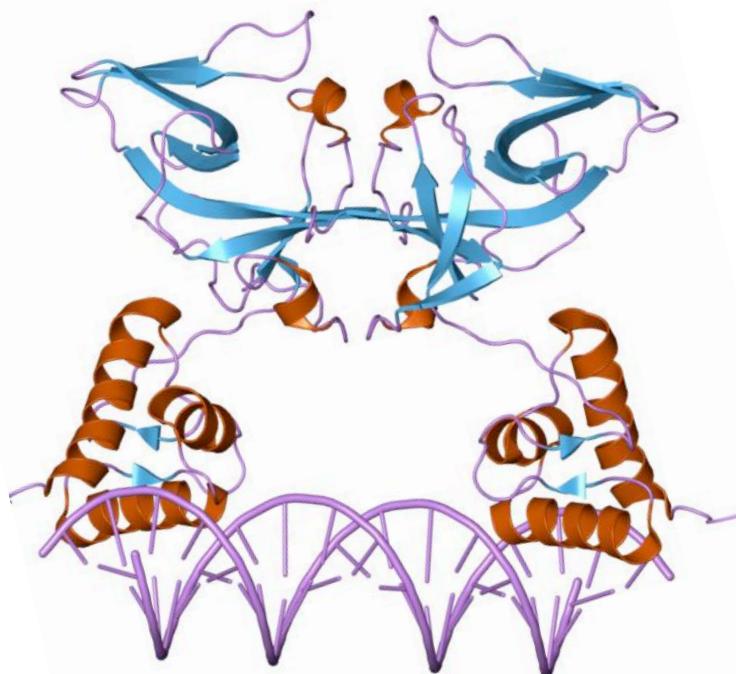


Regulatory Sequence Analysis

***Discovering phylogenetic footprints
in bacterial promoters***

DNA-protein binding interface



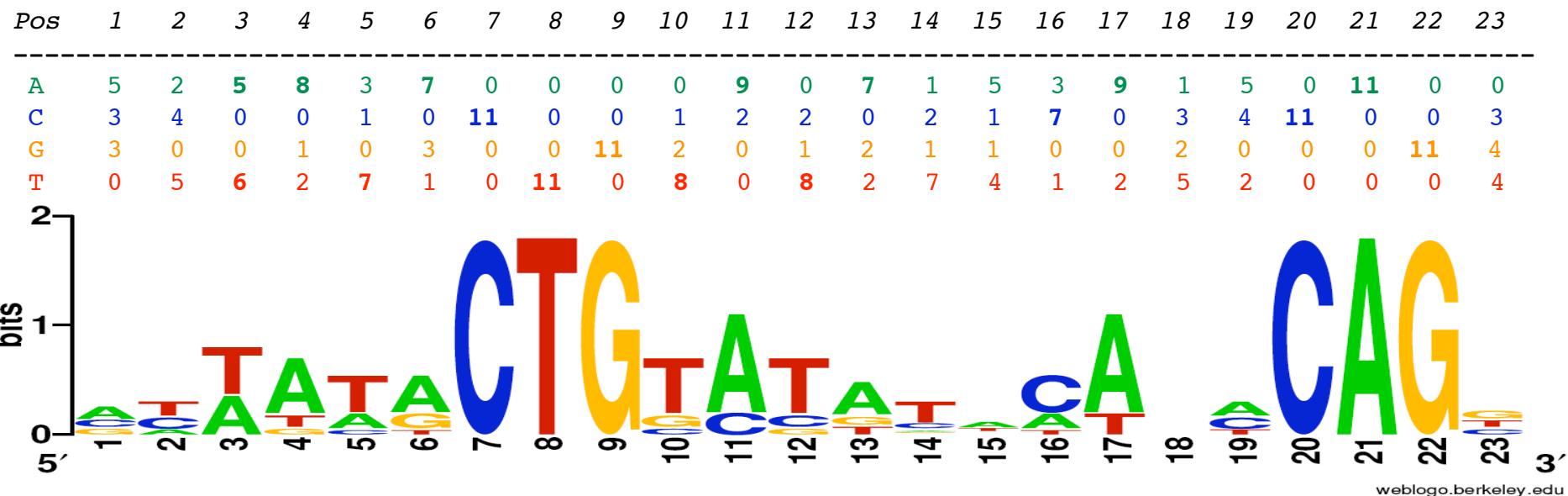
- Example of DNA-protein interface.
 - LexA homodimer on the *recA* promoter.
 - Note: this is only a model, the structure has still not been crystallized.
 - Source: <http://srs.ebi.ac.uk>

LexA binding sites (RegulonDB, oct 2006)

ECK120012770	recA	b2699	-21	gaagcaatta TACTGT TATGCTCATA CAGT Atcaagtgttt
ECK120012770	lexA	b4043	-10	aaatcgcctt TTGCTGT TATATACTCAC CAGC ataactgtat
ECK120012770	lexA	b4043	10	atactcacagCATA ACTGT TATATACACCC CAGggggcggaa
ECK120012770	recA	b2699	-21	gaagcaatta TACTGT TATGCTCATA CAGT Atcaagtgttt
ECK120012770	ssb	b4059	-46.5	gacacaaaatt GACCTGAATGAATATA CAGTattggaatgc
ECK120012770	sulA	b0958	-2	agccccctgtgAGTT ACTGT TATGGATGTAC CAGt acatccag
ECK120012770	uvrA	b4058	-31.5	gacacaaaatt GACCTGAATGAATATA CAGTattggaatgc
ECK120012770	uvrB	b0779	-21	ttatggtgat GAACTGT TTTTTTAT CCAGT ataatttgtt
ECK120012770	uvrD	b3813	11	taatcagcaa ATCTGT TATATAACC CAGCT ttttggcggaa
ECK120012770	rpsU	b3065	4.5	atttgaaaat AAGCTGGCGTTGATGCCAGC ggcaaaccga
ECK120012770	phr	b0708	25.5	ttatcctgac GCCTGGCTTCAGGGCAGCG tatttcgaa

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	CCAATACTGTATATTCAATTCAAGG
6	6	:	-6/15	GATGTACTGTACATCCATACAGT
7	7	:	-7/13	CCAATACTGTATATTCAATTCAAGG
8	8	:	-8/13	ATTATACTGGATAAAAAAACAGT
9	9	:	9/7	GCAAATCTGTATATATACCCAGC
10	10	:	10/8	AATAAGCTGGCGTTGATGCCAGC
11	11	:	-11/12	ATAACGCTGCCCTGAAAGCCAGG



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>).

