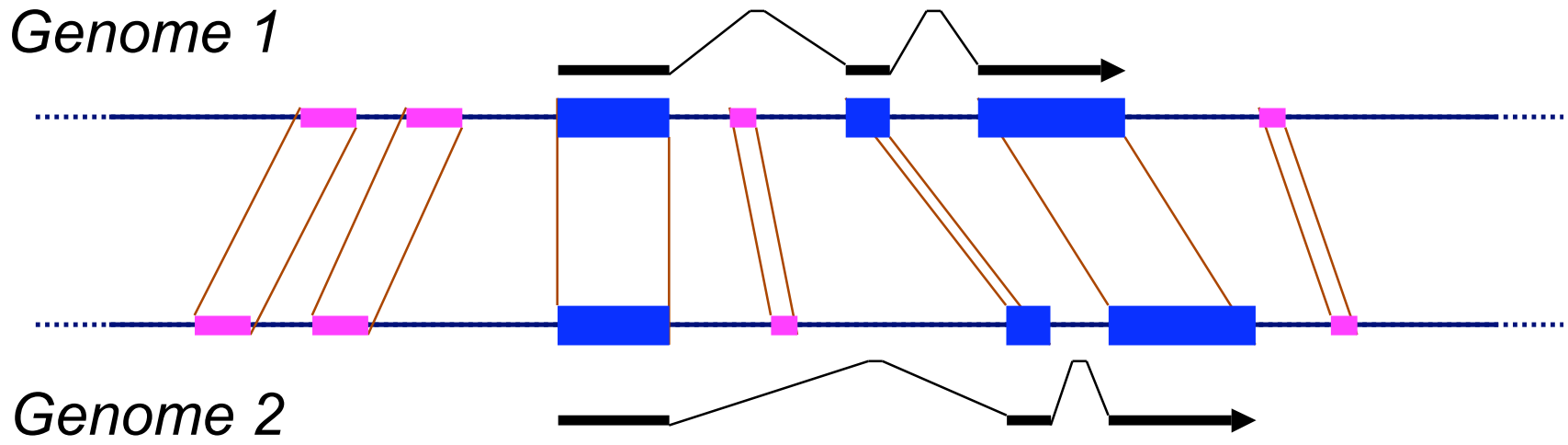
Applications of comparative genomics to the analysis of regulatory sequences

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

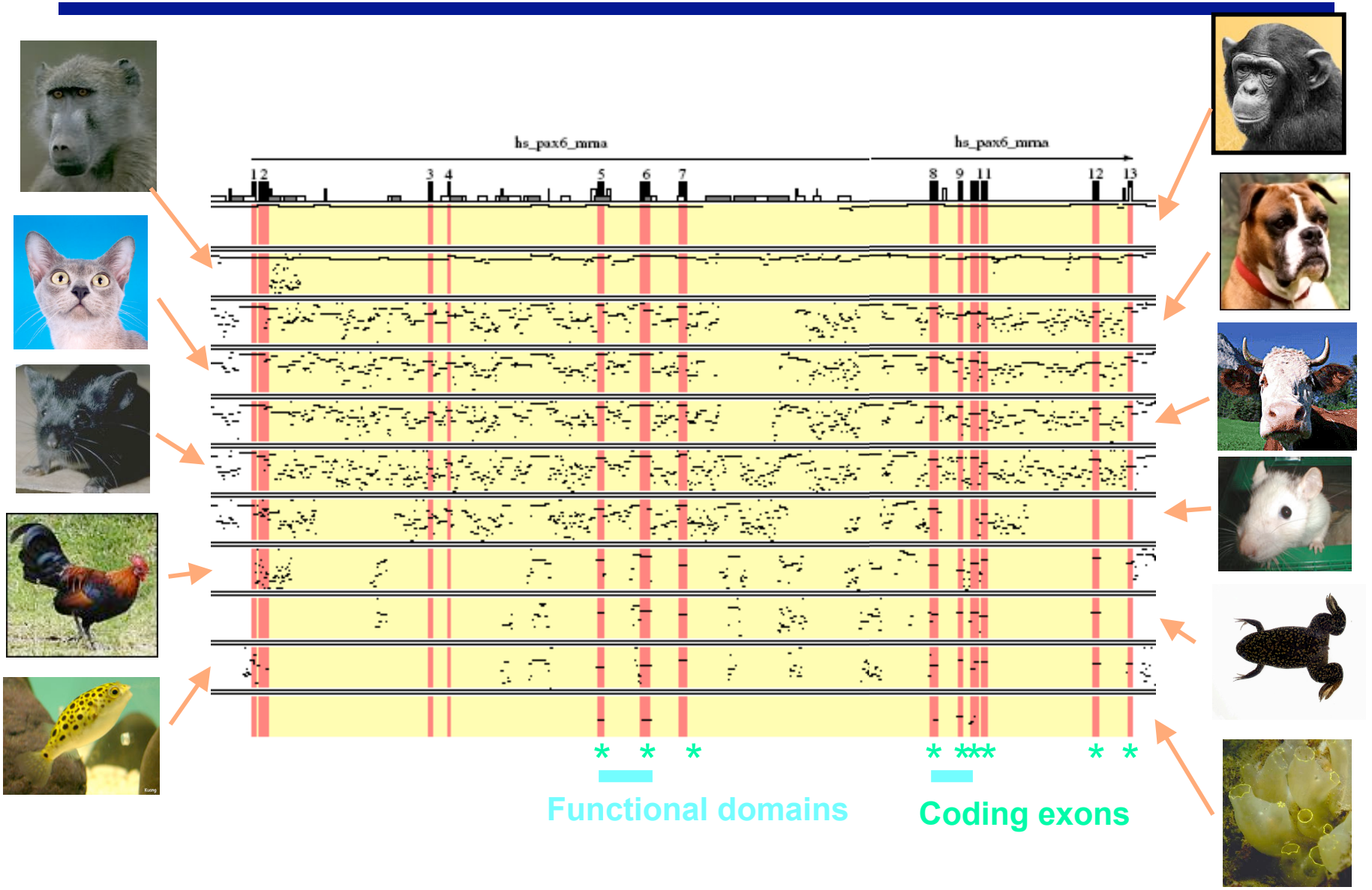
Phylogenetic footprinting to define regulatory regions



