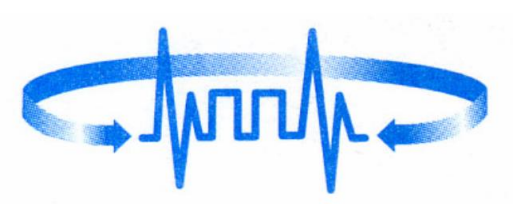


Evolution of transcriptional regulation in bacteria

Mikhail Gelfand

Institute of Cytology and Genetics, SB RAS

Novosibirsk, 2 XII 2015

1. Methods

- Position Weight Matrices (PWMs), or profiles
  - Usually not very specific
- Many close genomes (alignment of upstream regions possible)
  => phylogenetic footprinting (conservation of homologous binding sites)
- More distant genomes
  => consistency check (conservation of regulon content, presence of binding sites upstream of orthologous genes)

Consensus

```
codB        CCCACGAAAACGATTGCTTTTT
purE        GCCACGCAACCGTTTTCCTTGC
pyrD        GTTCGGAAAACGTTTGCGTTTT
purT        CACACGCAAACGTTTTCGTTTA
cvpA        CCTACGCAAACGTTTTCTTTTT
purC        GATACGCAAACGTGTGCGTCTG
purM        GTCTCGCAAACGTTTGCTTTCC
purH        GTTGCGCAAACGTTTTCGTTAC
purL        TCTACGCAAACGGTTTCGTCGG
consensus     ACGCAAACGTTTTCGT
```

Positional weight matrix

j	a	C	G	m	**A**	**A**	A	**C**	**G**	t	T	**T**	k	**C**	k	**T**
A	6	0	0	2	9	9	8	0	0	1	0	0	0	0	0	0
C	1	8	0	7	0	0	1	9	0	0	0	0	0	9	1	0
G	1	1	9	0	0	0	0	0	9	1	1	0	5	0	5	0
T	1	0	0	0	0	0	0	0	0	7	8	9	4	0	3	9

A	1.1	–1.0	–0.7	0.5	2.2	2.2	1.9	–0.7	–0.7	–0.1	–1.0	–0.7	–1.1	–0.7	–1.4	–0.7
C	–0.4	1.9	–0.7	1.6	–0.7	–0.7	0.1	2.2	–0.7	–1.2	–1.0	–0.7	–1.1	2.2	–0.3	–0.7
G	–0.4	0.1	2.2	–1.1	–0.7	–0.7	–1.0	–0.7	2.2	–0.1	–0.1	–0.7	1.2	–0.7	1.0	–0.7
T	–0.4	–1.0	–0.7	–1.1	–0.7	–0.7	–1.0	–0.7	–0.7	1.5	1.9	2.2	1.0	–0.7	0.6	2.2

$$W(b,j)=\ln(N(b,j)+0.5) - 0.25\sum_i \ln(N(i,j)+0.5)$$

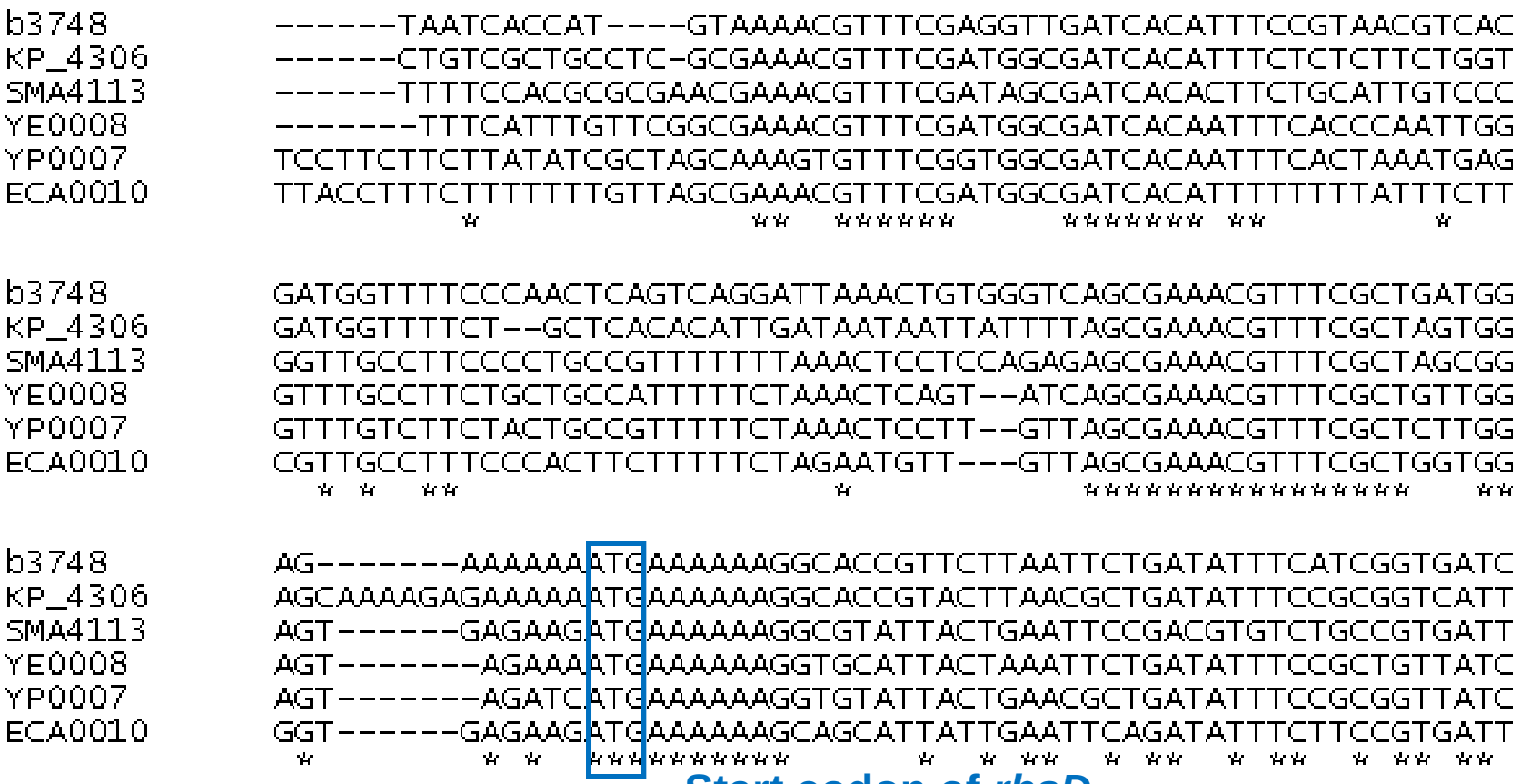
Positional weight matrix

A	1.1	−1.0	−0.7	0.5	2.2	2.2	1.9	−0.7	−0.7	−0.1	−1.0	−0.7	−1.1	−0.7	−1.4	−0.7
C	−0.4	1.9	−0.7	1.6	−0.7	−0.7	0.1	2.2	−0.7	−1.2	−1.0	−0.7	−1.1	2.2	−0.3	−0.7
G	−0.4	0.1	2.2	−1.1	−0.7	−0.7	−1.0	−0.7	2.2	−0.1	−0.1	−0.7	1.2	−0.7	1.0	−0.7
T	−0.4	−1.0	−0.7	−1.1	−0.7	−0.7	−1.0	−0.7	−0.7	1.5	1.9	2.2	1.0	−0.7	0.6	2.2
	A	**C**	**G**	**A**	**A**	**A**	**A**	**C**	**G**	**A**	**T**	**T**	**G**	**C**	**T**	**T**

S = 1.1+1.9+2.2+0.5+2.2+2.2+1.9+2.2+2.2–0.1+1.9+2.2+1.2+2.2+0.6+2.2

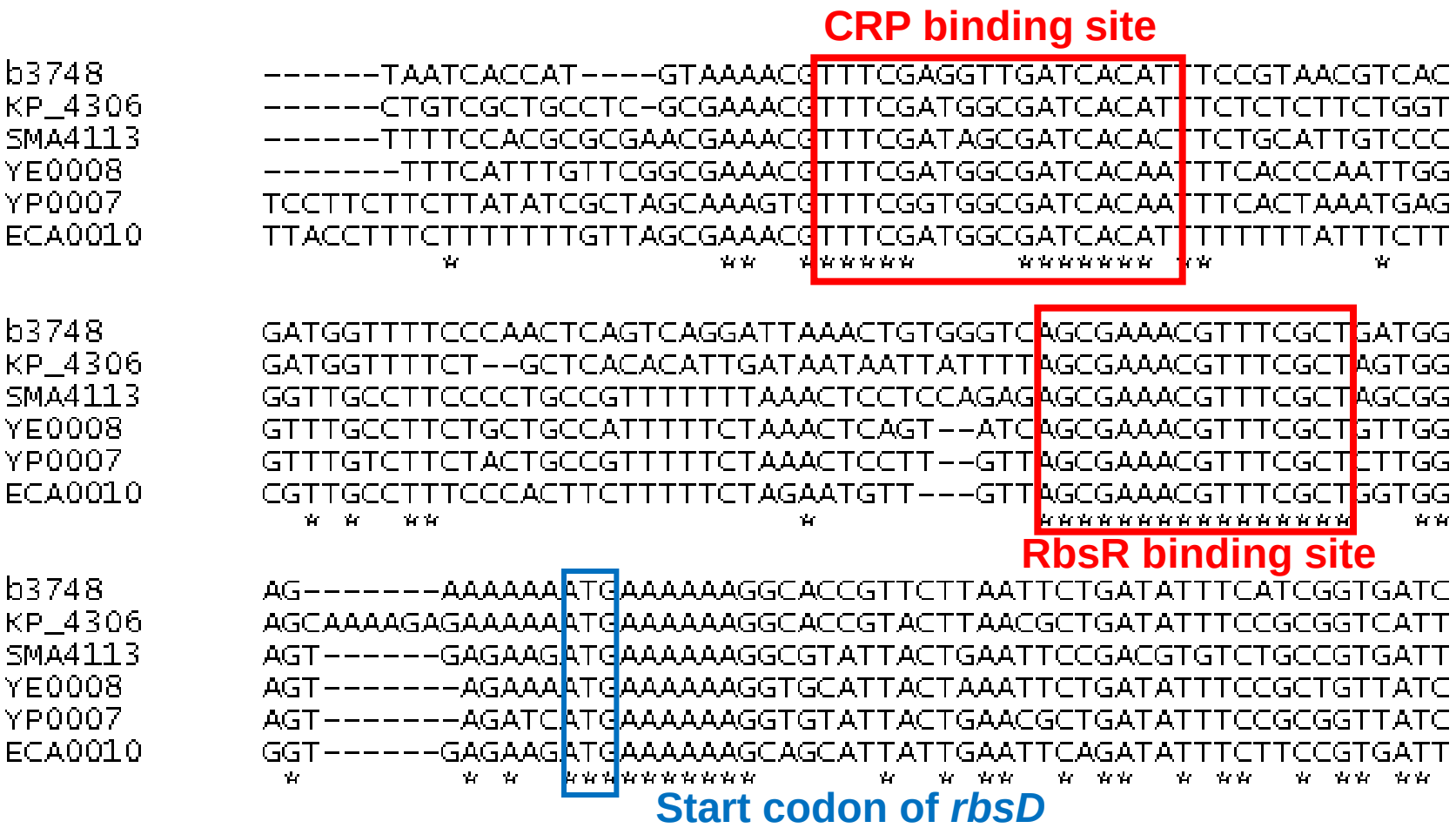
Phylogenetic footprinting

rbs operon in Enterobacteriaceae

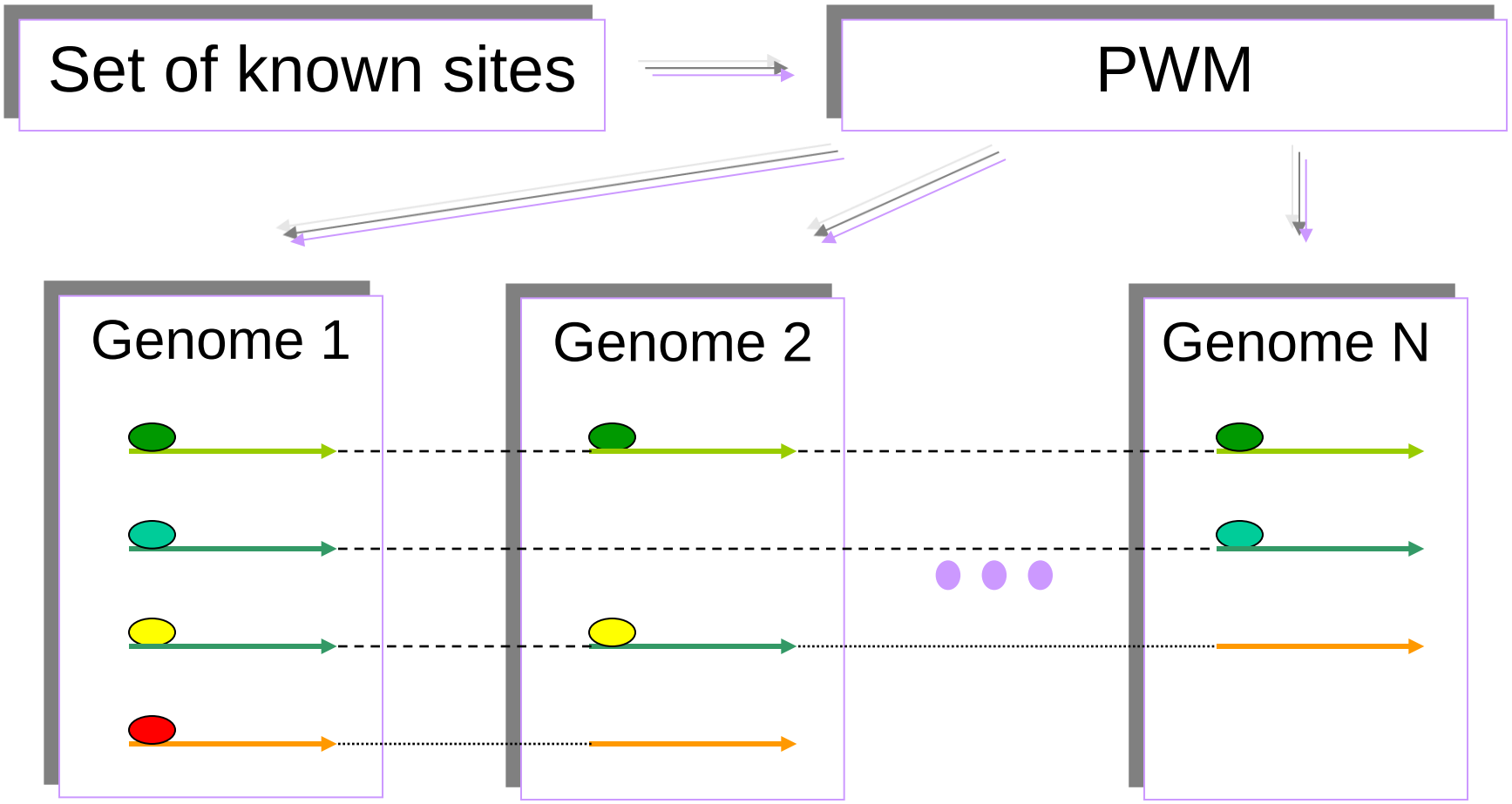


Phylogenetic footprinting

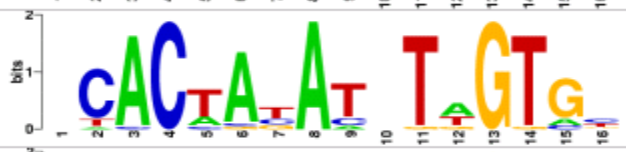
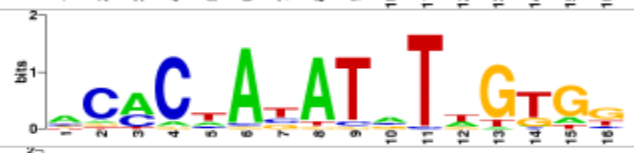
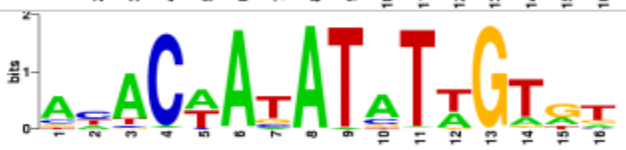
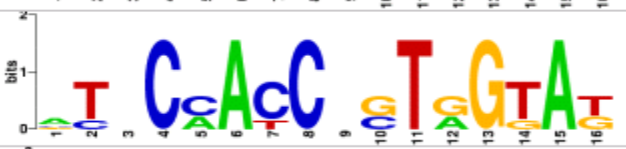
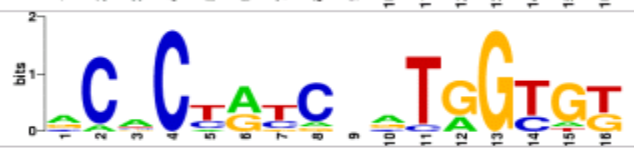
rbs operon in Enterobacteriaceae
… regulated by CRP and RbsR



