Discovering phylogenetic footprints in bacterial promoters

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

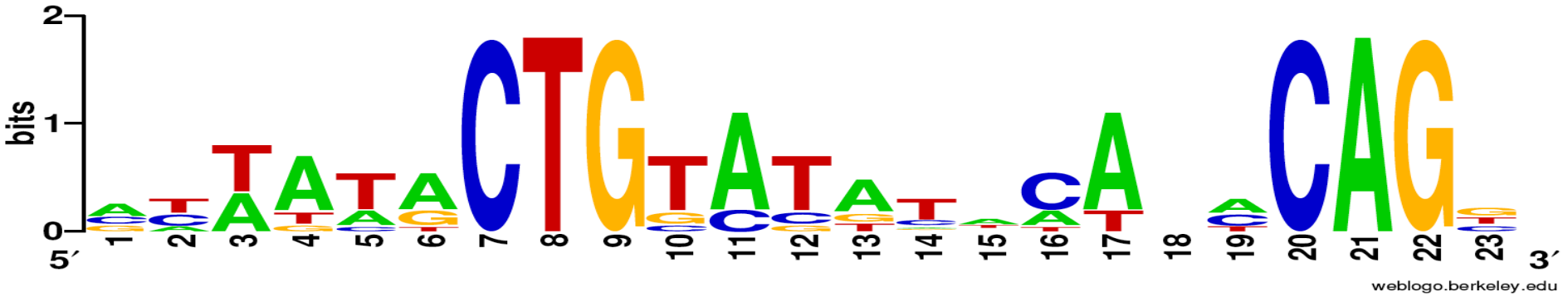
LexA binding sites (RegulonDB, oct 2006)

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**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	ATTATA**CTG**TATGCTCATA**CAG**T
2 | 2	:	2/8	CTTTTG**CTG**TATATACTCA**CAG**C
3 | 3	:	3/10	GCATAA**CTG**TATATACACC**CAG**G
4 | 4	:	4/7	ATTATA**CTG**TATGCTCATA**CAG**T
5 | 5	:	-5/13	CCAATA**CTG**TATATTCATT**CAG**G
6 | 6	:	-6/15	GATGTA**CTG**TACATCCATA**CAG**T
7 | 7	:	-7/13	CCAATA**CTG**TATATTCATT**CAG**G
8 | 8	:	-8/13	ATTATA**CTG**GATAAAAAAA**CAG**T
9 | 9	:	9/7	GCAAAT**CTG**TATATATACC**CAG**C
10 | 10	:	10/8	AATAAG**CTG**GCGTTGATGC**CAG**C
11 | 11	:	-11/12	ATAACG**CTG**CCCTGAAAGC**CAG**G

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

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	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_atrosepticaa	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	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
Nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
…. . .
```

Significantly over-represented dyads in promoters of PUTA orthologs

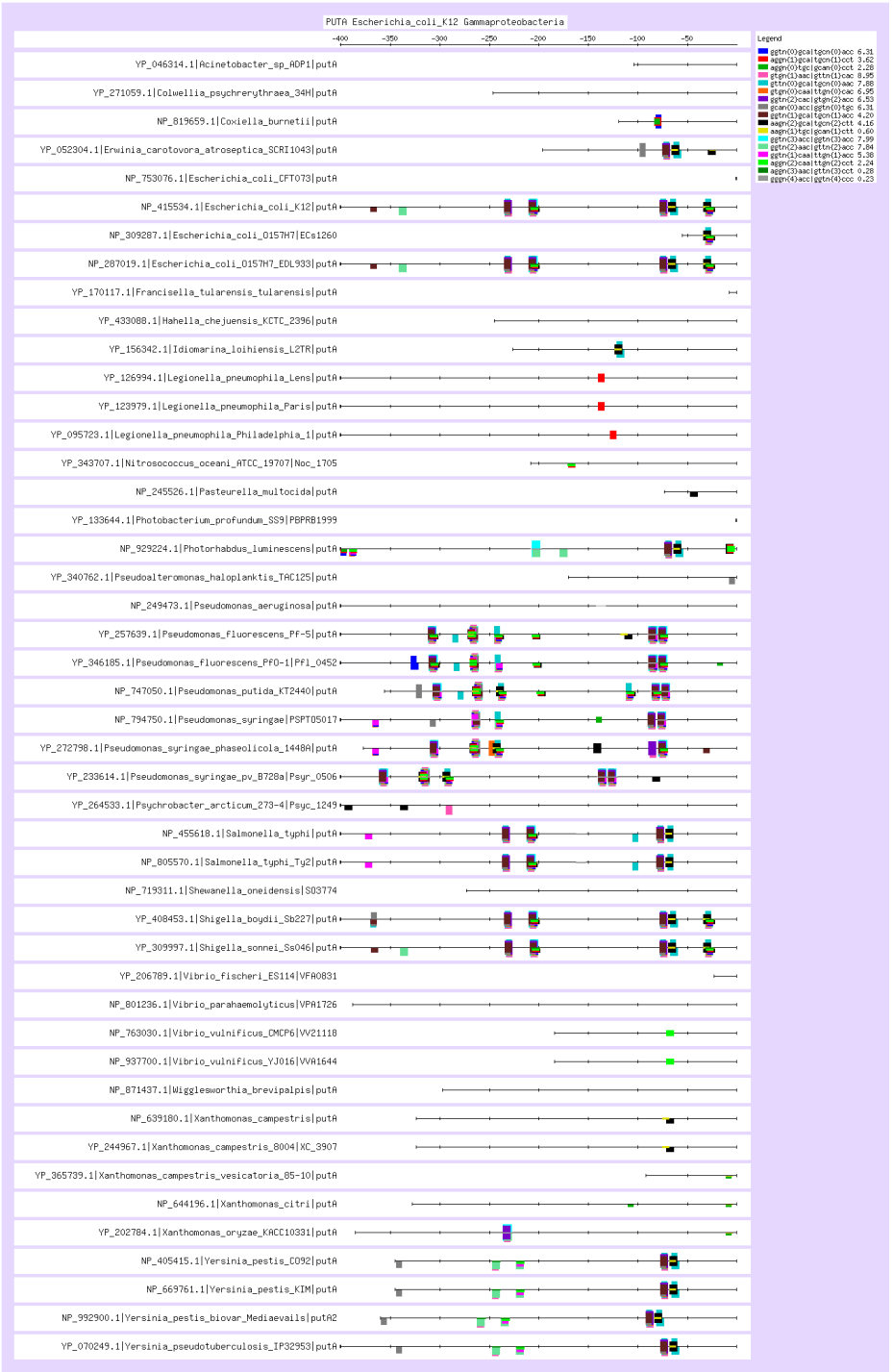
; column headers
; 1 sequence