Analysis of Regulatory Sequences

***Phylogenetic footprint discovery
in promoters of orthologous genes***

Orthologs of *Escherichia coli* K12 gene PPUTA in Gammaproteobacteria

- Orthologs of *Escherichia coli* K12 gene PPUTA in Gammaproteobacteria

#ref_gene	ref_org	query_gene	ident	e_value
NP_245526.1	<i>Pasteurella_multocida</i>	NP_415534.1	52.93	0.0
YP_340762.1	<i>Pseudoalteromonas_haloplanktis_TAC125</i>	NP_415534.1	45.21	0.0
YP_271059.1	<i>Colwellia_psychrerythraea_34H</i>	NP_415534.1	43.81	0.0
NP_719311.1	<i>Shewanella_oneidensis</i>	NP_415534.1	47.35	0.0
YP_156342.1	<i>Idiomarina_loihensis_L2TR</i>	NP_415534.1	45.44	0.0
YP_095723.1	<i>Legionella_pneumophila_Philadelphia_1</i>	NP_415534.1	51.25	0.0
YP_126994.1	<i>Legionella_pneumophila_Lens</i>	NP_415534.1	51.63	0.0
YP_123979.1	<i>Legionella_pneumophila_Paris</i>	NP_415534.1	51.25	0.0
NP_819659.1	<i>Coxiella_burnetii</i>	NP_415534.1	47.21	0.0
YP_133644.1	<i>Photobacterium_profundum_SS9</i>	NP_415534.1	46.63	0.0
YP_206789.1	<i>Vibrio_fischeri_ES114</i>	NP_415534.1	48.96	0.0
NP_937700.1	<i>Vibrio_vulnificus_YJ016</i>	NP_415534.1	47.40	0.0
NP_763030.1	<i>Vibrio_vulnificus_CMCP6</i>	NP_415534.1	47.60	0.0
NP_801236.1	<i>Vibrio_parahaemolyticus</i>	NP_415534.1	47.92	0.0
NP_639180.1	<i>Xanthomonas_campestris</i>	NP_415534.1	52.40	0.0
YP_202784.1	<i>Xanthomonas_oryzae_KACC10331</i>	NP_415534.1	52.69	0.0
YP_244967.1	<i>Xanthomonas_campestris_8004</i>	NP_415534.1	52.40	0.0
NP_644196.1	<i>Xanthomonas_citri</i>	NP_415534.1	52.59	0.0
YP_365739.1	<i>Xanthomonas_campestris Vesicatoria_85-10</i>	NP_415534.1	52.40	0.0
NP_805570.1	<i>Salmonella_typhi_Ty2</i>	NP_415534.1	91.29	0.0
NP_455618.1	<i>Salmonella_typhi</i>	NP_415534.1	91.29	0.0
YP_052304.1	<i>Erwinia_carotovora_atrosepticaSCRI1043</i>	NP_415534.1	74.24	0.0
YP_408453.1	<i>Shigella_boydii_Sb227</i>	NP_415534.1	99.09	0.0
YP_309997.1	<i>Shigella_sonnei_Ss046</i>	NP_415534.1	99.39	0.0
NP_753076.1	<i>Escherichia_coli_CFT073</i>	NP_415534.1	99.39	0.0
NP_415534.1	<i>Escherichia_coli_K12</i>	NP_415534.1	100.00	0.0
NP_309287.1	<i>Escherichia_coli_O157H7</i>	NP_415534.1	99.55	0.0
NP_287019.1	<i>Escherichia_coli_O157H7_EDL933</i>	NP_415534.1	99.55	0.0
NP_871437.1	<i>Wigglesworthia_brevipalpis</i>	NP_415534.1	54.57	0.0
NP_992900.1	<i>Yersinia pestis_biovar_Mediaevails</i>	NP_415534.1	79.44	0.0
NP_669761.1	<i>Yersinia pestis_KIM</i>	NP_415534.1	79.37	0.0
YP_070249.1	<i>Yersinia_pseudotuberculosis_IP32953</i>	NP_415534.1	79.37	0.0
NP_405415.1	<i>Yersinia pestis_CO92</i>	NP_415534.1	79.37	0.0
NP_929224.1	<i>Photorhabdus_luminescens</i>	NP_415534.1	77.37	0.0
YP_343707.1	<i>Nitrosococcus_oceani_ATCC_19707</i>	NP_415534.1	50.00	0.0
YP_170117.1	<i>Francisella_tularensis_tularensis</i>	NP_415534.1	50.34	0.0
YP_433088.1	<i>Hahella_chejuensis_KCTC_2396</i>	NP_415534.1	48.61	0.0
NP_794750.1	<i>Pseudomonas_syringae</i>	NP_415534.1	73.64	0.0
NP_747050.1	<i>Pseudomonas_putida_KT2440</i>	NP_415534.1	73.45	0.0
YP_257639.1	<i>Pseudomonas_fluorescens_Pf-5</i>	NP_415534.1	73.60	0.0
YP_346185.1	<i>Pseudomonas_fluorescens_PfO-1</i>	NP_415534.1	73.45	0.0
NP_249473.1	<i>Pseudomonas_aeruginosa</i>	NP_415534.1	48.56	0.0
YP_272798.1	<i>Pseudomonas_syringae_phaseolicola_1448A</i>	NP_415534.1	73.77	0.0
YP_233614.1	<i>Pseudomonas_syringae_pv_B728a</i>	NP_415534.1	73.80	0.0
YP_264533.1	<i>Psychrobacter_arcticum_273-4</i>	NP_415534.1	44.42	0.0
YP_046314.1	<i>Acinetobacter_sp_ADP1</i>	NP_415534.1	59.51	0.0