Conservation of regulation => consistency check



2. Example. Conserved motif upstream of *nrd* genes

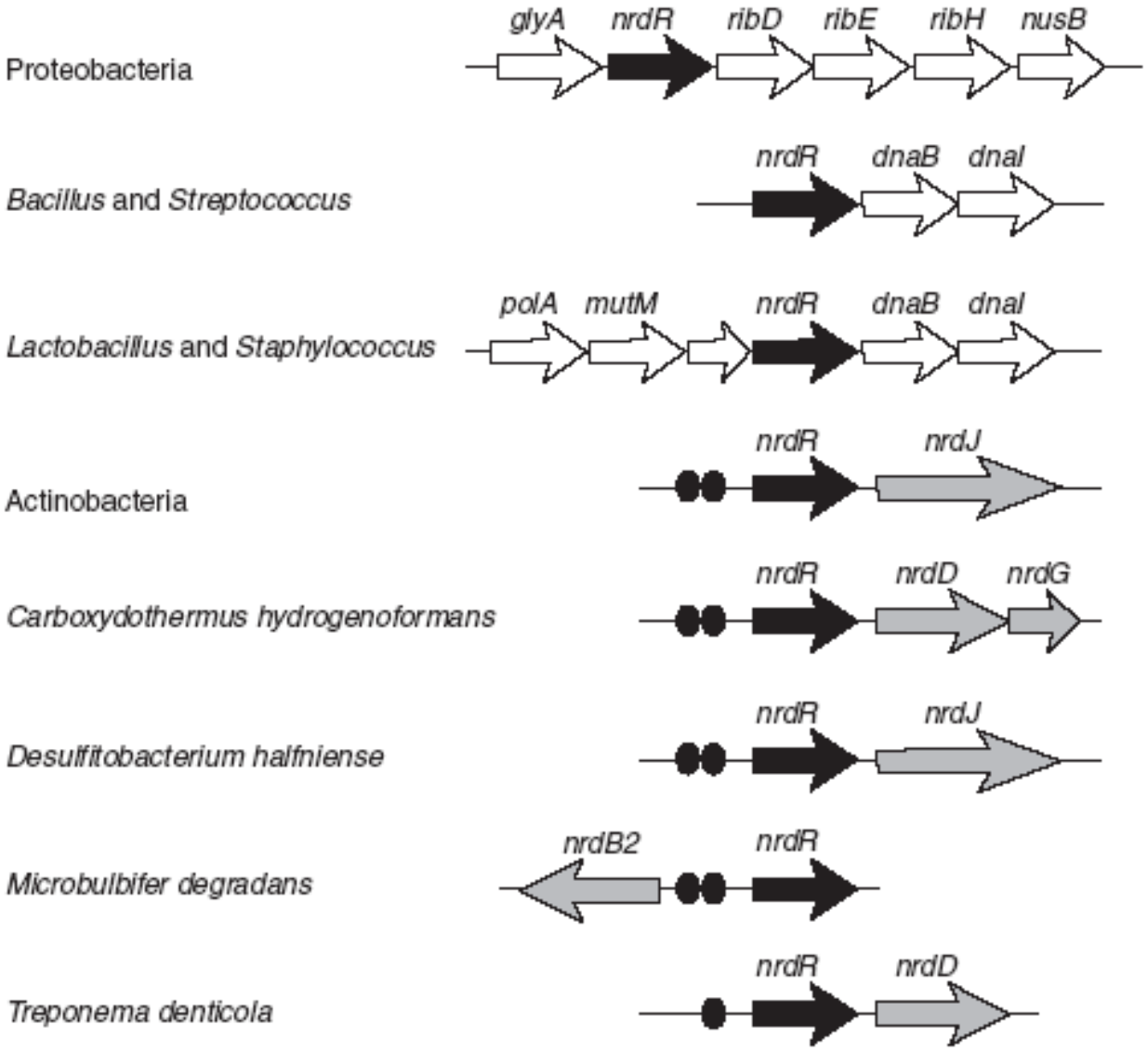
α-proteobacteria	
β-proteobacteria	
γ-proteobacteria	
δ-proteobacteria	
Bacillus/Clostridium	
Thermotogales	
Thermus/Deinococcus	
Chlamydiales	
Cyanobacteria	

ybaD is regulator of ribonucleotide reductases (*nrdR*)

- COG1327: exactly the same phylogenetic pattern with the signal
  - "large scale" on the level of major taxa
  - "small scale" within major taxa:
    - absent in small parasites among alpha- and gamma-proteobacteria
    - absent in *Desulfovibrio* spp. among delta-proteobacteria
    - absent in *Nostoc* sp. among cyanobacteria
    - absent in *Oenococcus* and *Leuconostoc* among Firmicutes
    - present only in *Treponema denticola* among four spirochetes
- Predicted transcriptional regulator, consists of a Zn-ribbon and ATP-cone domains
- *ybaD* in *E. coli*, renamed to *nrdR*

Additional evidence – 1

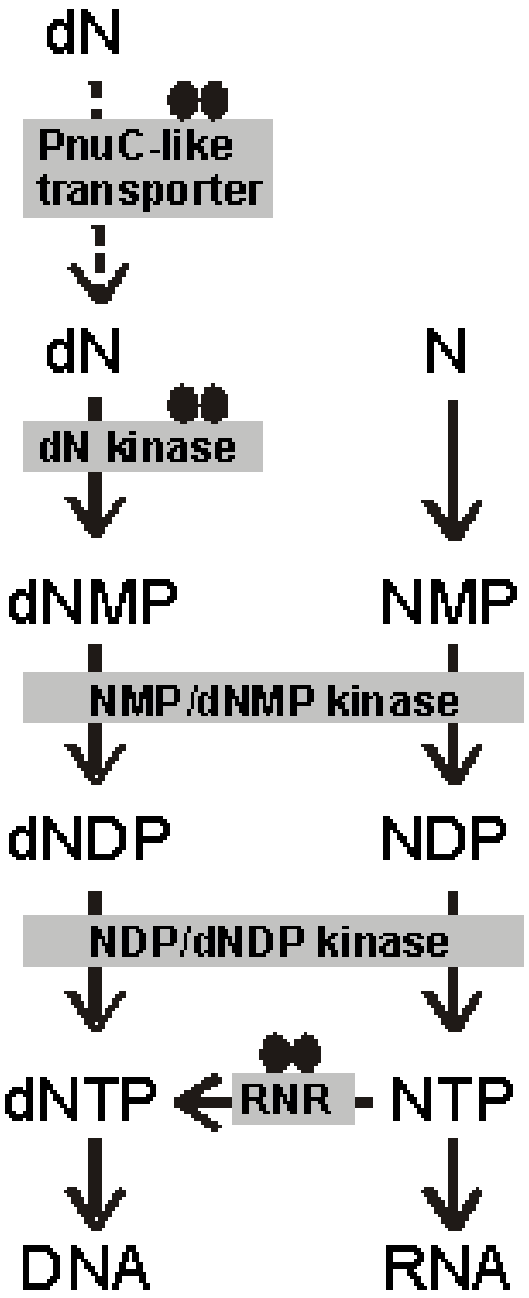
nrdR is sometimes clustered with *nrd* genes or with replication genes *dnaB, dnaI, polA*



Additional evidence – 2

In some genomes, candidate NrdR-binding sites are found upstream of other replication-related genes

- dNTP salvage
- topoisomerase I, replication initiator *dnaA*, chromosome partitioning, DNA helicase II

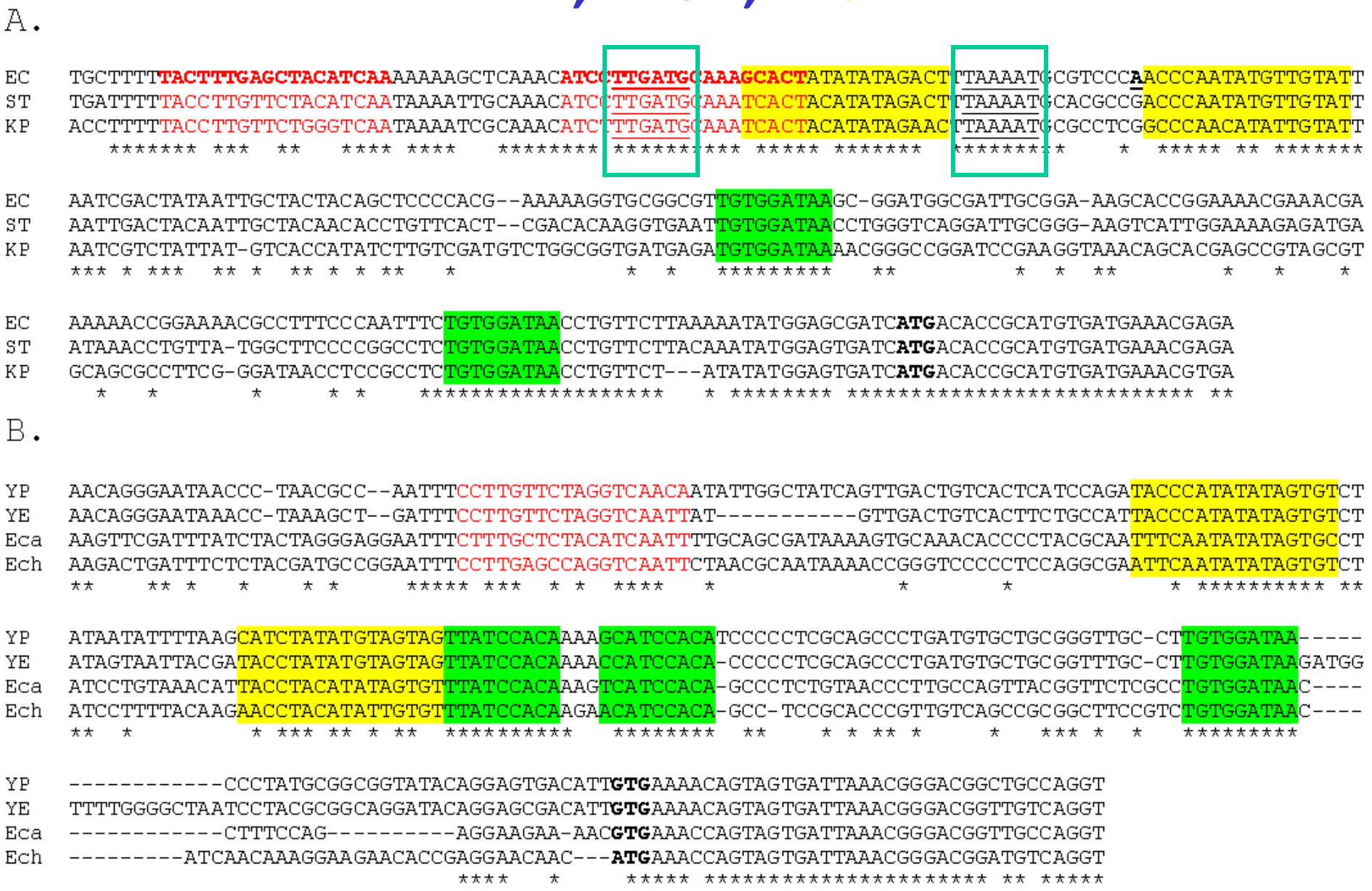


ybaD is regulator of ribonucleotide reductases (*nrdR*) and replication

- COG1327: Predicted transcriptional regulator, consists of a Zn-ribbon and ATP-cone domains
- exactly the same phylogenetic pattern with the signal
  - "large scale" on the level of major taxa
  - "small scale" within major taxa:
    - absent in small parasites among alpha- and gamma-proteobacteria
    - absent in *Desulfovibrio* spp. among delta-proteobacteria
    - absent in *Nostoc* sp. among cyanobacteria
    - absent in *Oenococcus* and *Leuconostoc* among Firmicutes
    - present only in *Treponema denticola* among four spirochetes
- sometimes clustered with *nrd* genes or with replication genes *dnaB, dnaI, polA*
- candidate signals upstream of other replication-related genes
  - dNTP salvage
  - topoisomerase I, replication initiator *dnaA*, chromosome partitioning, DNA helicase II

Multiple sites (*nrd* genes):

FNR, DnaA, NrdR



Mode of regulation

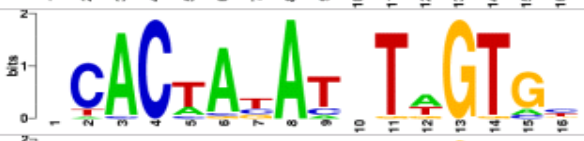
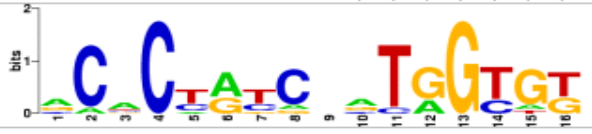
- Repressor (overlaps with promoters)
- Co-operative binding:
  - most sites occur in tandem (> 90% cases)
  - the distance between the copies (centers of palindromes) equals an integer number of DNA turns:
    - mainly (94%) 30-33 bp, in 84% 31-32 bp – 3 turns
    - 21 bp (2 turns) in *Vibrio* spp.
    - 41-42 bp (4 turns) in some Firmicutes

- experimental confirmation in *Streptomyces* (Borovok et al. 2004, Grinberg et al. 2006) and in *E. coli* (Grinberg et al. 2006)

3. Regulators and their motifs

- Cases of motif conservation at surprisingly large distances
- Subtle changes at close evolutionary distances
- Correlation between contacting nucleotides and amino acid residues
- Conserved non-consensus positions

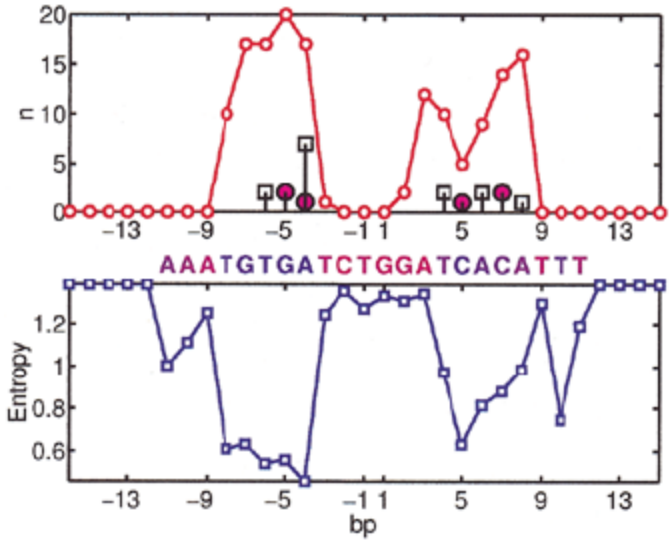
NrdR: conservation at large distances

α-proteobacteria	
β-proteobacteria	
γ-proteobacteria	
δ-proteobacteria	
Bacillus/Clostridium	
Thermotogales	
Thermus/Deinococcus	
Chlamydiales	
Cyanobacteria	

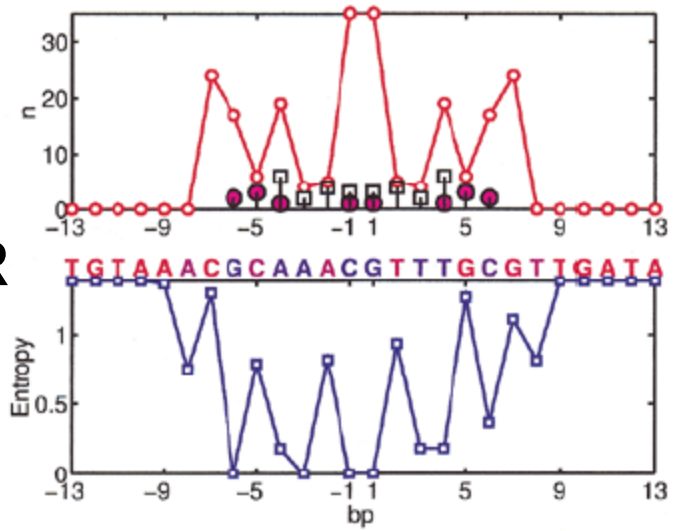
DNA motifs and protein-DNA interactions

Entropy at aligned sites and the number of contacts
(heavy atoms in a base pair at a distance <cutoff from a protein atom)