; 2 identifier
; 3 expected_freq
; 4 occ observed occurrences
; 5 exp_occ expected occurrences
; 6 occ_P occurrence probability (binomial)
; 7 occ_E E-value for occurrences (binomial)
; 8 occ_sig occurrence significance (binomial)
; 9 ovl_occ number of overlapping occurrences
; 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 pattern	50		
; number of input patter	17		
;			
;assembly # 1	**seed: gtgnaa**	**17 words**	**lengt**
; aligntt	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

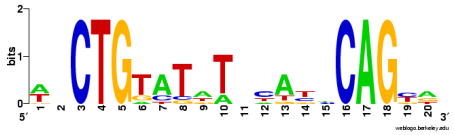
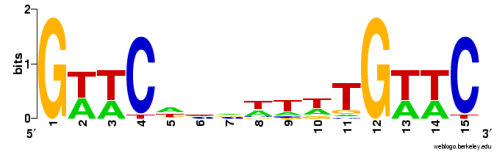
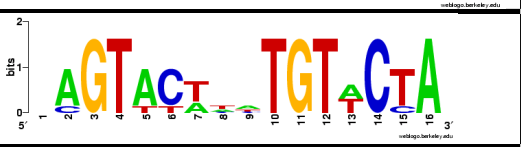
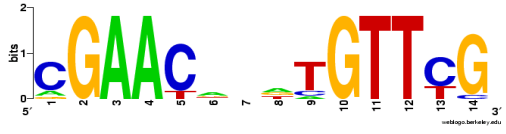
- For each gene, we applied the same pattern discovery approach
  - Identify orthologs
  - Retrieve upstream sequences
  - Detect over-represented dyads
- Questions
  - How good is this method in predicting cis-acting elements ?
    - Can we detect correct motifs ?
    - What is the rate of false positives ?
  - 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

Evolution of the auto-regulation of the LexA transcription factor

Study case: LexA auto-regulation

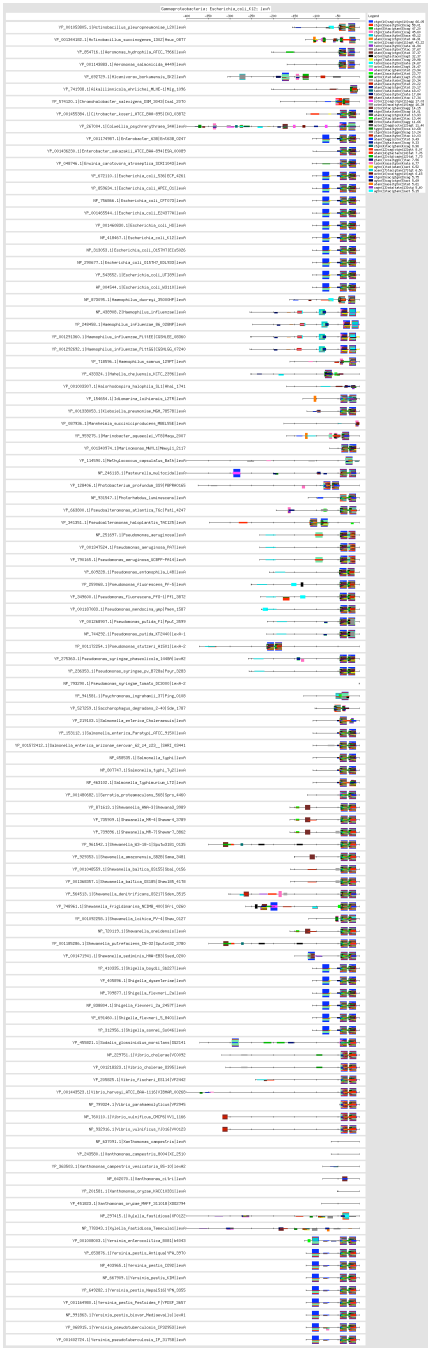
- The transcription factor LexA represses several genes involved in the SOS response
- The *lexA* gene is auto-regulated.
