From phylogenetic footprints to co-regulation networks

DNA-protein binding interface

- Example of DNA-protein interface.
  - LexA homodimer on the *recA* promoter.
  - Note: this is only a model, the structure has still not been crystallized.
  - Source: http://srs.ebi.ac.uk

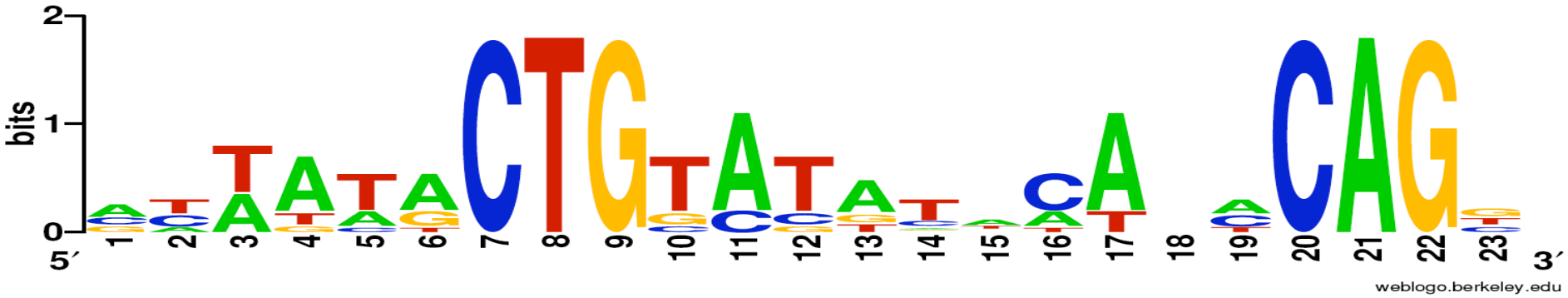
LexA binding sites (RegulonDB, oct 2006)

ECK120012770	recA	b2699	-21	gaagcaattaTA**CTG**TATGCTCATA**CAG**TAtcaagtgttt
ECK120012770	lexA	b4043	-10	aaatcgccttTTG**CTG**TATATACTCA**CAG**Cataactgtat
ECK120012770	lexA	b4043	10	atactcacagCATAA**CTG**TATATACACC**CAg**gggcggaa
ECK120012770	recA	b2699	-21	gaagcaattaTA**CTG**TATGCTCATA**CAG**TAtcaagtgttt
ECK120012770	ssb	b4059	-46.5	gacacaaattGAC**CTG**AATGAATATA**CAG**Tattggaatgc
ECK120012770	sulA	b0958	-2	agcccctgtgAGTTA**CTG**TATGGATGTA**CAg**tacatccag
ECK120012770	uvrA	b4058	-31.5	gacacaaattGAC**CTG**AATGAATATA**CAG**Tattggaatgc
ECK120012770	uvrB	b0779	-21	ttatggtgatGAA**CTG**TTTTTTTATC**CAG**Tataatttgtt
ECK120012770	uvrD	b3813	11	taatcagcaaAT**CTG**TATATATACC**CAG**CTttttggcgga
ECK120012770	rpsU	b3065	4.5	attttgaaatAAG**CTG**GCGTTGATGC**CAG**Cggcaaaccga
ECK120012770	phr	b0708	25.5	ttatcctgacGC**CTG**GCTTTCAGGG**CAG**CGttatttcgaa

LexA aligned binding sites and position-specific scoring matrix (PSSM)

```
 1|1    :    1/7     ATTATACTGTATGCTCATACAGT
 2|2    :    2/8     CTTTTGCTGTATATACTCACAGC
 3|3    :    3/10    GCATAACTGTATATACACCCAGG
 4|4    :    4/7     ATTATACTGTATGCTCATACAGT
 5|5    :   -5/13    CCAATACTGTATATTCATTCAGG
 6|6    :   -6/15    GATGTACTGTACATCCATACAGT
 7|7    :   -7/13    CCAATACTGTATATTCATTCAGG
 8|8    :   -8/13    ATTATACTGGATAAAAAAACAGT
 9|9    :    9/7     GCAAATCTGTATATATACCCAGC
10|10   :   10/8     AATAAGCTGGCGTTGATGCCAGC
11|11   : -11/12     ATAACGCTGCCCTGAAAGCCAGG
```

Pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
A	5	2	**5**	**8**	3	**7**	0	0	0	0	**9**	0	**7**	1	5	3	**9**	1	5	0	**11**	0	0
C	3	4	0	0	1	0	**11**	0	0	1	2	2	0	2	1	**7**	0	3	4	**11**	0	0	3
G	3	0	0	1	0	3	0	0	**11**	2	0	1	2	1	1	0	0	2	0	0	0	**11**	4
T	0	5	**6**	2	**7**	1	0	**11**	0	**8**	0	**8**	2	7	4	1	2	5	2	0	0	0	4



The site sequences were obtained from RegulonDB.
The alignment and matrix were created using consensus (Hertz & Stormo, 1999).
The logo was created with Web Logo (http://weblogo.berkeley.edu/logo.cgi).

Phylogenetic footprint discovery in promoters of orthologous genes

Orthologs of Escherichia coli K12 gene PUTA in Gammaproteobacteria

- Orthologs of Escherichia coli K12 gene PUTA in Gammaproteobacteria

#ref_gene	ref_org	query_gene	ident	e_value
NP_245526.1	Pasteurella_multocida	NP_415534.1	52.93	0.0
YP_340762.1	Pseudoalteromonas_haloplanktis_TAC125	NP_415534.1	45.21	0.0
YP_271059.1	Colwellia_psychrerythraea_34H	NP_415534.1	43.81	0.0
NP_719311.1	Shewanella_oneidensis	NP_415534.1	47.35	0.0
YP_156342.1	Idiomarina_loihiensis_L2TR	NP_415534.1	45.44	0.0
YP_095723.1	Legionella_pneumophila_Philadelphia_1	NP_415534.1	51.25	0.0
YP_126994.1	Legionella_pneumophila_Lens	NP_415534.1	51.63	0.0
YP_123979.1	Legionella_pneumophila_Paris	NP_415534.1	51.25	0.0
NP_819659.1	Coxiella_burnetii	NP_415534.1	47.21	0.0
YP_133644.1	Photobacterium_profundum_SS9	NP_415534.1	46.63	0.0
YP_206789.1	Vibrio_fischeri_ES114	NP_415534.1	48.96	0.0
NP_937700.1	Vibrio_vulnificus_YJ016	NP_415534.1	47.40	0.0
NP_763030.1	Vibrio_vulnificus_CMCP6	NP_415534.1	47.60	0.0
NP_801236.1	Vibrio_parahaemolyticus	NP_415534.1	47.92	0.0
NP_639180.1	Xanthomonas_campestris	NP_415534.1	52.40	0.0
YP_202784.1	Xanthomonas_oryzae_KACC10331	NP_415534.1	52.69	0.0
YP_244967.1	Xanthomonas_campestris_8004	NP_415534.1	52.40	0.0
NP_644196.1	Xanthomonas_citri	NP_415534.1	52.59	0.0
YP_365739.1	Xanthomonas_campestris_vesicatoria_85-10	NP_415534.1	52.40	0.0
NP_805570.1	Salmonella_typhi_Ty2	NP_415534.1	91.29	0.0
NP_455618.1	Salmonella_typhi	NP_415534.1	91.29	0.0
YP_052304.1	Erwinia_carotovora_atroseptica_SCRI1043	NP_415534.1	74.24	0.0
YP_408453.1	Shigella_boydii_Sb227	NP_415534.1	99.09	0.0
YP_309997.1	Shigella_sonnei_Ss046	NP_415534.1	99.39	0.0
NP_753076.1	Escherichia_coli_CFT073	NP_415534.1	99.39	0.0
NP_415534.1	Escherichia_coli_K12	NP_415534.1	100.00	0.0
NP_309287.1	Escherichia_coli_O157H7	NP_415534.1	99.55	0.0
NP_287019.1	Escherichia_coli_O157H7_EDL933	NP_415534.1	99.55	0.0
NP_871437.1	Wigglesworthia_brevipalpis	NP_415534.1	54.57	0.0
NP_992900.1	Yersinia_pestis_biovar_Mediaevails	NP_415534.1	79.44	0.0
NP_669761.1	Yersinia_pestis_KIM	NP_415534.1	79.37	0.0
YP_070249.1	Yersinia_pseudotuberculosis_IP32953	NP_415534.1	79.37	0.0
NP_405415.1	Yersinia_pestis_CO92	NP_415534.1	79.37	0.0
NP_929224.1	Photorhabdus_luminescens	NP_415534.1	77.37	0.0
YP_343707.1	Nitrosococcus_oceani_ATCC_19707	NP_415534.1	50.00	0.0
YP_170117.1	Francisella_tularensis_tularensis	NP_415534.1	50.34	0.0
YP_433088.1	Hahella_chejuensis_KCTC_2396	NP_415534.1	48.61	0.0
NP_794750.1	Pseudomonas_syringae	NP_415534.1	73.64	0.0
NP_747050.1	Pseudomonas_putida_KT2440	NP_415534.1	73.45	0.0
YP_257639.1	Pseudomonas_fluorescens_Pf-5	NP_415534.1	73.60	0.0
YP_346185.1	Pseudomonas_fluorescens_PfO-1	NP_415534.1	73.45	0.0
NP_249473.1	Pseudomonas_aeruginosa	NP_415534.1	48.56	0.0
YP_272798.1	Pseudomonas_syringae_phaseolicola_1448A	NP_415534.1	73.77	0.0
YP_233614.1	Pseudomonas_syringae_pv_B728a	NP_415534.1	73.80	0.0
YP_264533.1	Psychrobacter_arcticum_273-4	NP_415534.1	44.42	0.0
YP_046314.1	Acinetobacter_sp_ADP1	NP_415534.1	59.51	0.0

Upstream sequences of PUTA orthologs

```
>YP_046314.1|Acinetobacter_sp_ADP1|putA         YP_046314.1; upstream from -103 to -1; size: 103;
ACAAAATTTTCCTCTAAAAAATGAATCAATTATAGTCAGTTATTTGGTAATTATTCTGTA
ATGATAAATTATCTAAACCCCTTTAAACAATATACCTTAGAGT
>YP_271059.1|Colwellia_psychrerythraea_34H|putA YP_271059.1; upstream from -245 to -1; size: 245
ATAATAACCCACGAACACTCCCTACAAATTATAAAAAACGATTTGCAGCACTTTATACTG
TTGAATTCGTACTCCCCCATATAAAAGTGTTTTAAATCCTGAAAATAACCAGCACATCCT
GTGGTTGTTACCCTAAAATTCGCTCATAAATTTAATGTCGTCACCAACTAATAATATATG
TATTAGTGGAAAAAAAGACTATAACTAAAGCAGGATTTCTACCTGTCACACTTTGAGGAA
TGGTT
>NP_819659.1|Coxiella_burnetii|putA NP_819659.1; upstream from -118 to -1; size: 118;
GAAGTAGCCCGTATGAAGCGAAGCGAAATACGGGGAGGTGCACGTATTGTTCCCGTATTC
GCTTCGTTTCATACGGGCTACATCGCGGAAATGAAAATTAACTCCTTAATGAGGACAT
>YP_052304.1|Erwinia_carotovora_atroseptica_SCRI1043|putA   YP_052304.1; upstream from -195 to -1; size: 195;
TTAACTCTCCACATTTTTTTCGTGCCGCCGTCGCGGCGACGCTGTGTTTTATAGTAATCA
TTCAGGCCGGAACGAGGTCTGAAAATGATTATGGGCAGCAACCATTCCATTTGTTAACAA
GGTTGCACAAAGTTGCAACATGATTGATATTGACGGTATCCCGATGTGCATCTTCTCATT
ACAGGAGTGGACTCT
>NP_753076.1|Escherichia_coli_CFT073|putA       NP_753076.1; upstream from 0 to -1; size: 0;
>NP_415534.1|Escherichia_coli_K12|putA          NP_415534.1; upstream from -400 to -1; size: 400;
ATCGGCAATGTCGAAACTTGCCGTTATATCTGCCACCGGAACGGGGTAACAGAGTTTATG
TTTTACCAGGGCGACCGTATCCTGCCGGAAGCGCTGGTTATTCACAATCGATTTAACACA
CCATTTACATTAAATTTTAGTGCTCAGCGACACTATTTTTCATCAGGTTGCACTCTCTCA
CATTTTTTGCGGTTGCACCTTTCAAAAATGTTAACTGCCGCAGAGAAAAAGTCTGAGTTA
TTTTTTTAATCCCTGTCATATCGATTTCTTTTATTAACATTTCATTCATTTTTAAGCTTG
CTACGCATGTCACATTTAACATGGTTGCACAAAGTTGCAACATCATGGATATTTCACGAT
AACGTTAAGTTGCACCTTTCTGAACAACAGGAGTAATGGC
>NP_309287.1|Escherichia_coli_O157H7|ECs1260    NP_309287.1; upstream from -54 to -1; size: 54;
GGATATTTCACGATAACGTTAAGTTGCACCTTTCAGAACAACAGGAGTAATGGC
>NP_287019.1|Escherichia_coli_O157H7_EDL933|putA            NP_287019.1; upstream from -400 to -1; size: 400;
ATCGGCAATGTCGAAACTTGCCGTTATATCTGCCACCGGAACGGGGTAACAGAGTTTATG
TTTTACCAGGGCGACCGTATCCTGCCGGAAGCGCTGGTTATTCACAATCGATTTAACACA
CCATTTACATTAAATTTTAGTGCTCAGCGACACTATTTTTCATCAGGTTGCACTCTCTCA
CATTTTTTGCGGTTGCACCTTTCAAAAATGTTAACTGCCGCAGAGAAAAAGTCTGAGTTA
TTTTTTTAATCCCTGTCATATCGATTTCTTTTATTAACATTTCATTCATTTTTAAGCTTG
CTACGCAGGTCACATTTAACATGGTTGCACAAAGTTGCAACATCATGGATATTTCACGAT
AACGTTAAGTTGCACCTTTCAGAACAACAGGAGTAATGGC
…...
```