conserved exon *conserved non-coding region*

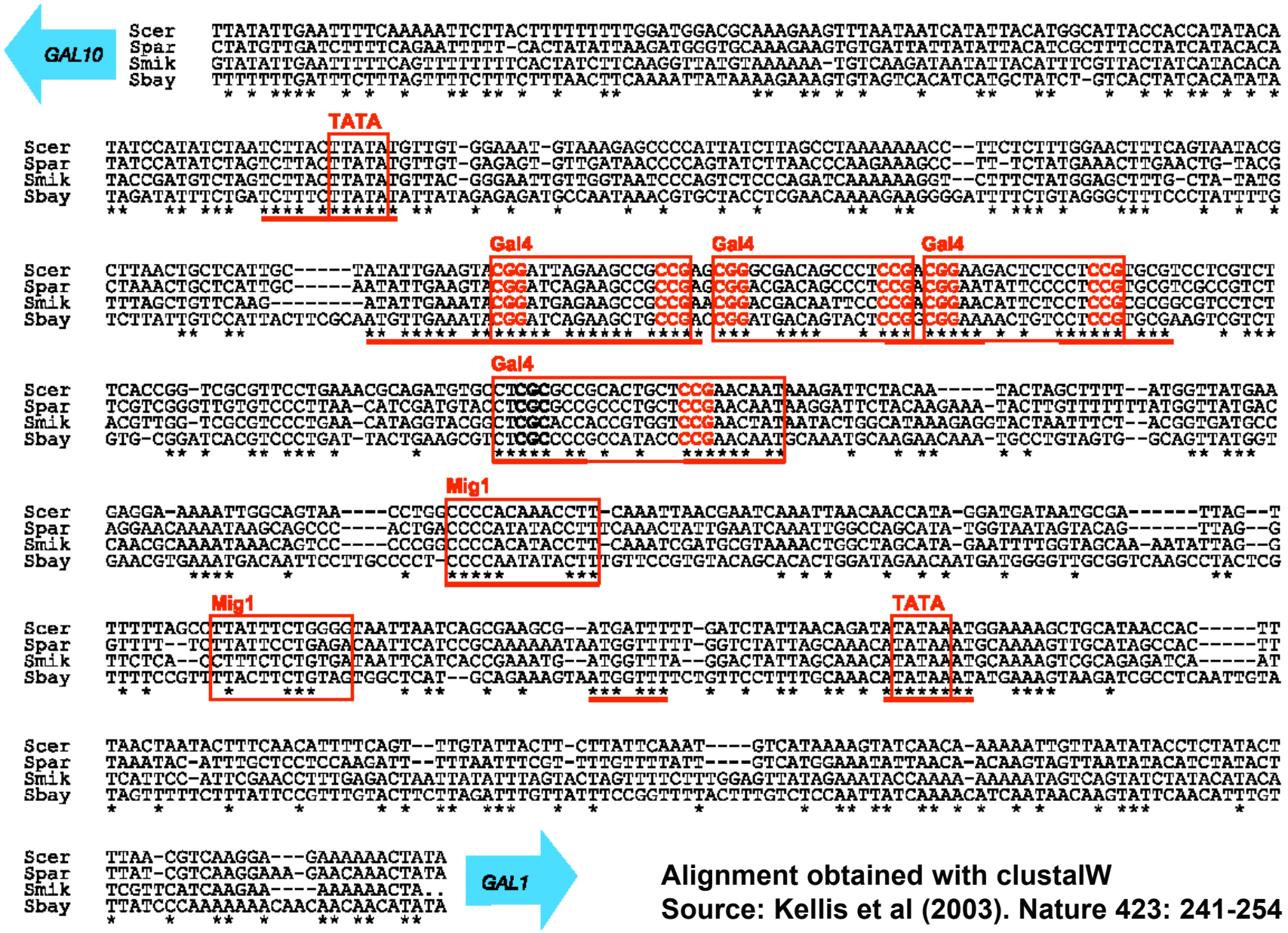
- Within non-coding sequences, regulatory elements evolve slower than their surrounding.
- Conserved non-coding sequences contain a high concentration in regulatory elements.