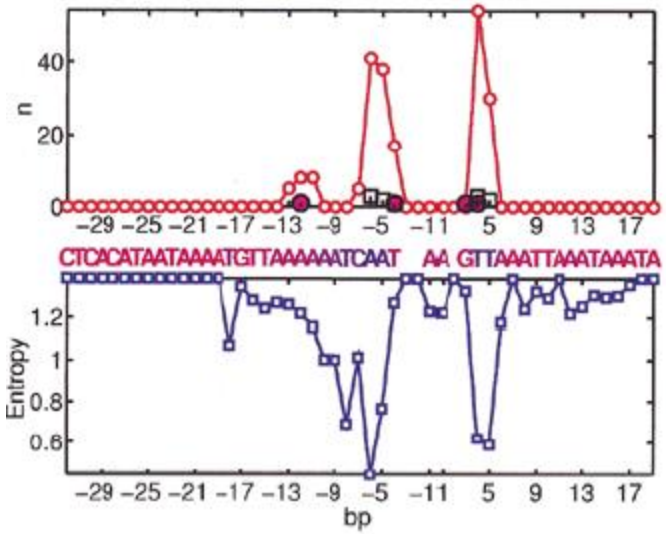
CRP



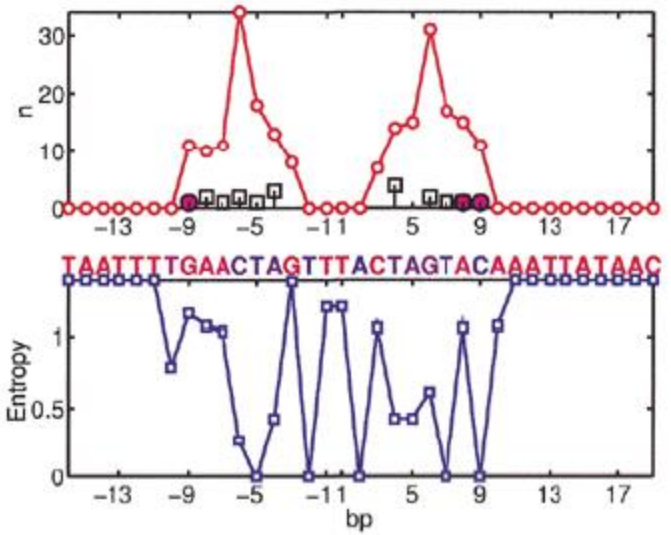
PurR



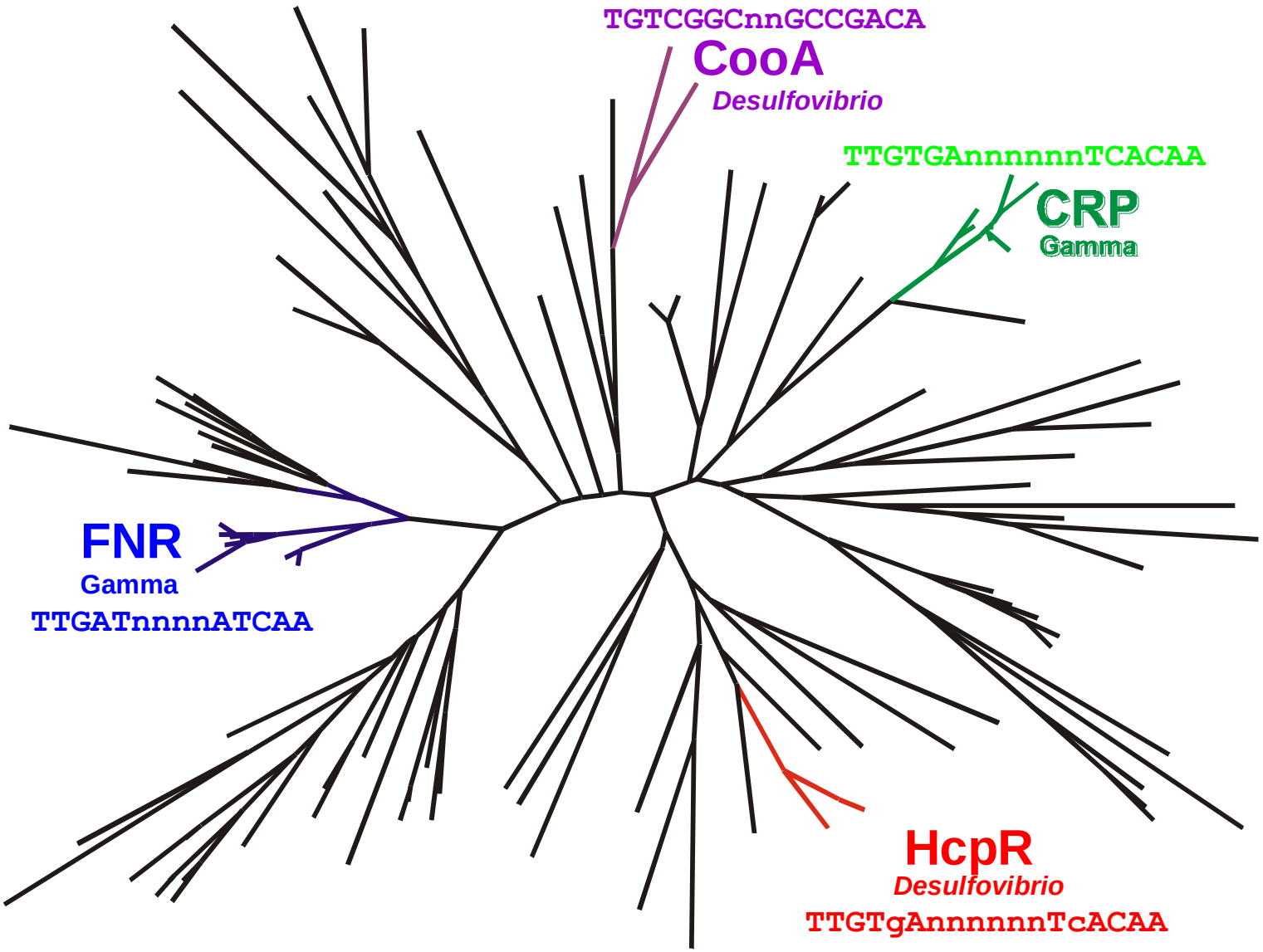
IHF



TrpR



Changes: CRP/FNR family

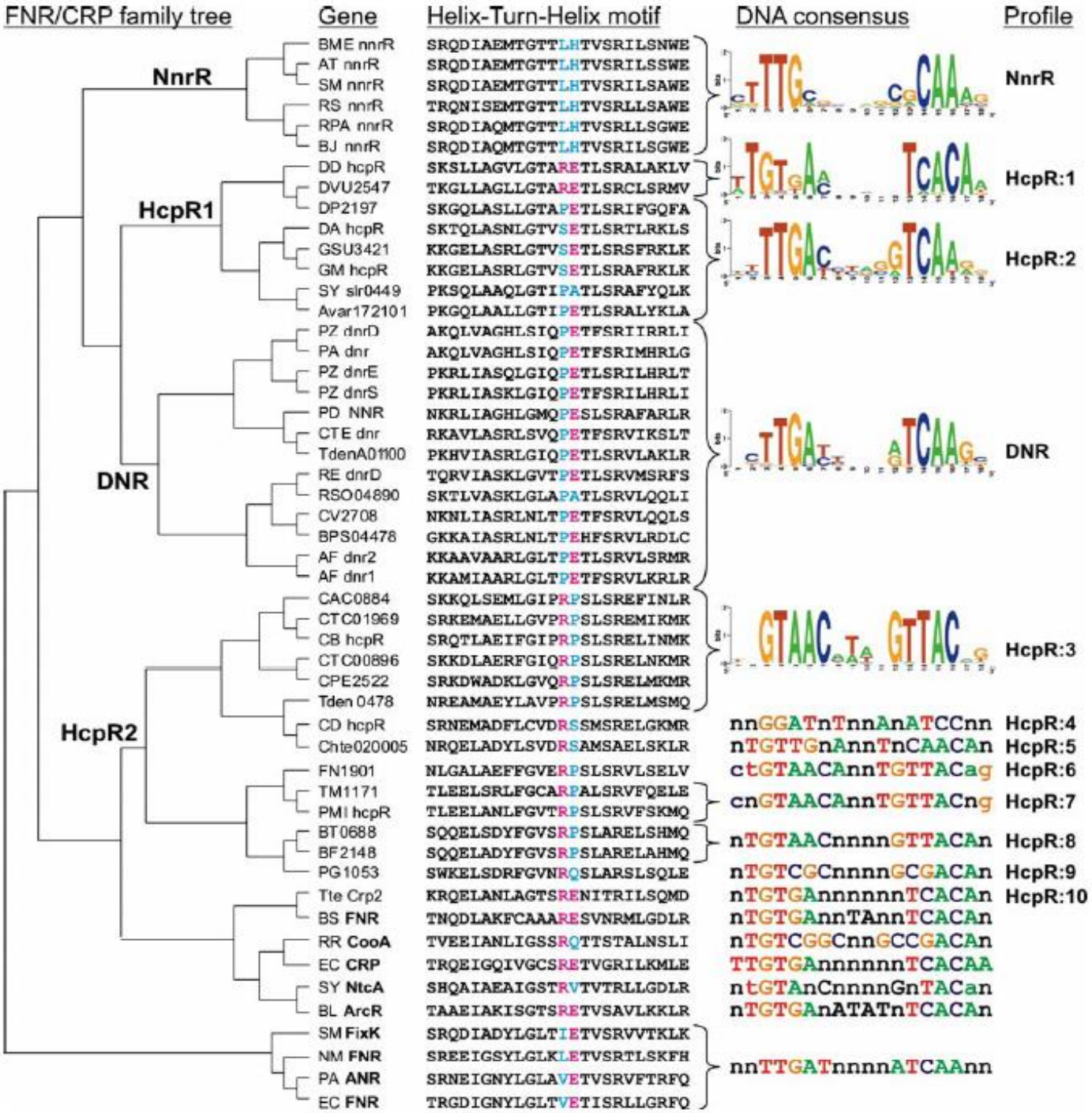


Correlation between contacting nucleotides and amino acid residues

- CooA in *Desulfovibrio* spp.
- CRP in Gamma-proteobacteria
- HcpR in *Desulfovibrio* spp.
- FNR in Gamma-proteobacteria

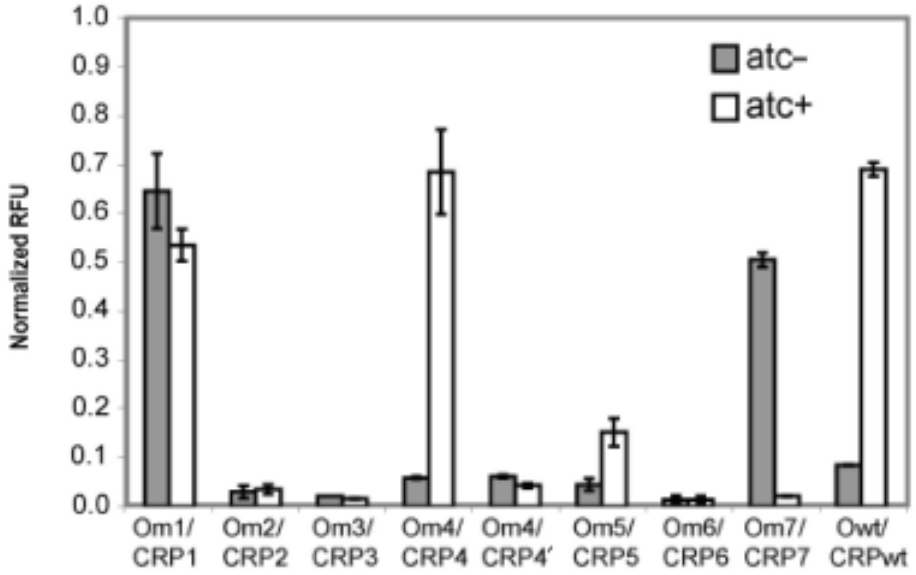
Contacting residues: REnnnR
TG: 1st arginine
GA: glutamate and 2nd arginine

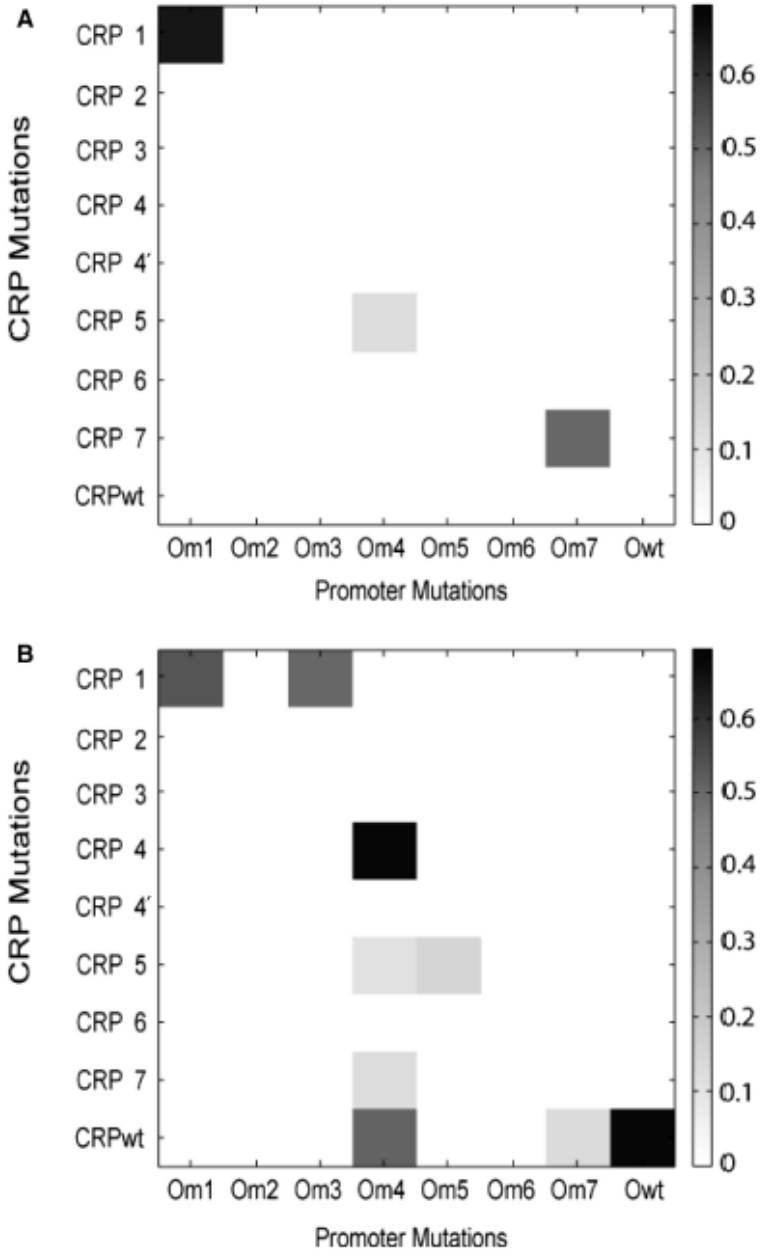
The correlation holds for other factors in the family



Engineering transcription factors with novel DNA-binding specificity using comparative genomics

Tasha A. Desai[1], Dmitry A. Rodionov[2,3], Mikhail S. Gelfand[3,4], Eric J. Alm[5,*] and Christopher V. Rao[1,*]



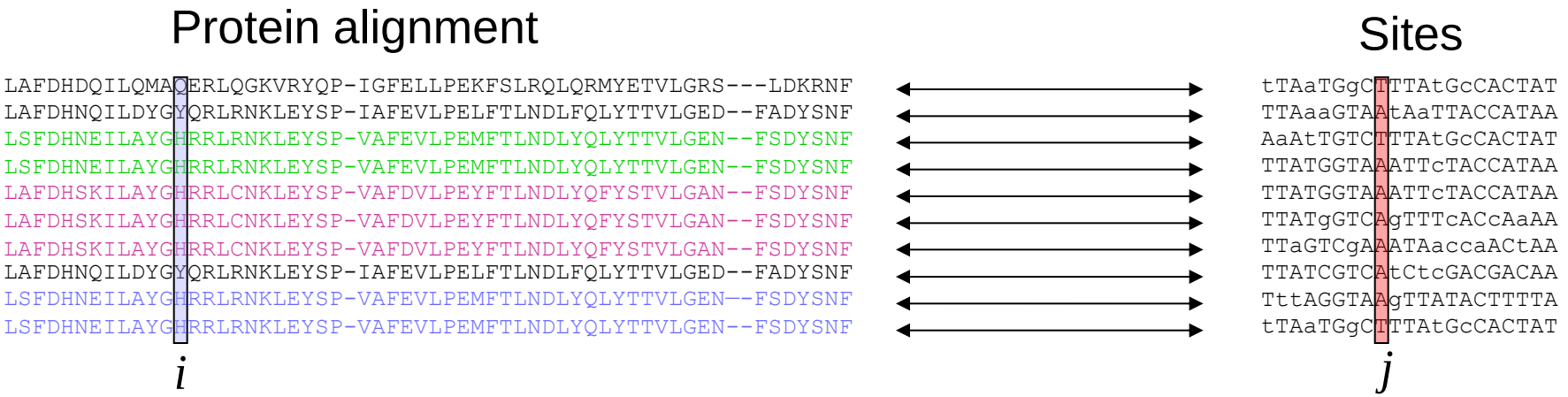


LacI family: systematic analysis

- 1369 DNA-binding domains in 200 orthologous rows <Id>=35%, <L>=71 a.o.
- 4484 binding sites, L=20н., <Id>=45%

- Calculate mutual information between columns of TF and site alignments
- Set threshold on mutual information of correlated pairs