Purged upstream sequences of PUTA orthologs

```
>YP_046314.1|Acinetobacter_sp_ADP1|putA
ACAAAATTTTCCTCTAAAAAATGAATCAATTATAGTCAGTTATTTGGTAATTATTCTGTA
ATGATAAATTATCTAAACCCCTTTAAACAATATACCTTAGAGT
>YP_271059.1|Colwellia_psychrerythraea_34H|putA
ATAATAACCCACGAACACTCCCTACAAATTATAAAAAACGATTTGCAGCACTTTATACTG
TTGAATTCGTACTCCCCCATATAAAAGTGTTTTAAATCCTGAAAATAACCAGCACATCCT
GTGGTTGTTACCCTAAAATTCGCTCATAAATTTAATGTCGTCACCAACTAATAATATATG
TATTAGTGGAAAAAAAGACTATAACTAAAGCAGGATTTCTACCTGTCACACTTTGAGGAA
TGGTT
>NP_819659.1|Coxiella_burnetii|putA
GAAGTAGCCCGTATGAAGCGAAGCGAAATACGGGGAGGTGCACGTATTGTTnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnATCGCGGAAATGAAAATTAACTCCTTAATGAGGACAT
>YP_052304.1|Erwinia_carotovora_atroseptica_SCRI1043|putA
TTAACTCTCCACATTTTTTTCGTGCCGCCGTCGCGGCGACGCTGTGTTTTATAGTAATCA
TTCAGGCCGGAACGAGGTCTGAAAATGATTATGGGCAGCAACCATTCCATTTGTTAACAA
GGTTGCACAAAGTTGCAACATGATTGATATTGACGGTATCCCGATGTGCATCTTCTCATT
ACAGGAGTGGACTCT
>NP_753076.1|Escherichia_coli_CFT073|putA
>NP_415534.1|Escherichia_coli_K12|putA
ATCGGCAATGTCGAAACTTGCCGTTATATCTGCCACCGGAACGGGGTAACAGAGTTTATG
TTTTACCAGGGCGACCGTATCCTGCCGGAAGCGCTGGTTATTCACAATCGATTTAACACA
CCATTTACATTAAATTTTAGTGCTCAGCGACACTATTTTTCATCAGGTTGCACTCTCTCA
CATTTTTTGCGGTTGCACCTTTCAAAAATGTTAACTGCCGCAGAGAAAAAGTCTGAGTTA
TTTTTTTAATCCCTGTCATATCGATTTCTTTTATTAACATTTCATTCATTTTTAAGCTTG
CTACGCATGTCACATnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnTCACGAT
AACGTTAAGTTGCACCTTTCTGAACAACAGGAGTAATGGC
>NP_309287.1|Escherichia_coli_O157H7|ECs1260
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
>NP_287019.1|Escherichia_coli_O157H7_EDL933|putA
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
Nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
…...
```

Significantly over-represented dyads in promoters of PUTA orthologs

; column headers

;	1	sequence	
;	2	identifier	
;	3	expected_freq	
;	4	occ	observed occurrences
;	5	exp_occ	expected occurrences
;	6	occ_P	occurrence probability (binomial)
;	7	occ_E	E-value for occurrences (binomial)
;	8	occ_sig	occurrence significance (binomial)
;	9	ovl_occ	number of overlapping occurrences
;	10	all_occ	number of non-overlapping + overlapping occurrences
;	11	rank	rank
;	12	ov_coef	overlap coefficient
;	13	remark	remark

;sequence	**identifier**	**expected_fre**	**occ**	**exp_occ**	**occ_P**	**occ_E**	**occ_sig**	**ovl_occ**	**all_occ**	**rank**	**ov_coef**	**remark**
gtgn{1}aac	gtgn{1}aac|gttn{1}cac	0.00058697	32	4.44	2.80E-14	1.10E-09	**8.95**	0	32	1	1.0166	
ggtn{3}acc	ggtn{3}acc|ggtn{3}acc	0.00016832	18	1.26	2.60E-13	1.00E-08	**7.99**	0	18	2	1.0166	inv_rep
gttn{0}gca	gttn{0}gca|tgcn{0}aac	0.00077131	35	5.87	3.30E-13	1.30E-08	**7.88**	5	40	3	1.0166	
ggtn{2}aac	ggtn{2}aac|gttn{2}acc	0.00048995	28	3.68	3.60E-13	1.40E-08	**7.84**	0	28	4	1.0166	
gtgn{0}caa	gtgn{0}caa|ttgn{0}cac	0.00070077	32	5.34	2.80E-12	1.10E-07	**6.95**	0	32	5	1.0166	
ggtn{2}cac	ggtn{2}cac|gtgn{2}acc	0.00040329	24	3.03	7.40E-12	2.90E-07	**6.53**	0	24	6	1.0166	
ggtn{0}gca	ggtn{0}gca|tgcn{0}acc	0.00052995	27	4.04	1.20E-11	4.90E-07	**6.31**	0	27	7	1.0166	
gcan{0}acc	gcan{0}acc|ggtn{0}tgc	0.00052995	27	4.04	1.20E-11	4.90E-07	**6.31**	0	27	8	1.0166	
ggtn{1}caa	ggtn{1}caa|ttgn{1}acc	0.00058494	27	4.42	1.10E-10	4.20E-06	**5.38**	0	27	9	1.0166	
ggtn{1}gca	ggtn{1}gca|tgcn{1}acc	0.00052995	24	4.01	1.60E-09	6.40E-05	**4.20**	0	24	10	1.0166	
aagn{2}gca	aagn{2}gca|tgcn{2}ctt	0.00062177	26	4.67	1.70E-09	7.00E-05	**4.16**	1	27	11	1.0166	
aggn{1}gca	aggn{1}gca|tgcn{1}cct	0.0004801	22	3.63	6.10E-09	2.40E-04	**3.62**	0	22	12	1.0166	
aggn{0}tgc	aggn{0}tgc|gcan{0}cct	0.0004801	20	3.66	1.30E-07	5.20E-03	**2.28**	0	20	13	1.0166	
aggn{2}caa	aggn{2}caa|ttgn{2}cct	0.00052992	21	3.98	1.40E-07	5.80E-03	**2.24**	0	21	14	1.0166	
aagn{1}tgc	aagn{1}tgc|gcan{1}ctt	0.00062177	20	4.7	6.30E-06	2.50E-01	**0.60**	1	21	15	1.0166	
aggn{3}aac	aggn{3}aac|gttn{3}cct	0.00044387	16	3.32	1.30E-05	5.30E-01	**0.28**	0	16	16	1.0166	
gggn{4}acc	gggn{4}acc|ggtn{4}ccc	0.00025997	12	1.93	1.50E-05	5.90E-01	**0.23**	4	16	17	1.0166	

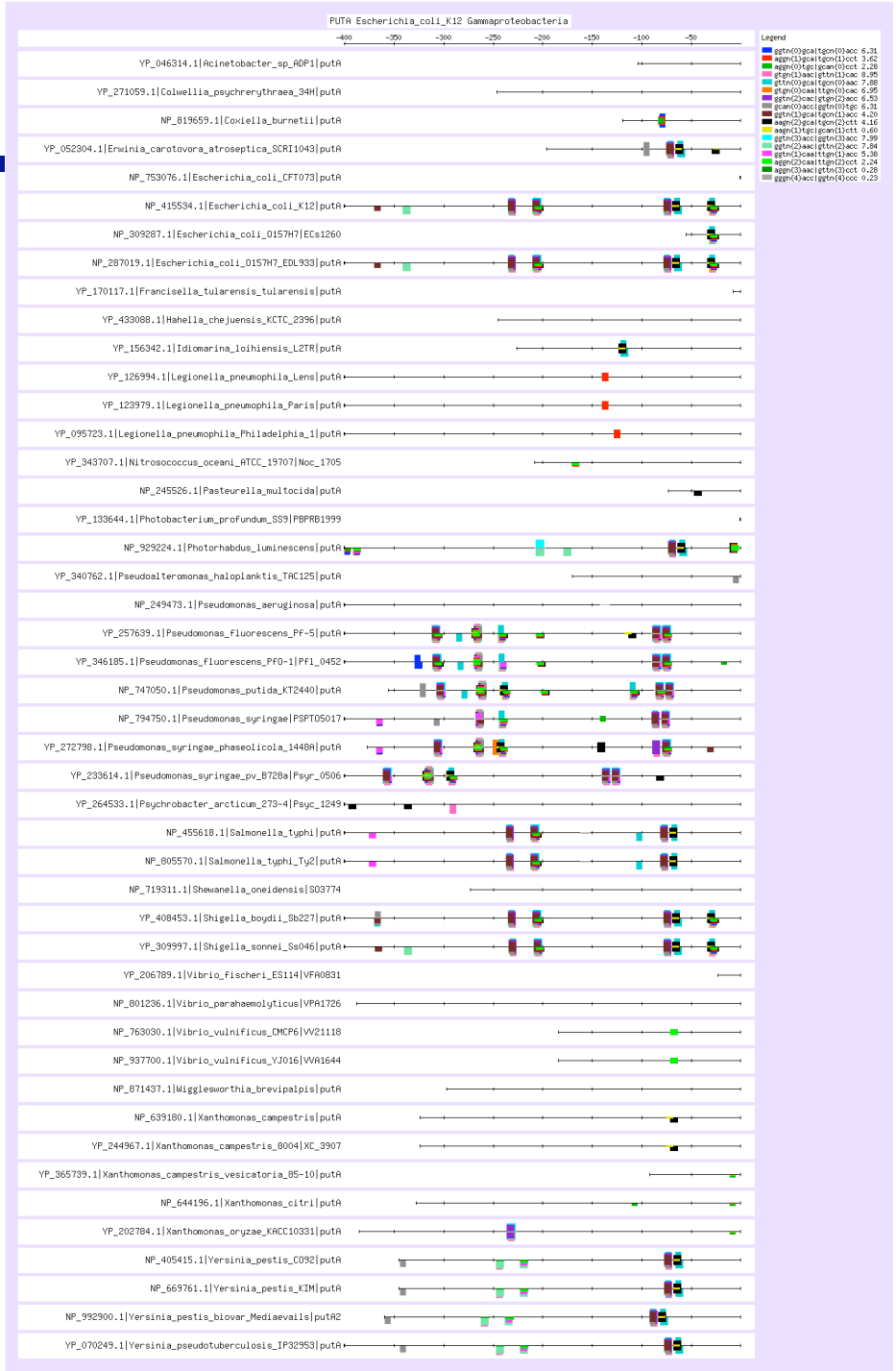
;Job started 13/02/06 21:31:01 CET
;Job done 13/02/06 21:31:28 CET

Significantly over-represented dyads in promoters of PUTA orthologs

```
; pattern-assembly  -v 1 -i /home/jvanheld/research/collabo
; Input score column                  8
; Output score column                 0
; two strand assembly
; max flanking bases                  1
; max substitutions                   0
; max assembly size                  50
; max number of patterr              50
; number of input patter             17
;
```

;assembly # 1	**seed: gtgnaa 17 words**	**lengt**
; alignt	rev_cpl	score
aagnngca....	....tgcnnctt	4.16
aagntgc.....	.....gcanctt	0.6
.aggngca....	....tgcncct.	3.62
.aggtgc.....	.....gcacct.	2.28
.aggnncaa...	...ttgnncct.	2.24
.aggnnnaac..	..gttnnncct.	0.28
..ggtnnnacc.	.ggtnnnacc..	7.99
..ggtnnaac..	..gttnnacc..	7.84
..ggtgca....	....tgcacc..	6.31
..ggtncaa...	...ttgnacc..	5.38
..ggtnnnnccc	gggnnnnacc..	0.23
...gtgnaac..	..gttncac...	8.95
...gtgcaa...	...ttgcac...	6.95
...gtgnnacc.	.ggtnncac...	6.53
....tgcaac..	..gttgca....	7.88
....tgcnacc.	.ggtngca....	4.2
.....gcaacc.	.ggttgc.....	6.31
aaggtgcaaccc	**gggttgcacctt**	**8.95 best**

```
;Job started 13/02/06 21:31:29 CET
;Job done    13/02/06 21:31:31 CET
```



Questions

- For each gene, we applied the same pattern discovery approach
  - Identify orthologs
  - Retrieve upstream sequences
  - Detect over-represented dyads
- Questions
  - How good is this method in predicting cis-acting elements ?
    - Can we detect correct motifs ?
    - What is the rate of false positives ?
  - Can we learn something about the evolution of cis-acting elements ?
  - On the basis of the discovered motifs, can we regroup the co-regulated genes ?
    - Detect pair-wise associations between genes
    - Detect clusters of genes regulated by the same transcription factor

Detailed analysis of a study case: auto-regulation of the LexA transcription factor

Study case: LexA auto-regulation

- The transcription factor LexA represses several genes involved in the SOS response
- The *lexA* gene is auto-regulated.
- *lexA* auto-regulation has been characterized in details in several bacterial species.
- Note that this is an easy case:
  - LexA binding motif is highly conserved in Gammaproteobacteria
  - This motif is easily detected by most pattern discovery algorithms.
- We will start by this example, and then generalize the evaluation to other transcription factors.

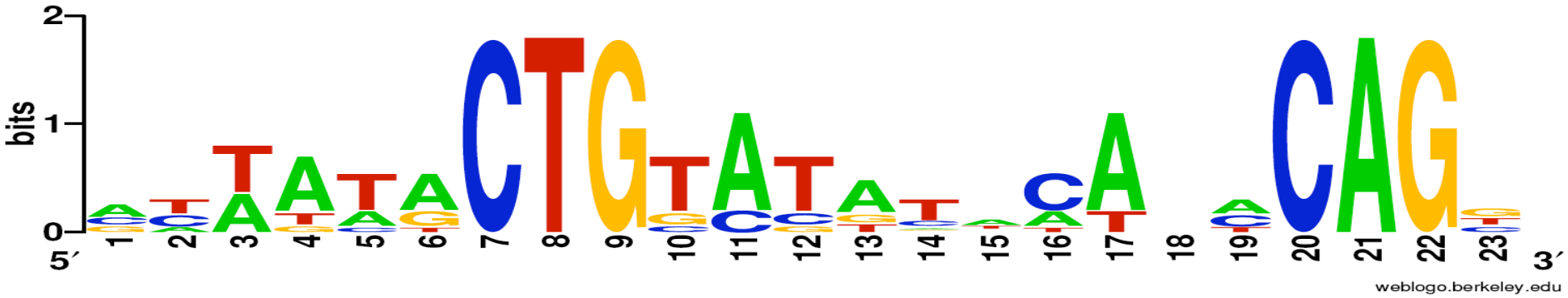
LexA binding sites (RegulonDB, oct 2006)

ECK120012770	recA	b2699	-21	gaagcaattaTA**CTG**TATGCTCATA**CAG**TAtcaagtgttt
ECK120012770	lexA	b4043	-10	aaatcgccttTTG**CTG**TATATACTCA**CAG**Cataactgtat
ECK120012770	lexA	b4043	10	atactcacagCATAA**CTG**TATATACACC**CAg**ggggcggaa
ECK120012770	recA	b2699	-21	gaagcaattaTA**CTG**TATGCTCATA**CAG**TAtcaagtgttt
ECK120012770	ssb	b4059	-46.5	gacacaaattGAC**CTG**AATGAATATA**CAG**Tattggaatgc
ECK120012770	sulA	b0958	-2	agcccctgtgAGTTA**CTG**TATGGATGTA**CAg**tacatccag
ECK120012770	uvrA	b4058	-31.5	gacacaaattGAC**CTG**AATGAATATA**CAG**Tattggaatgc
ECK120012770	uvrB	b0779	-21	ttatggtgatGAA**CTG**TTTTTTTATC**CAG**Tataatttgtt
ECK120012770	uvrD	b3813	11	taatcagcaaAT**CTG**TATATATACC**CAG**CTttttggcgga
ECK120012770	rpsU	b3065	4.5	attttgaaatAAG**CTG**GCGTTGATGC**CAG**Cggcaaaccga
ECK120012770	phr	b0708	25.5	ttatcctgacGC**CTG**GCTTTCAGGG**CAG**CGttatttcgaa