- LexA auto-regulation has been characterized in details in several bacterial species.
  - LexA protein has evolved and recognizes different motifs in different taxa.
- Note that 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.

Evolution of lexA binding motifs
Motifs characterized experimentally in different bacterial taxa

Taxonomic level	Species	Consensus	Weblogo
Gamma-proteobacteria	*E.coli*	CTGTNNNNNNNNACAG	
Xanthomonadales	*X.oryzae* *X.campestris* *X.citri* *Xylella fastidiosa*	TTAGTArwawTACTAa	
Alpha proteobacteria	*R.etli* *R.meliloti* *A.tumefasciens* *Rhodobacter sphaeroides* *Sinorhizobium meliloti*	GTTCNNNNNNNGTTC GAACNNNNNNNGAAC	
Delta proteobacteria	*Myxococcus xanthus*	CTRHAMRYBYGTTCAGS	
	Geobacter sulfurreducens	GGTTNNCNNNNGNNNACC	
Cyanobacteria	*Anabaena PCC7120*	TAGTACTWATGTTCTA	
Gram+ bacteria	*B.subtilis* *Mycobacterium tuberculosis* *Mycobacterium smegmatis*	CGAACRNRYGTTYC	

Janky, R. and van Helden, J. (2008). Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution. BMC Bioinformatics 9, 37.

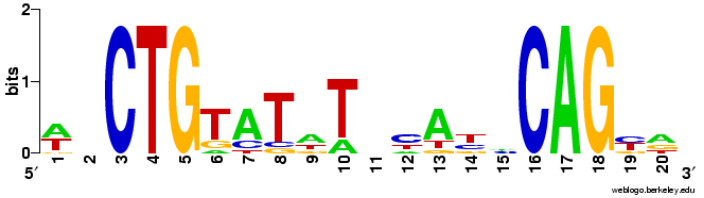
Significant dyads in promoters of lexA orthologs in Gammaproteobacteria



- In Gammaproteobacteria, the most significant motif is `aaCTGtatacatataCAGtt`
- This corresponds to the LexA motif annotated in RegulonDB

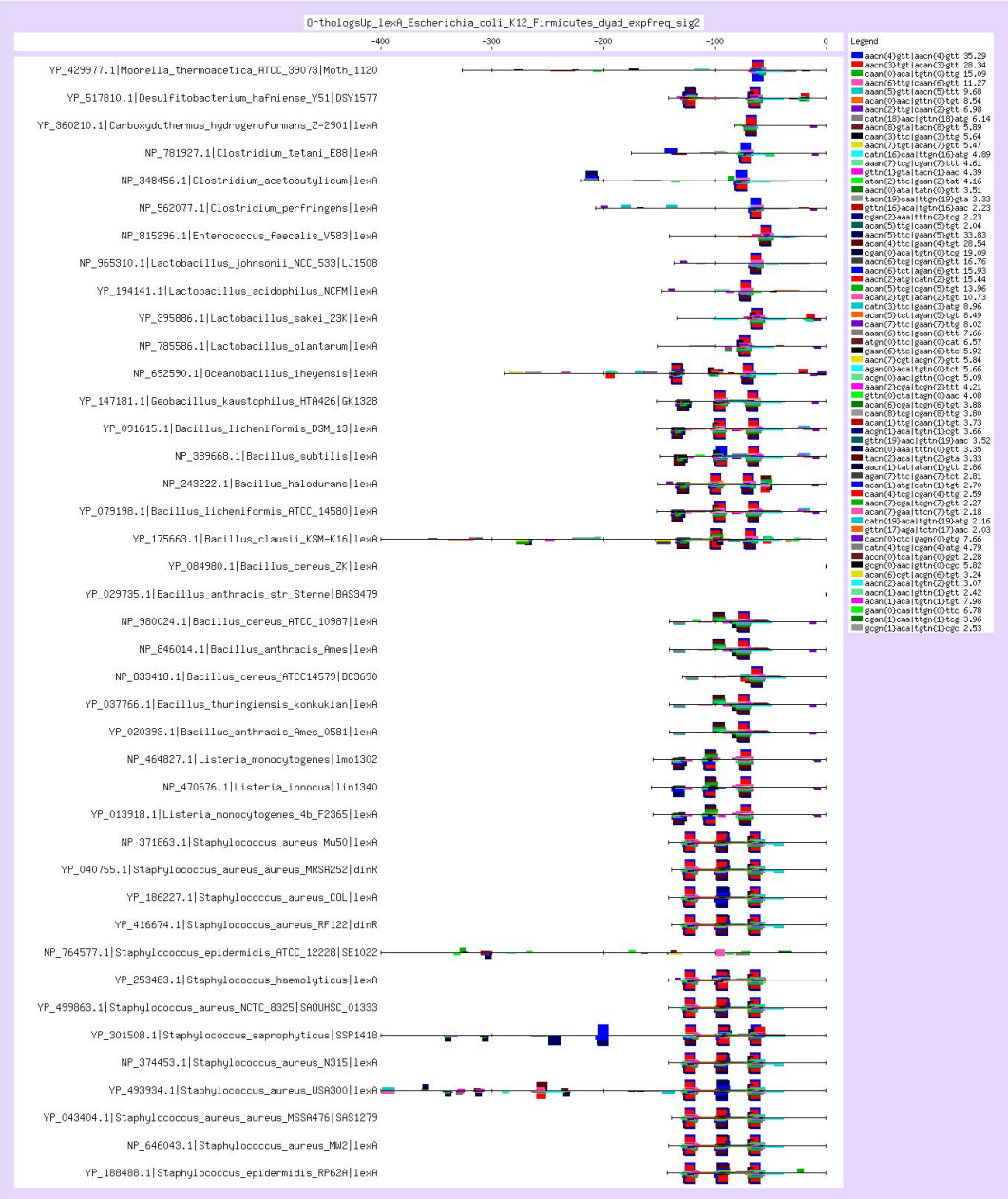
Discovered motif *(footprint-discovery)*

Annotated LexA binding motif (RegulonDB)



Footprint-discovery result with lexA in Gammaproteobacteria, March 2009

Significant dyads in promoters of lexA orthologs in Firmicutes

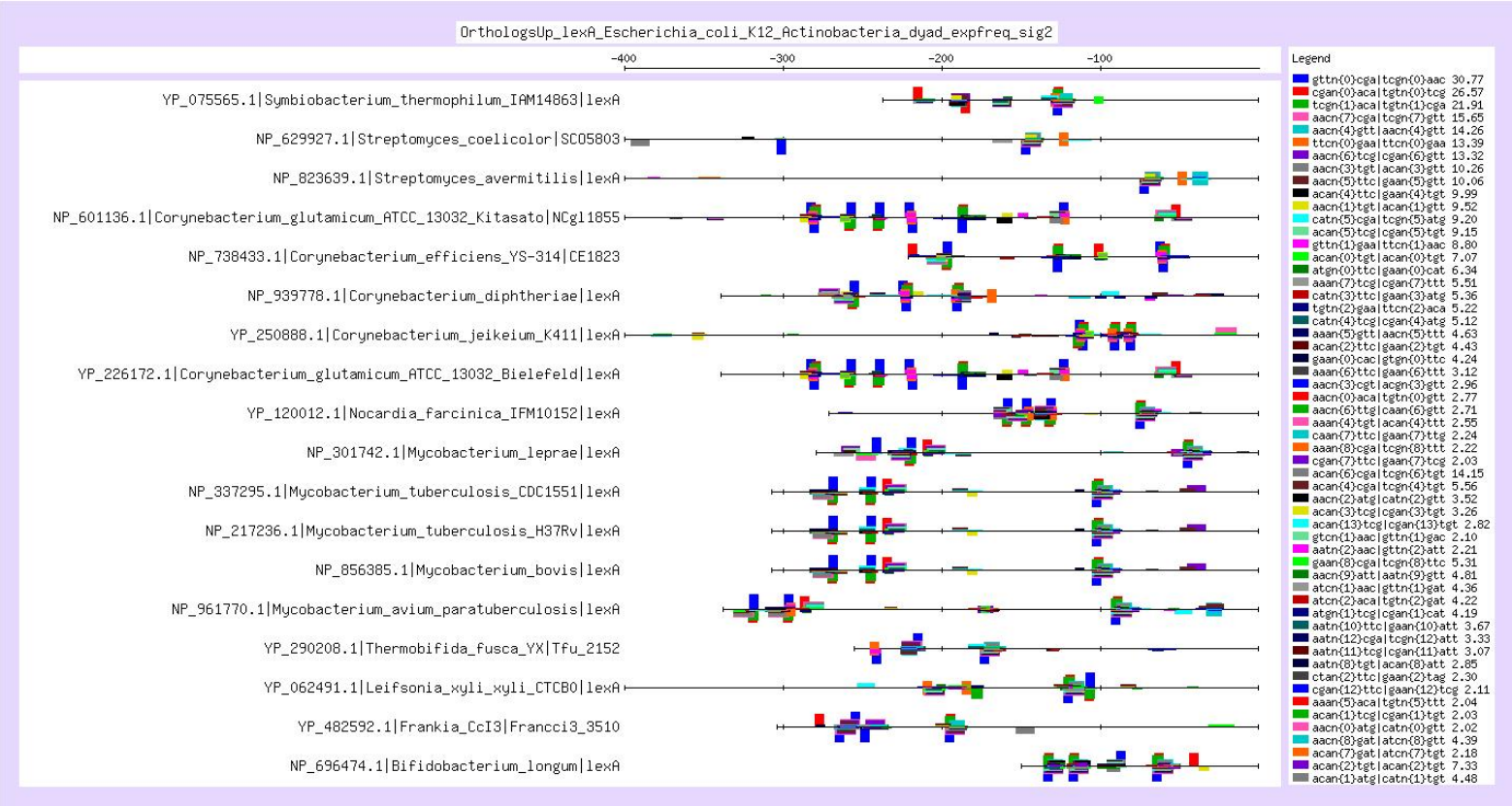


- In Firmicutes, the most significant motif is
  TAC**GAAC**ATAT**GTTC**GTA
- This corresponds to the “Cheo box”
  GAACn_4**GTTC**

Janky, R. and van Helden, J. Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution *BMC Bioinformatics* **9**, 37 (2008).