Definitions



Mutual information

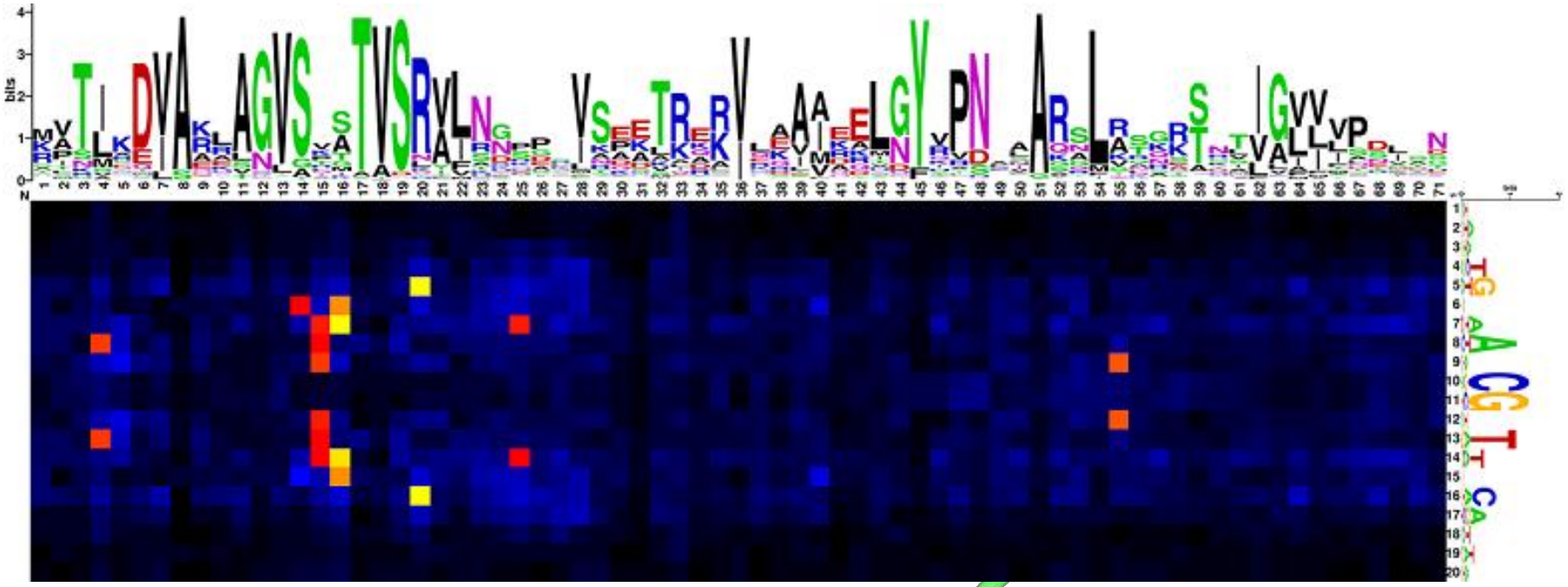
$$I(i,j) = \sum_{n=1}^{4}\sum_{a=1}^{20} p_{i,j}(a,n)\,\frac{\log p_{i,j}(a,n)}{p_i(a)\,p_j(n)}$$

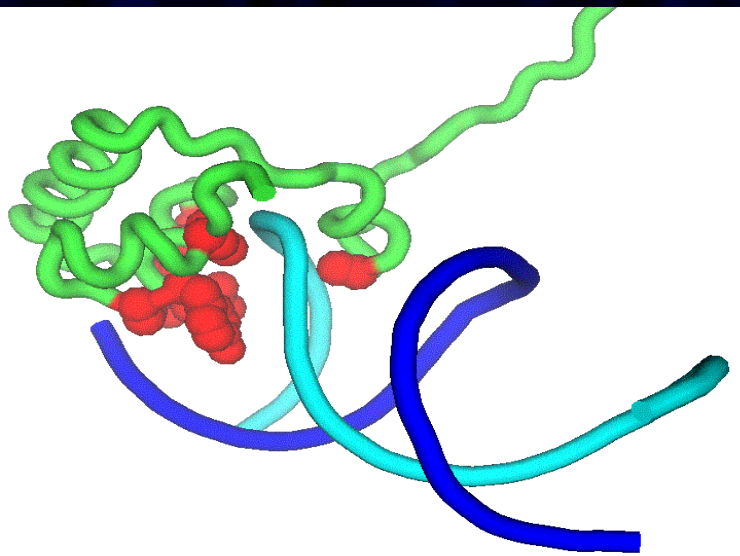
$\tilde{I}_{i,j}$

$$Z_{i,j} = \frac{I_{i,j} - E(\tilde{I}_{i,j})}{\sigma(\tilde{I}_{i,j})}$$

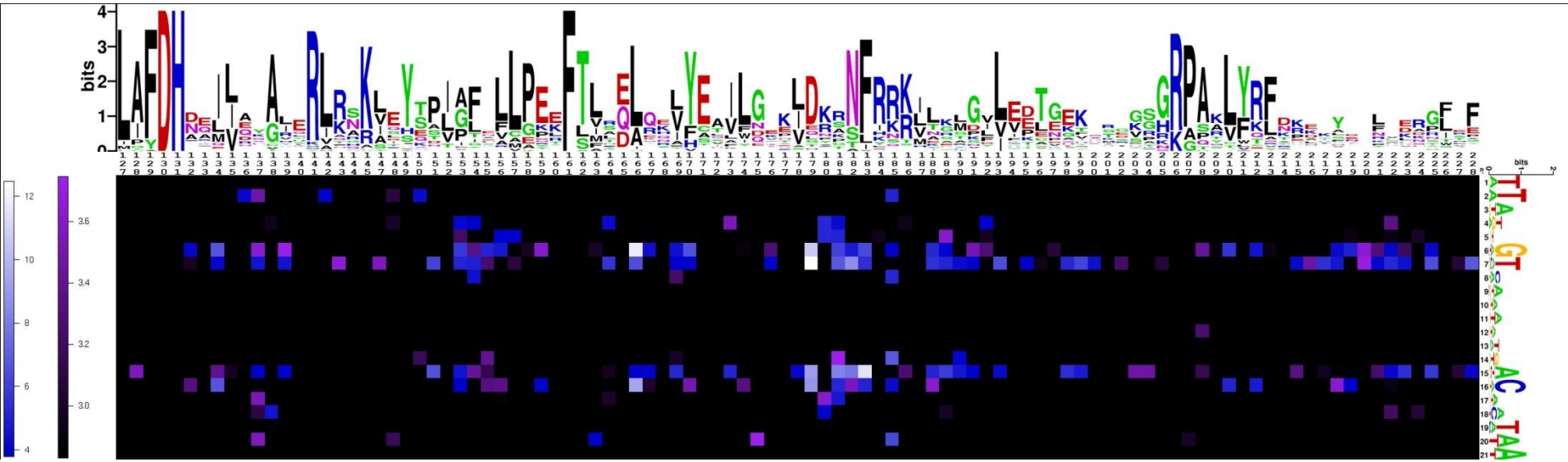
Z-score

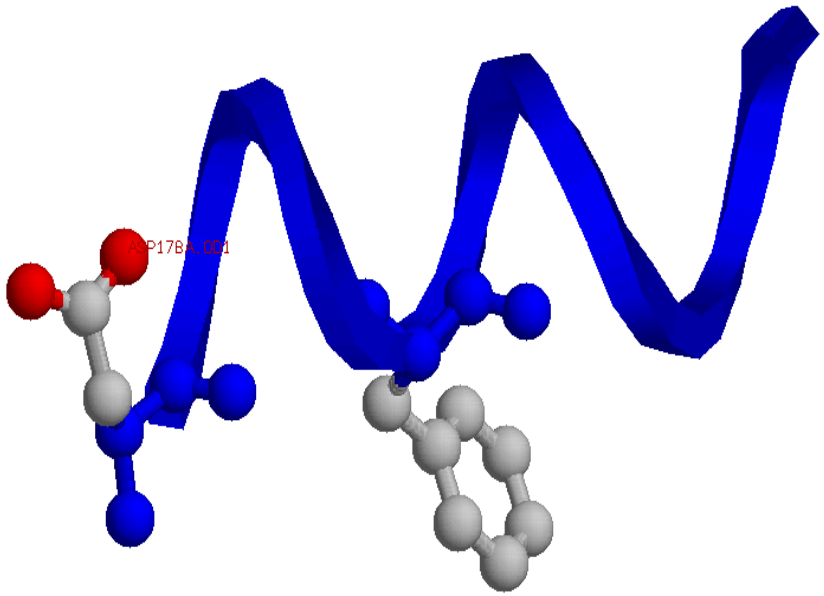
Correlated pairs





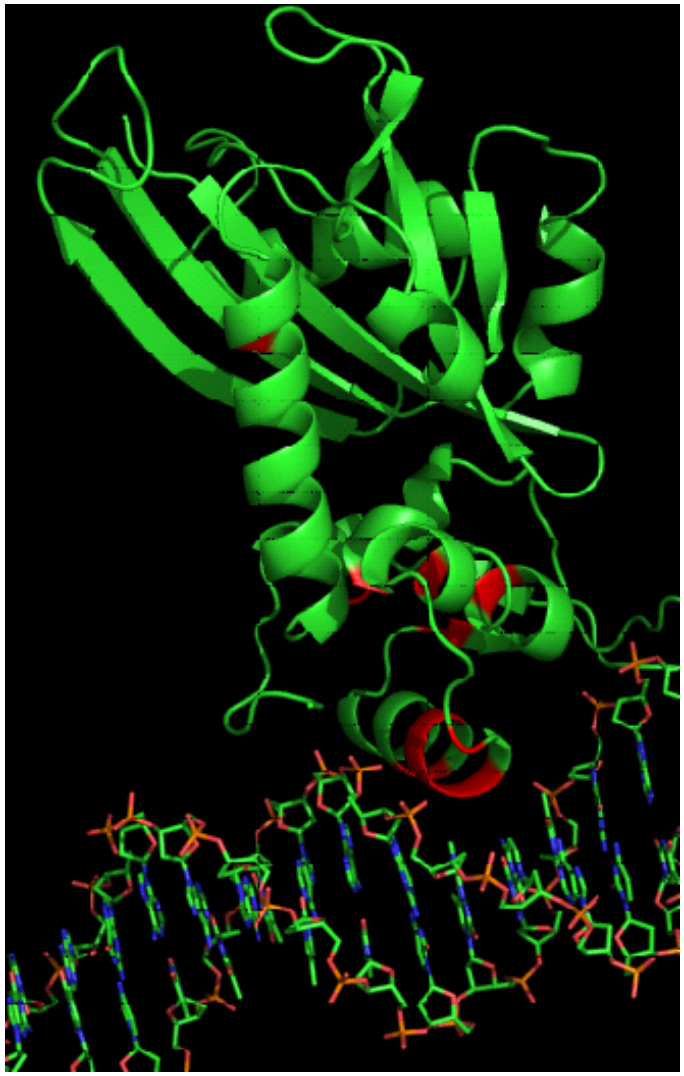
NrtR (regulator of NAD metabolism)

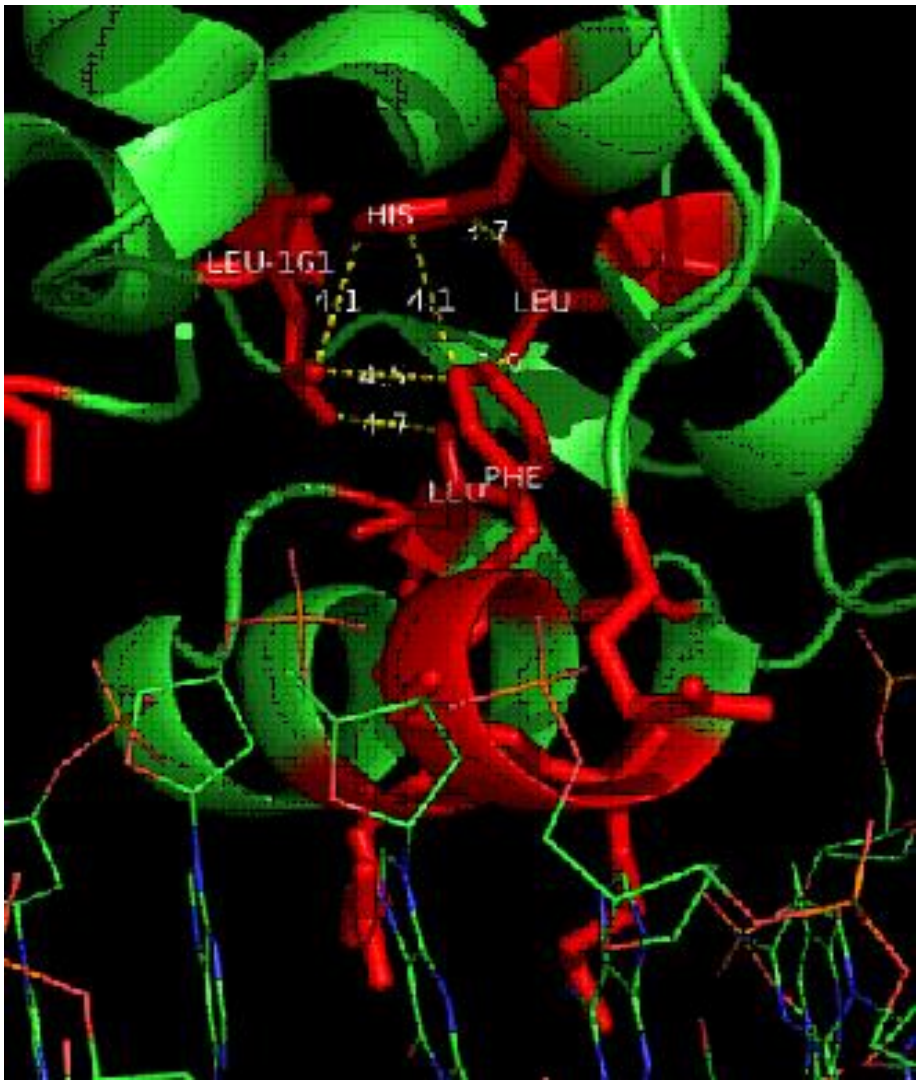




Comparison with the recently solved structure:

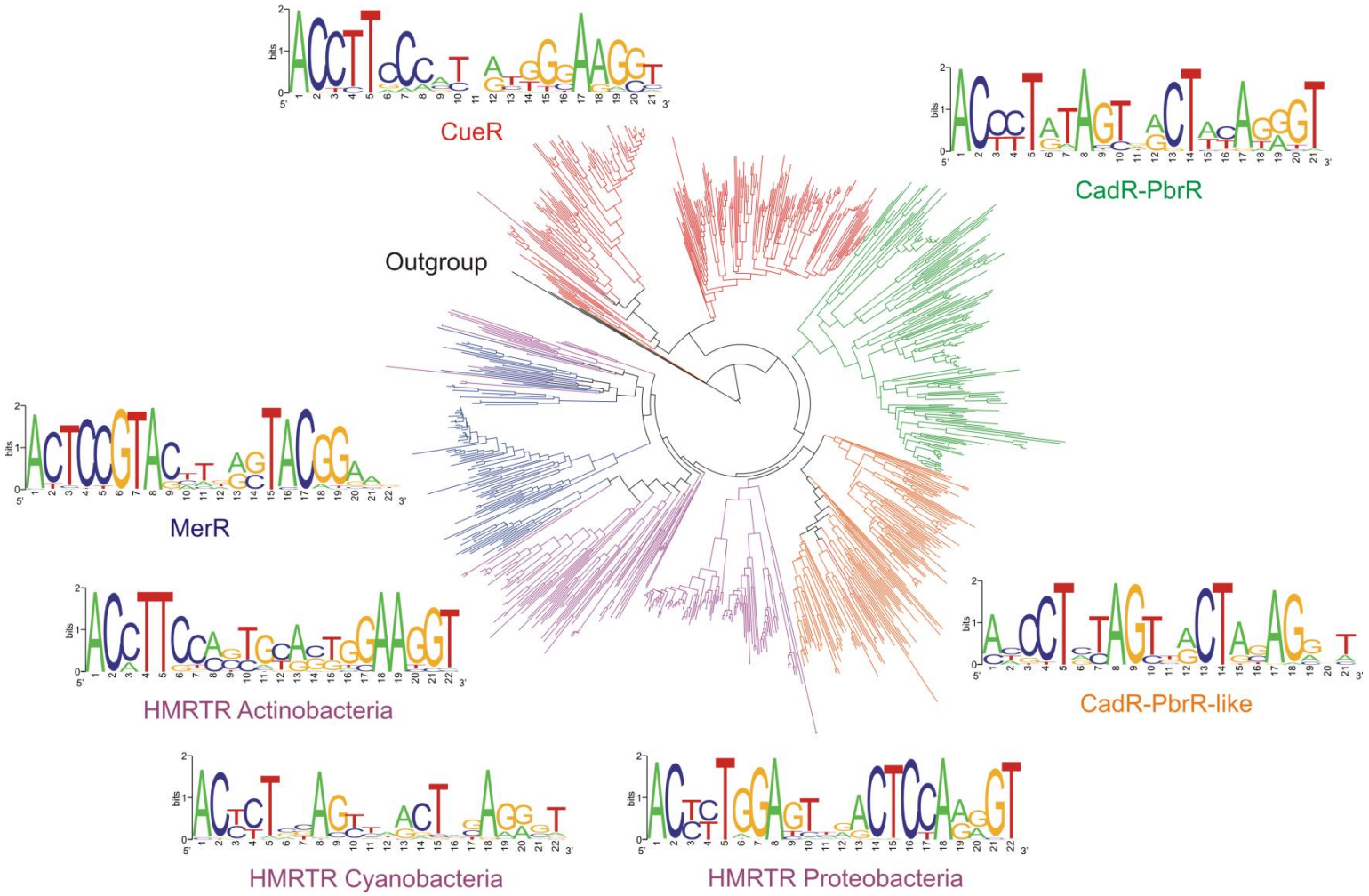
correlated positions indeed bind the DNA
(more exactly, form a hydrophobic cluster)