LexA aligned binding sites and position-specific scoring matrix (PSSM)

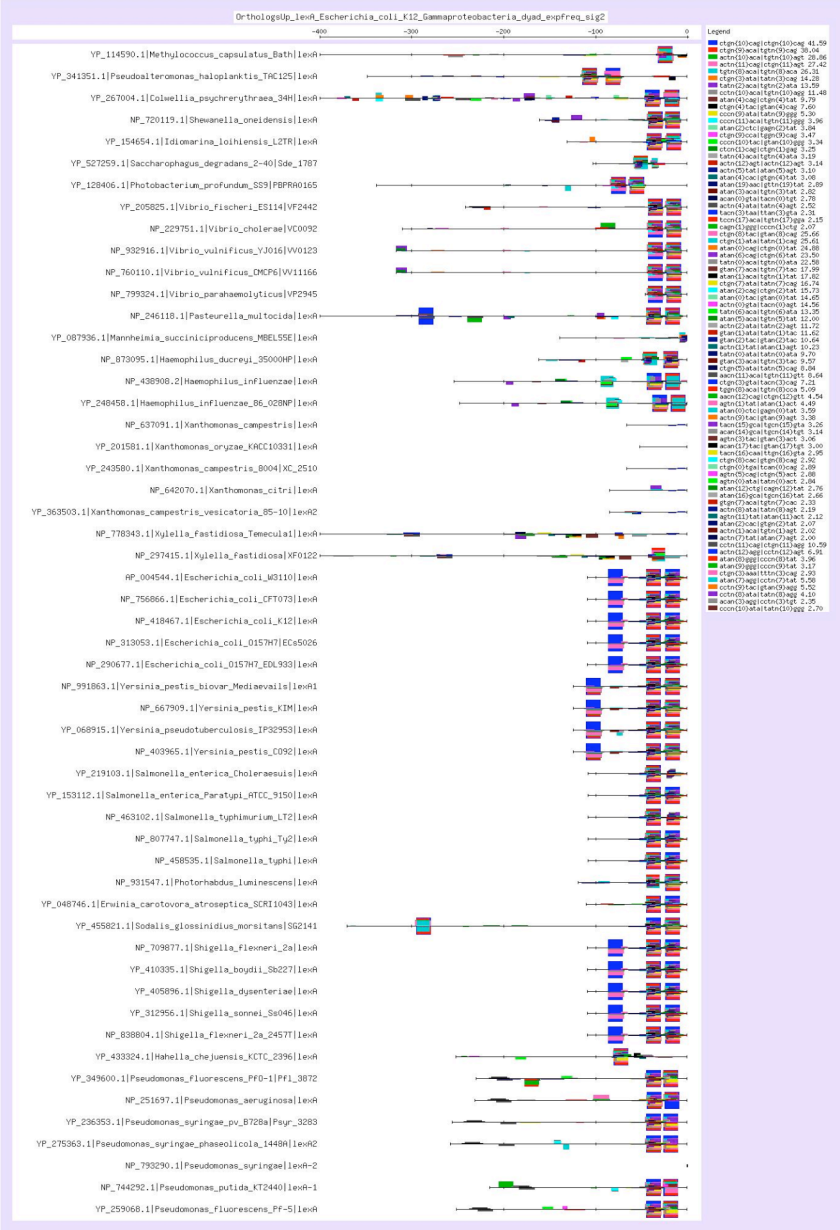
```
 1|1   :   1/7     ATTATACTGTATGCTCATACAGT
 2|2   :   2/8     CTTTTGCTGTATATACTCACAGC
 3|3   :   3/10    GCATAACTGTATATACACCCAGG
 4|4   :   4/7     ATTATACTGTATGCTCATACAGT
 5|5   :  -5/13    CCAATACTGTATATTCATTCAGG
 6|6   :  -6/15    GATGTACTGTACATCCATACAGT
 7|7   :  -7/13    CCAATACTGTATATTCATTCAGG
 8|8   :  -8/13    ATTATACTGGATAAAAAAACAGT
 9|9   :   9/7     GCAAATCTGTATATATACCCAGC
10|10  :  10/8     AATAAGCTGGCGTTGATGCCAGC
11|11  : -11/12    ATAACGCTGCCCTGAAAGCCAGG
```

Pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
A	5	2	**5**	**8**	3	**7**	0	0	0	0	**9**	0	**7**	1	5	3	**9**	1	5	0	**11**	0	0
C	3	4	0	0	1	0	**11**	0	0	1	2	2	0	2	1	**7**	0	3	4	**11**	0	0	3
G	3	0	0	1	0	3	0	0	**11**	2	0	1	2	1	1	0	0	2	0	0	0	**11**	4
T	0	5	**6**	2	**7**	1	0	**11**	0	**8**	0	**8**	2	7	4	1	2	5	2	0	0	0	4



The site sequences were obtained from RegulonDB.
The alignment and matrix were created using consensus (Hertz & Stormo, 1999).
The logo was created with Web Logo (http://weblogo.berkeley.edu/logo.cgi).

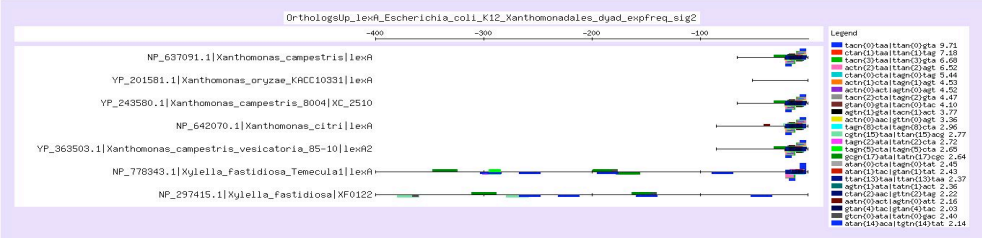
Significant dyads in promoters of lexA orthologs in Gammaproteobacteria



- When all the Gammaproteobacterial promoters are analyzed together, a large number of dyads are detected as significant.
- Most of these dyads are however mutually overlapping, and reveal different fragments of the same binding sites.
- These sites correspond to the annotated *lexA* cis-acting elements.
- The most significant dyad **CTG**n_{10}**CAG**, corresponds to the most conserved residues of the LexA binding sites (see PSSM in previous slides), and to the residues at the DNA-factor interface.

Janky & van Helden, submitted

Significant dyads in promoters of lexA orthologs in Xanthomonadales



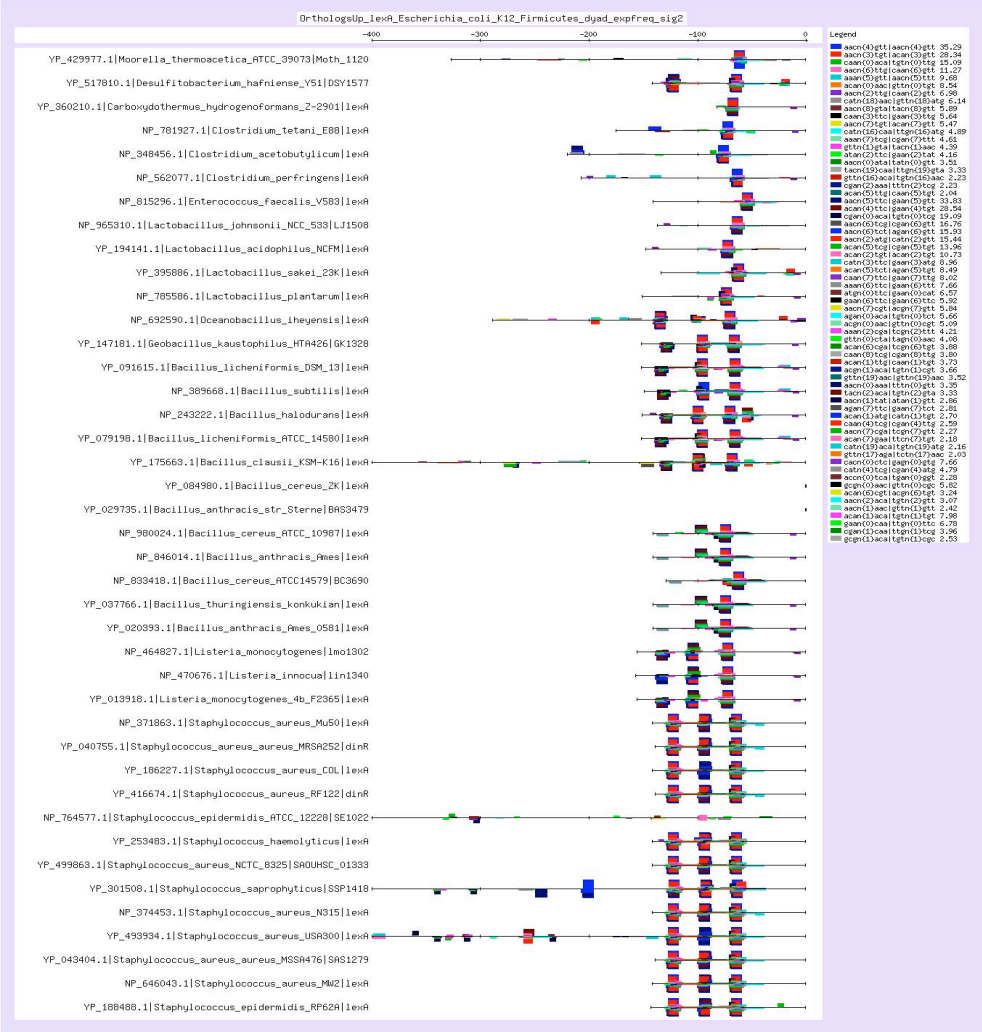
- The LexA binding site detected at the level of Gammaproteobacteria is absent from a subset of species, belonging to the Xanthomonadales.
- When the analysis is restricted to Xanthomonadales, a distinct conserved motif is detected.
- The assembly of the Xanthomonadales-specific dyads forms the consensus, including a reverse paindromic pair of hexanucleotides.

  G**TTAGTA**ATAC**TACTAA**C
- This predicted motif matches the experimentally validated consensus

  .TTAGTARWAWTACTA

Janky & van Helden, submitted

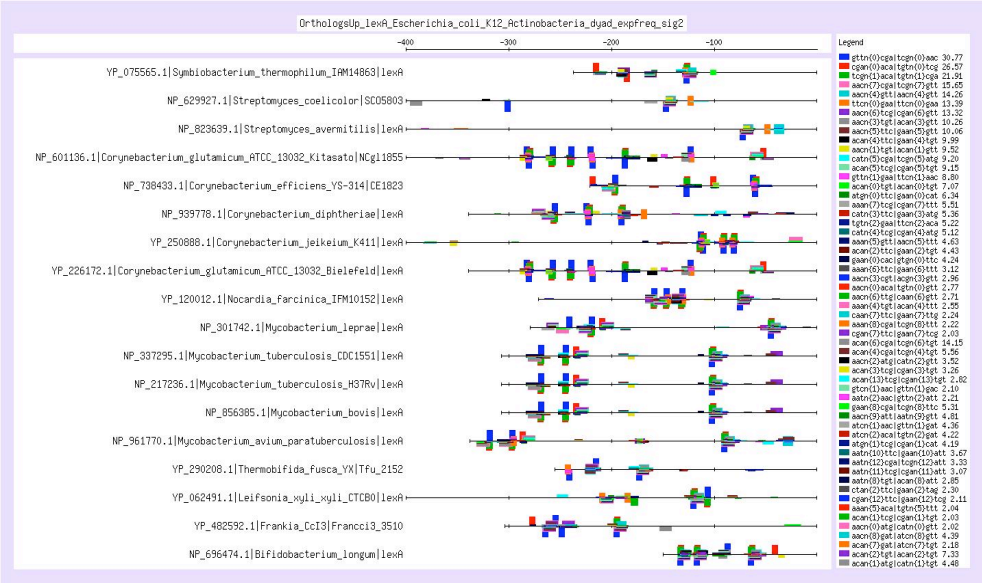
Significant dyads in promoters of lexA orthologs in Firmicutes



- In Firmicutes, the most significant motif is TAC**GAAC**ATAT**GTTC**GTA
- This corresponds to the “Cheo box” $GAACn_4GTTC$.

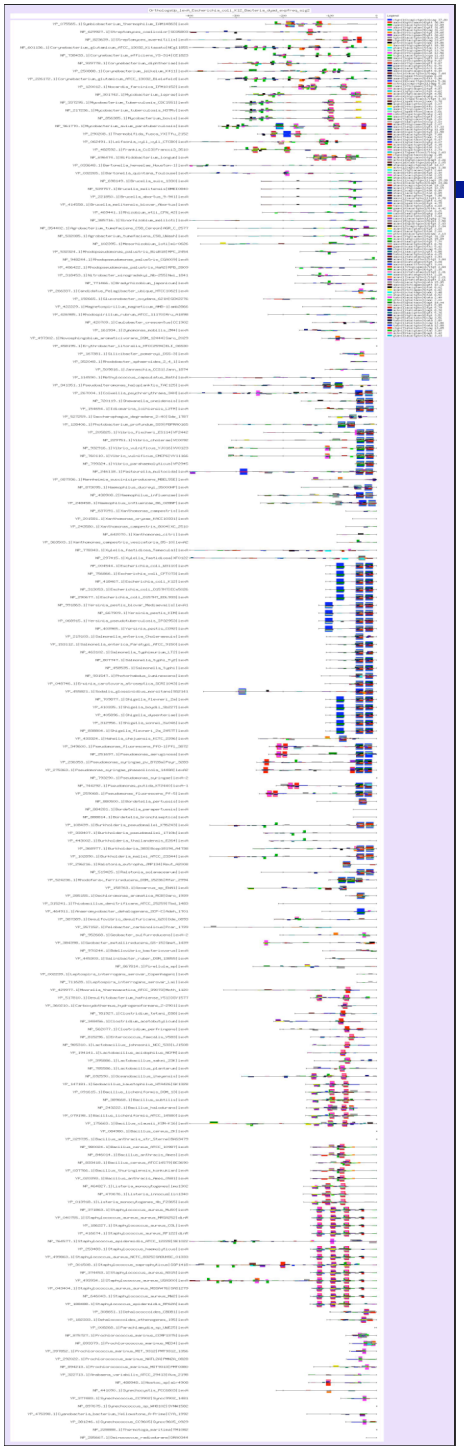
Janky & van Helden, submitted

Significant dyads in promoters of lexA orthologs in Actinobacteria



- In Actinobacteria, the most significant motif is TCGAACA.
- This shows an almost perfect match to one half of the Cheo box motif detected in Firmicutes (TA**CGAACA**TA**TGTTCG**TA)
- In addition, this larger Cheo box is also detected in Actinobacteria, albeit with a lower significance than the half site.