Upstream sequences of PUTA orthologs

```
>YP_046314.1|Acinetobacter_sp_ADPI|putA          YP_046314.1; upstream from -103 to -1; size: 103;  
ACAAAATTTCTCTAAAAAATGAATCAATTATAGTCAGTTATTGGTAATTATTCTGTA  
ATGATAAAATTATCTAACCCCTTAAACAATATACCTTAGAGT  
>YP_271059.1|Colwellia_psychrerythraea_34H|putA YP_271059.1; upstream from -245 to -1; size: 245  
ATAATAACCCACGAACACTCCCTACAAATTATAAAAACGATTGCAGCACTTATACTG  
TTGAATTCTGTACTCCCCATATAAAAGTGTAACTCCTGAAAATAACCAGCACATCCT  
GTGGTTGTTACCTAAATCGCTCATAAATTAAATGTCGTACCAACTAATAATATG  
TATTAGTGGAAAAAAAGACTATAACTAAAGCAGGATTCTACCTGTCACACTTGAGGAA  
TGGTT  
>NP_819659.1|Coxiella_burnetii|putA NP_819659.1; upstream from -118 to -1; size: 118;  
GAAGTAGCCCGTATGAAGCGAAGCGAAATACGGGGAGGTGCACGTATTGTTCCCGTATTG  
GCTTCGTTTCATACGGGCTACATCGCGGAAATGAAAATTAACTCCTTAATGAGGACAT  
>YP_052304.1|Erwinia_carotovora_atrosepticaSCRI1043|putA YP_052304.1; upstream from -195 to -1; size: 195;  
TTAACTCTCCACATTTTCTGTGCCCGTGCAGCGACGCTGTGTTTATAGTAATCA  
TTCAGGCCGGAACGAGGTCTGAAAATGATTATGGGCAGCAACCATTCCATTGTTAACAA  
GGTTGCACAAAGTTGCAACATGATTGATATTGACGGTATCCGATGTGCATCTTCATT  
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;  
ATCGGAATGTCGAAACTGCCGTTATATCTGCCACCGGAACGGGTAACAGAGTTTATG  
TTTACCGGGCGACCGTATCCTGCCGGAAGCGCTGGTTATTCAAATCGATTAAACACA  
CCATTACATTAATTTAGTGCAGCGACACTATTTTATCAGGTTGCCTCTCA  
CATTTTTGCGGTTGCACCTTCAAAATGTTAACTGCCGAGAGAAAAAGTCTGAGTTA  
TTTTTTAATCCCTGTATCGATTCTTTATTAACATTCAATTAAAGCTTG  
CTACGCATGTCACATTAAACATGGTTGCACAAAGTTGCAACATCATGGATATTCA  
AACGTTAAGTTGCACCTTCTGAACACAGGAGTAATGGC  
>NP_309287.1|Escherichia_coli_O157H7|ECs1260   NP_309287.1; upstream from -54 to -1; size: 54;  
GGATATTTCACGATAACGTTAAGTTGCACCTTCAGAACACAGGAGTAATGGC  
>NP_287019.1|Escherichia_coli_O157H7_EDL933|putA      NP_287019.1; upstream from -400 to -1; size: 400;  
ATCGGAATGTCGAAACTGCCGTTATATCTGCCACCGGAACGGGTAACAGAGTTTATG  
TTTACCGGGCGACCGTATCCTGCCGGAAGCGCTGGTTATTCAAATCGATTAAACACA  
CCATTACATTAATTTAGTGCAGCGACACTATTTTATCAGGTTGCCTCTCA  
CATTTTTGCGGTTGCACCTTCAAAATGTTAACTGCCGAGAGAAAAAGTCTGAGTTA  
TTTTTTAATCCCTGTATCGATTCTTTATTAACATTCAATTAAAGCTTG  
CTACGCAGGTCACATTAAACATGGTTGCACAAAGTTGCAACATCATGGATATTCA  
AACGTTAAGTTGCACCTTCAGAACACAGGAGTAATGGC  
.....
```

Purged upstream sequences of PUTA orthologs

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_freq	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}cct	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