Phylogenetic footprints for the pax6 gene



Slide from Philippe Gautier

Global alignment of intergenic regions



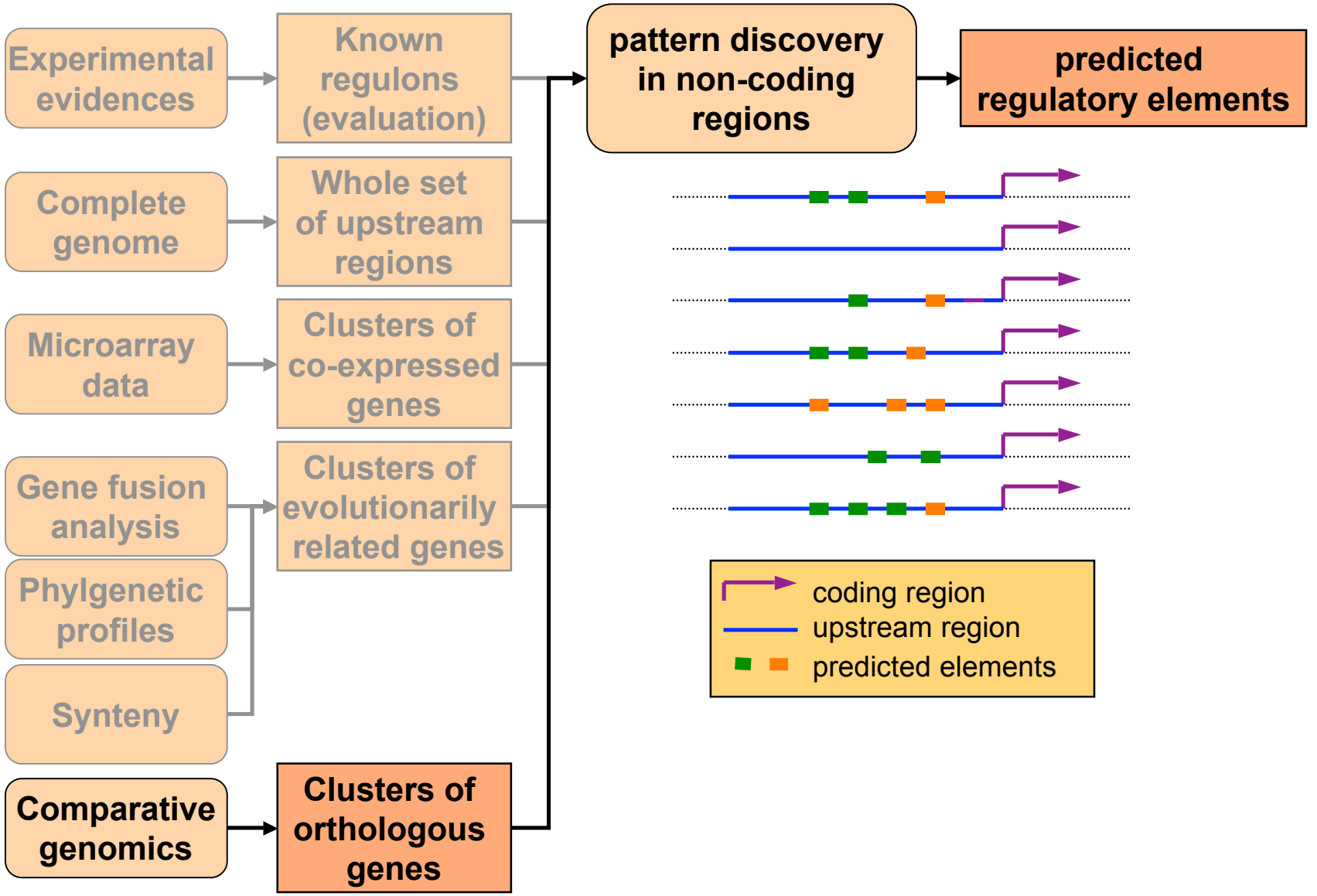
Alignment obtained with clustalW

Source: Kellis et al (2003). Nature 423: 241-254

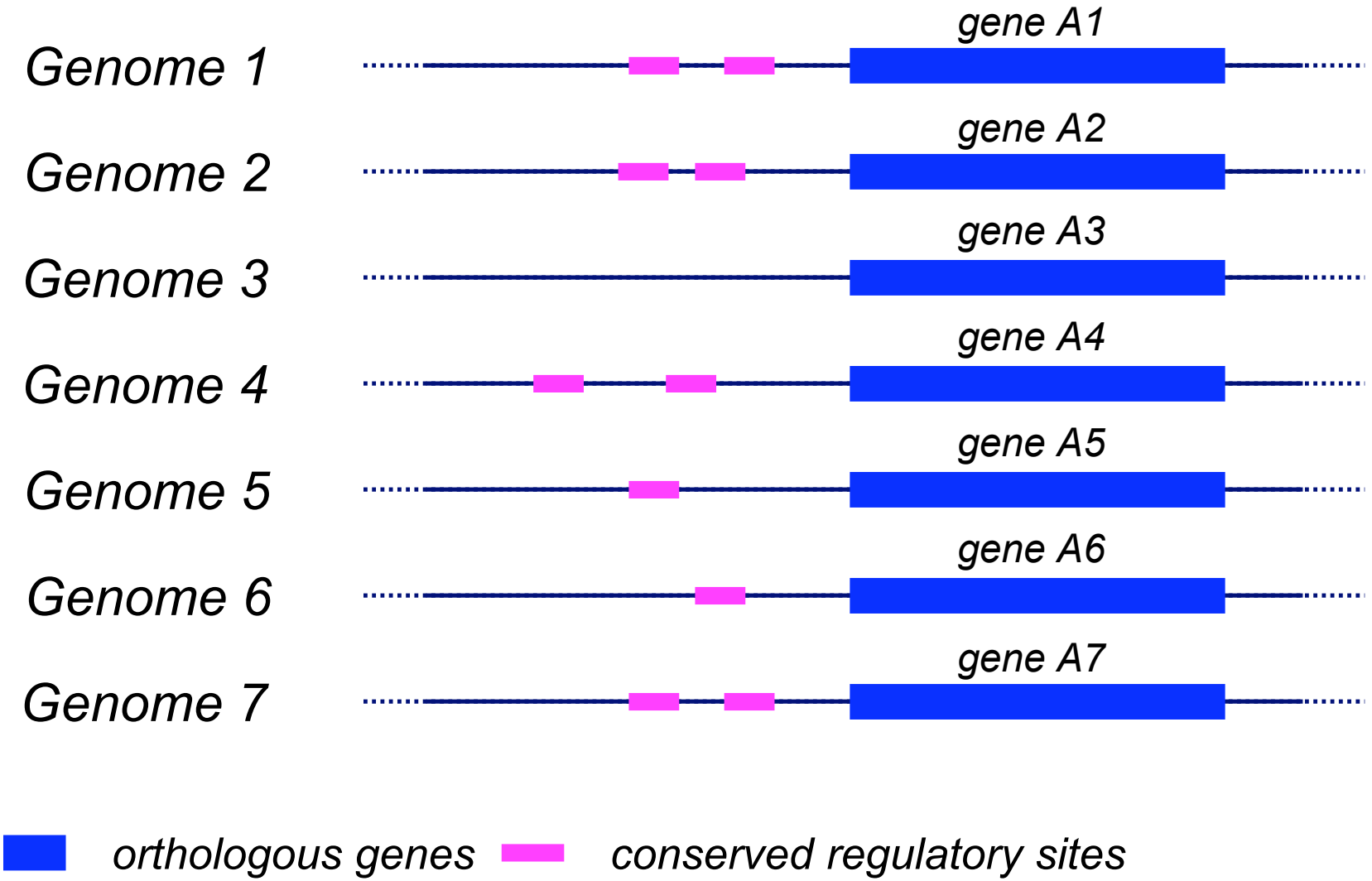
Another alignment in the same genomes