Janky & van Helden, submitted

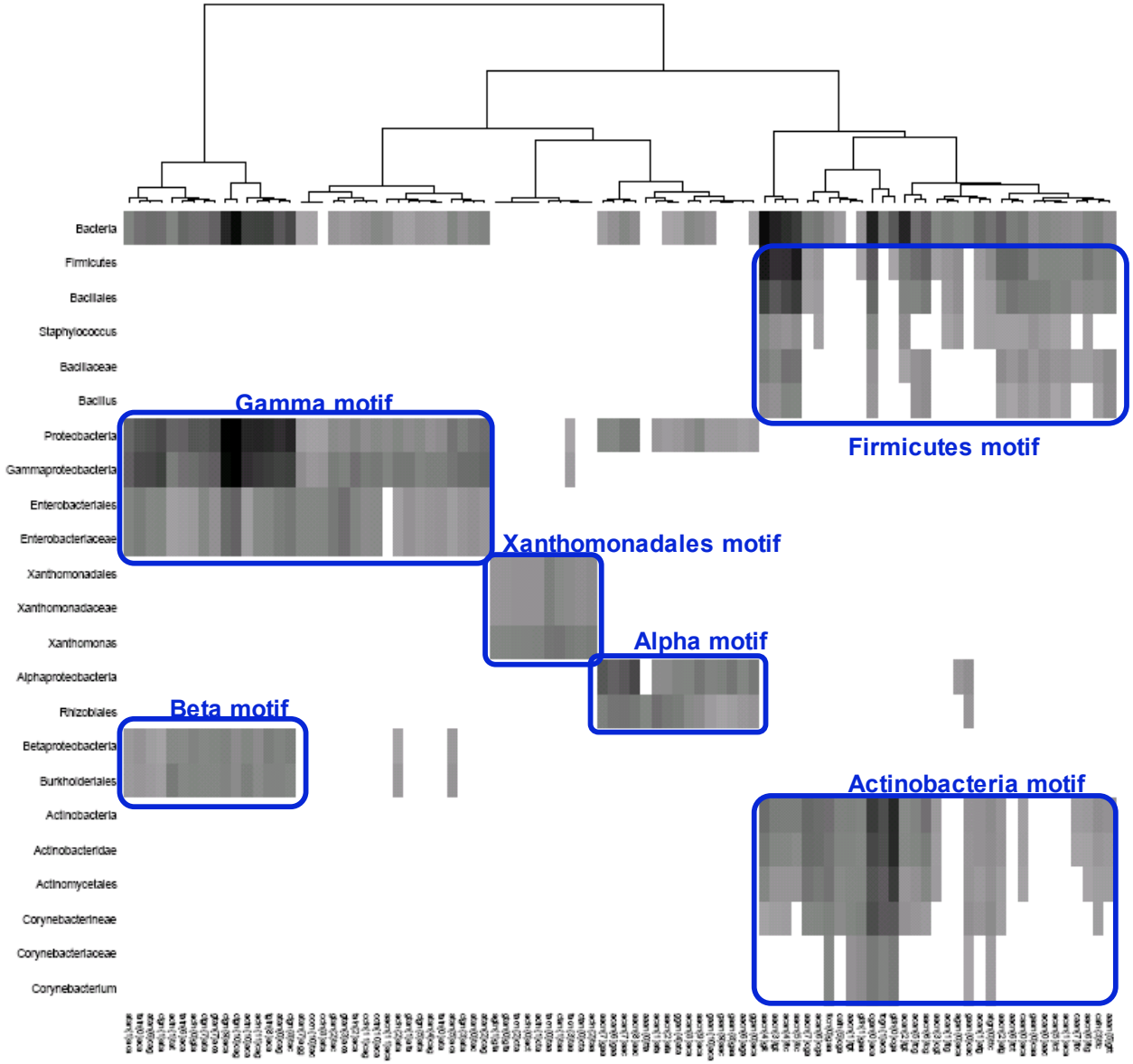


Significant dyads in promoters of lexA orthologs in Bacteria

- When all the bacterial promoters are analyzed together, the program dyad-analysis detects most of taxon-specific motifs discussed before, and the feature-map highlights their taxon-specific locations.
- This illustrates the robustness of the method: the motifs can be detected even if present in a subset of the sequences only.
- The significance is however lower when all sequences are analyzed together than with the taxon-per-taxon analysis.

Janky & van Helden, submitted

Significance map of the motifs discovered in promoters of lexA orthologs at all taxonomical levels



- The heat map illustrates the most significant motifs (sig >= 8) found at different taxonomical levels.
- Each column corresponds to one dyad, each row to one taxon.
- The grey level indicates the significance.

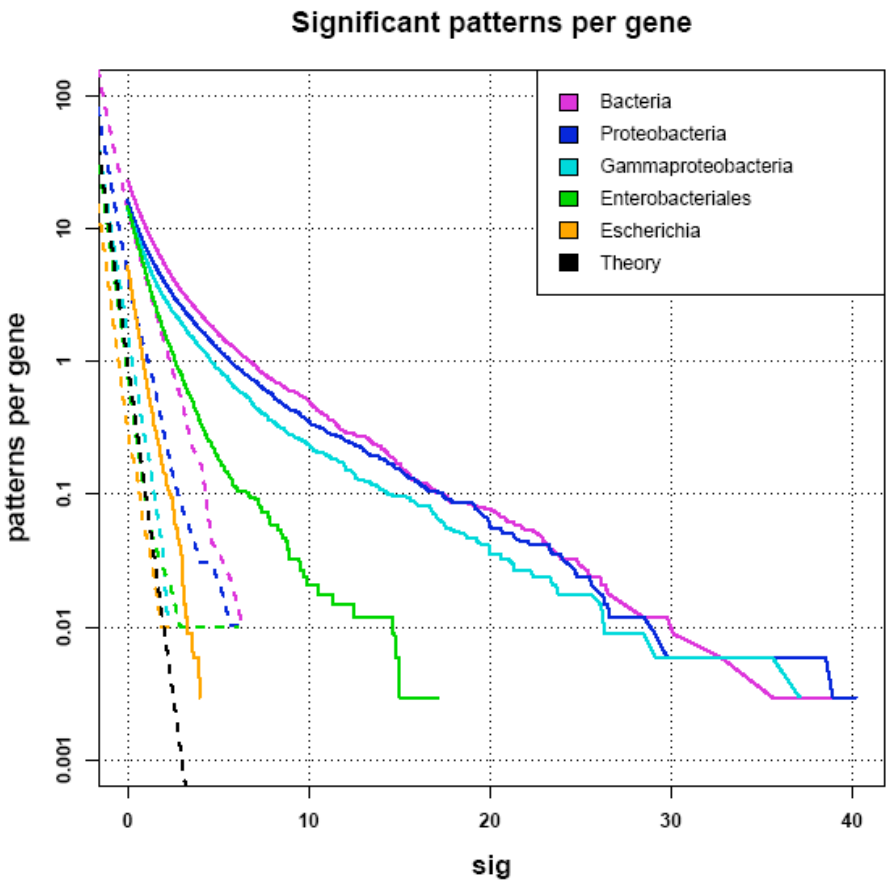
Janky & van Helden, submitted

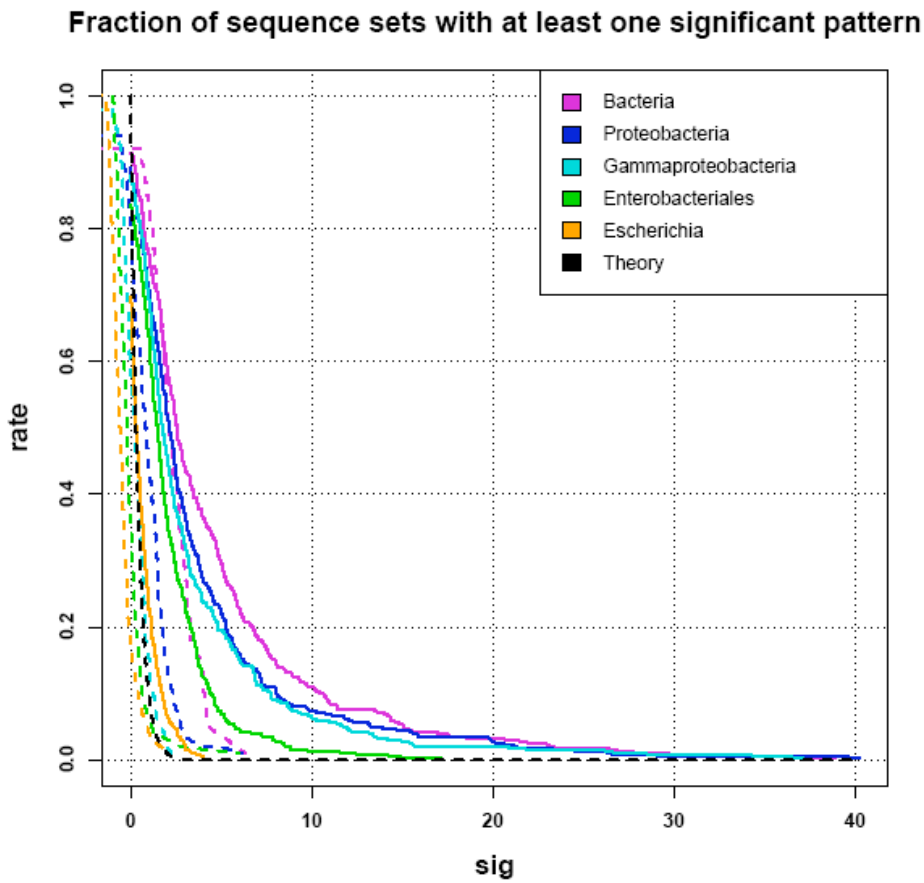
Systematic evaluation of phylogenetic footprints discovery

Principle of the evaluation

- Choice of the sequence sets:
  - Positive set: promoters of orthologs for all the genes having at least one annotated site in RegulonDB
  - Negative set: random selections of promoters.
- Evaluation criteria
  - Distribution of significance scores: does this score allow us to discriminate groups of co-regulated promoters from random selections?
  - Motif correctness: do the motifs predicted in the positive set correspond to the annotated sites ?

Distribution of significance scores





Correctness of the discovered motifs

- How to compare a discovered motif with a collection of annotated binding sites ?
  - The annotated binding sites can be considered as a set of oligonucleotides (“words”).
  - The discovered motif is a set of dyads.
  - We can compare two sets of words with the program compare-patterns.

Discovered motifs in promoters of glpD orthologs from Enterobacteriales

Dyad	exp_freq	occ	exp_occ	occ_P	occ_E	sig	rank
atgn{1}tcg|cgan{1}cat	0.00069	11	1.02	1.30E-08	2.40E-04	3.62	1
aacn{0}att|aatn{0}gtt	0.00131	13	1.95	1.60E-07	3.00E-03	2.52	2
cgan{0}aca|tgtn{0}tcg	0.00056	9	0.84	2.90E-07	5.20E-03	2.28	3
agtn{10}tcg|cgan{10}act	0.00026	6	0.35	2.20E-06	3.90E-02	1.41	4
acan{13}acg|cgtn{13}tgt	0.00044	7	0.6	3.50E-06	6.40E-02	1.19	5
aacn{2}aca|tgtn{2}gtt	0.00082	9	1.21	5.60E-06	1.00E-01	0.99	6
aatn{2}tcg|cgan{2}att	0.00089	9	1.31	1.00E-05	1.90E-01	0.73	7
atgn{0}ttc|gaan{0}cat	0.00094	9	1.4	1.70E-05	3.10E-01	0.52	8
acgn{2}cat|atgn{2}cgt	0.00054	7	0.79	2.10E-05	3.80E-01	0.42	9
aacn{3}cat|atgn{3}gtt	0.00101	9	1.47	2.60E-05	4.60E-01	0.33	10
aacn{8}aag|cttn{8}gtt	0.00080	8	1.11	2.60E-05	4.70E-01	0.33	11
cctn{4}cac|gtgn{4}agg	0.00023	5	0.34	2.90E-05	5.30E-01	0.28	12
aacn{0}gaa|ttcn{0}gtt	0.00106	9	1.58	4.40E-05	7.90E-01	0.1	13
actn{7}aag|cttn{7}agt	0.00027	5	0.37	4.90E-05	8.90E-01	0.05	14

;assembly # 1	seed: atgntcg	13 words
aaangtt....	....aacnttt	0.82
.aatgtt....	....aacatt.	3.38
.aatnntcg..	..cgannatt.	2.1
..atgntcg..	..cgancat..	7.55
..atgttc...	...gaacat..	3.29
..atgnnngtt	aacnnncat..	1.63
..atgnncgt.	.acgnncat..	1.46
...tgttcg..	..cgaaca...	4.58
...tgtnngtt	aacnnaca...	2.4
...tgtncgt.	.acgnaca...	1.32
....gttngtt	aacnaac....	1.05
....gttcgt.	.acgaac....	0.47
.....ttcgtt	aacgaa.....	1.53
aaatgttcgtt	aacgaacattt	7.55
;assembly # 2	seed: atgnnnnnnntcg	19 words
aaangtt.........	.........aacnttt	0.82
.aatgtt.........	.........aacatt.	3.38
.aatnntcg.......	.......cgannatt.	2.1
..atgntcg.....	**.....cgancat..**	**7.55**
..atgttc........	........gaacat..	3.29
..atgnnnnnnntcg.	.cgannnnnncat..	1.82
..atgnnngtt.....	.....aacnnncat..	1.63
..atgnncgt......	......acgnncat..	1.46
..atgnnnnnnncga	tcgnnnnnnncat..	0.93
...tgttcg.......	.......cgaaca...	4.58
...tgtnngtt.....	.....aacnnaca...	2.4
...tgtncgt......	......acgnaca...	1.32
...tgtnnnnnntcg.	.cgannnnnnaca...	0.74
....gttngtt.....	.....aacnaac....	1.05
....gttcgt......	......acgaac....	0.47
....gttnnnnntcg.	.cgannnnnaac....	0.35
.....ttcgtt.....	.....aacgaa.....	1.53
......tcgnnntcg.	.cgannncga......	0.6
......tcgnnnncga	tcgnnnncga......	0.06
aaatgttcgttntcga	tcganaacgaacattt	7.55
;assembly # 3	seed: agtnnnnnnnnntcg	3 words
agtnnnnnnnnntcg	**cgannnnnnnnnact**	**1.52**
.gttnnnnnnnntcg	cgannnnnnnnnaac.	0.9
..ttcnnnnnnntcg	cgannnnnnnngaa..	0.78
agttcnnnnnnntcg	cgannnnnnngaact	1.52
;assembly # 4	seed: acannnnnnnnnnnnnacg	3 words
acannnnnnnnnnnnnacg.	**.cgtnnnnnnnnnnnntgt**	**1.26**
acannnnnnnnnnnnnncga	tcgnnnnnnnnnnnnntgt	0.48
acannnnnnnnnnnnaac..	..gttnnnnnnnnnnntgt	0.01
acannnnnnnnnnnnaacga	tcgttnnnnnnnnnnntgt	1.26