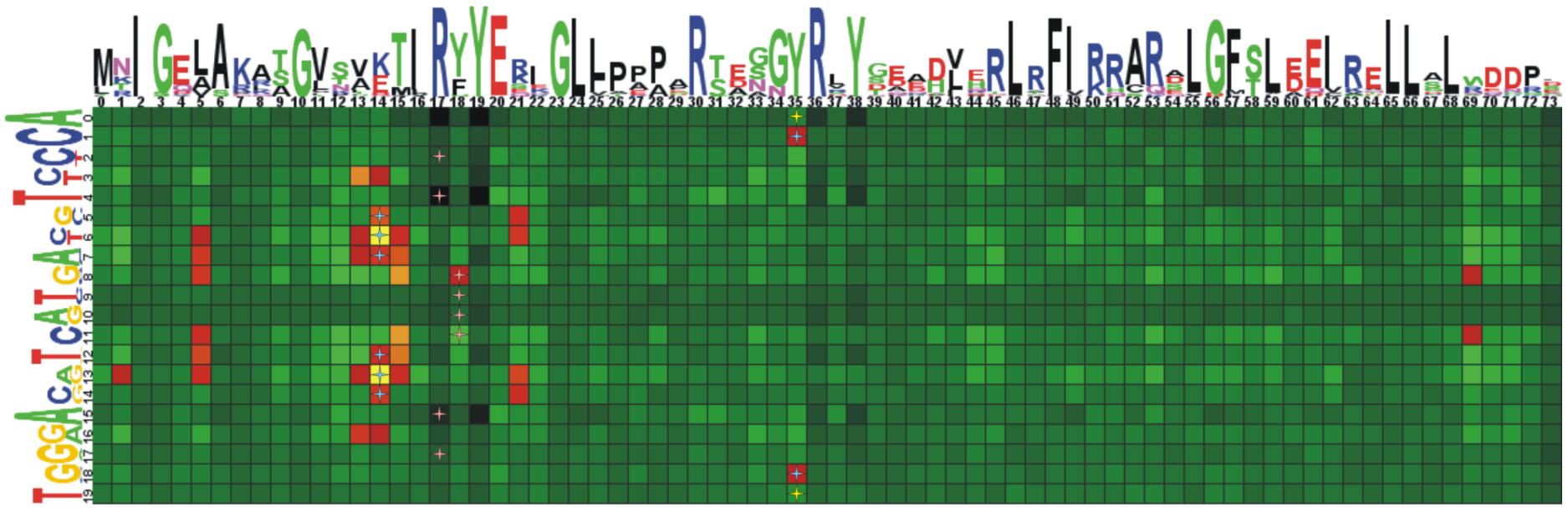
MerR family

Phylogenetic tree of HMR transcriptional regulators from MerR family



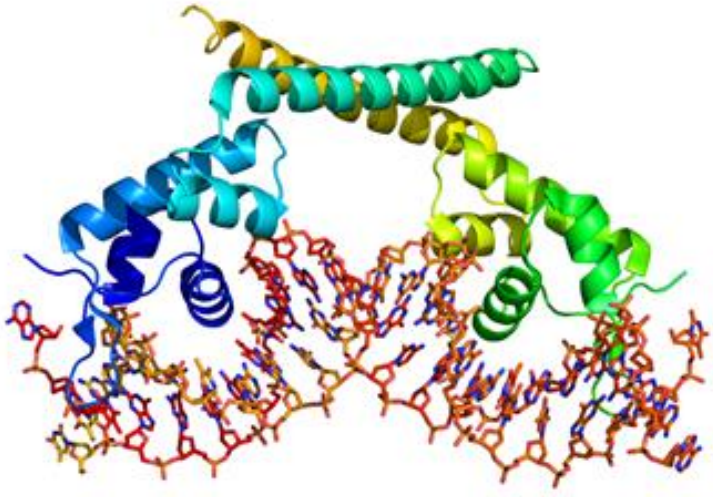
First 3 positions in sequence logos are the 3' end of 10 promoter boxes.

Correlations and structure



Uncorrelated pairs (below threshold)
Correlated pairs (above threshold)

Protein-DNA contacts from crystal structures
hydrogen bonds
water bridges
Van der Waals contac



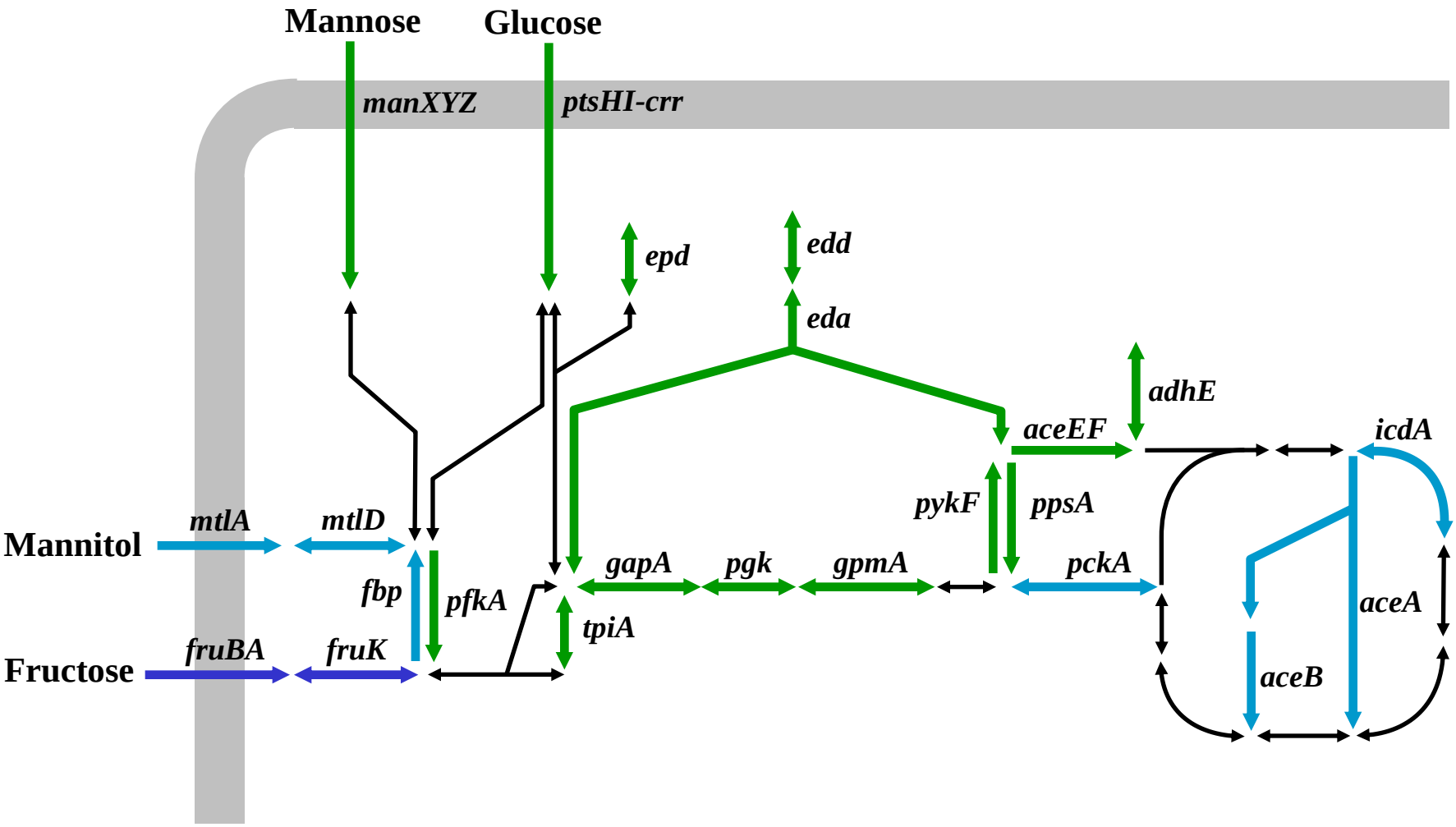
4. Evolution of regulatory networks

- Expansion and contraction of regulons
- New regulators (where from?)
- Duplications of regulators with or without regulated loci
- Loss of regulators with or without regulated loci
- Re-assortment of regulators and structural genes
- … especially in complex systems
- Horizontal transfer
- Birth of new sites
  - positions under selection in intergenic regions
  - conservation of sites

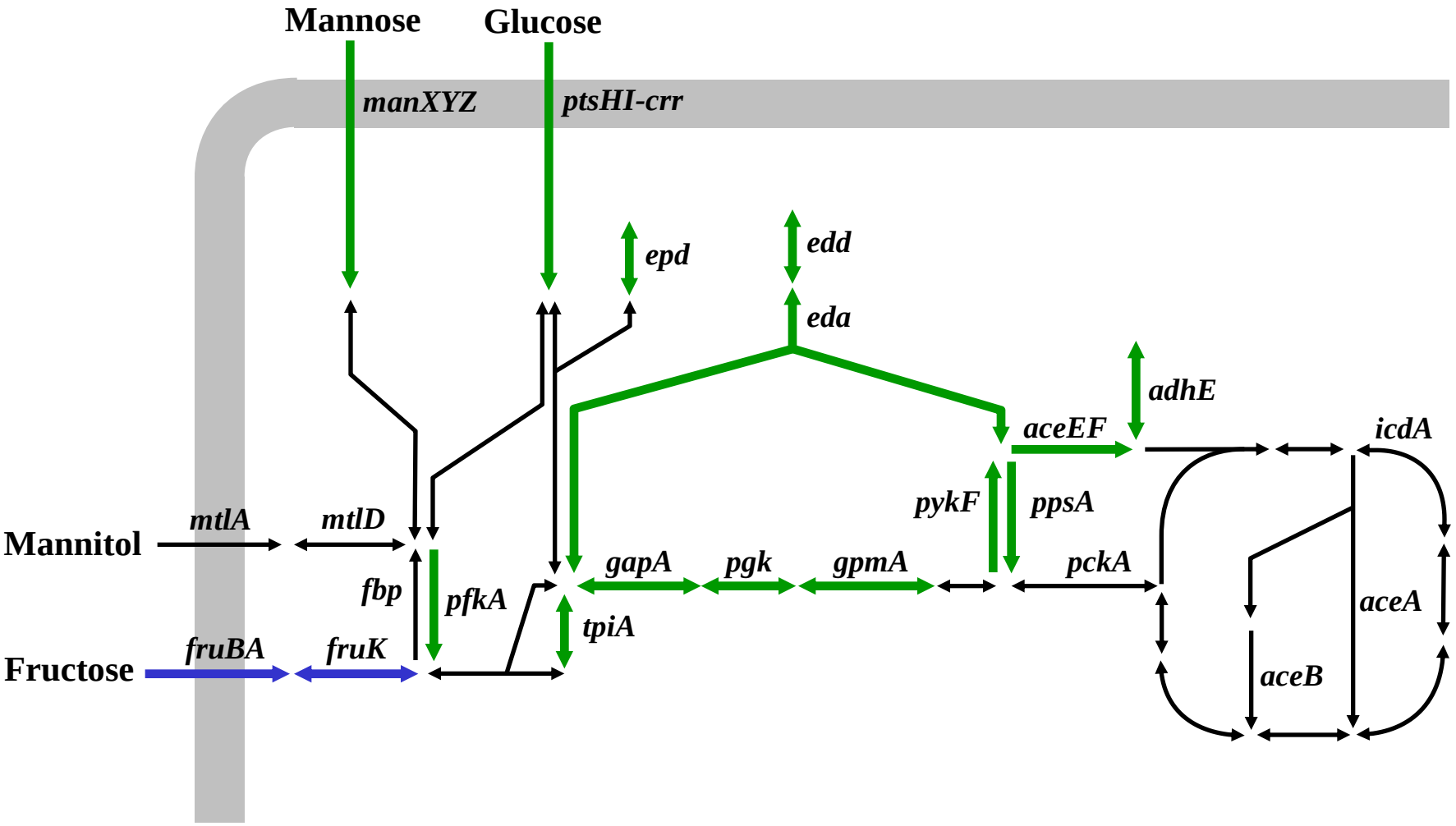
Regulon expansion, or how FruR has become CRA

- CRA (a.k.a. FruR) in *Escherichia coli*:
  - global regulator
  - well-studied in experiment (many regulated genes known)
- **Going back in time:** looking for candidate CRA/FruR sites upstream of (orthologs of) genes known to be regulated in *E.coli*