```
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            ............................................................
                * ********** ******** **** *   **** *    * ** * ***********
```

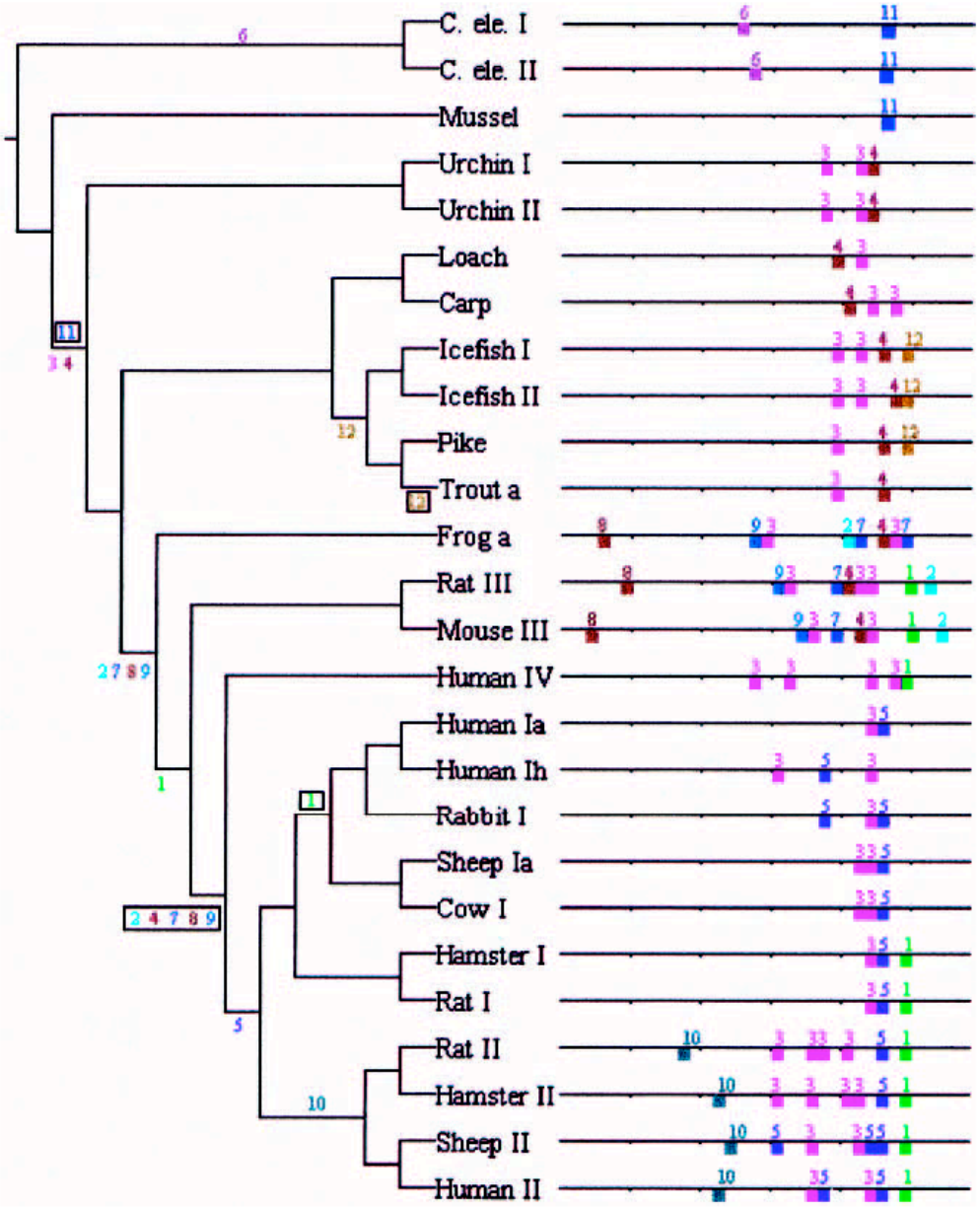
Motifs in clusters of orthologous genes (COGs)