Matching list - discovered dyads against annotated sites in glpD promoters

Annotated sites for the gene glpD (Escherichia coli K12)

Site sequence	***Factor***	***site ID***
gataaacgccATAATGTTATACATATCACTCTaaaatgtttt	CRP	ECK120014013
tctttgctaaTATGTTCGATAACGAACATTtatgagcttt	GlpR	ECK120012732
taacgaacatTTATGAGCTTTAACGAAAGTgaatgagggc	GlpR	ECK120012734
gggatcaactGGTTTGCGTTTTGGCGCAAAttcagtgtta	GlpR	ECK120013968
aaaccggaaaTTAAGCGCGGATTCGAATATtctgactgtt	GlpR	ECK120013970

Perfect matches between detected dyads and annotated sites

Site ID	matching bases	strand	offset	weight	**seq1**	**seq2**
CRP_glpD_ECK120014013	6	R	30	6	**AAAACATTTTAGAGTGATATGTATAACATTATGGCGTTTATC**	**aacn{0}att**
GlpR_glpD_ECK120012732	7	D	11	6	**tctttgctaaTATGTTCGATAACGAACATTtatgagcttt**	**atgn{1}tcg**
GlpR_glpD_ECK120012732	6	D	24	6	**tctttgctaaTATGTTCGATAACGAACATTtatgagcttt**	**aacn{0}att**
GlpR_glpD_ECK120012732	6	D	22	6	**tctttgctaaTATGTTCGATAACGAACATTtatgagcttt**	**cgan{0}aca**
GlpR_glpD_ECK120012732	8	D	20	6	**tctttgctaaTATGTTCGATAACGAACATTtatgagcttt**	**aacn{2}aca**
GlpR_glpD_ECK120012732	8	R	20	6	**AAAGCTCATAAATGTTCGTTATCGAACATATTAGCAAAGA**	**aatn{2}tcg**
GlpR_glpD_ECK120012732	6	D	11	6	**tctttgctaaTATGTTCGATAACGAACATTtatgagcttt**	**atgn{0}ttc**
GlpR_glpD_ECK120012732	8	D	21	6	**tctttgctaaTATGTTCGATAACGAACATTtatgagcttt**	**acgn{2}cat**
GlpR_glpD_ECK120012732	9	D	20	6	**tctttgctaaTATGTTCGATAACGAACATTtatgagcttt**	**aacn{3}cat**
GlpR_glpD_ECK120012732	6	D	20	6	**tctttgctaaTATGTTCGATAACGAACATTtatgagcttt**	**aacn{0}gaa**
GlpR_glpD_ECK120012734	7	R	0	6	**GCCCTCATTCACTTTCGTTAAAGCTCATAAATGTTCGTTA**	**atgn{1}tcg**
GlpR_glpD_ECK120012734	6	D	5	6	**taacgaacatTTATGAGCTTTAACGAAAGTgaatgagggc**	**aacn{0}att**
GlpR_glpD_ECK120012734	6	D	3	6	**taacgaacatTTATGAGCTTTAACGAAAGTgaatgagggc**	**cgan{0}aca**
GlpR_glpD_ECK120012734	19	D	6	6	**taacgaacatTTATGAGCTTTAACGAAAGTgaatgagggc**	**acan{13}acg**
GlpR_glpD_ECK120012734	8	D	1	6	**taacgaacatTTATGAGCTTTAACGAAAGTgaatgagggc**	**aacn{2}aca**
GlpR_glpD_ECK120012734	8	R	1	6	**GCCCTCATTCACTTTCGTTAAAGCTCATAAATGTTCGTTA**	**aatn{2}tcg**
GlpR_glpD_ECK120012734	6	R	0	6	**GCCCTCATTCACTTTCGTTAAAGCTCATAAATGTTCGTTA**	**atgn{0}ttc**
GlpR_glpD_ECK120012734	8	D	2	6	**taacgaacatTTATGAGCTTTAACGAAAGTgaatgagggc**	**acgn{2}cat**
GlpR_glpD_ECK120012734	9	D	1	6	**taacgaacatTTATGAGCTTTAACGAAAGTgaatgagggc**	**aacn{3}cat**
GlpR_glpD_ECK120012734	10	R	28	6	**GCCCTCATTCACTTTCGTTAAAGCTCATAAATGTTCGTTA**	**cctn{4}cac**
GlpR_glpD_ECK120012734	6	D	1	6	**taacgaacatTTATGAGCTTTAACGAAAGTgaatgagggc**	**aacn{0}gaa**
GlpR_glpD_ECK120012734	13	R	20	6	**GCCCTCATTCACTTTCGTTAAAGCTCATAAATGTTCGTTA**	**actn{7}aag**
GlpR_glpD_ECK120013970	8	R	21	6	**AACAGTCAGAATATTCGAATCCGCGCTTAATTTCCGGTTT**	**aatn{2}tcg**
GlpR_glpD_ECK120013970	14	D	1	6	**aaaccggaaaTTAAGCGCGGATTCGAATATtctgactgtt**	**aacn{8}aag**

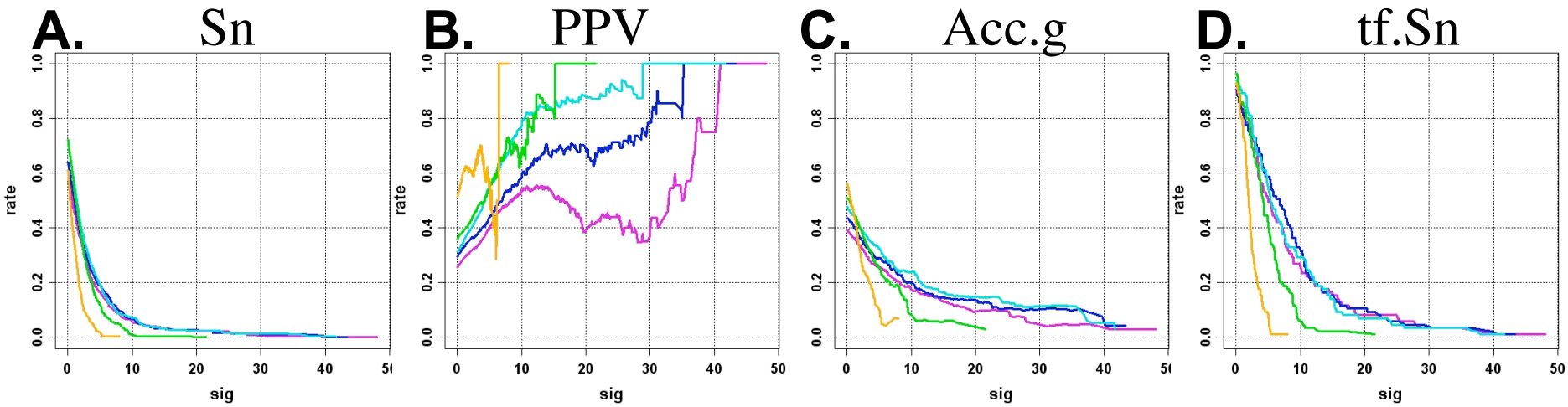
Matching table - discovered dyads against annotated sites in glpD promoters

; sequence	gataaacgccATAATGTTATACATATCACTCTaaaatgtttt	tctttgctaaTATGTTCGATAACGAACATTTatgagcttt	taacgaacatTTATGAGCTTTAACGAAAGTgaatgagggc	gggatcaactGGTTTGCGTTTTGGCGCAAAttcagtgtta	aaaccggaaaTTAAGCGCGGATTCGAATATAttctgactgtt
atgn{1}tcg	.	6	6	.	.
aacn{0}att	6	6	6	.	.
cgan{0}aca	.	6	6	.	.
agtn{10}tcg	.	.	.	.	.
acan{13}acg	.	.	6	.	.
aacn{2}aca	.	6	6	.	.
aatn{2}tcg	.	6	6	.	6
atgn{0}ttc	.	6	6	.	.
acgn{2}cat	.	6	6	.	.
aacn{3}cat	.	6	6	.	.
aacn{8}aag	.	.	.	.	6
cctn{4}cac	.	.	6	.	.
aacn{0}gaa	.	6	6	.	.
actn{7}aag	.	.	6	.	.

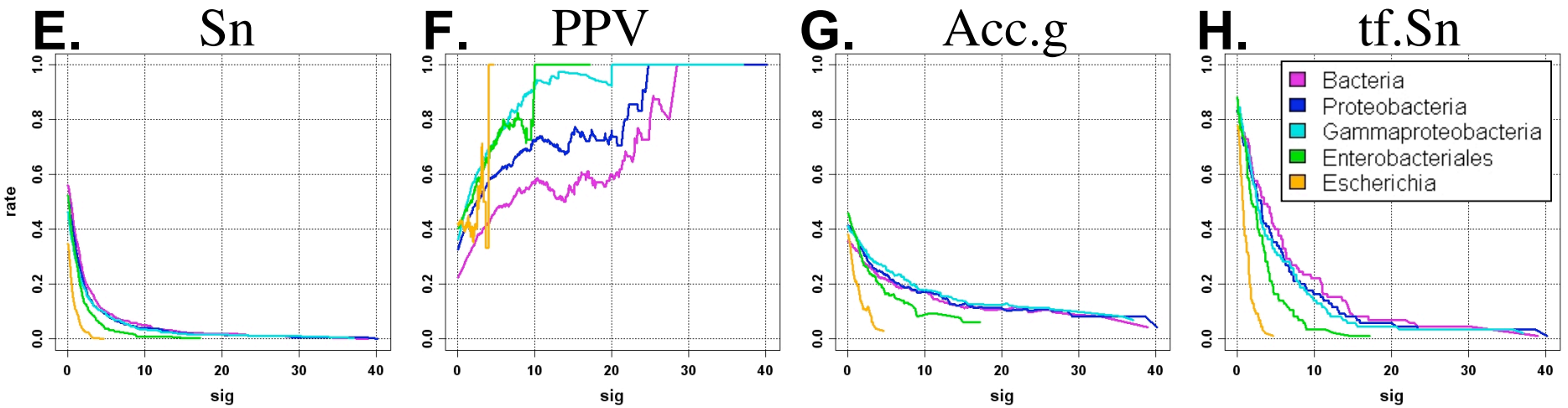
- The dyad-site cross-table allows us to define correspondence statistics.
- Sensitivity (Sn)
  - fraction of annotated sites matched by at least one discovered dyad
- Positive Predictive Value (PPV)
  - Fraction of discoverd dyads matching at least one annotated site
- Accuracy (geometric)
  - Geometric mean of Sn and PPV
  - Acc=sqrt(Sn*PPV)

Correctness statistics for all the genes having at least one annotated site in RegulonDB