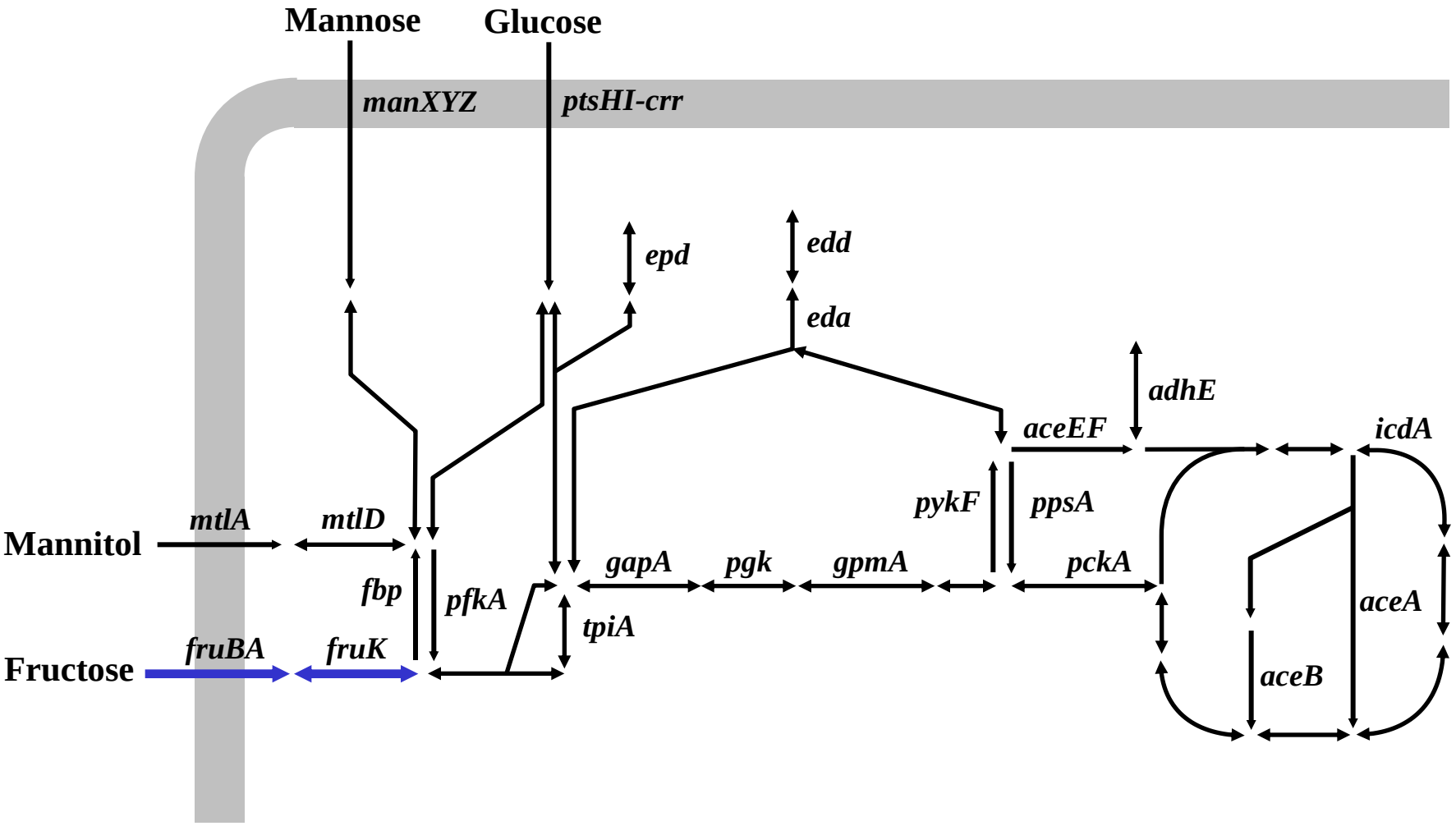
Experimental data in *E. coli*



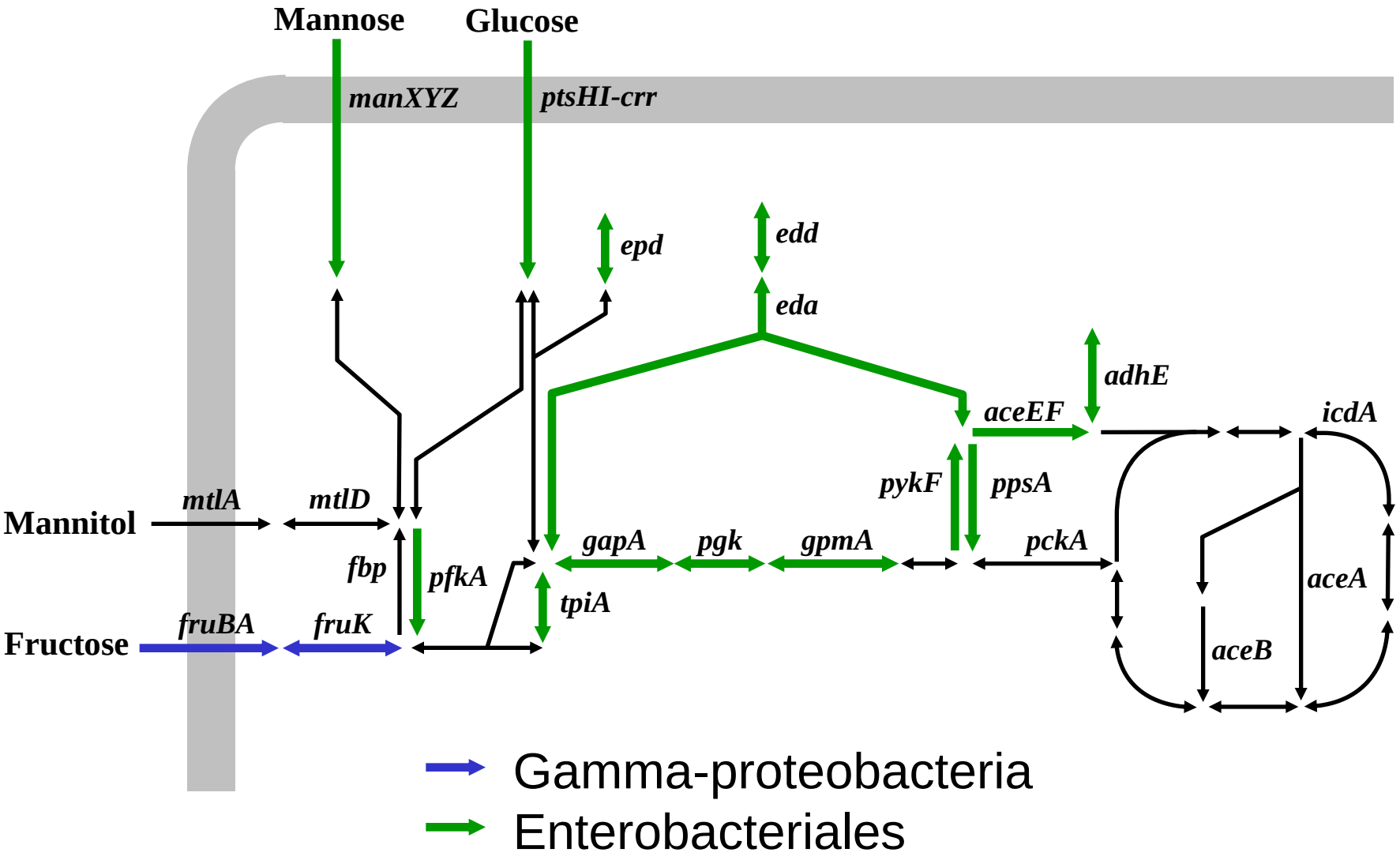
Sites conserved in Enterobacteriales



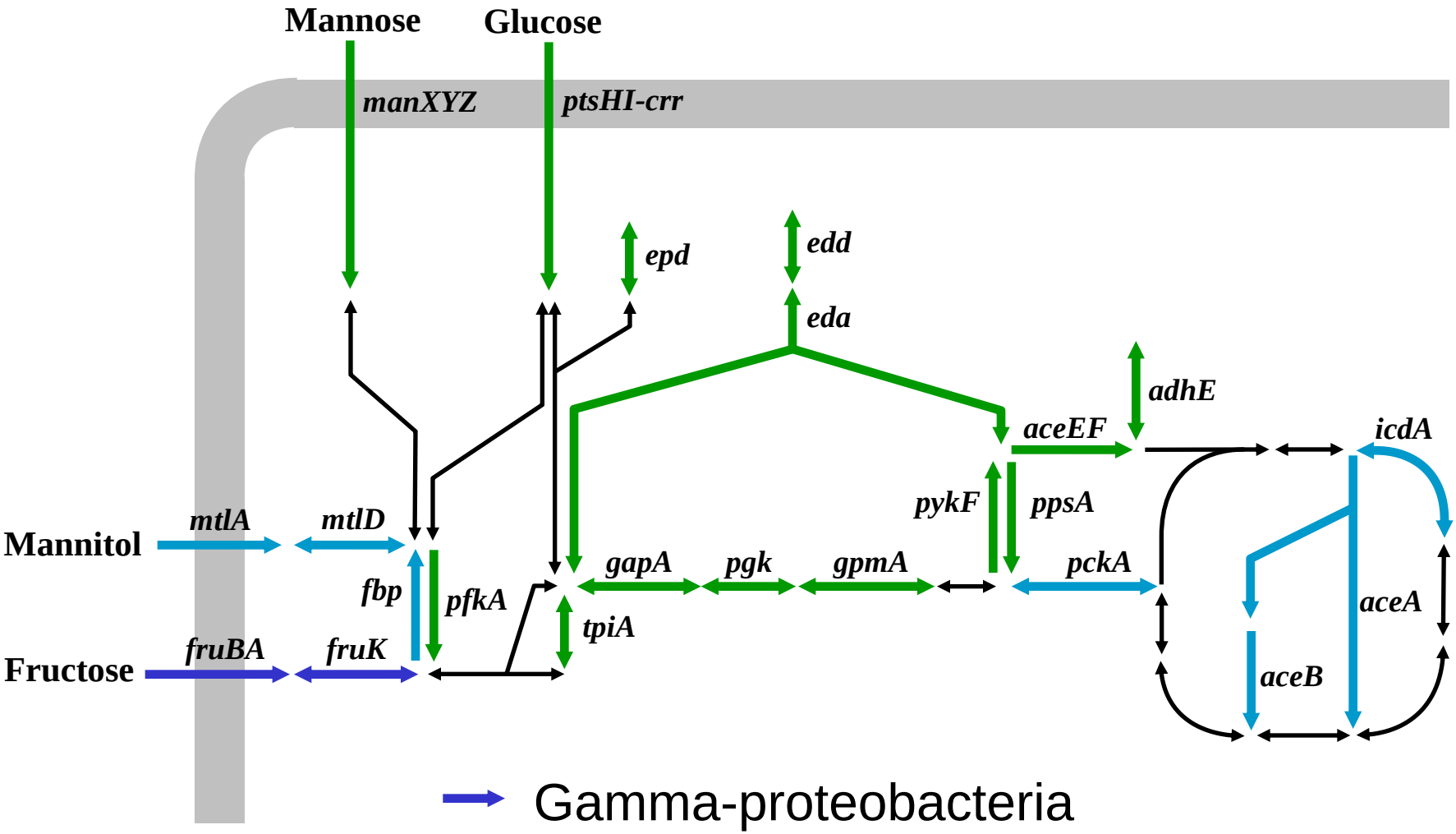
Common ancestor of Enterobacteriales and Vibrionales



Common ancestor of Enterobacteriales



Common ancestor of *Escherichia* and *Salmonella*



→ Gamma-proteobacteria

→ Enterobacteriales

→ *E. coli* and *Salmonella* spp.

Regulation of iron homeostasis (the *Escherichia coli* paradigm)

Iron:

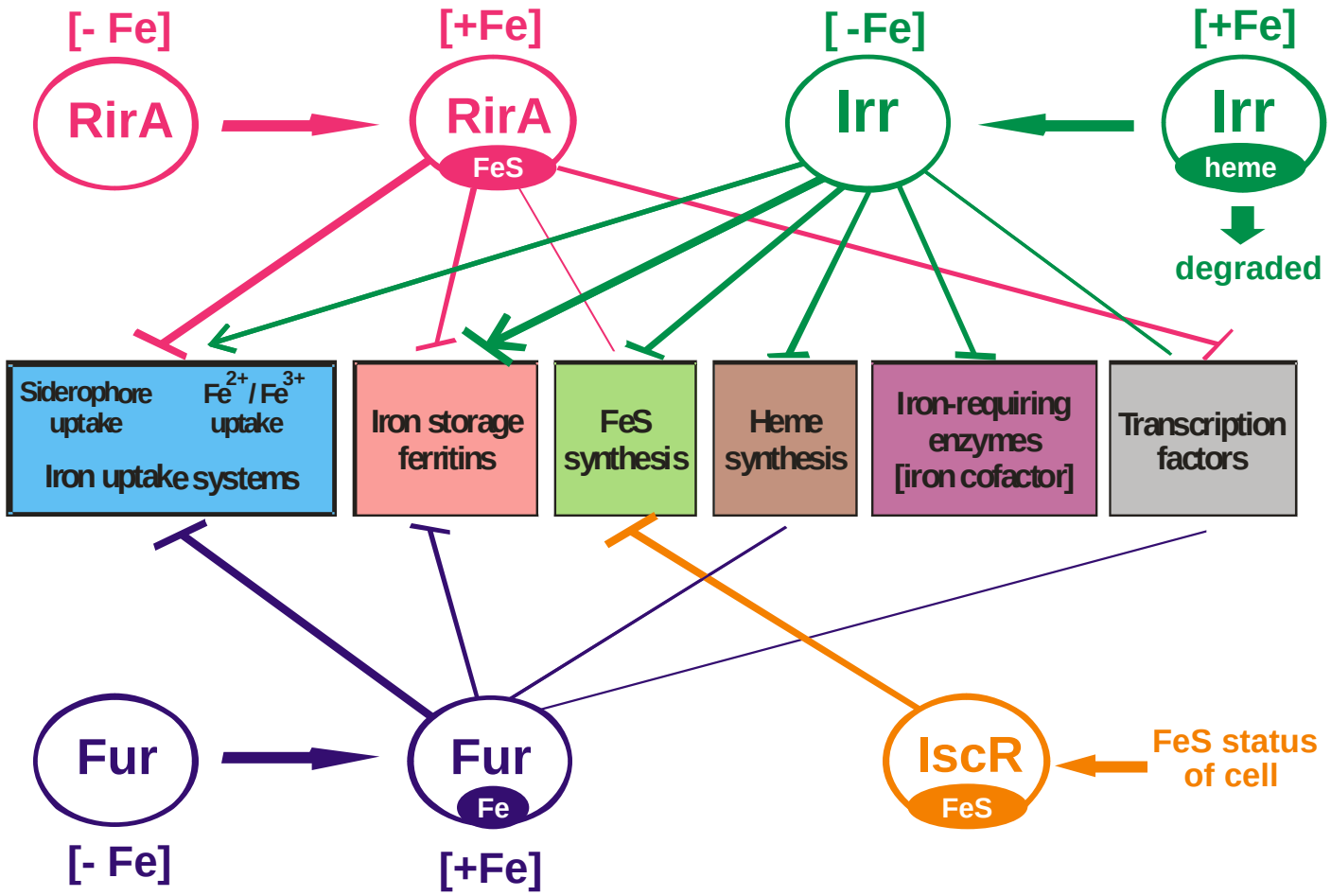
- essential cofactor (limiting in many environments)
- dangerous at large concentrations

FUR (responds to iron):

- synthesis of siderophores
- transport (siderophores, heme, Fe^{2+}, Fe^{3+})
- storage
- iron-dependent enzymes
- synthesis of heme
- synthesis of Fe-S clusters

Similar in *Bacillus subtilis*

Regulation of iron homeostasis in α-proteobacteria

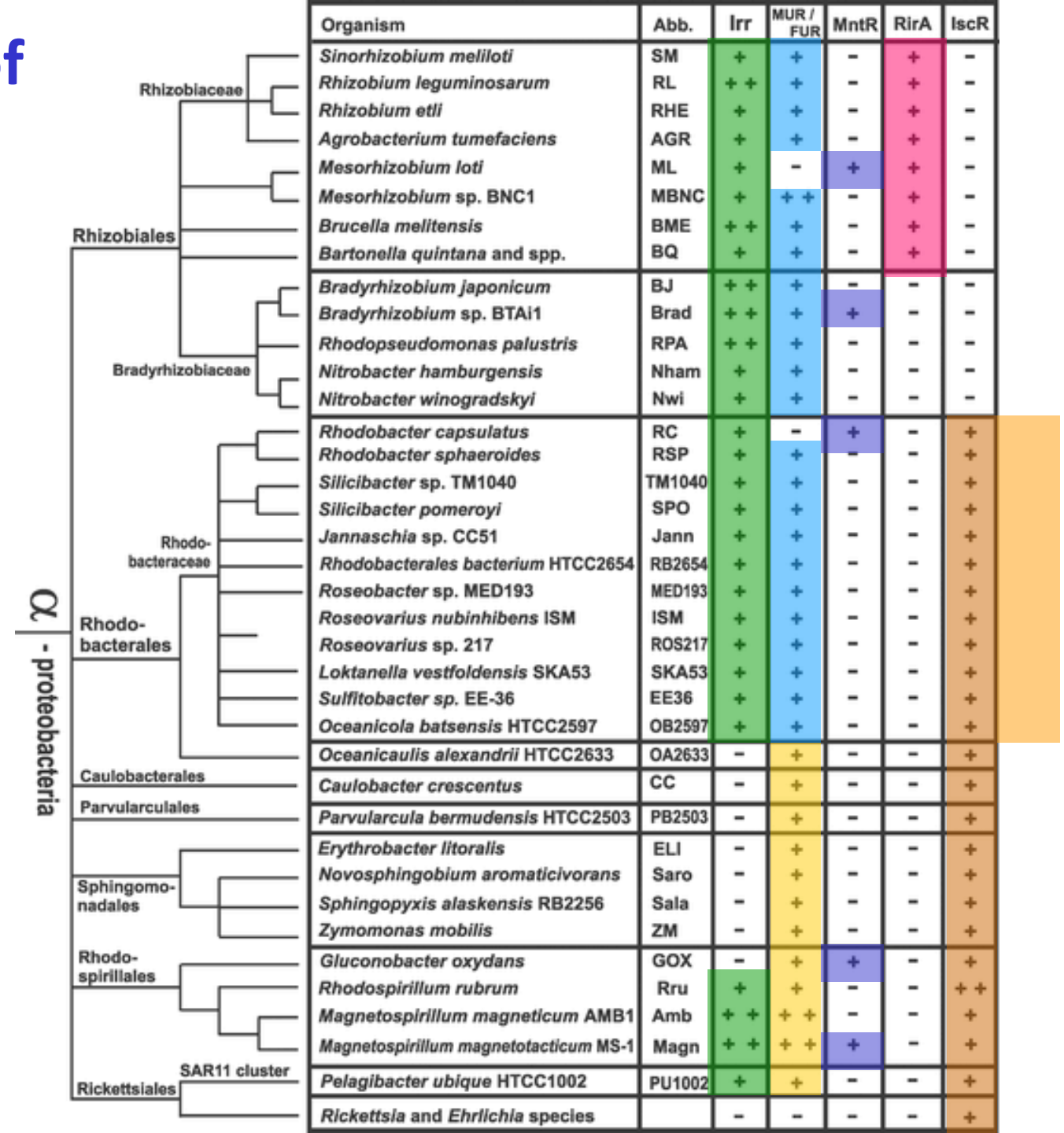


Experimental studies:

- **FUR/MUR**: *Bradyrhizobium*, *Rhizobium* and *Sinorhizobium*
- **RirA** (Rrf2 family): *Rhizobium* and *Sinorhizobium*
- **Irr** (FUR family): *Bradyrhizobium*, *Rhizobium* and *Brucella*

Distribution of transcription factors in genomes

Search for candidate motifs and binding sites using standard comparative genomic techniques



Organism	Abb.	Irr	MUR / FUR	MntR	RirA	IscR
Sinorhizobium meliloti	SM	+	+	–	+	–
Rhizobium leguminosarum	RL	+ +	+	–	+	–
Rhizobium etli	RHE	+	+	–	+	–
Agrobacterium tumefaciens	AGR	+	+	–	+	–
Mesorhizobium loti	ML	+	–	+	+	–
Mesorhizobium sp. BNC1	MBNC	+	+ +	–	+	–
Brucella melitensis	BME	+ +	+	–	+	–
Bartonella quintana and spp.	BQ	+	+	–	+	–
Bradyrhizobium japonicum	BJ	+ +	+	–	–	–
Bradyrhizobium sp. BTAi1	Brad	+ +	+	+	–	–
Rhodopseudomonas palustris	RPA	+ +	+	–	–	–
Nitrobacter hamburgensis	Nham	+	+	–	–	–
Nitrobacter winogradskyi	Nwi	+	+	–	–	–
Rhodobacter capsulatus	RC	+	–	+	–	+
Rhodobacter sphaeroides	RSP	+	+	–	–	+
Silicibacter sp. TM1040	TM1040	+	+	–	–	+
Silicibacter pomeroyi	SPO	+	+	–	–	+
Jannaschia sp. CC51	Jann	+	+	–	–	+
Rhodobacterales bacterium HTCC2654	RB2654	+	+	–	–	+
Roseobacter sp. MED193	MED193	+	+	–	–	+
Roseovarius nubinhibens ISM	ISM	+	+	–	–	+
Roseovarius sp. 217	ROS217	+	+	–	–	+
Loktanella vestfoldensis SKA53	SKA53	+	+	–	–	+
Sulfitobacter sp. EE-36	EE36	+	+	–	–	+
Oceanicola batsensis HTCC2597	OB2597	+	+	–	–	+
Oceanicaulis alexandrii HTCC2633	OA2633	–	+	–	–	+
Caulobacter crescentus	CC	–	+	–	–	+
Parvularcula bermudensis HTCC2503	PB2503	–	+	–	–	+
Erythrobacter litoralis	ELI	–	+	–	–	+
Novosphingobium aromaticivorans	Saro	–	+	–	–	+
Sphingopyxis alaskensis RB2256	Sala	–	+	–	–	+
Zymomonas mobilis	ZM	–	+	–	–	+
Gluconobacter oxydans	GOX	–	+	+	–	+
Rhodospirillum rubrum	Rru	+	+	–	–	+ +
Magnetospirillum magneticum AMB1	Amb	+ +	+ +	–	–	+
Magnetospirillum magnetotacticum MS-1	Magn	+ +	+ +	+	–	+
Pelagibacter ubique HTCC1002	PU1002	+	+	–	–	+
Rickettsia and *Ehrlichia* species		–	–	–	–	+