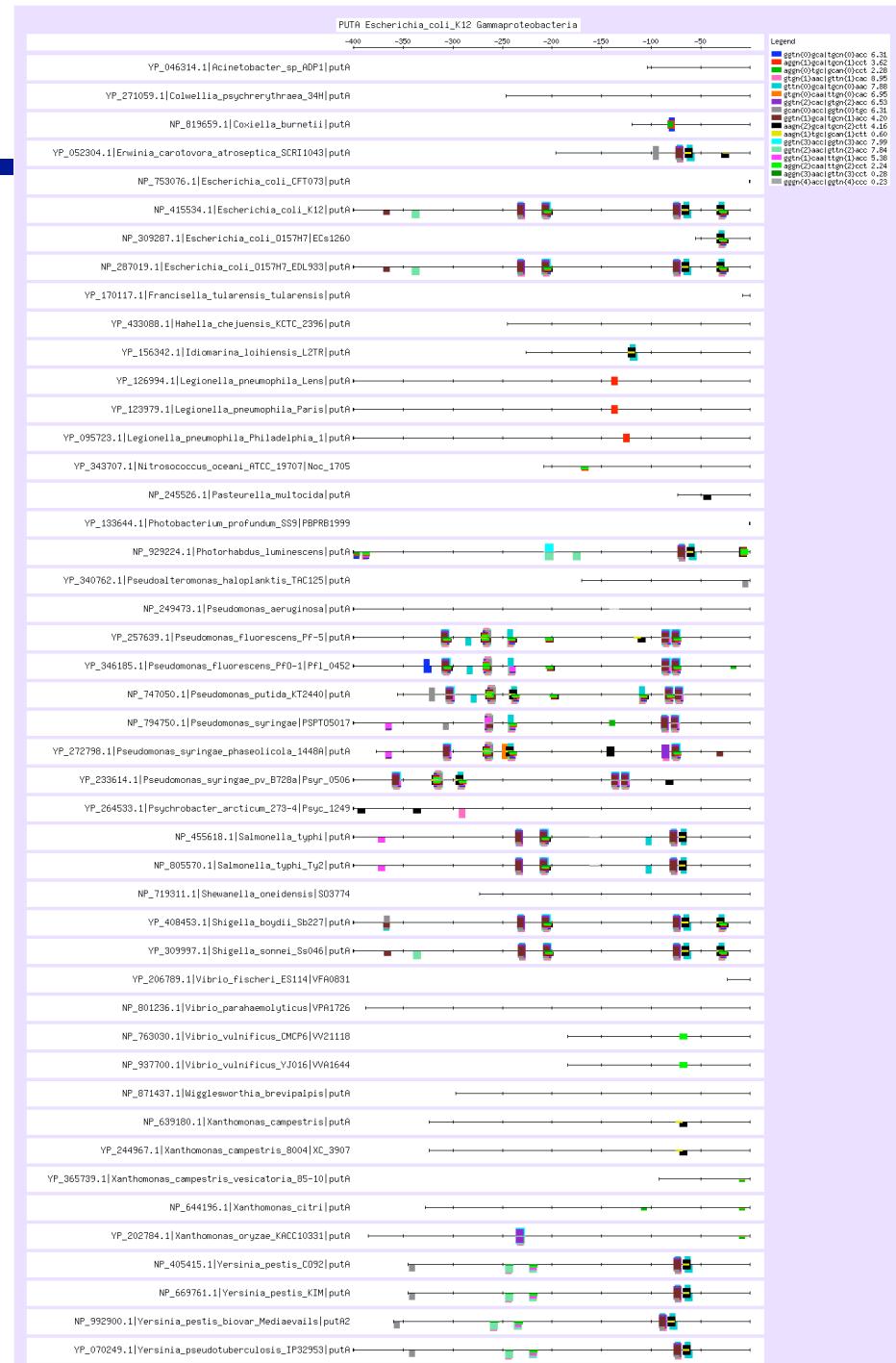
;Job done 13/02/06 21:31:28 CET

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 pattern    17
;
```

	seed: gtgnaa 17 words	length
;	rev_cpl	score
aagnngca....tgcnctt	4.16
aagntgc.....gcanctt	0.6
.aggnngca....tgcnccct.	3.62
.aggtgc.....gcacct.	2.28
.aggnncaa...	...ttgnncct.	2.24
.aggnnnnaac..	..gttnnnccct.	0.28
..ggtnnnnacc.	.ggtnnnnacc..	7.99
..ggtnnaac..	..gttnnacc..	7.84
..ggtgca....tgcac..	6.31
..ggtncaa...	...ttgnacc..	5.38
..ggtnnnnnccc	gggnnnnnacc..	0.23
...gtgnaac..	..gttncac...	8.95
...gtgcaa...	...ttgcac...	6.95
...gtgnnacc.	.ggtnncac...	6.53
....tgcaac..	..gttgca....	7.88
....tgcnacc.	.ggtnrgca....	4.2
.....gcaacc.	.ggttgc.....	6.31
aagggtgcaaccc	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
 1. How good is this method in predicting cis-acting elements ?
 - a) Can we detect correct motifs ?
 - b) What is the rate of false positives ?
 2. Can we learn something about the evolution of cis-acting elements ?
 3. On the basis of the discovered motifs, can we regroup the co-regulated genes ?
 - a) Detect pair-wise associations between genes
 - b) Detect clusters of genes regulated by the same transcription factor

Analysis of Regulatory Sequences

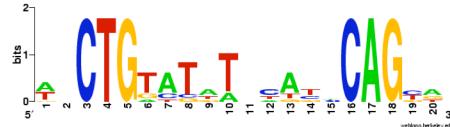
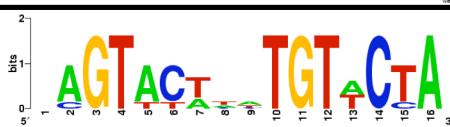
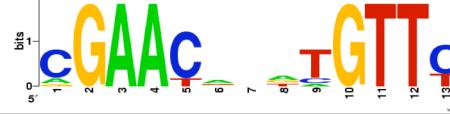
***Detailed analysis of a study case:
auto-regulation of
the LexA transcription factor
(analysis performed by Rekin's Janky, PhD Student)***

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

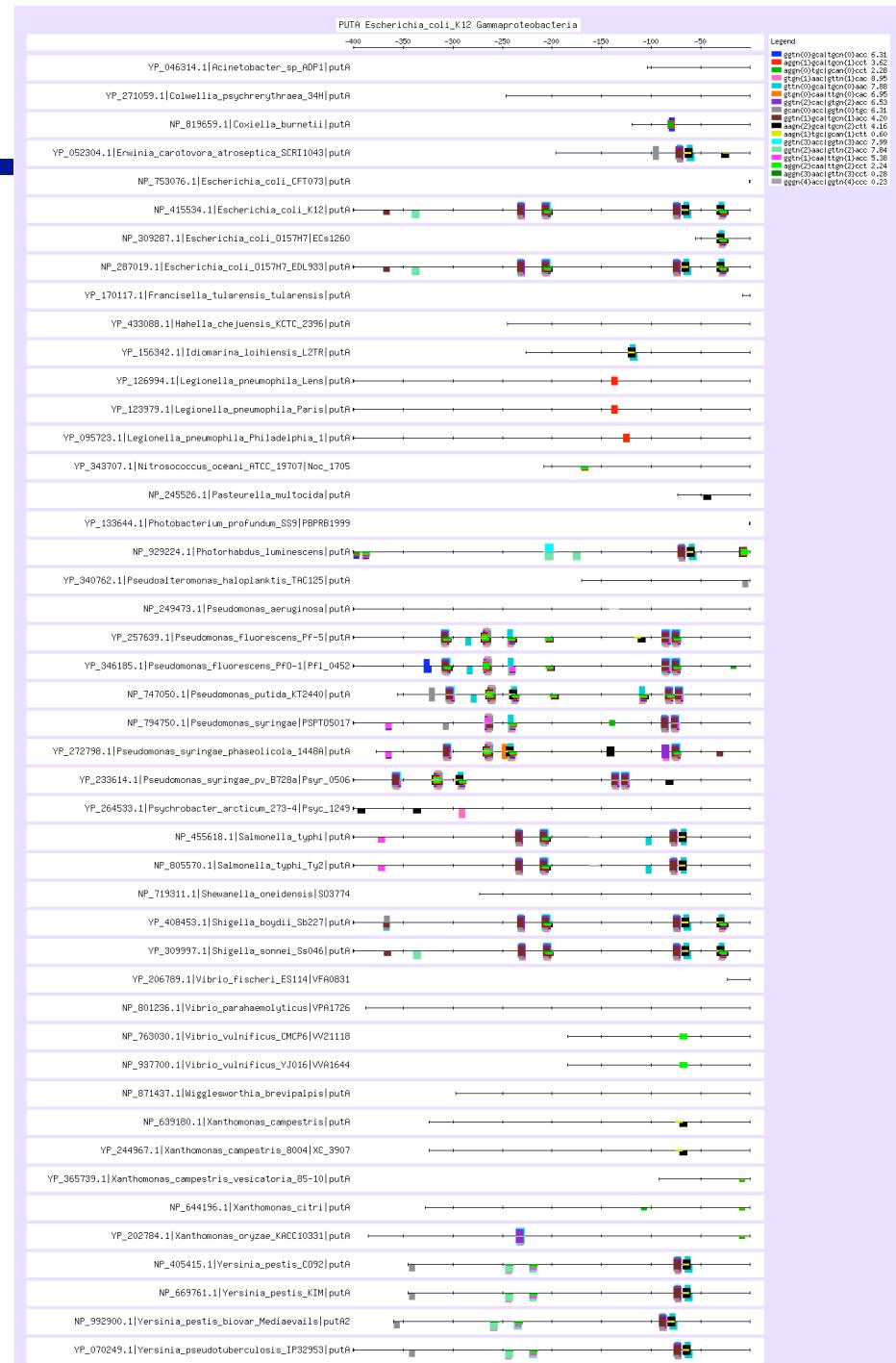
Taxonomic level	Species	Consensus	Weblogo
Gamma-proteobacteria	<i>E. coli</i>	CTGTNNNNNNNNACAG	
Xanthomonadales	<i>X. oryzae</i> <i>X. campestris</i> <i>X. citri</i> <i>Xylella fastidiosa</i>	TTAGTArwawTACTAa	
Alpha proteobacteria	<i>R. etli</i> <i>R. meliloti</i> <i>A. tumefasciens</i> <i>Rhodobacter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

Significantly over-represented dyads in promoters of PUTA orthologs

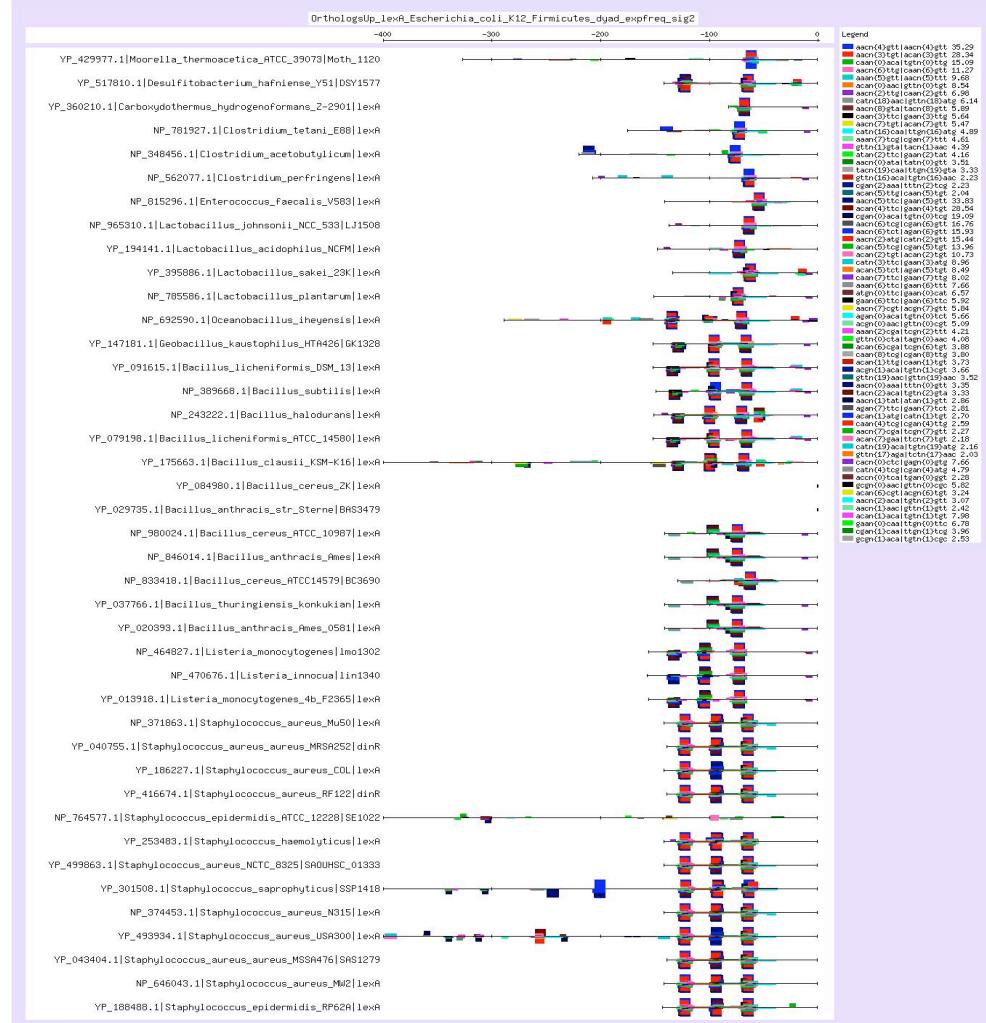
```
; pattern-assembly -v 1 -i /home/jvanheld/research/collabo
; Input score column          8
; Output score column         0
; two strand assembly
; max flanking bases        1
; max substitutions          0
; max assembly size          50
; max number of pattern      50
; number of input pattern 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