Phylogenetic footprinting to predict regulatory sites



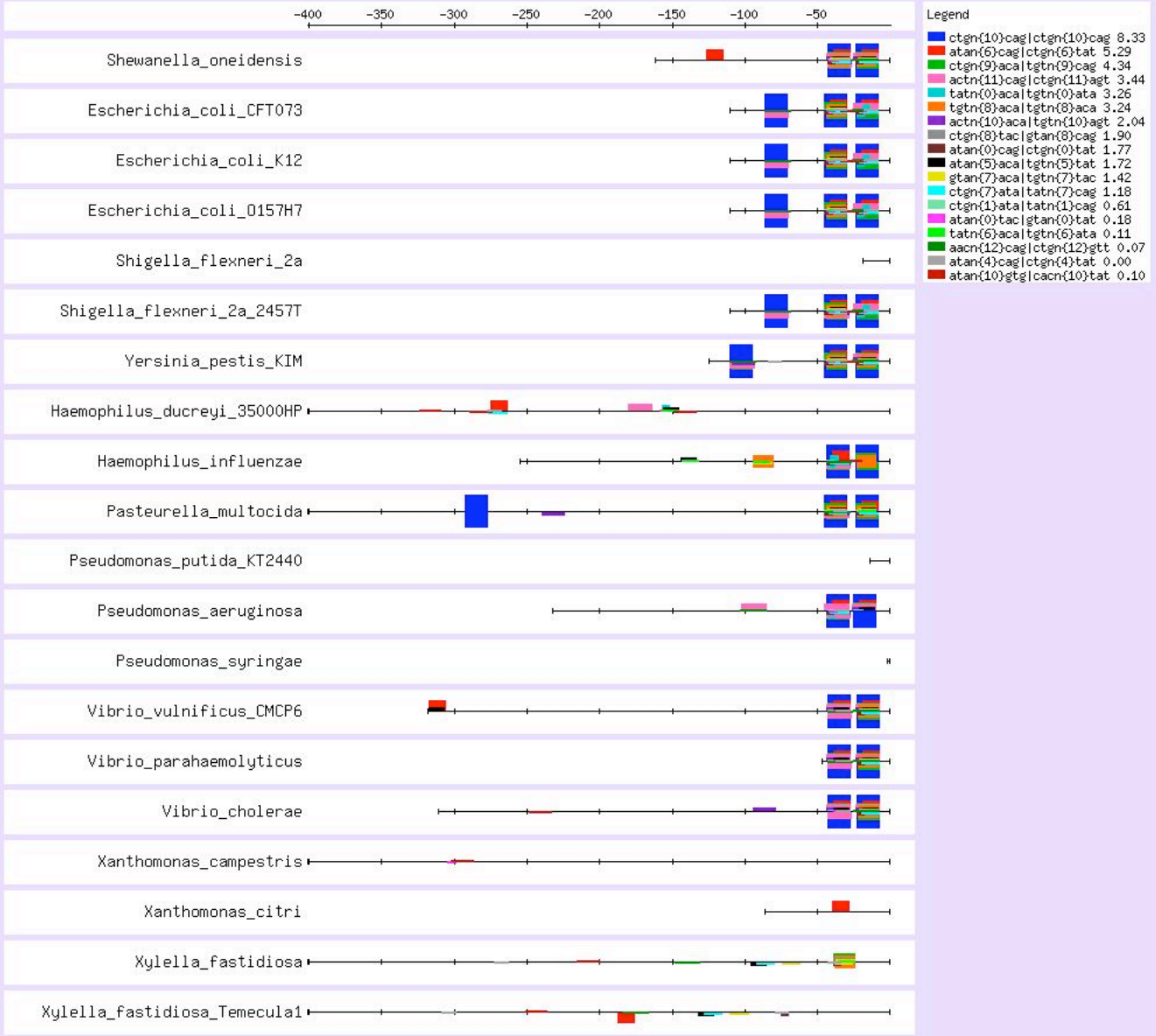
Footprinter example metallothionein



- 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.

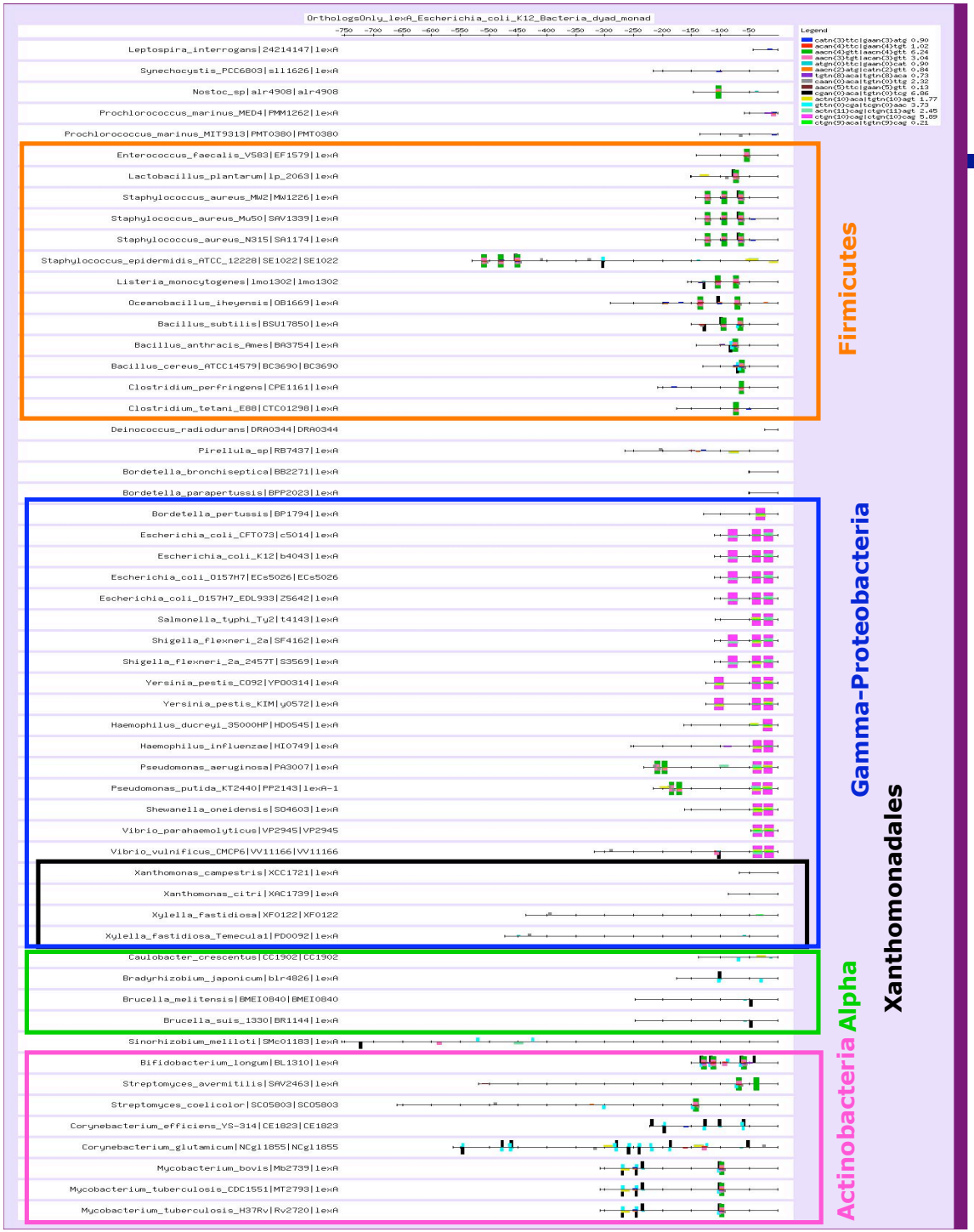
Source: Blanchette and Tompa (2002). Genome Research. 12, 739–748.