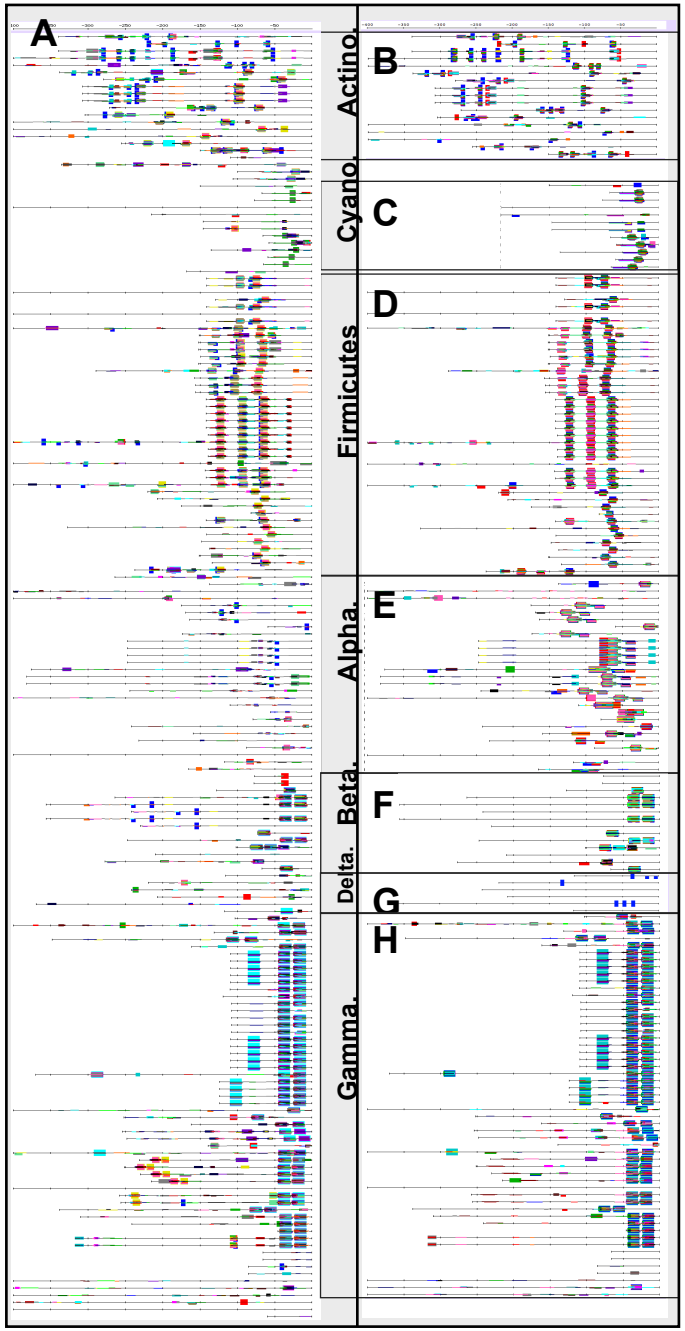
Significant dyads in promoters of lexA orthologs in Actinobacteria

- In Actinobacteria, the most significant motif is T**CGAACA**.
- 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, R. and van Helden, J. Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution *BMC Bioinformatics* **9**, 37 (2008).