TAXFREQ



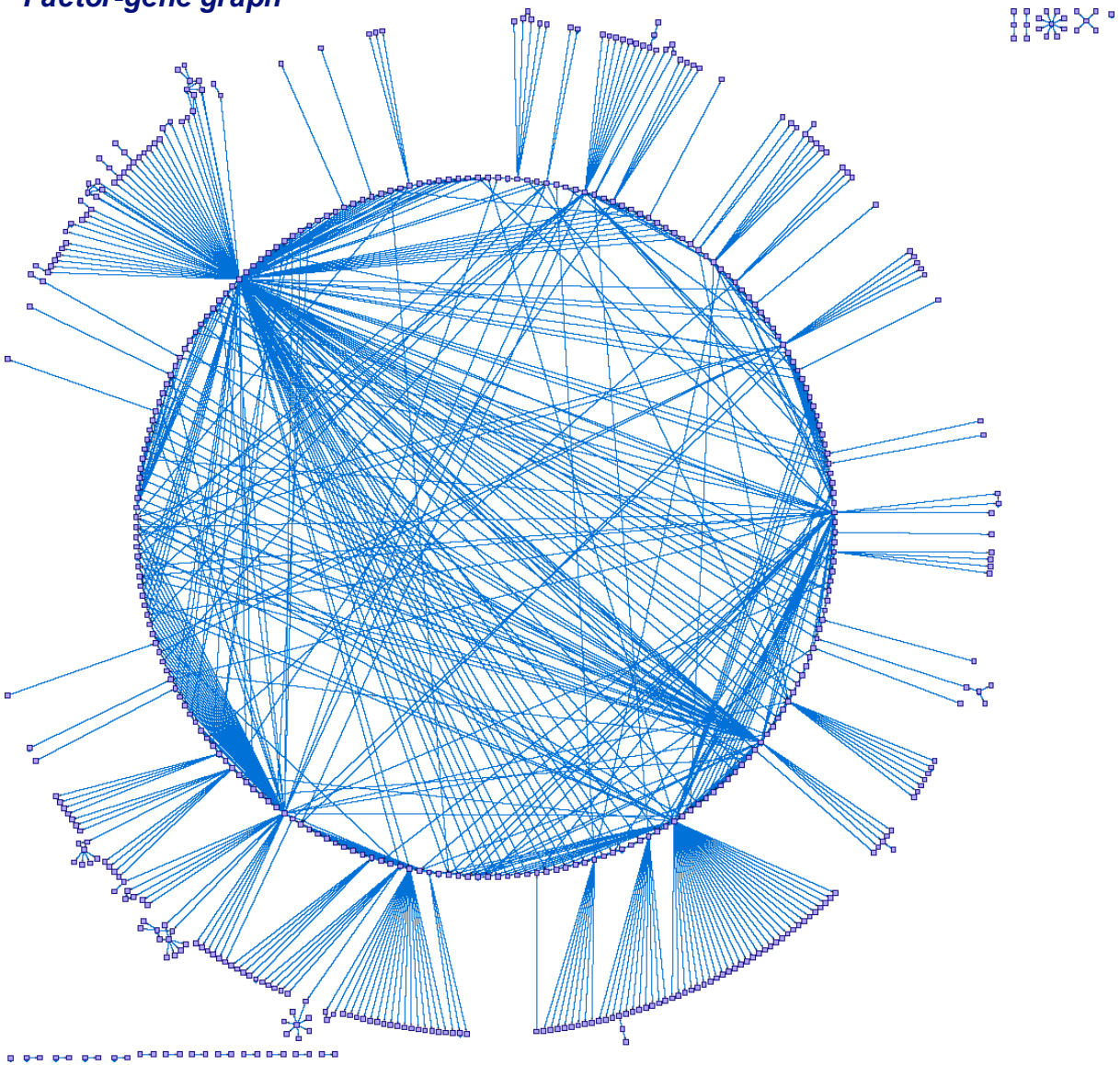
MONAD



Inferring co-regulation network from phylogenetic footprints

RegulonDB factor -> gene network

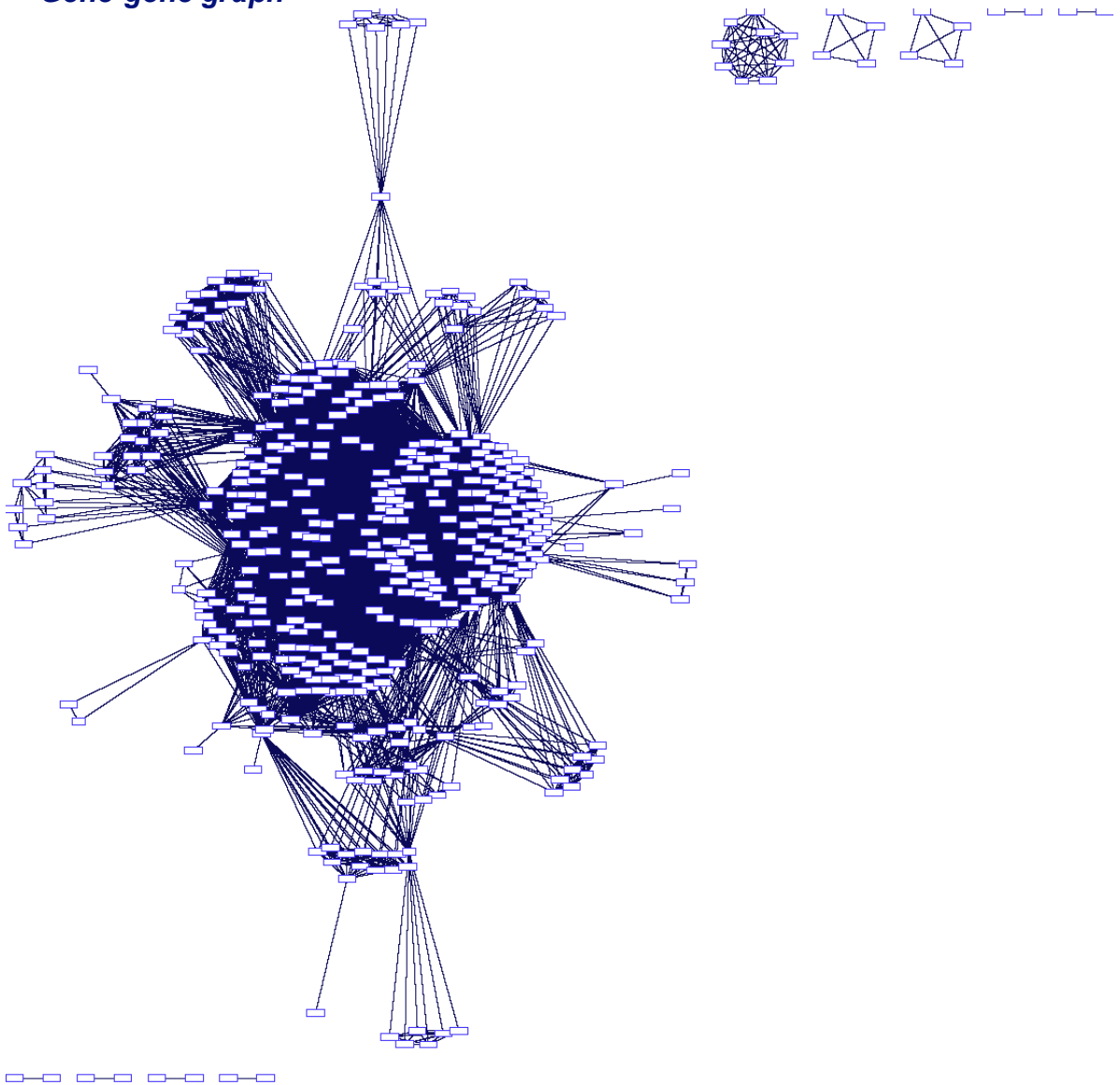
Factor-gene graph



- RegulonDB (Oct. 2005 version)
- The graph represents the relationships between factors and their target genes
  (factor -> gene graph)
  - 125 transcription factors
  - 467 target genes
  - 847 factor->gene interactions
  - 45 self-regulations
  - Note: CRP alone regulates 132 target genes.

RegulonDB gene <-> gene network

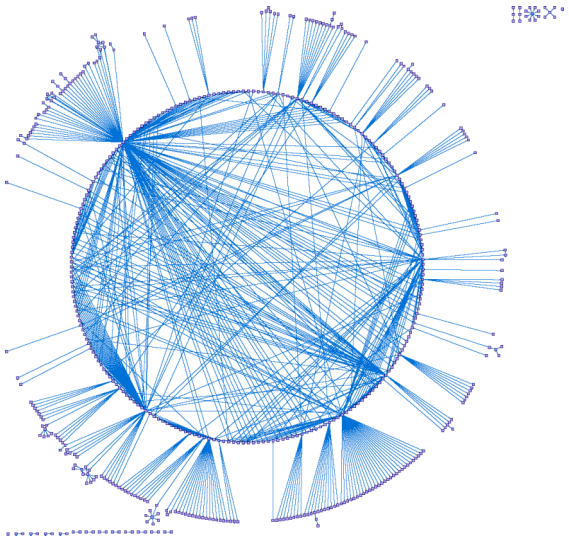
Gene-gene graph



- The factor -> gene graph can be converted in a gene <-> gene graph, with an edge joining each pair of co-regulated genes (bottom).
- The gene-gene co-regulation graph contains
  - 458 nodes (genes).
    - Note: some genes are the only target of a given factor, they are thus not present in this graph.
  - 14,939 arcs (co-regulation between a pair of genes).
  - Note : (132*131)/2=8,646 of these arcs correspond to links via CRP.

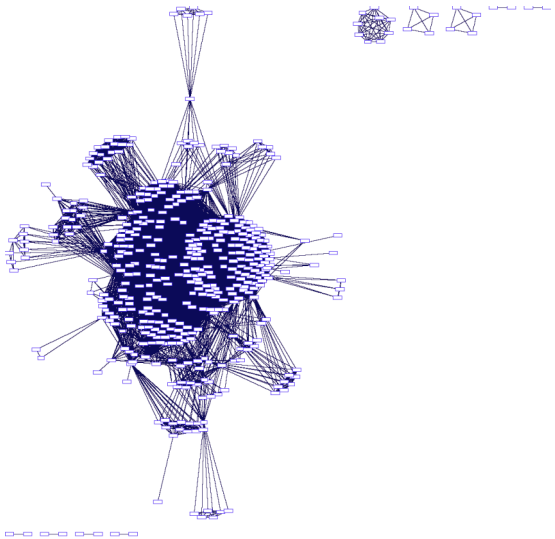
RegulonDB networks

Factor-gene graph



- RegulonDB (Oct. 2005 version)
- The top figure displays the interaction graph between factors and genes (factor -> gene graph)
  - 125 transcription factors
  - 467 target genes
  - 847 factor->gene interactions
  - 45 self-regulations
  - Note: CRP alone regulates 132 target genes.

Gene-gene graph



- The factor -> gene graph can be converted in a gene <-> gene graph, with an edge joining each pair of co-regulated genes (bottom).
- The gene-gene co-regulation graph contains
  - 458 nodes (genes).
    - Note: some genes are the only target of a given factor, they are thus not present in this graph.
  - 14,939 arcs (co-regulation between a pair of genes).
  - Note : (132*131)/2=8,646 of these arcs correspond to links via CRP.

Testing the principle with RegulonDB

Gene	Factor	GI
metN	MetJ	1786398
ahpC	MetJ	1786822
metK	MetJ	1789311
metC	MetJ	1789383
metB	MetJ	1790375
metF	MetJ	1790377
metA	MetJ	1790443
glyA	MetR	1788902
metH	MetR	1790450
mhpA	MhpR	1786543
purE	PurR	1786734
pyrD	PurR	1787177
pyrC	PurR	1787301
hflD	PurR	1787377
purR	PurR	1787948
cvpA	PurR	1788652
purC	PurR	1788820
purM	PurR	1788845
guaB	PurR	1788855
glyA	PurR	1788902
glnB	PurR	1788904
gcvT	PurR	1789272
purH	PurR	1790439
purL	PurR	48994899
putA	PutA	1787250

- Test
  - Discover motifs in upstream sequences of its orthologous genes of each gene annotated in regulonDB.
  - Link genes with similar discovered motifs.
  - Compare the inferred links with the regulon membership
- Small test case
  - Genes regulated by methionine (MetJ and MetR targets), purine (PurR targets) and proline (PutR, PutA targets)
- Complete test case
  - All the genes annotated as target gene or TF-coding in regulonDB (537 genes)

Dyad-profiles

- We can collect the significant dyads found for each gene, and display this information as “dyad profiles”:
  - Each row represents one gene
  - Each column represents one dyad found significant in at least one gene
  - Dots indicate that a dyad was not significant for a given gene (occ_sig < 0)
  - Note that the majority of the cells are empty.

	aaan{0}aat	aagn{0}ggg	aagn{1}gga	aagn{1}tgc	aagn{2}gag	aagn{2}gca	aagn{3}agc	aagn{4}gcc	aagn{6}cca	acgn{0}caa	acgn{0}tct	acgn{1}cta	acgn{1}ttg	acgn{2}aac	acgn{2}tgc	acgn{3}acg	acgn{4}cgt	actn{9}gag	agan{10}tct	agan{1}gtc	agan{5}aga	…
META	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.7	2.7	…
METB	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2.6	2.2	.	…
METF	.	.	.	.	.	0.0	.	.	.	.	.	0.5	.	.	.	.	.	.	.	.	.	…
METH	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	…
METJ	.	.	.	.	.	.	.	.	.	.	2.7	2.1	.	.	.	.	.	.	1.4	3.5	.	…
METK	.	.	.	.	.	.	.	.	.	.	.	5.5	.	.	.	.	.	.	.	.	.	…
METN	.	.	.	.	.	.	.	.	.	.	.	0.5	.	.	.	.	.	.	.	.	.	…
METR	.	.	.	.	.	.	.	.	.	.	.	0.1	.	.	.	.	.	.	.	.	.	…
PURA	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	…
PURC	.	.	1.2	.	0.9	.	0.9	0.7	0.1	.	.	.	.	.	.	.	.	0.6	.	.	.	…
PURE	.	.	.	.	.	.	.	.	.	.	.	.	0.1	.	1.5	.	0.1	.	.	.	.	…
PURH	.	.	.	.	.	.	.	.	.	.	.	.	.	0.2	.	0.9	2.5	.	.	.	.	…
PURL	.	.	.	.	.	.	.	.	.	1.0	.	.	.	.	.	.	.	.	.	.	.	…
PURM	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	…
PURR	.	0.2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	…
PUTA	.	.	.	0.6	.	4.2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	…
PUTP	1.8	.	.	.	.	2.1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Compare-profiles

- The program ***compare-profiles*** was originally designed to compare binary profiles.
- It takes as input a profile file and compares each pair of profiles Q and R.
  - Profile intersection (QR),
  - union (QvR),
  - differences (Q!R, R!Q),
  - common exclusion (!Q!R)
  - Jaccard similarity = intersection/union
  - Significance of the intersection QR
    - hypergeometric significance test on the right tail : P-value + E-value + sig
  - Significance of thecommon exclusion !Q!R
    - hypergeometric significance test on the left tail : P-value + E-value + sig
  - Mutual information
- This program can be used to detect co-occurrence or mutual exclusion of genes across phylogenetic profiles.
- We can also use it to compare dyad profiles
  - Convert significance scores to a Boolean value

Converting dyad profiles to dyad classes

- Analysis of RegulonDB (537 genes)
  - 422 genes with at least one motif
  - 4944 dyads found significant in at least one gene
  - The profile table contains 422 rows x 4944 columns = 2,086,368 cells.
  - No more than 6,709 of these 2,086,368 cells are non-empty.
- Complete analysis of Escherichia coli (4,200 genes)
  - 2,844 genes with at least one significant motif
  - 2,844 x 17,153 dyads = 48,783,132 cells
  - No more than 33,370 of these 48,783,132 cells are non-empty.
- The profile profile matrices are very sparse.
- A (much) cheapest structure to store the same information is a 3 column file
  - dyad
  - gene
  - significance score

dyad	gene	occ_sig
gacn{0}gtc	META	0.16
agan{1}gtc	META	0.65
atcn{5}gtc	META	0.75
agan{5}aga	META	2.74
agan{1}gtc	METB	2.17
agan{10}tct	METB	2.56
gacn{0}gtc	METB	2.96
aagn{2}gca	METF	0.01
caan{3}gca	METF	0.35
gcan{0}agc	METF	0.35
acgn{1}cta	METF	0.52
agcn{2}caa	METF	0.64
gacn{0}gtc	METF	1.64
tagn{5}taa	METF	1.95
agcn{3}aag	METF	2.15
tgan{7}tca	METF	2.83
agcn{1}gca	METF	3.02
tgan{7}tca	METH	0.11
agan{6}gac	METJ	0.15
agan{9}gtc	METJ	0.15
cgtn{0}cta	METJ	0.48
cgtn{2}aga	METJ	1.04
agan{10}tct	METJ	1.35
gacn{2}cta	METJ	1.9
acgn{1}cta	METJ	2.05
acgn{0}tct	METJ	2.73
gacn{0}gtc	METJ	3.1
agan{1}gtc	METJ	3.46
taan{0}aaa	METK	0.1
gacn{0}gtc	METK	0.44
cgtn{2}aga	METK	0.51
…	…	…

Compare-classes

- The program ***compare-classes*** was designed to compare clusters
  - Clusters of co-expressed genes versus GO classes
  - Clusters of co-expressed genes versus annotated regulons
  - Annotated regulons versus annotated regulons (to detect synergy between transcription factors)
  - Chip-on-chip data versus chip-on-chip data

- We will now use compare-classes to compare the sets of dyads between genes.
- For each pair of genes Q and R, compare-classes calculates
  - number of dyads found significant in Q
  - number of dyads found significant in R
  - number of dyads in the intersection (QR)
  - number of dyads in the union (QvR)
  - Jaccard similarity = intersection/union
  - dot product (see next slide)
  - significance test on this intersection (hypergeometric, right tail)
    - P-value, E-value, sig