Pattern discovery in upstream regions of COGs



- Sequences
  - Upstream sequences of the cluster of orthologs for the gene lexA in all Gammaproteobacteria
- Pattern discovery
  - dyad-analysis
- A very highly significant signal is detected. It is found in two conserved positions in most genes of the group, and in 3 positions for certain bacteria.
- This motif corresponds to the known lexA binding site

Janky & van Helden, in prep.

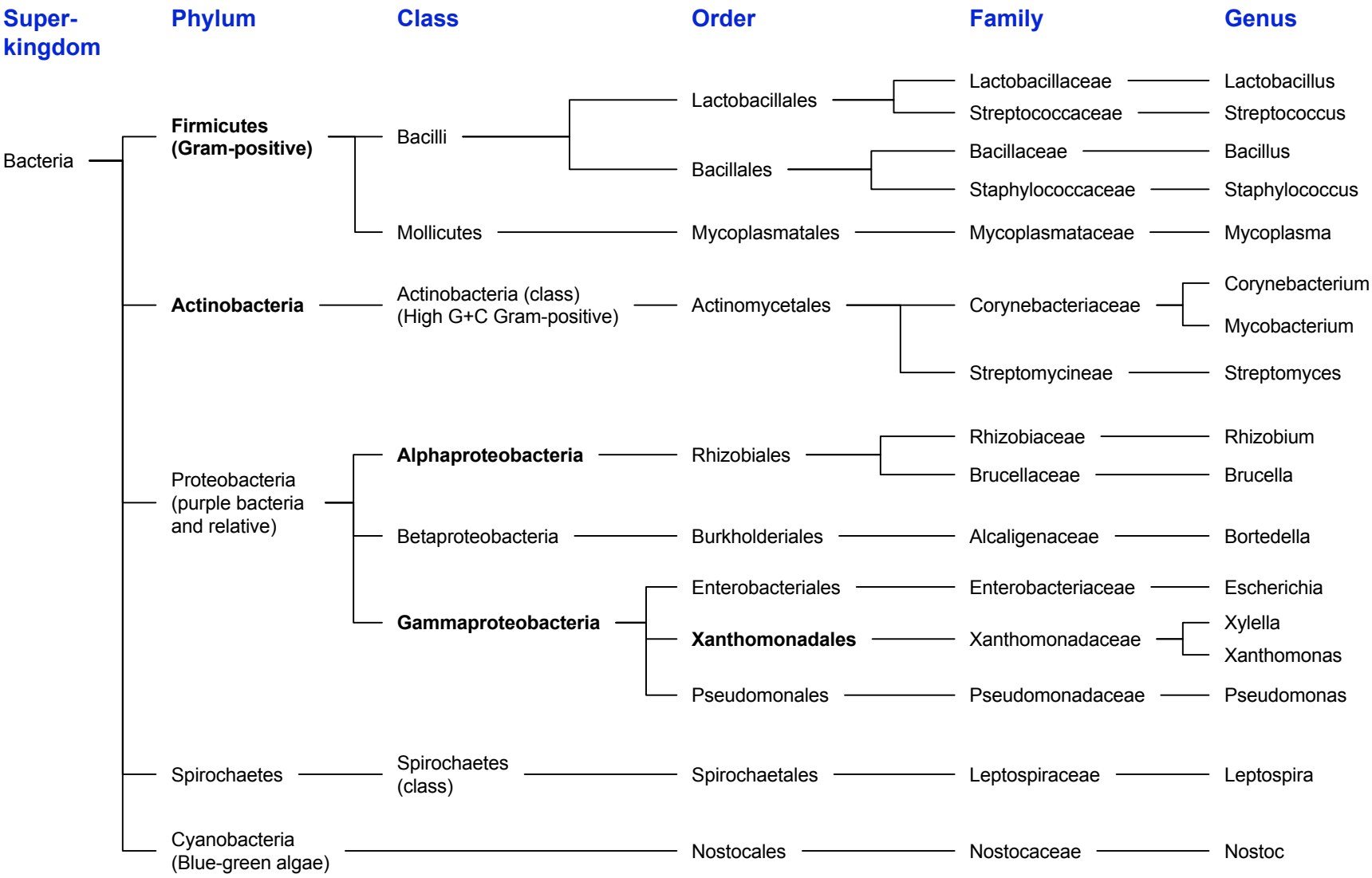


Evolution of cis-acting elements

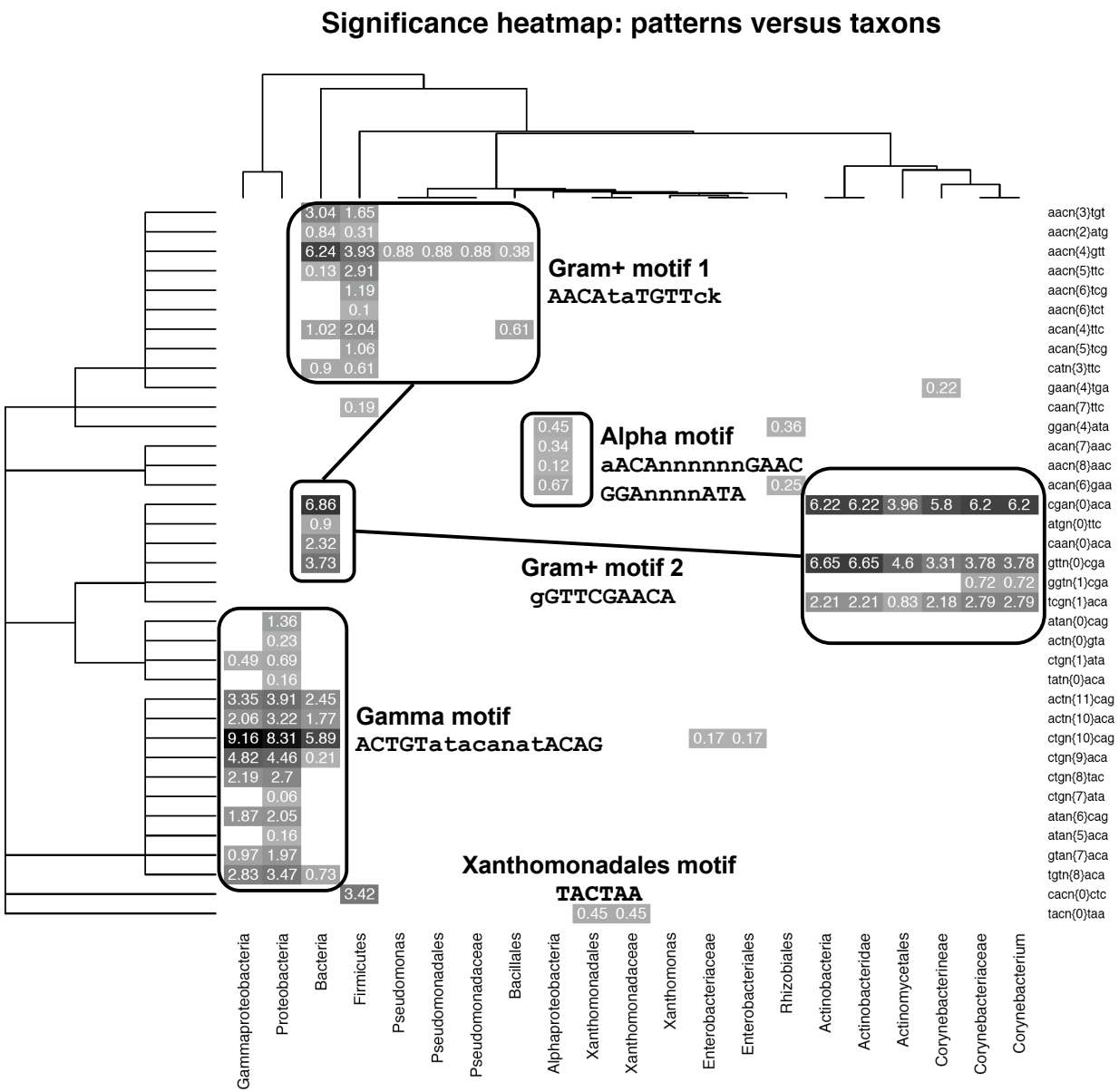
- Pattern discovery was applied
  - to upstream sequences the of 56 bacterial orthologs of lexA
  - at all the taxonomical levels.
- Note: the figure only shows the motifs disovered at one level (all bacterai analyzed together).
- Several motifs are discovered.
- Several taxon-specific motifs are discovered
  - Gamma-prtoeobacteria
    - Xanthomonadales
  - Gram-positive bacteria
    - Firmicutes
    - Actinobacteria
  - Alpha-proteobacteria
- These motifs correspond to those documented in the biological literature.

Janky & van Helden, in prep.