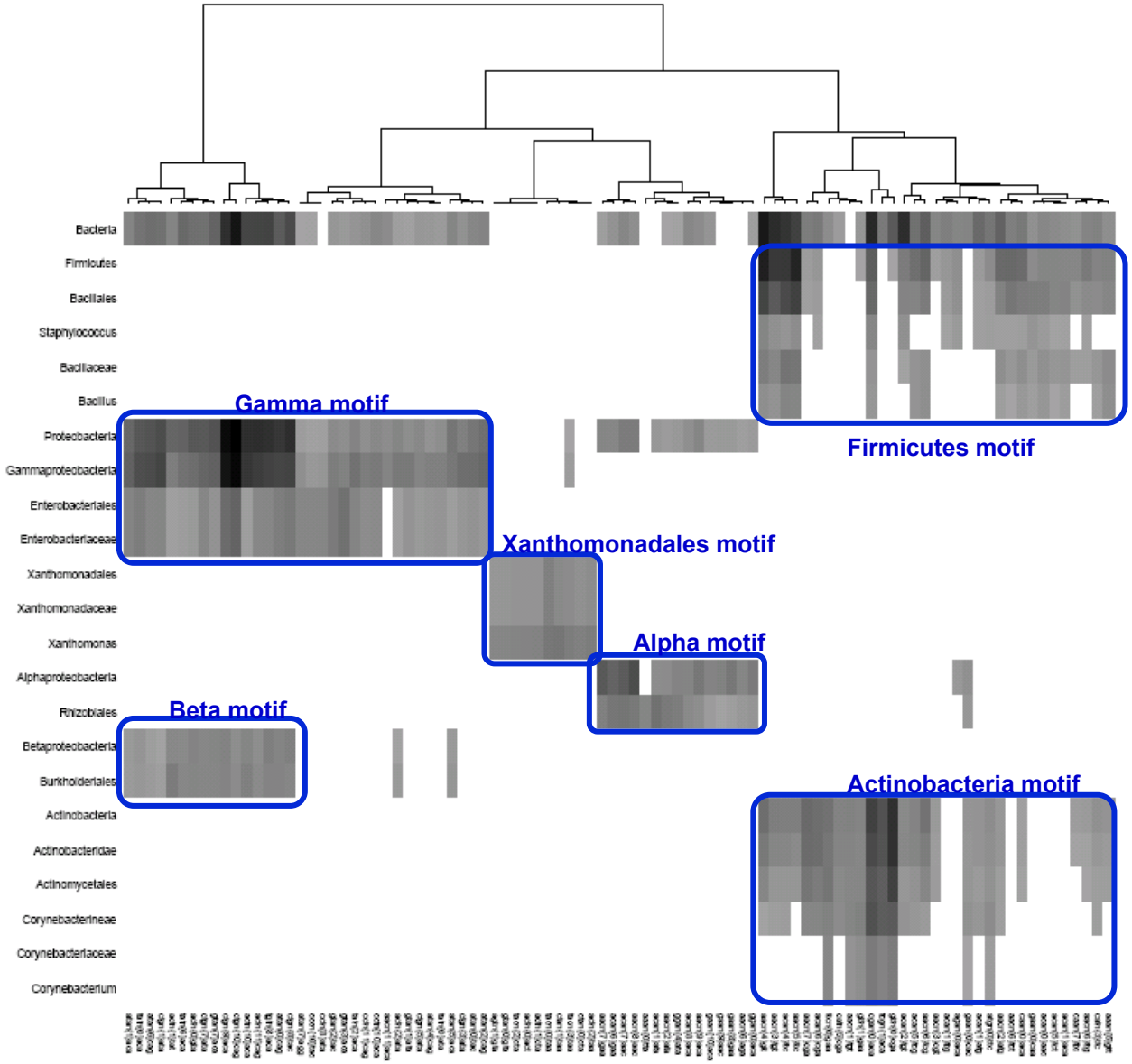
Significant dyads in promoters of lexA orthologs in Bacteria



- 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, R. and van Helden, J. Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution *BMC Bioinformatics* **9**, 37 (2008).

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.

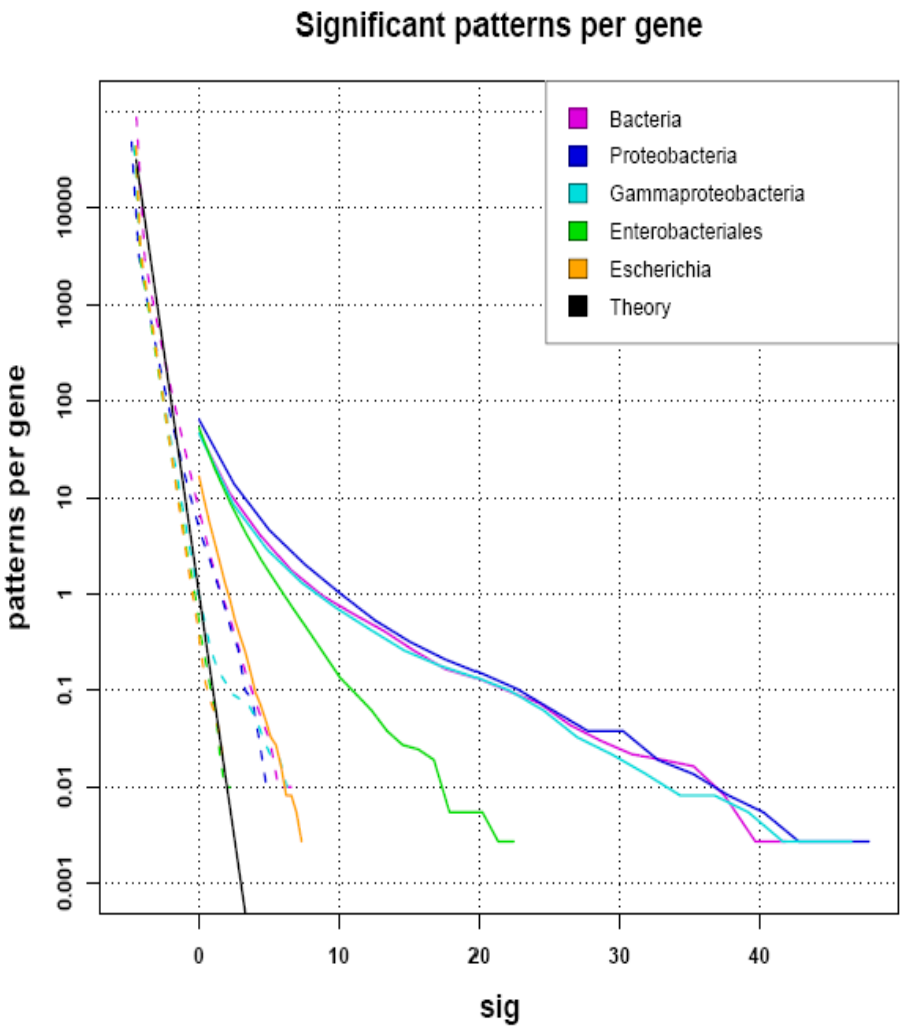
Janky, R. and van Helden, J. *BMC Bioinformatics* **9**, 37 (2008).