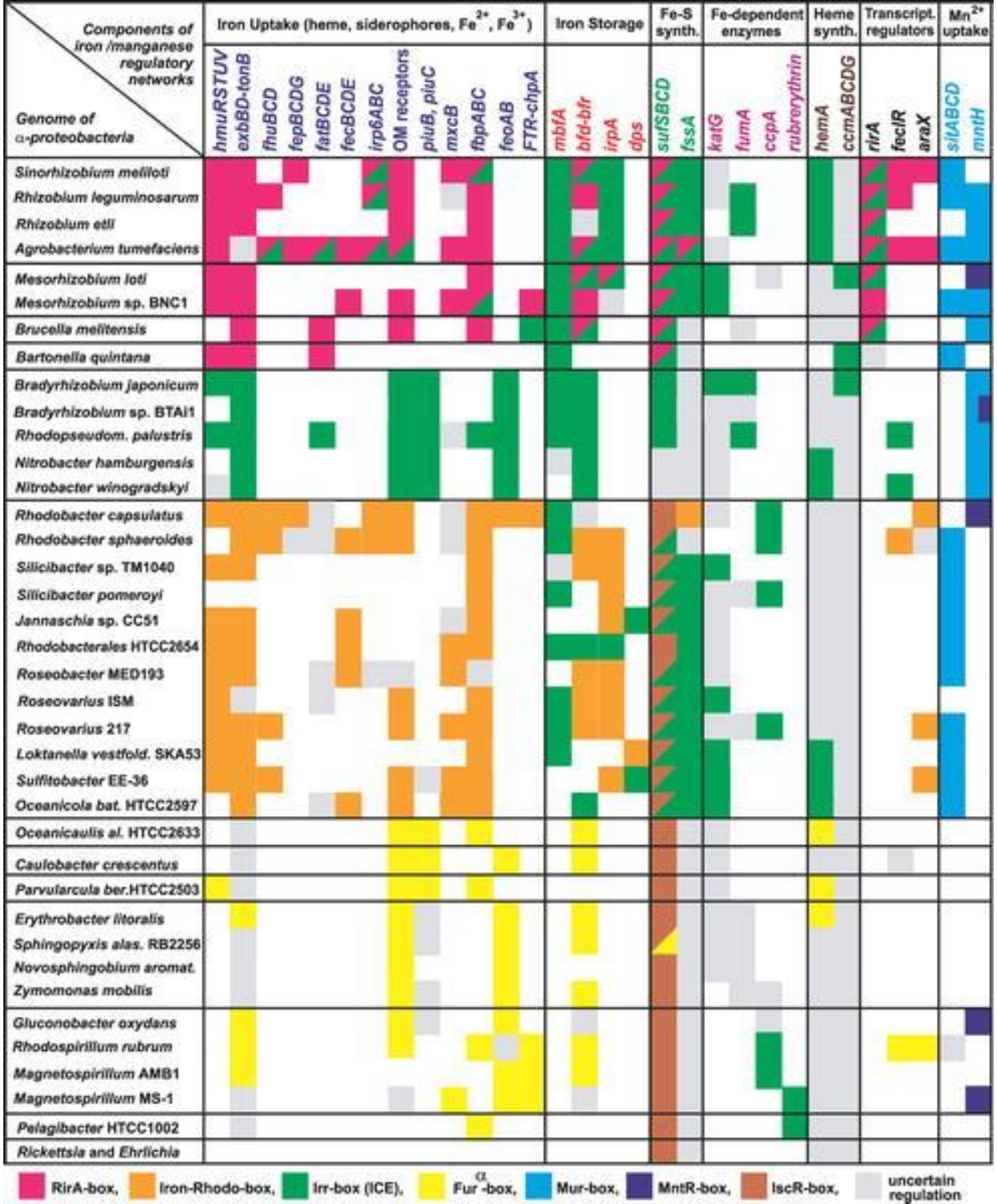
Regulation of genes in functional subsystems

Rhizobiales

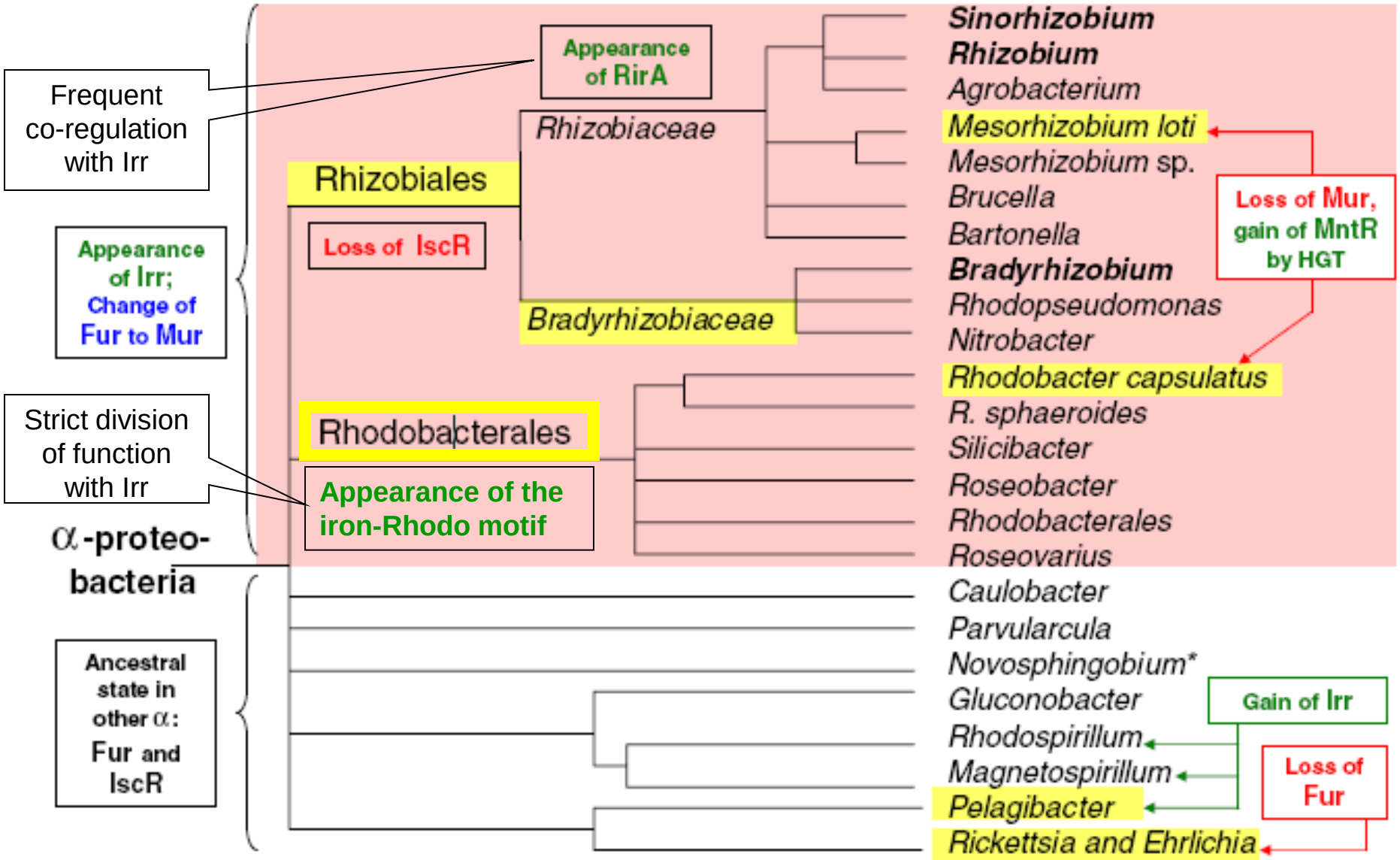
Bradyrhizobiaceae

Rhodobacteriales

The Zoo (likely ancestral state)

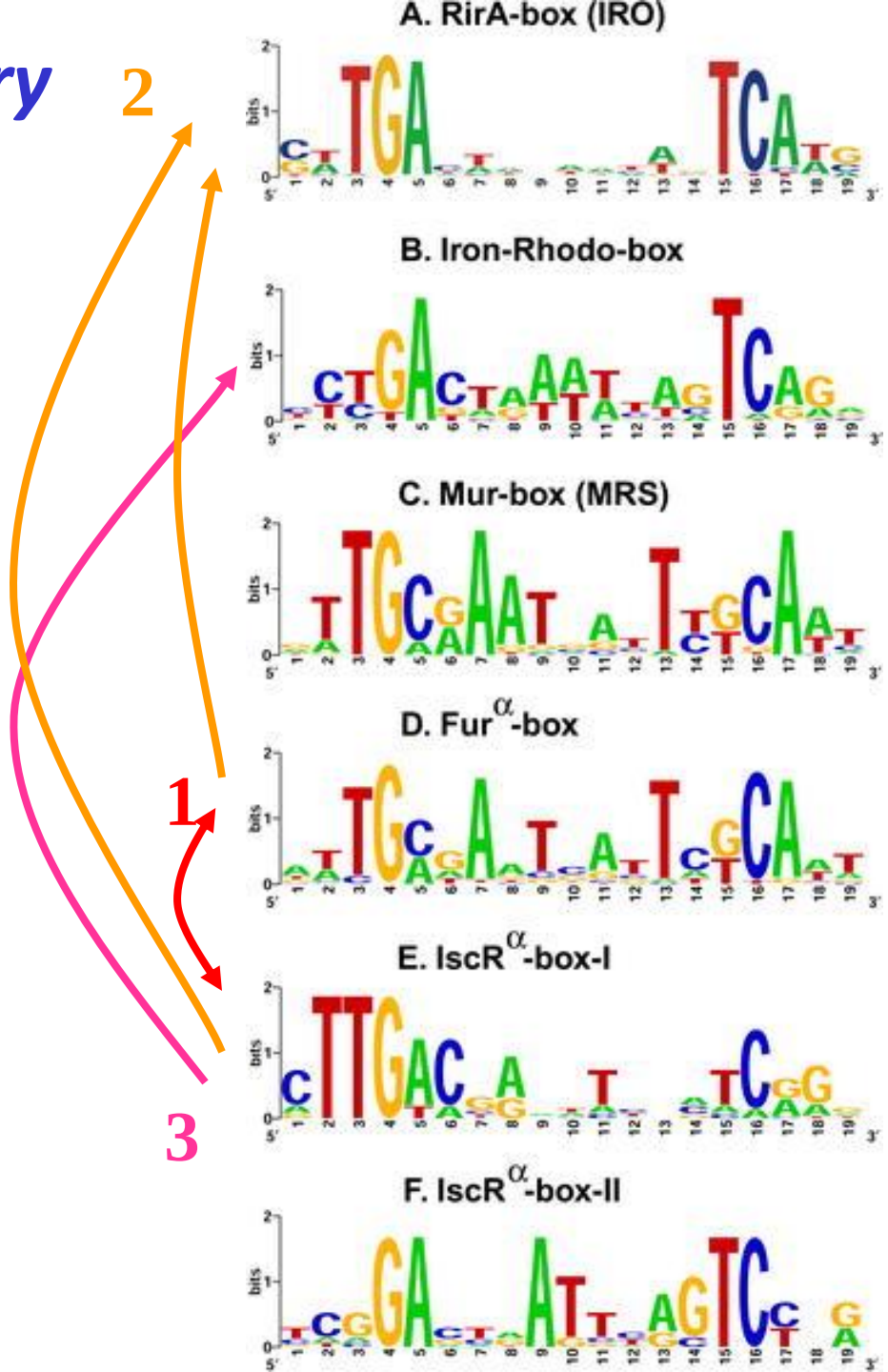


Reconstruction of history

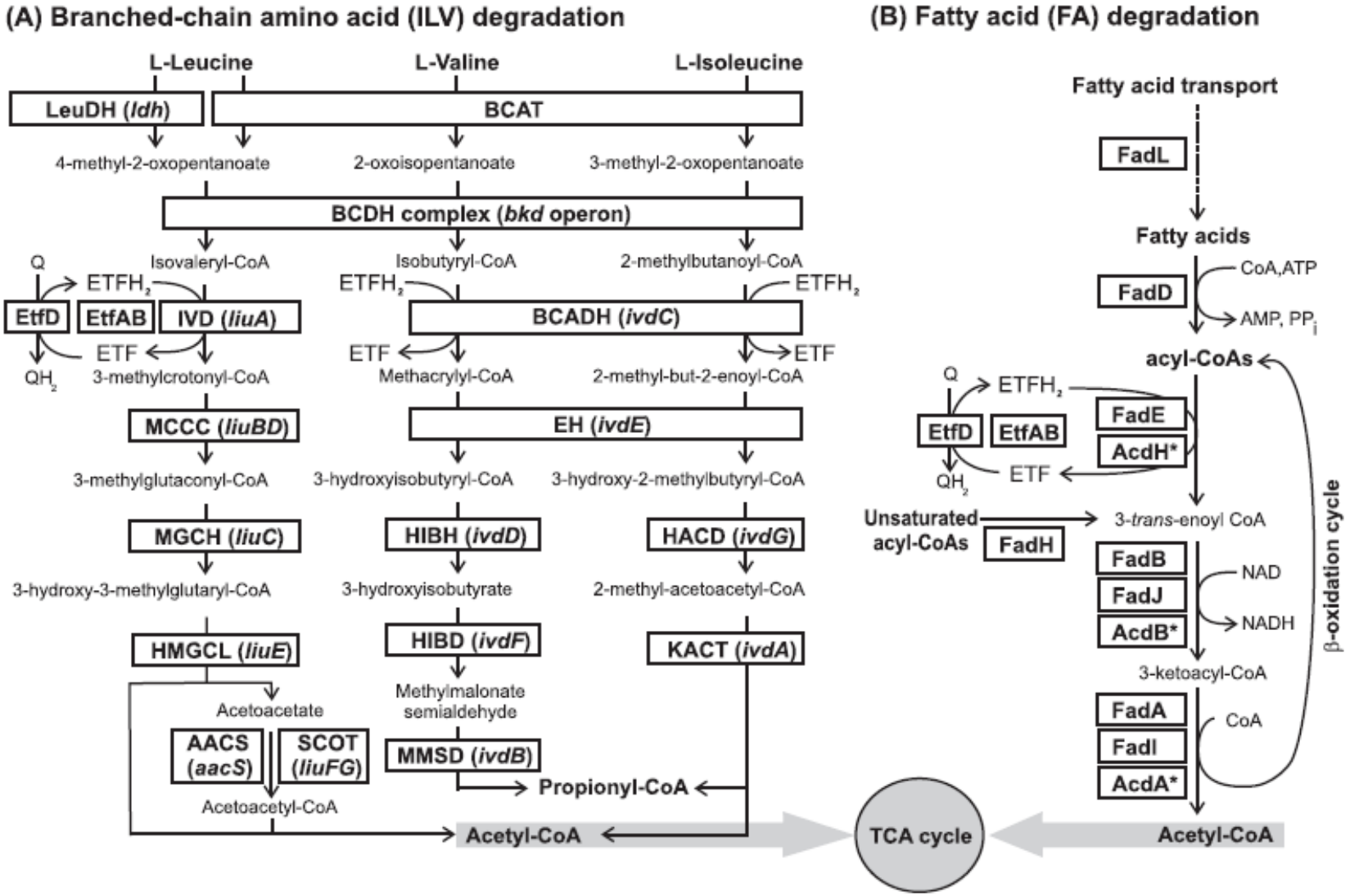


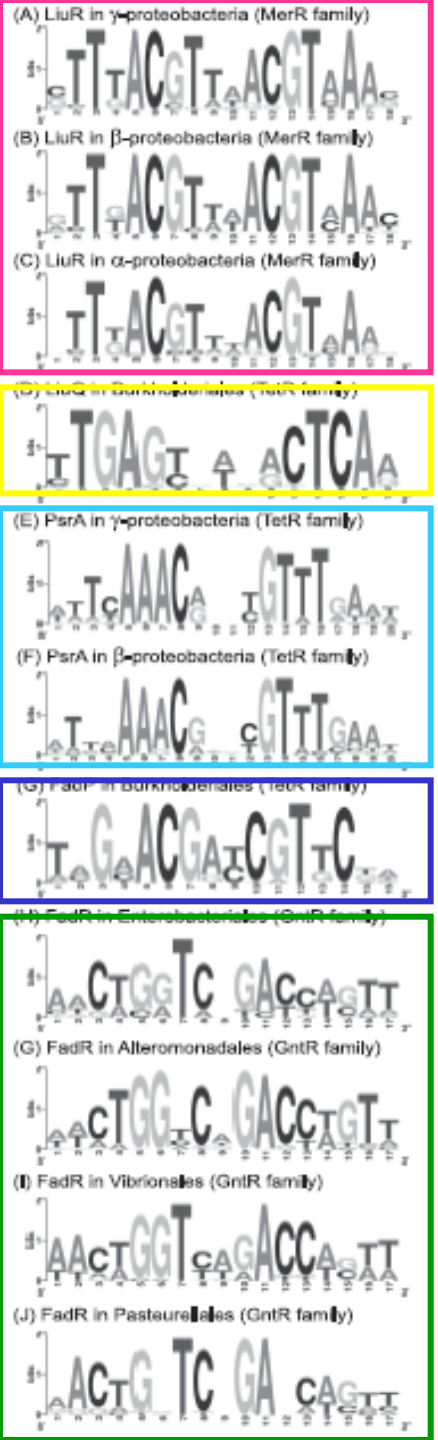
All logos and *Some Very Tempting Hypotheses*:

1. Cross-recognition of FUR and IscR motifs in the ancestor.
2. When FUR had become MUR, and IscR had been lost in Rhizobiales, emerging RirA (from the Rrf2 family, with a rather different general consensus) took over their sites.
3. Iron-Rhodo boxes are recognized by IscR: *directly testable*
   - **Update**: seems to be correct



Large-scale restructuring: Catabolism of branched chain amino acids and fatty acids in gamma- and beta-proteobacteria





MerR

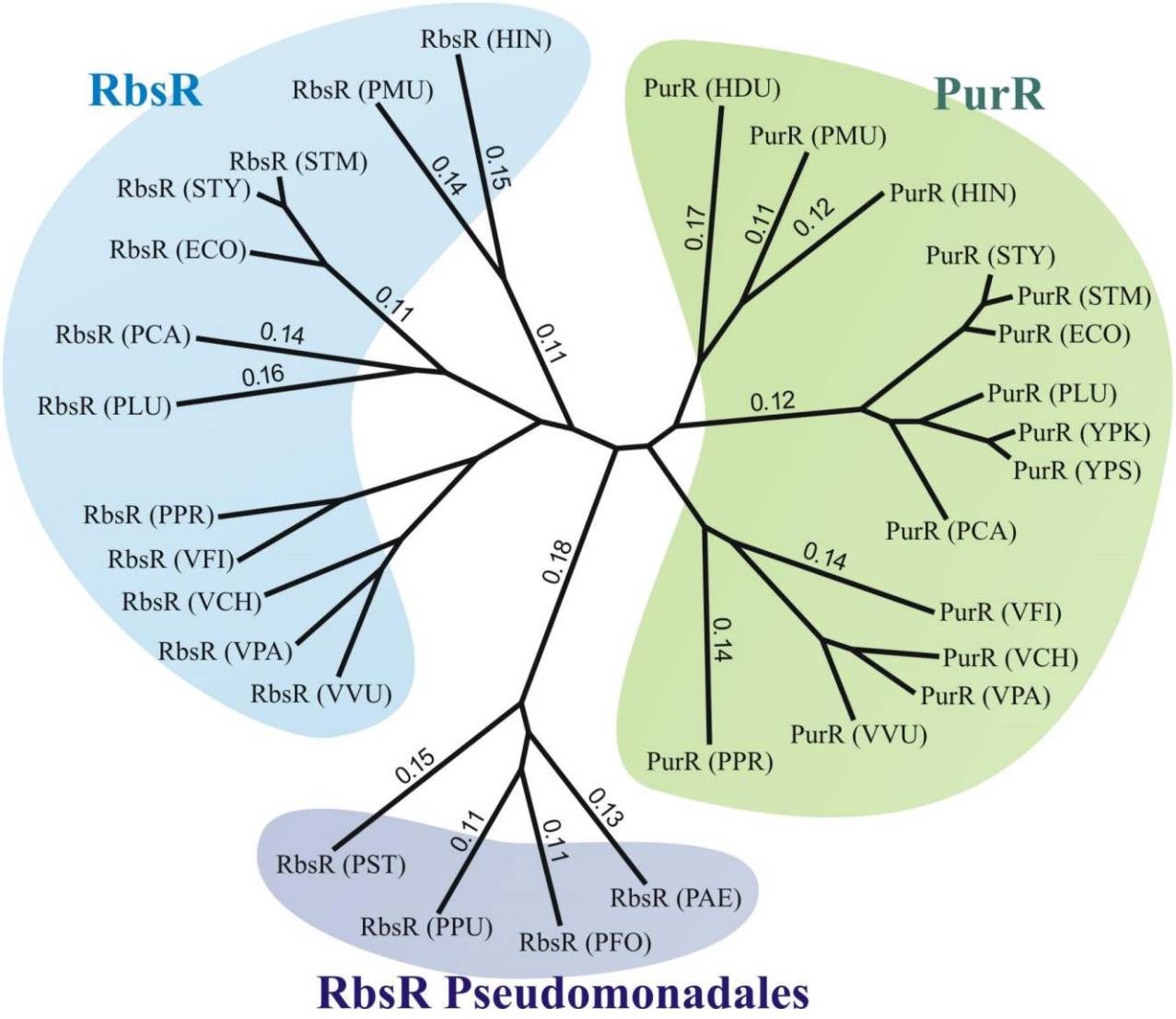
TetR

GntR

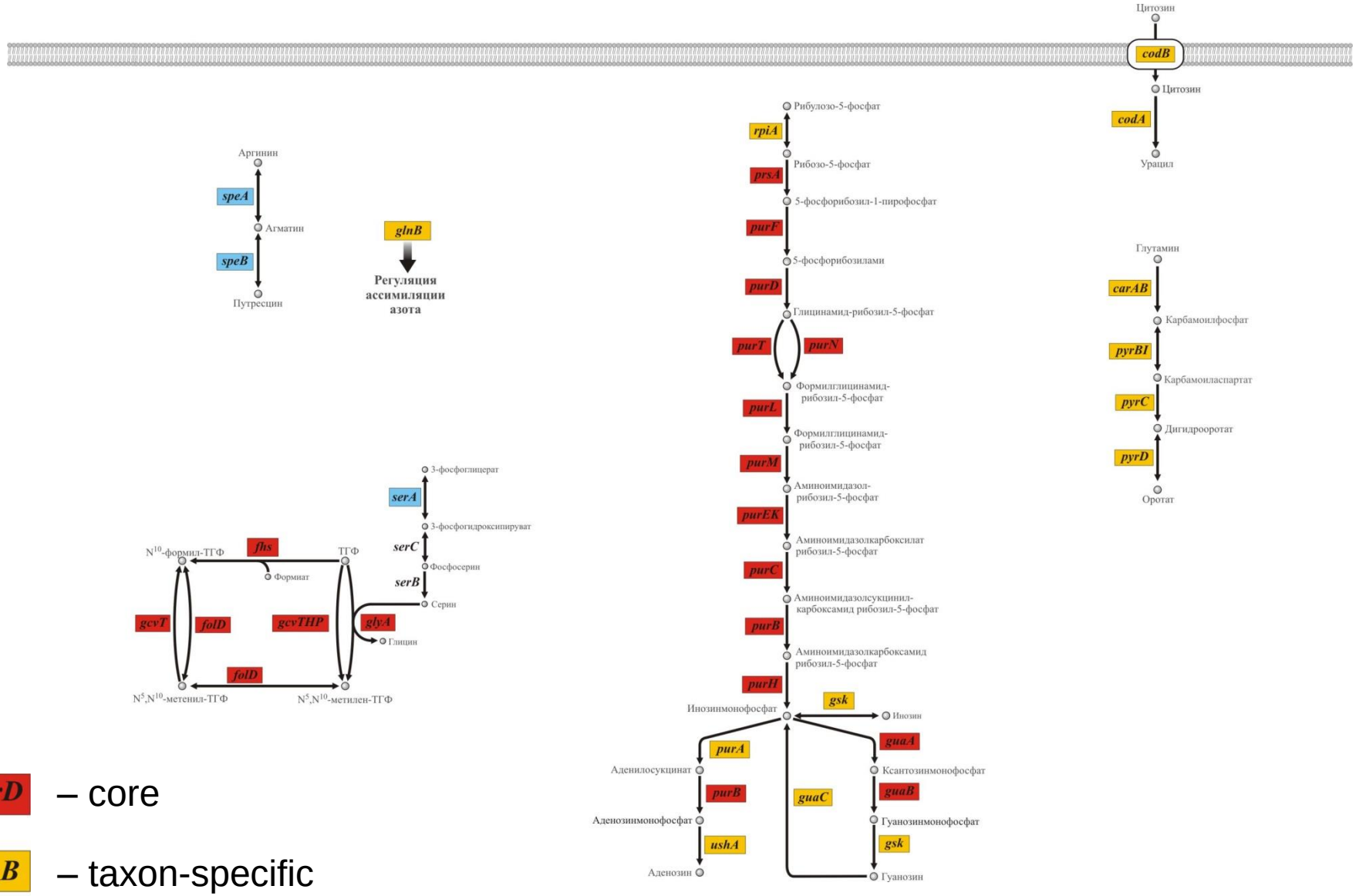
Components of ILV / FA degradation regulatory networks / Genome of γ-, or β-proteobacteria	ILV degradation									ETF		FA degradation							
	liuA	*liuBCDE*	*liuFG*	*aacS*	*ivdA*	*ivdC*	*ivdG*	*ivdBDEF*	*bkd*	*ldh*	*etfBA*	*etfD*	*fadBA*	*fadH*	*fadD*	*fadIJ*	*fadE*	*fadL*	*acdAB*	*acdH*
Enterobacteriales (5 species)																				
Pasteurellales (7 species)																				
Shewanella spp.																				
Idiomarina loihiensis																				
Colwellia psychrerythraea																				
Pseudoalteromonas haloplanktis																				
Pseudoalteromonas atlantica																				
Saccharophagus degradans																				
Vibrio cholerae																				
Vibrio fischeri																				
Vibrio parahaemolyticus																				
Vibrio vulnificus																				
Photobacterium profundum																				
Pseudomonas aeruginosa																				
Pseudomonas putida																				
Pseudomonas fluorescens																				
Pseudomonas syringae																				
Pseudomonas entomophila																				
Xanthomonadales (3 species)																				
Hahella chejuensis																				
Alcanivorax borkumensis																				
Chromohalobacter salexigens																				
Chromobacterium violaceum																				
Dechloromonas aromatica																				
Azoarcus sp.																				
Bordetella (3 species)																				
Ralstonia solanacearum																				
Ralstonia eutropha																				
Ralstonia metallidurans																				
Burkholderia xenovorans																				
Burkholderia (4 species)																				
Methylibium petroleiphilum																				
Polaromonas sp.																				
Rhodoferax ferrireducens																				

LiuR LiuQ FadR PsrA FadP unknown regulation

RbsR and PurR: duplication and subsequent change of specificity



PurR regulon



purD – core

glnB – taxon-specific

speA – *E.coli*-specific (experimental)

Evolution of DNA motifs

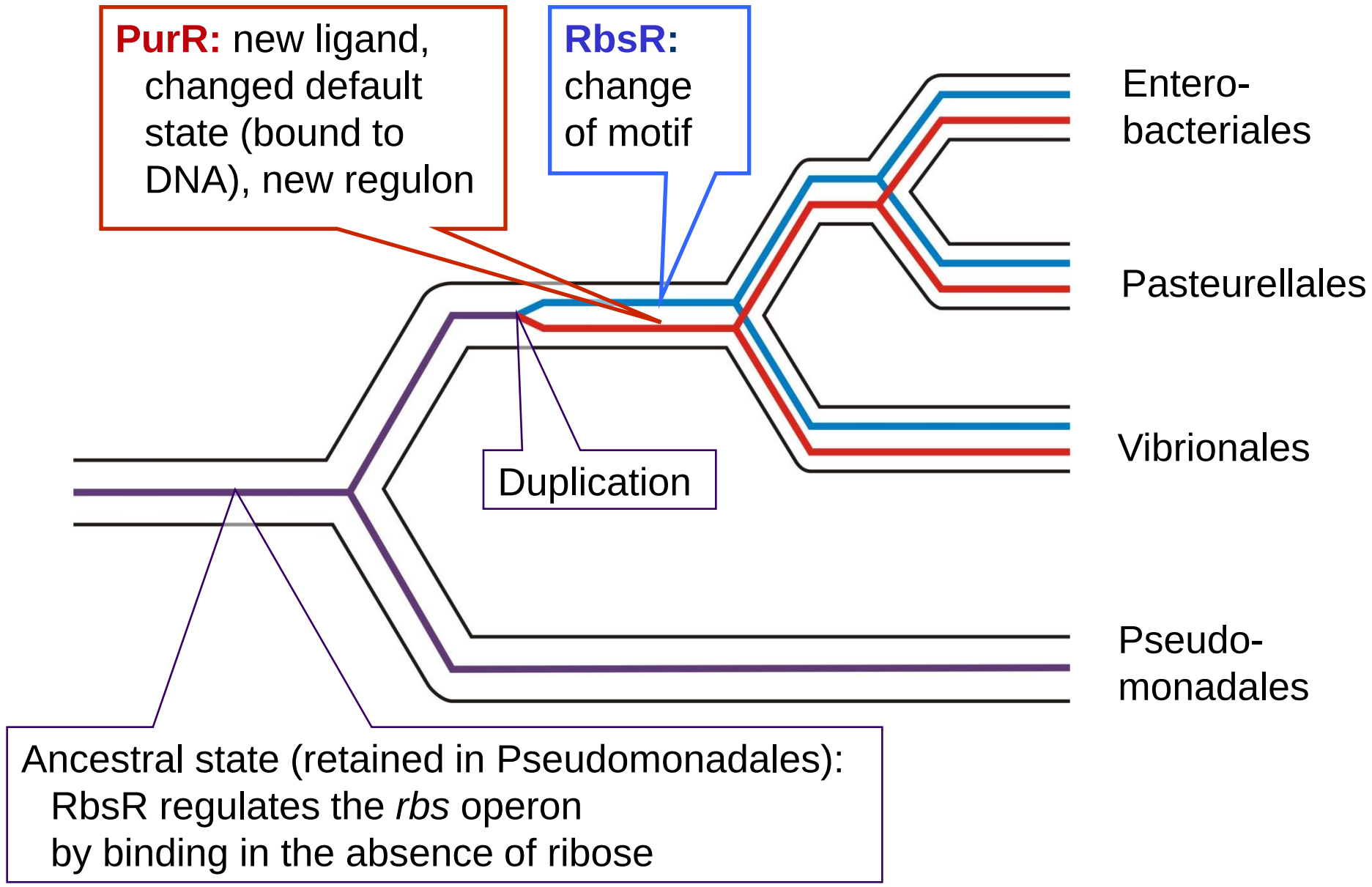
Different Conserved Different

PurR

RbsR
Pseudomonadales

RbsR

History


PurR: new ligand, changed default state (bound to DNA), new regulon
RbsR: change of motif
Entero-bacteriales
Pasteurellales
Vibrionales
Pseudo-monadales
Duplication
Ancestral state (retained in Pseudomonadales):
RbsR regulates the *rbs* operon
by binding in the absence of ribose

Summary and open problems

- Regulatory systems are very flexible
  - easily lost
  - easily expanded (in particular, by duplication)
  - may change specificity
  - rapid turnover of regulatory sites
- … yielding significant changes in genome functioning
- With more stories like these, can we start thinking about a general theory?
  - catalog of elementary events; how frequent?
  - mechanisms (duplication, birth e.g. from enzymes, horizontal transfer)
  - conserved (regulon core) and non-conserved (regulon peryphery) genes in relation to metabolic and functional subsystems/roles
  - (TF family-specific) protein-DNA recognition code

Учебно-Научный Центр

Биоинформатика

- **Andrei A. Mironov** – software, algorithms

- Alexei Kazakov (IITP, LANL)– branched chain amino acids and fatty acids
- * Olga Kalinina (Saarbrucken University) – SDP
- Yuri Korostelev – protein-DNA correlations
- * Olga Laikova – LacI
- * Alexandra Rakhmaninova – SDP, protein-DNA correlations
- * Dmitry Ravcheev (University of Luxembourg) – CRA/FruR, PurR/RbsR
- Dmitry Rodionov (IITP, Burnham Institute) – NrdR, iron, fatty acids etc.
- Olga Tsoy – CRA/FruR
- Ilya Zharov – MerR

- Andy Jonson (U. of East Anglia) – experimental validation (iron)
- Eric Alm (MIT) – experimental validation (CRP/FNR)
- Leonid Mirny (MIT) – protein-DNA, SDP

- Russian Foundation of Basic Research
- Russian Academy of Sciences, program “Molecular and Cellular Biology”
- Russian Science Foundation

ЭВОЛЮЦИЯ

просветительский
фонд

В любой
непонятной ситуации

– ЭВОЛЮЦИОНИРУЙ!

- http://evolutionfund.ru/
- https://planeta.ru/campaigns/evolutionhelp

template

- text