	seed: gtgnaa 17 words	length
;	rev_cpl	score
aagnngca....tgcnnc tt	4.16
aagntgc.....gcanctt	0.6
.aggnngca....tgcncc t.	3.62
.aggtgc.....gcacct.	2.28
.aggnncaa...	...ttgnncct.	2.24
.aggnnnac..	..gttnnncc t.	0.28
..ggtnnnacc.	.ggtnnnacc..	7.99
..ggtnnaac..	..gttnnacc..	7.84
..ggtgca....tgcac ..	6.31
..ggtncaa...	...ttgnacc..	5.38
..ggtnnnnncc	gggnnnnnacc..	0.23
...gtgnaac..	..gttncac...	8.95
...gtgcaa...	...ttgcac...	6.95
...gtgnnacc.	.ggtnncac...	6.53
....tgcaac..	..gttgca....	7.88
....tgcnacc.	.ggtn gca....	4.2
.....gcaacc.	.ggttgc.....	6.31
aagg tgca accc	gggttgcac ctt	8.95 best !

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

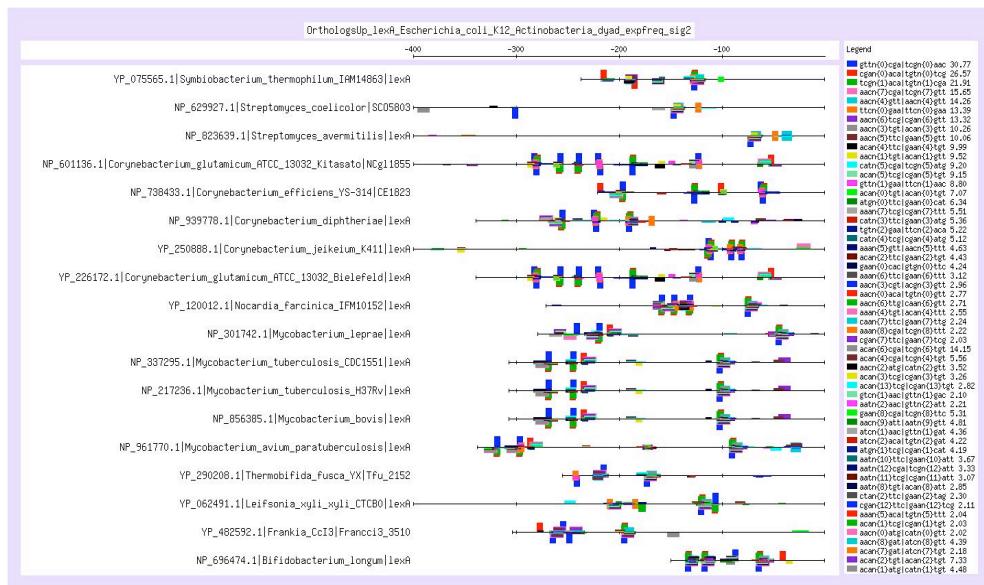


Significant dyads in promoters of *lexA* orthologs in Firmicutes



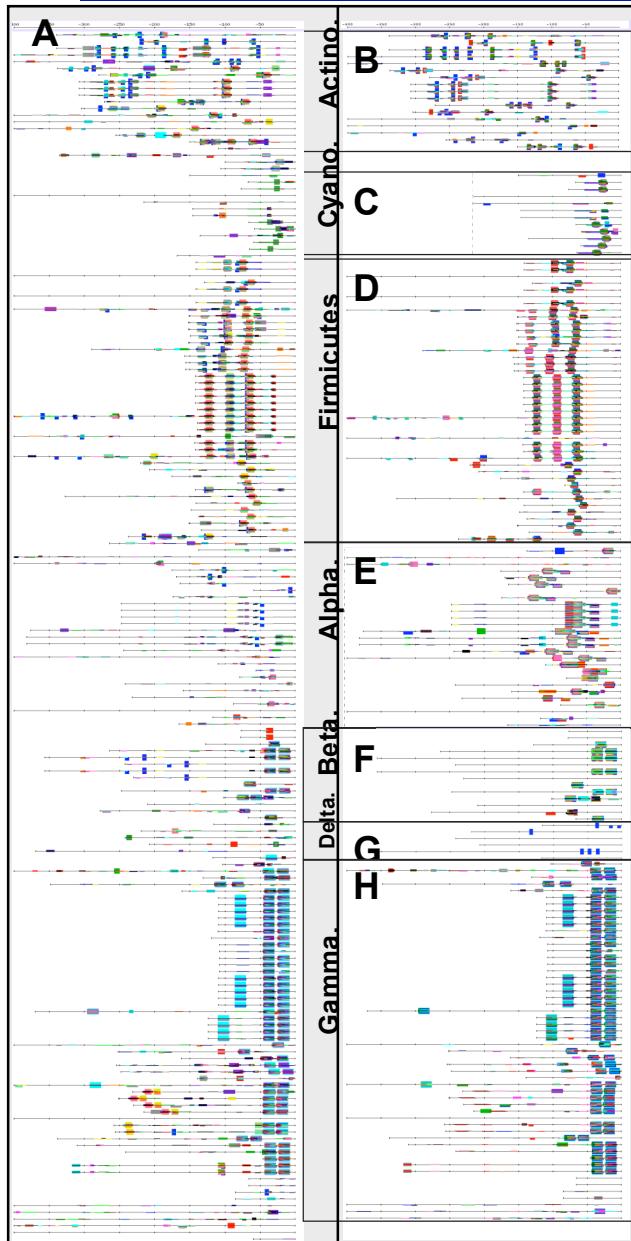
- In Firmicutes, the most significant motif is **TACGAACATATGTTCGTA**
- This corresponds to the “Cheo box” **GAAC N_4 GTTC**.