Bacteria taxonomy

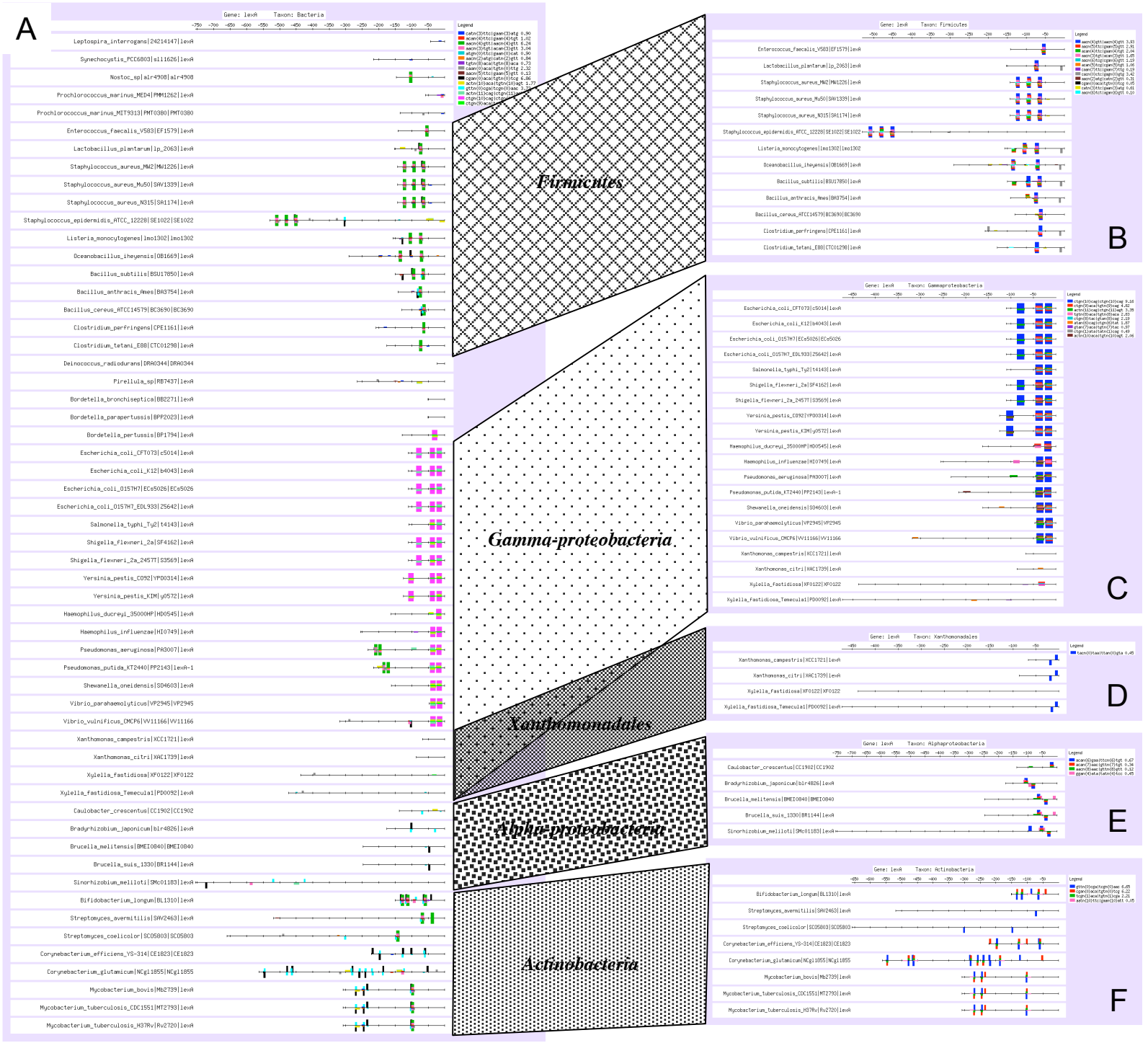


Heat map - pattern sinificance per taxon



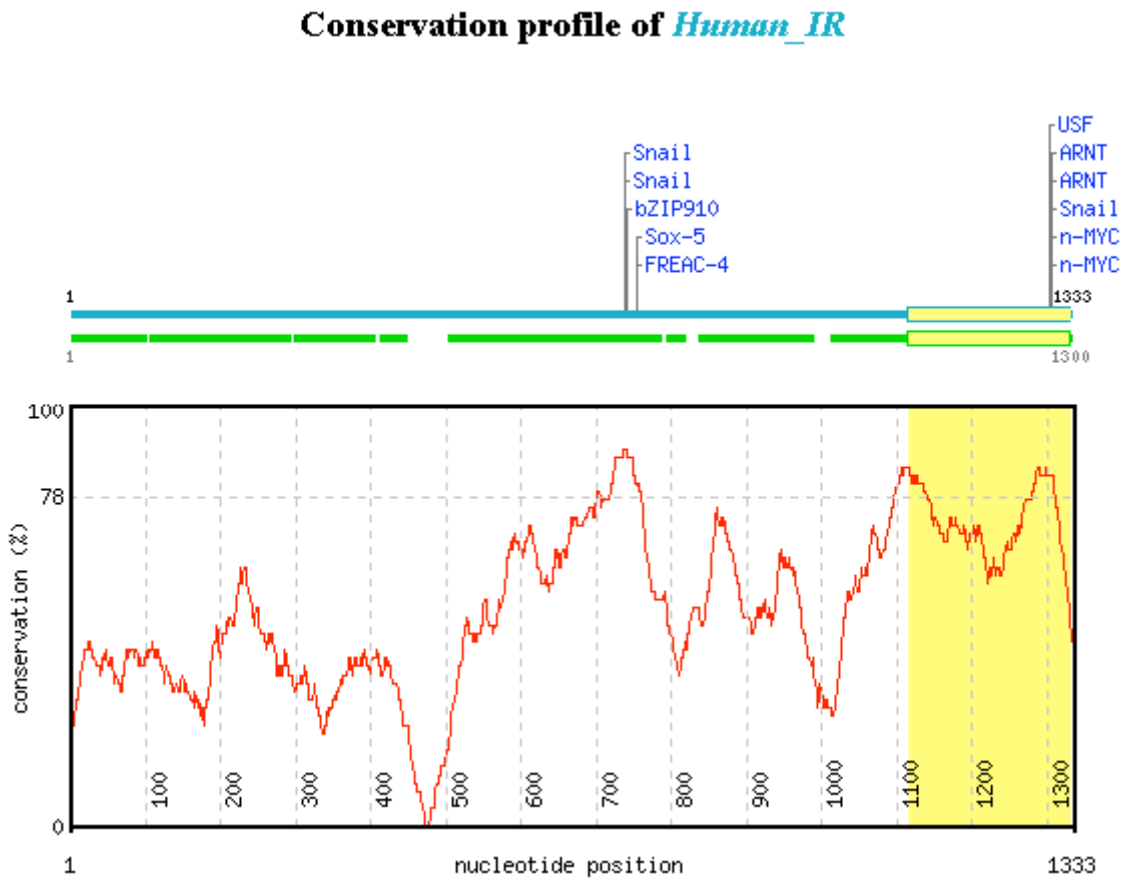
- The color in each cell indicates the significance of one pattern (row) in one taxon (column)
  - Rows are clustered by similarity between patterns.
  - Taxons are clustered by similarity between profiles of pattern significance.
- Several coherent groups of patterns are detected

Janky, R. & van Helden, J., in prep.


A
Gene: lexA Taxon: Bacteria
Leptospira_interrogans|24214147|lexA
Synechocystis_PCC6803|sll1626|lexA
Nostoc_sp|alr4908|alr4908
Prochlorococcus_marinus_MED4|PMM1262|lexA
Prochlorococcus_marinus_MIT9313|PMT0380|PMT0380
Enterococcus_faecalis_V583|EF1579|lexA
Lactobacillus_plantarum|lp_2063|lexA
Staphylococcus_aureus_MW2|MW1226|lexA
Staphylococcus_aureus_Mu50|SAV1339|lexA
Staphylococcus_aureus_N315|SA1174|lexA
Staphylococcus_epidermidis_ATCC_12228|SE1022|SE1022
Listeria_monocytogenes|lmo1302|lmo1302
Oceanobacillus_iheyensis|OB1669|lexA
Bacillus_subtilis|BSU17850|lexA
Bacillus_anthracis_Ames|BA3754|lexA
Bacillus_cereus_ATCC14579|BC3690|BC3690
Clostridium_perfringens|CPE1161|lexA
Clostridium_tetani_E88|CTC01298|lexA
Deinococcus_radiodurans|DRA0344|DRA0344
Pirellula_sp|RB7437|lexA
Bordetella_bronchiseptica|BB2271|lexA
Bordetella_parapertussis|BPP2023|lexA
Bordetella_pertussis|BP1794|lexA
Escherichia_coli_CFT073|c5014|lexA
Escherichia_coli_K12|b4043|lexA
Escherichia_coli_O157H7|ECs5026|ECs5026
Escherichia_coli_O157H7_EDL933|Z5642|lexA
Salmonella_typhi_Ty2|t4143|lexA
Shigella_flexneri_2a|SF4162|lexA
Shigella_flexneri_2a_2457T|S3569|lexA
Yersinia_pestis_CO92|YPO0314|lexA
Yersinia_pestis_KIM|y0572|lexA
Haemophilus_ducreyi_35000HP|HD0545|lexA
Haemophilus_influenzae|HI0749|lexA
Pseudomonas_aeruginosa|PA3007|lexA
Pseudomonas_putida_KT2440|PP2143|lexA-1
Shewanella_oneidensis|SO4603|lexA
Vibrio_parahaemolyticus|VP2945|VP2945
Vibrio_vulnificus_CMCP6|VV11166|VV11166
Xanthomonas_campestris|XCC1721|lexA
Xanthomonas_citri|XAC1739|lexA
Xylella_fastidiosa|XF0122|XF0122
Xylella_fastidiosa_Temecula1|PD0092|lexA
Caulobacter_crescentus|CC1902|CC1902
Bradyrhizobium_japonicum|blr4826|lexA
Brucella_melitensis|BMEI0840|BMEI0840
Brucella_suis_1330|BR1144|lexA
Sinorhizobium_meliloti|SMc01183|lexA
Bifidobacterium_longum|BL1310|lexA
Streptomyces_avermitilis|SAV2463|lexA
Streptomyces_coelicolor|SCO5803|SCO5803
Corynebacterium_efficiens_YS-314|CE1823|CE1823
Corynebacterium_glutamicum|NCgl1855|NCgl1855
Mycobacterium_bovis|Mb2739|lexA
Mycobacterium_tuberculosis_CDC1551|MT2793|lexA
Mycobacterium_tuberculosis_H37Rv|Rv2720|lexA
Firmicutes
Gamma-proteobacteria
Xanthomonadales
Alpha-proteobacteria
Actinobacteria
B
Gene: lexA Taxon: Firmicutes
C
Gene: lexA Taxon: Gammaproteobacteria
D
Gene: lexA Taxon: Xanthomonadales
E
Gene: lexA Taxon: Alphaproteobacteria
F
Gene: lexA Taxon: Actinobacteria