The dot product

- Problem
  - The hypergeometric test only compares group memberships: a dyad is considered to be significant or not, but its score is not taken into consideration.
  - We would however like to use the significance score returned by dyad analysis.
- For this, I implemented a dot product

$$dp = \sum_{i=1}^{p} \left(x_{Ai} \cdot x_{Bi} \right)$$

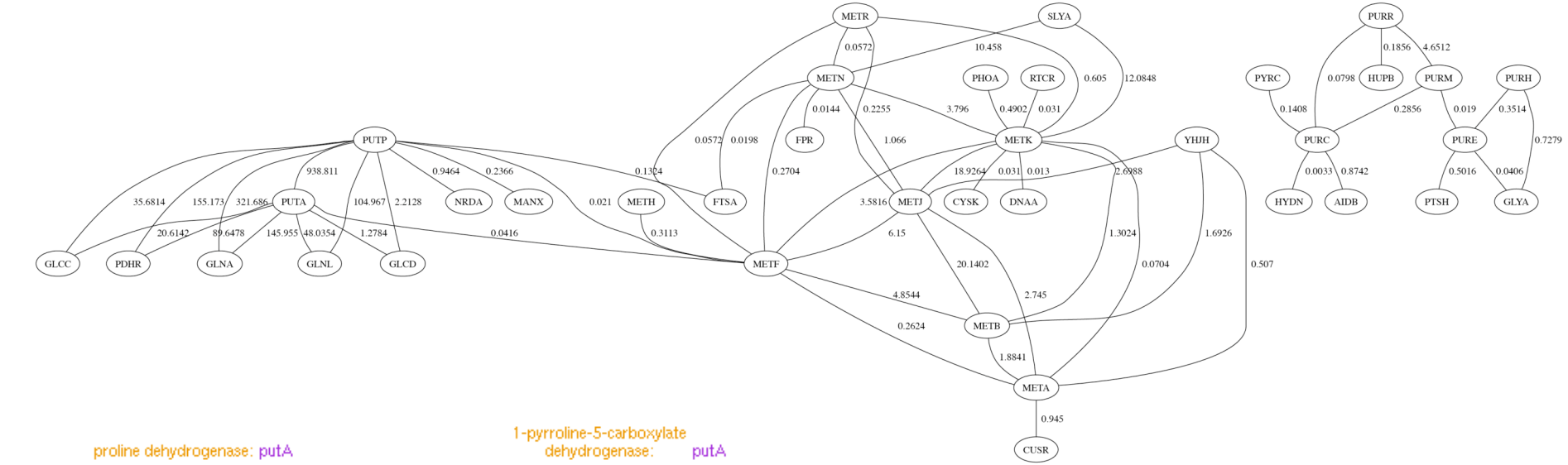
p	is the number of variables (dyads)
x_{Ai}	is the score of significance returned for dyad i in gene A
x_{Bi}	is the score of significance returned for dyad i in gene B

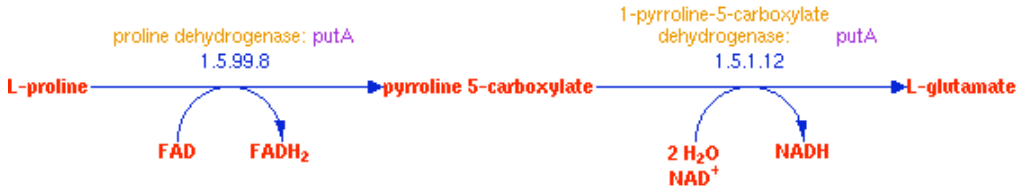
Attention : negative values of x_{Ai} and x_{Bi} have to be omitted (set to 0)

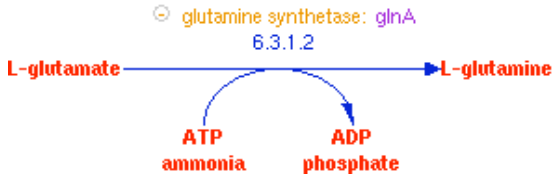
Gene pairs with all the regulonDB genes
Top scoring by dot product

Rank	gene R	gene Q	R	Q	QR	!Q!R	dotprod	P_val	E_val	sig
1	PUTP	PUTA	38	37	25	4894	2542.1	7.10E-49	6.30E-44	43.2
2	RECA	LEXA	57	101	14	4800	1889.8	3.30E-12	2.90E-07	6.53
3	GLNL	GLNA	51	88	32	4837	1713.11	7.00E-46	6.20E-41	40.21
4	UDP	MTLA	15	54	11	4886	1043.88	1.20E-19	1.10E-14	13.98
5	UVRD	LEXA	41	101	11	4813	831.49	2.80E-10	2.50E-05	4.6
6	TYRP	AROF	25	32	11	4898	815.729	5.10E-19	4.50E-14	13.34
7	PUTP	GLNA	38	88	3	4821	628.522	0.02936	2608.27	-3.42
8	MTLA	MLC	54	11	5	4884	448.632	5.70E-08	0.00503	2.3
9	PSPF	PSPA	64	64	64	4880	437.011	7.20E-148	6.40E-143	142.19
10	PUTA	GLNA	37	88	3	4822	431.504	0.02739	2432.93	-3.39
11	ILVY	ILVC	93	75	51	4827	386.541	9.00E-77	8.00E-72	71.1
12	SULA	RECA	41	57	11	4857	358.099	4.20E-13	3.70E-08	7.43
13	UVRD	RECA	41	57	10	4856	347.532	1.60E-11	1.40E-06	5.86
14	UVRA	LEXA	12	101	5	4836	346.642	2.30E-06	0.20236	0.69
15	PUTP	PDHR	38	59	1	4848	330.398	0.36741	32637.392	-4.51
16	POLA	DNAA	39	32	15	4888	296.35	6.80E-25	6.10E-20	19.22
17	BIOB	BIOA	31	36	22	4899	287.556	4.70E-44	4.20E-39	38.38
18	METJ	METB	45	23	19	4895	286.685	1.70E-37	1.50E-32	31.81
19	MTLA	CYDA	54	19	3	4874	281.577	0.00105	93.699	-1.97
20	PUTA	PDHR	37	59	1	4849	269.359	0.35971	31953.518	-4.5
21	NAGE	MTLA	19	54	6	4877	227.65	3.10E-08	0.00276	2.56
22	UDP	MLC	15	11	5	4923	200.044	5.60E-11	5.00E-06	5.3
23	MTLA	ADHE	54	8	3	4885	197.458	6.60E-05	5.896	-0.77
24	MALK	MALE	21	73	19	4869	197.254	2.70E-34	2.40E-29	28.62
25	NRDD	NRDA	39	63	20	4862	192.165	2.60E-29	2.30E-24	23.63
26	YFID	MTLA	29	54	5	4866	183.569	1.30E-05	1.113	-0.05
27	YFID	CYDA	29	19	6	4902	165.44	6.10E-10	5.40E-05	4.27
28	SULA	LEXA	41	101	4	4806	164.636	0.00931	826.926	-2.92
29	MTLA	CDD	54	14	4	4880	162.407	1.20E-05	1.043	-0.02
30	UVRA	RECA	12	57	5	4880	160.535	1.30E-07	0.01127	1.95
31	SSB	LEXA	22	101	2	4823	159.836	0.07326	6507.387	-3.81
32	CUER	COPA	43	47	29	4883	157.892	2.40E-53	2.20E-48	47.67
33	UHPT	MTLA	12	54	3	4881	157.165	0.00025	22.457	-1.35
34	MTLA	AGAR	54	31	8	4867	154.984	7.70E-10	6.80E-05	4.17
35	PUTP	GLNL	38	51	2	4857	150.447	0.05794	5147.255	-3.71
36	TYRR	AROF	29	32	6	4889	147.935	1.90E-08	0.0017	2.77
37	XSEA	GUAB	29	37	13	4891	139.955	1.30E-21	1.20E-16	15.92
38	ULAA	MTLA	27	54	3	4866	132.93	0.00299	265.909	-2.42
39	MTLA	GLPA	54	15	5	4880	130.283	3.60E-07	0.03161	1.5
40	PSTS	PHOB	28	31	7	4892	130.01	2.00E-10	1.80E-05	4.75
41	GLNK	GLNA	14	88	6	4848	125.801	7.20E-08	0.00637	2.2
42	MALP	MALE	19	73	11	4863	124.547	2.30E-16	2.00E-11	10.69
43	ARGE	ARGC	38	24	14	4896	122.146	3.10E-25	2.70E-20	19.56
44	PDHR	GLCC	59	13	7	4879	118.735	3.90E-11	3.40E-06	5.46
45	TYRR	TYRP	29	25	6	4896	114.252	3.90E-09	0.00034	3.47
46	METN	FPR	79	24	13	4854	113.54	3.40E-18	3.10E-13	12.51
47	MTLA	MALE	54	73	8	4825	112.418	9.20E-07	0.08208	1.09
48	PDHR	FADB	59	14	5	4876	110.459	3.80E-07	0.03338	1.48
49	TREB	MTLA	36	54	3	4857	107.764	0.00682	605.69	-2.78
50	PUTA	GLNL	37	51	2	4858	103.141	0.05525	4907.985	-3.69

Inferred co-regulation graph (selection)

