatactaag	cttagtatatattatatatactaag	10.4
Mycoplasma_pneumoniae	MP1	taataataagnnnnnaataa	ttattnnnnncttattatta	10.4
Mycoplasma_pneumoniae	MP1	ctaatattatt	aataatattag	10.4
Mycoplasma_pneumoniae	MP1	taataataagnnnnnaataa	ttattnnnnncttattatta	10.4
Mycoplasma_pneumoniae	MP1	aataatattattnnngtactattataataag	cttattataatagtacnnnaataatattatt	10.4
Mycoplasma_pneumoniae	MP1	aataatattatc	gataatattatt	10.4

gene clusters from McGuire et al. (2000). *Genome Res* 10(6), 744-57.

Phylogenetic footprinting resources

- CORG: a database for COmparative Regulatory Genomics
  - Dieterich et al. (2003), Nucleic Acids Res. 31:55-57.
  - http://corg.molgen.mpg.de
  - Systematic alignment of 15Kb upstream regions for each pair of mouse-human homologous genes (18.674 pairs).
  - 10.793 significant alignments ($P < 0.001$), containing 293.503 conserved non-coding blocks (CNB), covering 8% of the upstream sequences (http://corg.molgen.mpg.de/stats.html).

Summary - phylogenetic approaches

- Matching conserved sites for known transcription factors
  - Consite: http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite/
  - Lenhard et al. (2003). J.Biology 2:13.
- Global alignment of promoters of orthologous genes
  - clustalW
  - e.g.: Kellis et al (2003). Nature 423: 241-254.
- Pattern discovery in promoters of orthologous genes
  - Footprinter: http://bio.cs.washington.edu/software.html
  - Blanchette and Tompa (2002). Genome Research. 12, 739–748.