*Significant dyads in promoters of *lexA* orthologs in Actinobacteria*



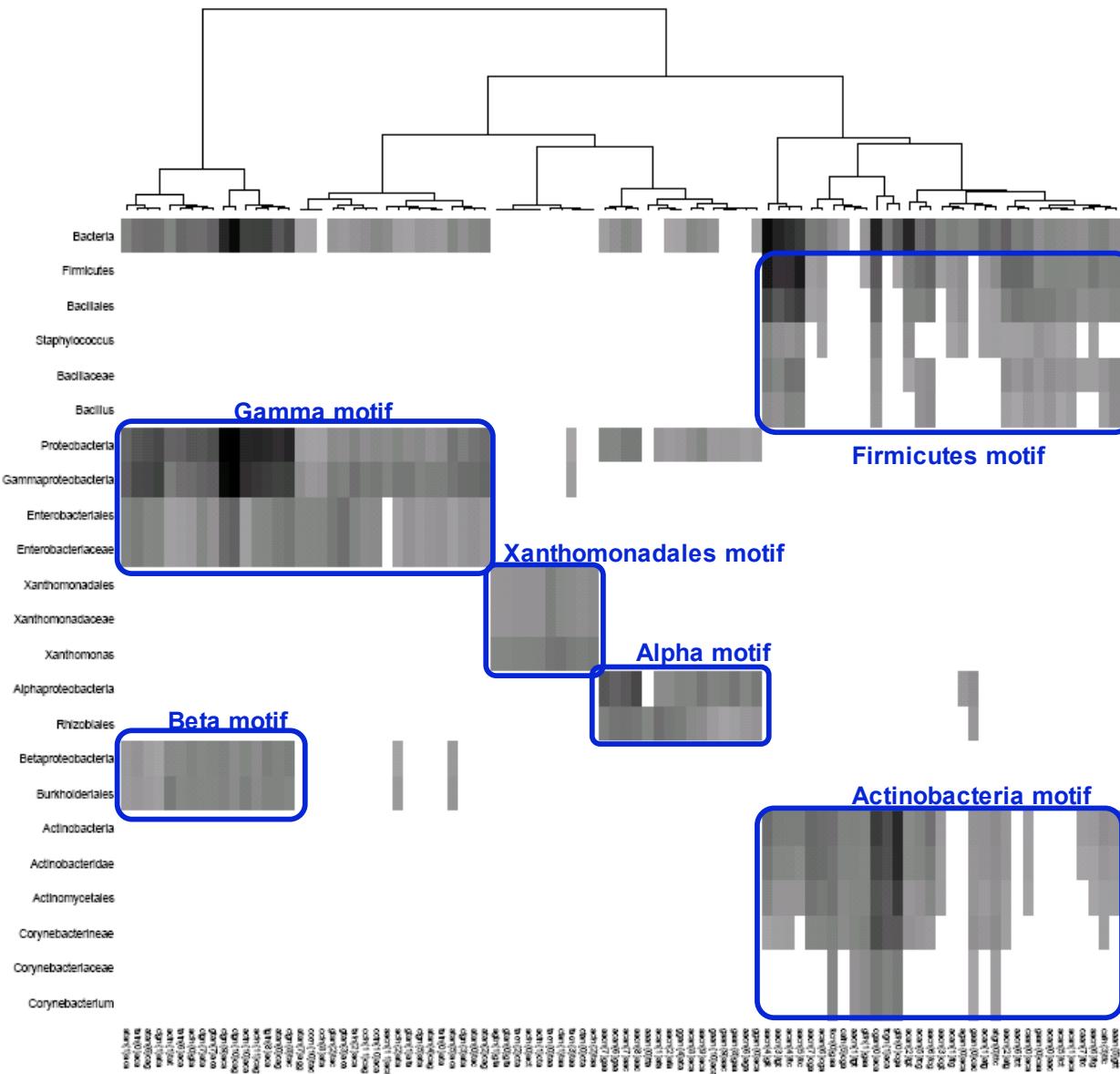
- In Actinobacteria, the most significant motif is TCGAACCA.
 - This shows an almost perfect match to one half of the Cheo box motif detected in Firmicutes
(TACGAACATATGTTCGTA)
 - In addition, this larger Cheo box is also detected in Actinobacteria, albeit with a lower significance than the half site.

*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.