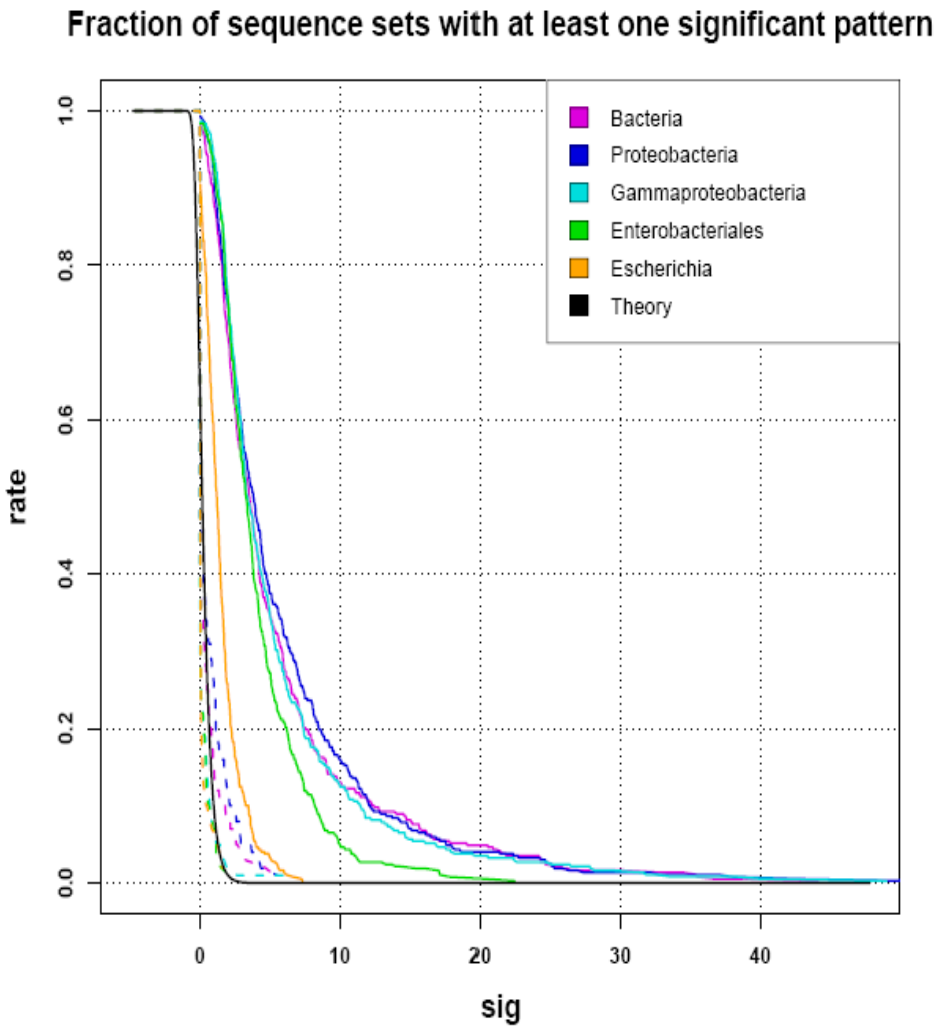
Systematic evaluation of phylogenetic footprints discovery

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 ?

Evaluation of significance scores:
distribution in promoters of orthologs versus random selections of genes





Janky, R. and van Helden, J. *BMC Bioinformatics* **9**, 37 (2008).

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.