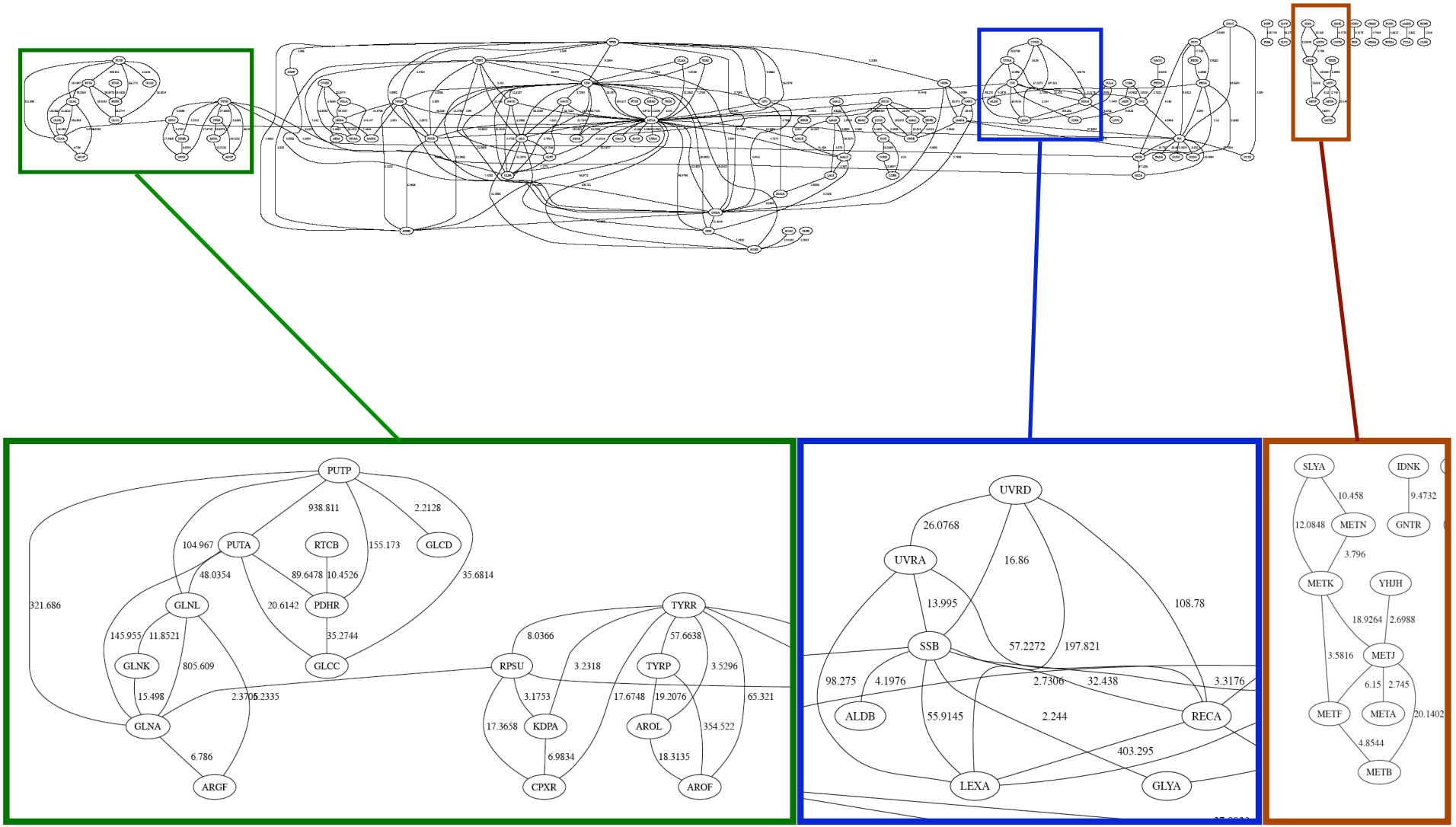






Metabolic pathways from http://biocyc.org/

All the pairs of dyad profiles with dp >= 2

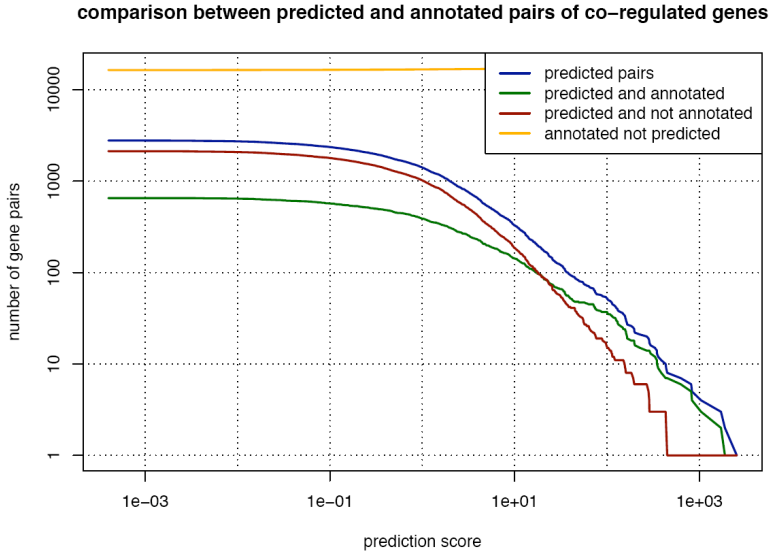


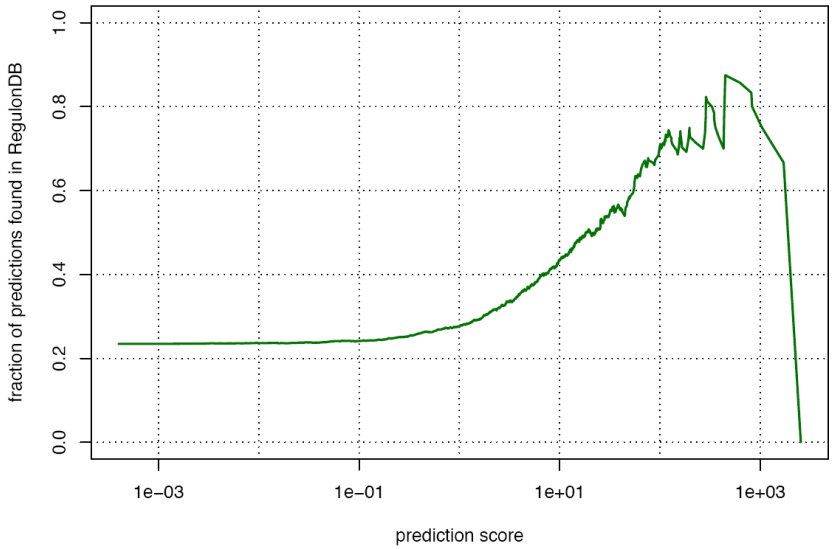
Predicted versus annotated gene pairs

Gene pair	Annot	Score	Sum pred	Sum annot	P.and.A	P.not.A	A.not.P	PPV	Sn
PUTP_PUTA	0	2542.1	1	0	0	1	17076	0.000	0.000
RECA_LEXA	1	1889.8	2	1	1	1	17075	0.500	0.000
GLNL_GLNA	1	1713.1	3	2	2	1	17074	0.667	0.000
UDP_MTLA	1	1043.9	4	3	3	1	17073	0.750	0.000
UVRD_LEXA	1	831.5	5	4	4	1	17072	0.800	0.000
TYRP_AROF	1	815.7	6	5	5	1	17071	0.833	0.000
PUTP_GLNA	1	628.5	7	6	6	1	17070	0.857	0.000
MTLA_MLC	1	448.6	8	7	7	1	17069	0.875	0.000
PSPF_PSPA	0	437.0	9	7	7	2	17069	0.778	0.000
PUTA_GLNA	0	431.5	10	7	7	3	17069	0.700	0.000
ILVY_ILVC	1	386.5	11	8	8	3	17068	0.727	0.001
SULA_RECA	1	358.1	12	9	9	3	17067	0.750	0.001
UVRD_RECA	1	347.5	13	10	10	3	17066	0.769	0.001
UVRA_LEXA	1	346.6	14	11	11	3	17065	0.786	0.001
PUTP_PDHR	1	330.4	15	12	12	3	17064	0.800	0.001
POLA_DNAA	1	296.4	16	13	13	3	17063	0.813	0.001
BIOB_BIOA	1	287.6	17	14	14	3	17062	0.824	0.001
METJ_METB	0	286.7	18	14	14	4	17062	0.778	0.001
MTLA_CYDA	0	281.6	19	14	14	5	17062	0.737	0.001
PUTA_PDHR	0	269.4	20	14	14	6	17062	0.700	0.001
NAGE_MTLA	1	227.7	21	15	15	6	17061	0.714	0.001
UDP_MLC	1	200.0	22	16	16	6	17060	0.727	0.001
MTLA_ADHE	1	197.5	23	17	17	6	17059	0.739	0.001
MALK_MALE	1	197.3	24	18	18	6	17058	0.750	0.001
NRDD_NRDA	0	192.2	25	18	18	7	17058	0.720	0.001
YFID_MTLA	0	183.6	26	18	18	8	17058	0.692	0.001
YFID_CYDA	1	165.4	27	19	19	8	17057	0.704	0.001
SULA_LEXA	1	164.6	28	20	20	8	17056	0.714	0.001
MTLA_CDD	1	162.4	29	21	21	8	17055	0.724	0.001
UVRA_RECA	1	160.5	30	22	22	8	17054	0.733	0.001
SSB_LEXA	1	159.8	31	23	23	8	17053	0.742	0.002
CUER_COPA	0	157.9	32	23	23	9	17053	0.719	0.002
UHPT_MTLA	1	157.2	33	24	24	9	17052	0.727	0.002
MTLA_AGAR	0	155.0	34	24	24	10	17052	0.706	0.002
PUTP_GLNL	0	150.4	35	24	24	11	17052	0.686	0.002
TYRR_AROF	1	147.9	36	25	25	11	17051	0.694	0.002
XSEA_GUAB	1	140.0	37	26	26	11	17050	0.703	0.002
ULAA_MTLA	1	132.9	38	27	27	11	17049	0.711	0.002
MTLA_GLPA	1	130.3	39	28	28	11	17048	0.718	0.002
PSTS_PHOB	1	130.0	40	29	29	11	17047	0.725	0.002
GLNK_GLNA	1	125.8	41	30	30	11	17046	0.732	0.002
MALP_MALE	1	124.5	42	31	31	11	17045	0.738	0.002
ARGE_ARGC	1	122.1	43	32	32	11	17044	0.744	0.002
PDHR_GLCC	0	118.7	44	32	32	12	17044	0.727	0.002
TYRR_TYRP	1	114.3	45	33	33	12	17043	0.733	0.002
METN_FPR	0	113.5	46	33	33	13	17043	0.717	0.002
MTLA_MALE	1	112.4	47	34	34	13	17042	0.723	0.002
PDHR_FADB	0	110.5	48	34	34	14	17042	0.708	0.002
TREB_MTLA	1	107.8	49	35	35	14	17041	0.714	0.002

- Among the 537*536/2=~143,916 gene pairs considered here (between 537 genes annotated in regulonDB), 2,790 show a common motif.
- Among the 50 top-scoring predicted gene pairs, 35 are annotated as co-regulated in RegulonDB.
- Among the 15 top-scoring predicted which are non-annotated, at least 8 are actually involved in related pathways
  - Some are already known to be co-regulated, but the annotation is not yet entered in RegulonDB.
  - Some others are very likely to be co-regulated.
- Whole genome analysis (4200 genes)
  - Among the ~16M possible gene pairs 54,099 show a common motif

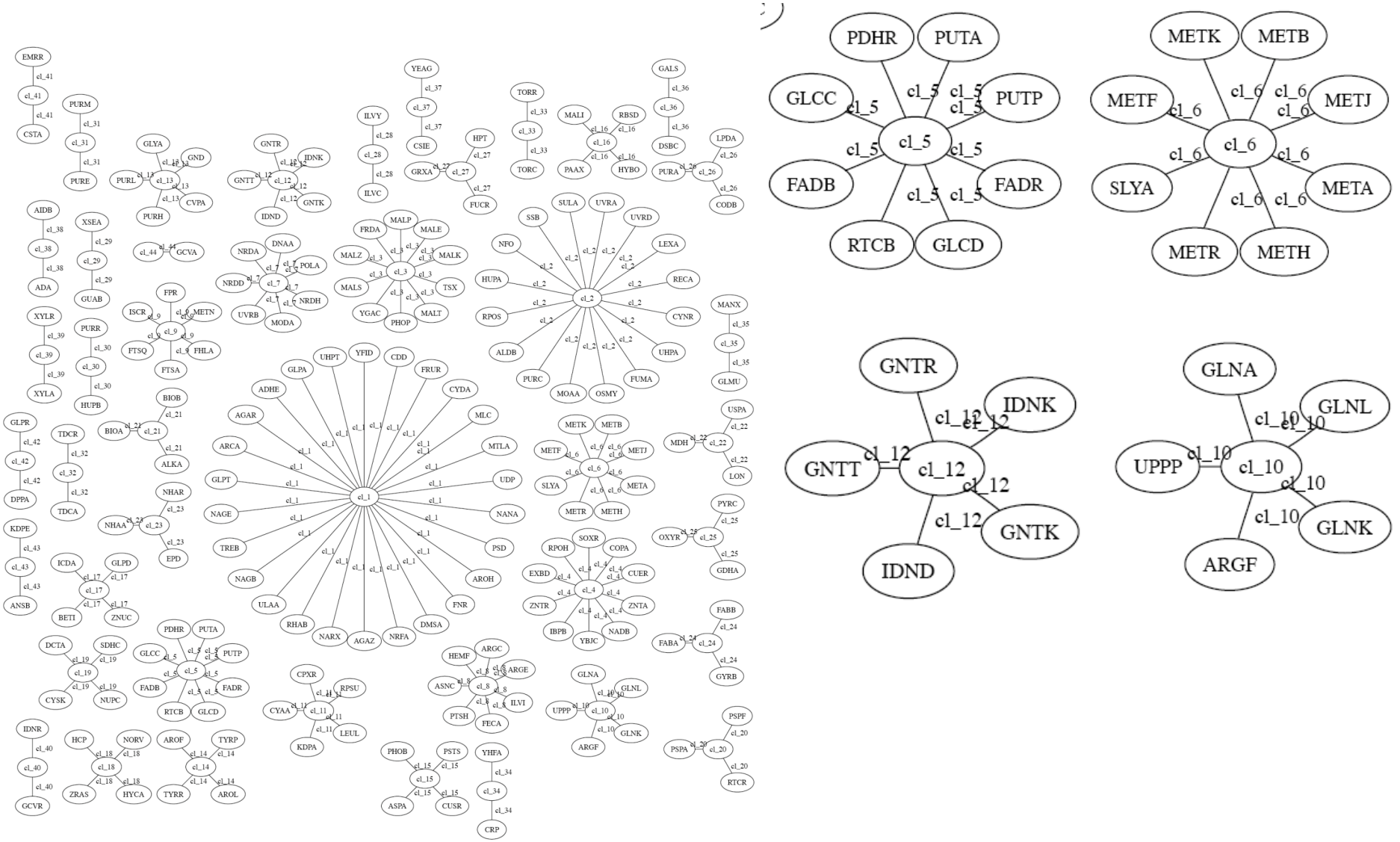
Predicted versus annotated gene pairs





- We performed a systematic estimation of the fraction of predictions found in RegulonDB, as a function of the prediction score (dot product).
- More than 40% of the predictions with a score dp >= 10 are annotated in RegulonDB.
- Note that predicted pairs which are not found in RegulonDB might nevertheless be correct, it is hard to estimate the rate of “false positive” in such a case.
- The sensitivity can however be estimated, and it is quite low.
  - Among the 14,939 gene-gene co-regulated pairs in RegulonDB, the method detects no more than 652 (4.4%).
  - Note that more than 50% of the annotated gene pairs are linked via CRP.
- To do : analyze the sensitivity per TF, to check whether some regulons are preferentially detected.

Clustering


PDHR
PUTA
GLCC
PUTP
cl_5
FADB
FADR
RTCB
GLCD
METK
METB
METF
METJ
cl_6
SLYA
META
METR
METH
GNTR
IDNK
cl_12
GNTT
GNTK
IDND
GLNA
GLNL
cl_10
UPPP
GLNK
ARGF

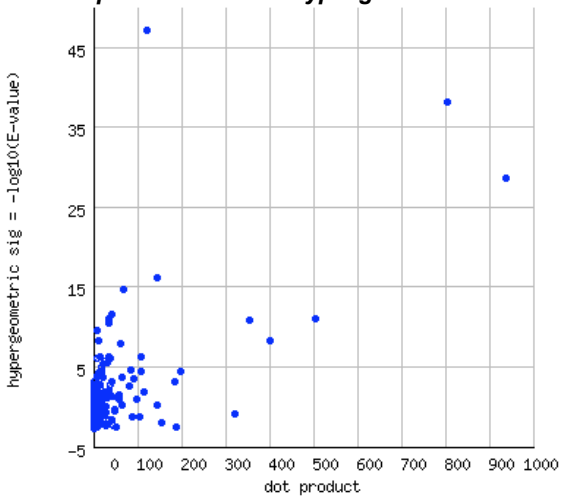
Supplementary material

Clusters versus regulons

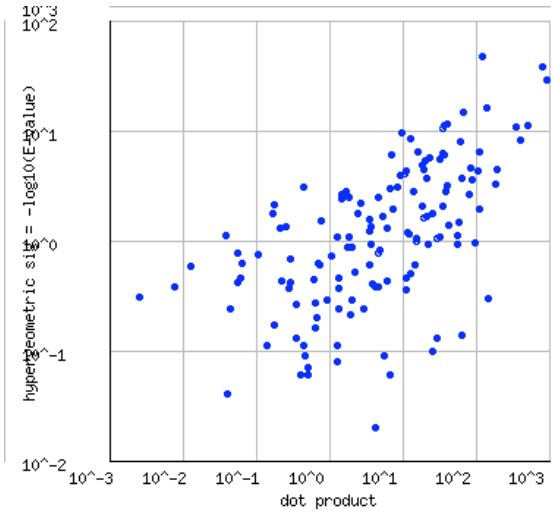
cluster	12	11	10	6	4	2	8	9	7	3	5	1	clust
CRP	22	4	10	1	6					4			47
FNR	4	1	1	7	2	3							18
IHF	5	2	1		1	2	2						13
ARCA	5	1	1	2									9
FIS	4			1		1				1		1	8
LEXA		1							1		6		8
NARL	4			1		3							8
PURR	7												7
FLHD	3				1	1		1					6
LRP	3	1									1		5
MALT			5										5
METJ								5					5
MODE	2			1		2							5
DNAA	1								2			1	4
MLC	2	1			1								4
NARP	3					1							4
TYRR							4						4
CPXR	2								1				3
CYTR	1		1							1			3
FRUR	1			1						1			3
GADX	2	1											3
GLPR	1				2								3
GNTR	3												3
MARA		1		1				1					3
NAGC		1			1					1			3
NTRC		3											3
PHOB							3						3
ARGP									1			1	2
BIRA							2						2
FADR	2												2
FHLA	2												2
GADE	1								1				2
GLCC		2											2
H-NS	1			1									2
IDNR	2												2
ILVY	2												2
MALI	2												2
METR	1							1					2
NAC	2												2

Compare-classes result (selection)

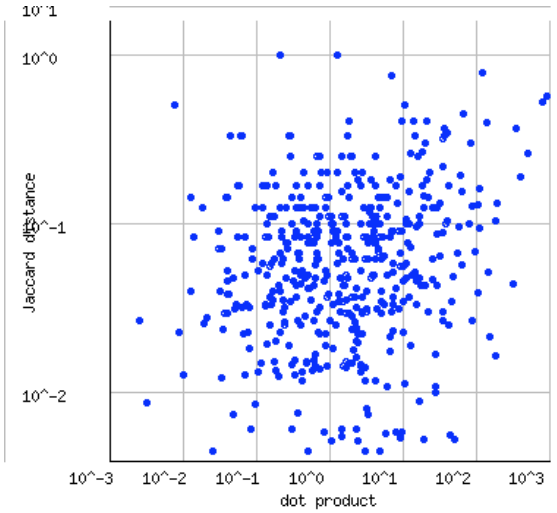
Dot product versus hypergeometric



Dot product versus hypergeometric (log)



Dot product versus Jaccard (log)



Hypergeometricc versus Jaccard (log)