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



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

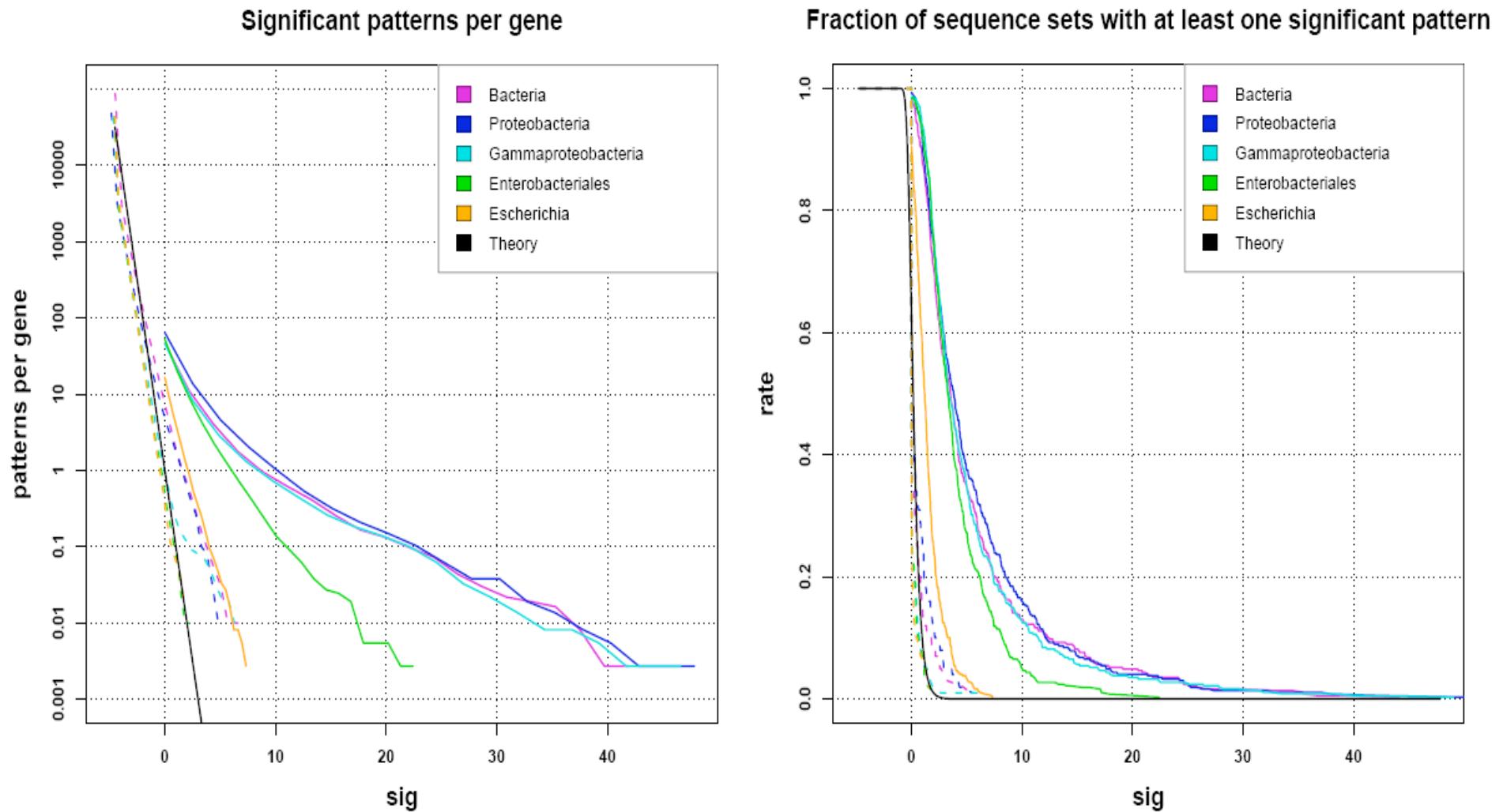
Analysis of Regulatory Sequences

***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



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
gataaaacgccATAATGTTATACATATCACTCTaaaatgttt	CRP	ECK120014013
tcttgctaaTATGTTCGATAACGAACATTtatgagctt	GlpR	ECK120012732
taacgaacatTTATGAGCTTTAACGAAAGTgaatgagggc	GlpR	ECK120012734
gggatcaactGGTTTGCCTTGGCGCAAAttcagtgtta	GlpR	ECK120013968
aaaccggaaaTTAAGCGCGATTCAATATTctgactgtt	GlpR	ECK120013970

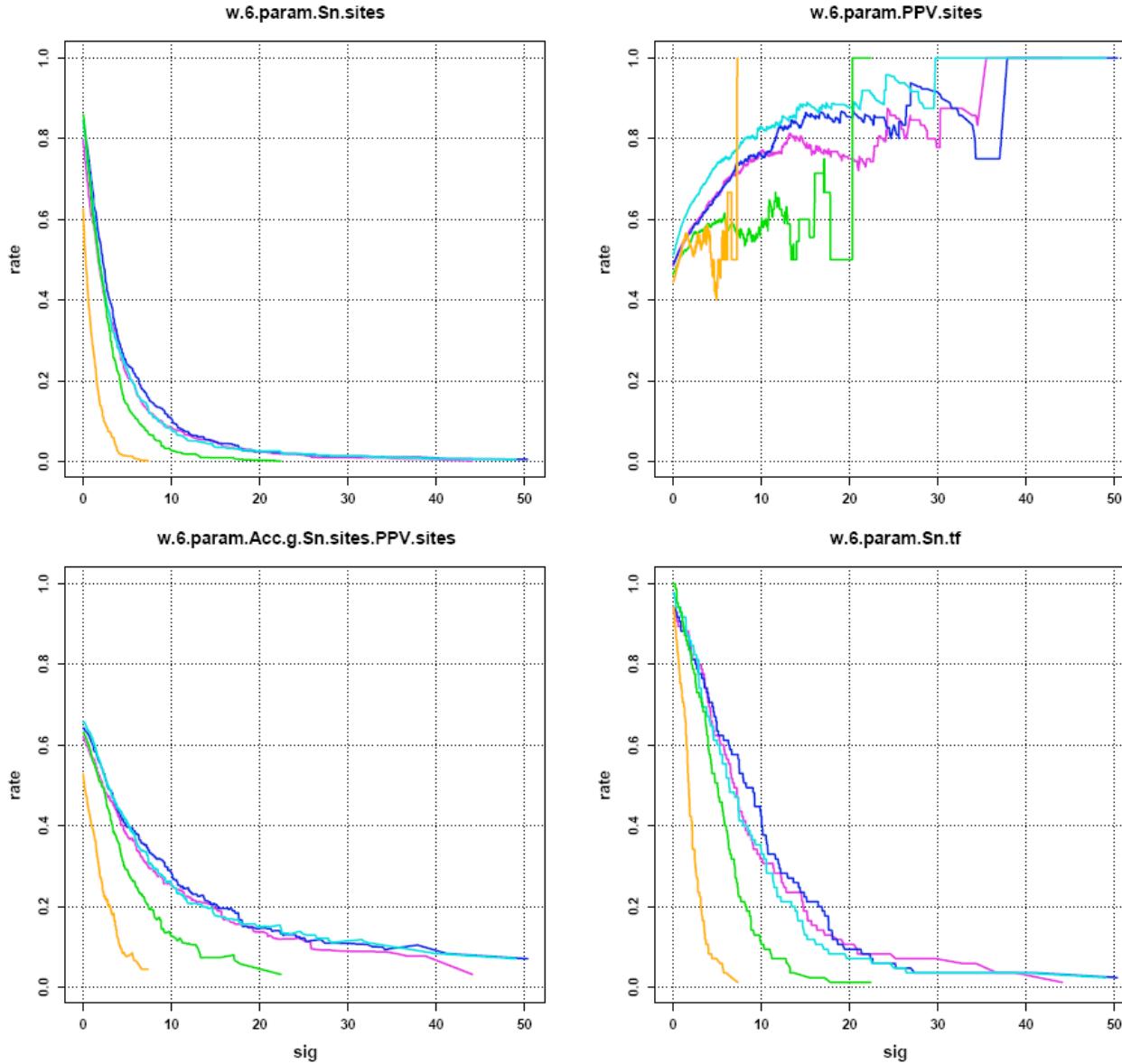
Perfect matches between detected dyads and annotated sites