Cross-matches in promoters of orthologous genes

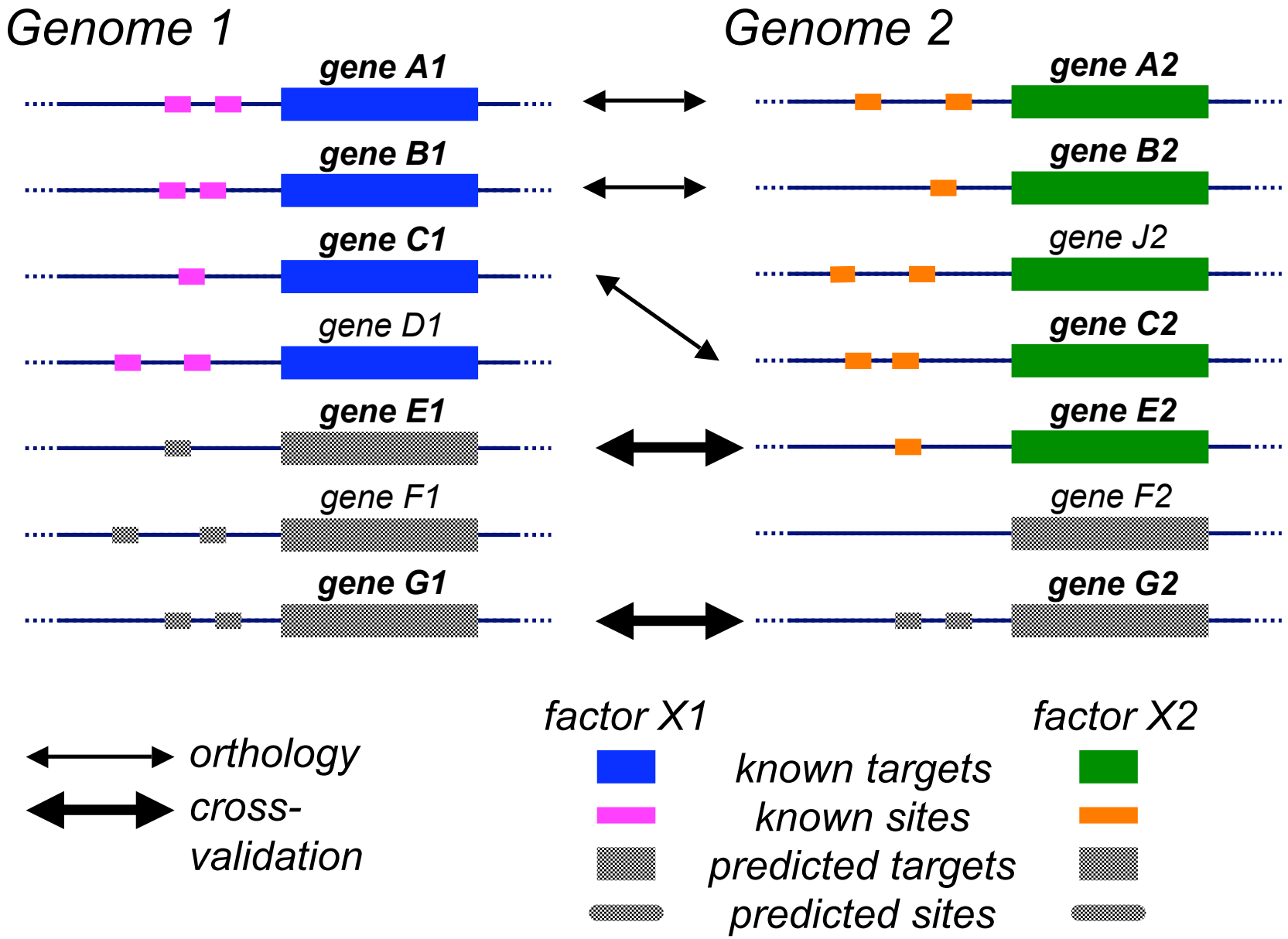
- Lenhard et al. (2003). J.Biology 2:13.
- 100 PSSM for known mammal transcription factors
- Searching for conserved matches in Human and mouse increases the selectivity by 85%.
- **Consite**: http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite/



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

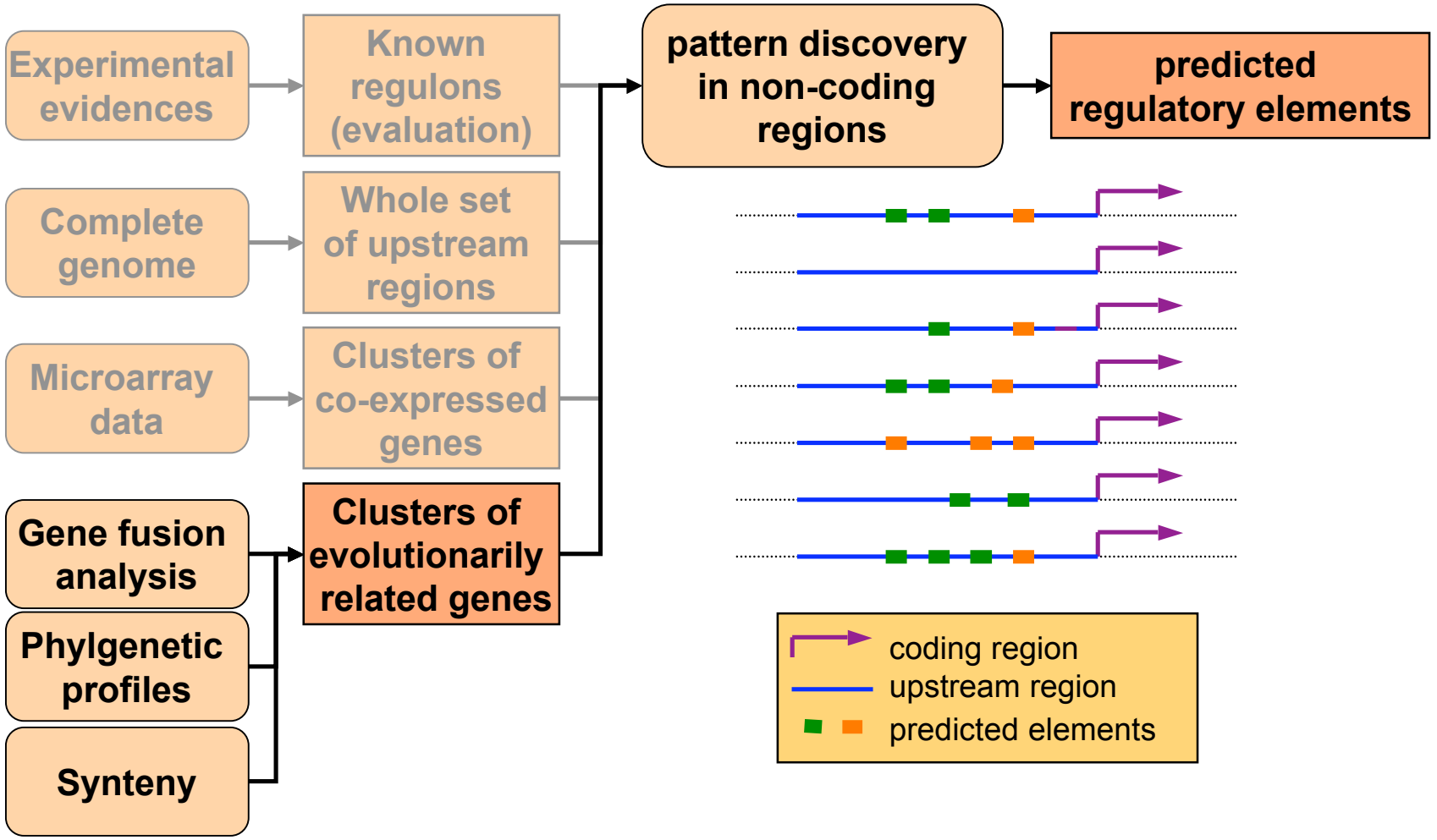
Cross-validation of pattern matching



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	ccccctcaccccctt	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.