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

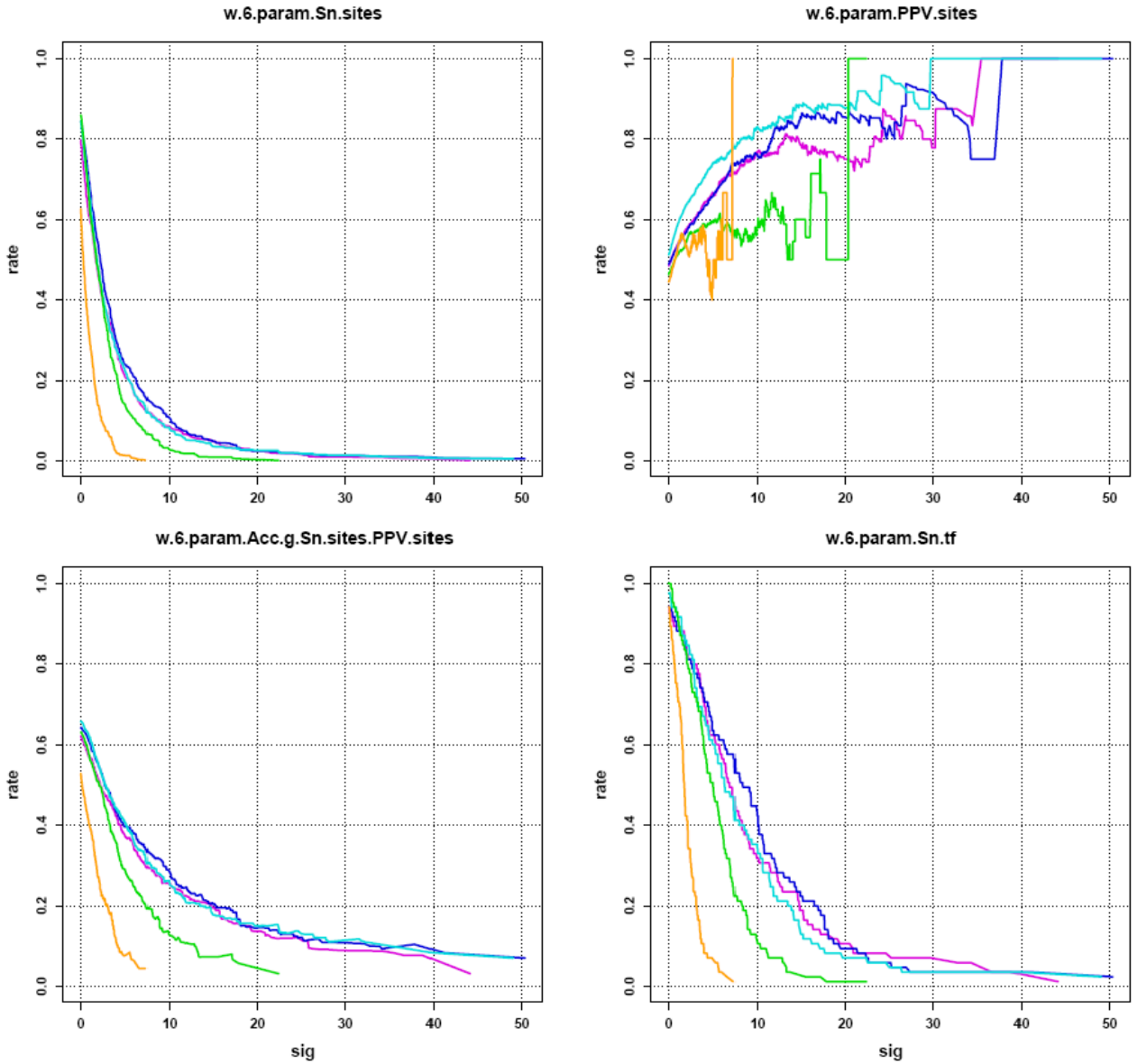
Site ID	matching bases	strand	offset	weight	seq1	seq2
CRP_glpD_ECK120014013	6	R	30	6	AAAACATTTTAGAGTGATATGTATAACATTATGGCGTTTATC	aacn{0}att
GlpR_glpD_ECK120012732	7	D	11	6	tctttgctaaTATGTTCGATAACGAACATTtatgagcttt	atgn{1}tcg
GlpR_glpD_ECK120012732	6	D	24	6	tctttgctaaTATGTTCGATAACGAACATTtatgagcttt	aacn{0}att
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

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

; sequence	gataaacgccATAATGTTATACATATCACTCTaaaatgtttt	tctttgctaaTATGTTCGATAACGAACATTtatgagcttt	taacgaacatTTATGAGCTTTAACGAAAGTgaatgagggc	gggatcaactGGTTTGCGTTTTGGCGCAAAttcagtgtta	aaaccggaaaTTAAGCGCGGATTCGAATATTctgactgtt
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	.	.

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



Supplementary material

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.	.cgannnnnnncat..	1.82
..atgnnngtt.....	.....aacnnncat..	1.63
..atgnncgt......	......acgnncat..	1.46
..atgnnnnnnnnnga	tcgnnnnnnnnncat..	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: agtnnnnnnnnnntcg	3 words
agtnnnnnnnnnntcg	**cgannnnnnnnnnact**	**1.52**
.gttnnnnnnnnntcg	cgannnnnnnnnaac.	0.9
..ttcnnnnnnnntcg	cgannnnnnnngaa..	0.78
agttcnnnnnnnntcg	cgannnnnnnngaact	1.52
;assembly # 4	seed: acannnnnnnnnnnnnacg	3 words
acannnnnnnnnnnnnacg.	**.cgtnnnnnnnnnnnntgt**	**1.26**
acannnnnnnnnnnnnncga	tcgnnnnnnnnnnnnntgt	0.48
acannnnnnnnnnnnaac..	..gttnnnnnnnnnnntgt	0.01
acannnnnnnnnnnnaacga	tcgttnnnnnnnnnnntgt	1.26