Site ID	matching					seq2
	bases	strand	offset	weight	seq1	
CRP_glpD_ECK120014013	6 R		30	6	AAAACATTTAGAGTGATATGTATAACATTATGGCGTTATC	aacn{0}att
GlpR_glpD_ECK120012732	7 D		11	6	tcttgctaaTATGTTCGATAACGAACATTtatgagctt	atgn{1}tcg
GlpR_glpD_ECK120012732	6 D		24	6	tcttgctaaTATGTTCGATAACGAACATTtatgagctt	aacn{0}att
GlpR_glpD_ECK120012732	6 D		22	6	tcttgctaaTATGTTCGATAACGAACATTtatgagctt	cgan{0}aca
GlpR_glpD_ECK120012732	8 D		20	6	tcttgctaaTATGTTCGATAACGAACATTtatgagctt	aacn{2}aca
GlpR_glpD_ECK120012732	8 R		20	6	AAAGCTCATAAATGTCGTTATCGAACATATTAGCAAAGA	aatn{2}tcg
GlpR_glpD_ECK120012732	6 D		11	6	tcttgctaaTATGTTCGATAACGAACATTtatgagctt	atgn{0}ttc
GlpR_glpD_ECK120012732	8 D		21	6	tcttgctaaTATGTTCGATAACGAACATTtatgagctt	acgn{2}cat
GlpR_glpD_ECK120012732	9 D		20	6	tcttgctaaTATGTTCGATAACGAACATTtatgagctt	aacn{3}cat
GlpR_glpD_ECK120012732	6 D		20	6	tcttgctaaTATGTTCGATAACGAACATTtatgagctt	aacn{0}gaa
GlpR_glpD_ECK120012734	7 R		0	6	GCCCTATTCACTTTCGTTAAAGCTCATAAATGTTCGTTA	atgn{1}tcg
GlpR_glpD_ECK120012734	6 D		5	6	taacgaacatTTATGAGCTTAACGAAAGTgaatgagggc	aacn{0}att
GlpR_glpD_ECK120012734	6 D		3	6	taacgaacatTTATGAGCTTAACGAAAGTgaatgagggc	cgan{0}aca
GlpR_glpD_ECK120012734	19 D		6	6	taacgaacatTTATGAGCTTAACGAAAGTgaatgagggc	acan{13}acg
GlpR_glpD_ECK120012734	8 D		1	6	taacgaacatTTATGAGCTTAACGAAAGTgaatgagggc	aacn{2}aca
GlpR_glpD_ECK120012734	8 R		1	6	GCCCTATTCACTTTCGTTAAAGCTCATAAATGTTCGTTA	aatn{2}tcg
GlpR_glpD_ECK120012734	6 R		0	6	GCCCTATTCACTTTCGTTAAAGCTCATAAATGTTCGTTA	atgn{0}ttc
GlpR_glpD_ECK120012734	8 D		2	6	taacgaacatTTATGAGCTTAACGAAAGTgaatgagggc	acgn{2}cat
GlpR_glpD_ECK120012734	9 D		1	6	taacgaacatTTATGAGCTTAACGAAAGTgaatgagggc	aacn{3}cat
GlpR_glpD_ECK120012734	10 R		28	6	GCCCTATTCACTTTCGTTAAAGCTCATAAATGTTCGTTA	cctn{4}cac
GlpR_glpD_ECK120012734	6 D		1	6	taacgaacatTTATGAGCTTAACGAAAGTgaatgagggc	aacn{0}gaa
GlpR_glpD_ECK120012734	13 R		20	6	GCCCTATTCACTTTCGTTAAAGCTCATAAATGTTCGTTA	actn{7}aag
GlpR_glpD_ECK120013970	8 R		21	6	AACAGTCAGAATATTCGAATCCGGCTTAATTCCGGTTT	aatn{2}tcg
GlpR_glpD_ECK120013970	14 D		1	6	aaaccggaaaTTAAGCGCGATTCAATATTctgactgtt	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)
 - fraction of annotated sites matched by at least one discovered dyad
 - $Sn = TP/(TP+FN)$
 - Positive Predictive Value (PPV)
 - Fraction of discovered dyads matching at least one annotated site
 - $PPV = TP/(TP+FP)$
 - Accuracy (geometric)
 - Geometric mean of Sn and PPV
 - $Acc = \sqrt{Sn \cdot PPV}$

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....aacnntt	0.82
.aatgtt....aacatt.	3.38
.aatnntcg..	..cgannatt.	2.1
..atgntcg..	..cgancat..	7.55
..atgttc...	...gaacat..	3.29
..atgnnnngt	aacnnncat..	1.63
..atgnncgt.	.acgnncat..	1.46
...tgttcg..	..cgaaca...	4.58
...tgtnnngt	aacnnaca...	2.4
...tgtncgt.	.acgnaca...	1.32
....gttnngt	aacnaac....	1.05
....gttcgt.	.acgaac....	0.47
.....ttcgtt	aacgaa.....	1.53
aaatgttcgtt	aacgaacattt	7.55
;assembly # 2	seed: atgnnnnnnnntcg	19 words
aaangtt.....aacnntt	0.82
.aatgtt.....aacatt.	3.38
.aatnntcg.....cgannatt.	2.1
..atgntcg.....cgancat..	7.55
..atgttc.....gaacat..	3.29
..atgnnnnnnnntcg.	.cgannnnnnncat..	1.82
..atgnnnngt....aacnnncat..	1.63
..atgnncgt....acgnncat..	1.46
..atgnnnnnnnnncga	tcgnnnnnnnnncat..	0.93
...tgttcg.....cgaaca...	4.58
...tgtnnngt....aacnnaca...	2.4
...tgtncgt....acgnaca...	1.32
...tgtnnnnnnntcg.	.cgannnnnnnaca...	0.74
....gttnngt....aacnaac....	1.05
....gttcgt....acgaac....	0.47
....gttnnnnnntcg.	.cgannnnnaac....	0.35
.....ttcgtt....aacgaa.....	1.53
.....tcgnnntcg.	.cgannncga.....	0.6
.....tcgnnnnncga	tcgnnnnncga.....	0.06
aaatgttcgttntcg	tcganaacgaacattt	7.55
;assembly # 3	seed: agtnnnnnnnnnntcg	3 words
atgnnnnnnnnnntcg	cgannnnnnnnnnact	1.52
.gtttnnnnnnnntcg	cgannnnnnnnnnnaac.	0.9
..ttcnnnnnnnntcg	cgannnnnnnnnngaa..	0.78
agttcnnnnnnnntcg	cgannnnnnnnnngaact	1.52
;assembly # 4	seed: acannnnnnnnnnnnnacg	3 words
acannnnnnnnnnnnnacg.	.cgtnnnnnnnnnnnntgt	1.26
acannnnnnnnnnnnncga	tcgtnnnnnnnnnnnntgt	0.48
acannnnnnnnnnnnaac..	..gttnnnnnnnnnntgt	0.01
acannnnnnnnnnnnaacga	tcgttnnnnnnnnnntgt	1.26