

Учебно-Научный Центр



Биоинформатика



# Эволюция бактериальных систем регуляции транскрипции

*Михаил Гельфанд*

ИПТТИ РАН и ФББ МГУ

Казань, 22 XI 2013

# 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	GCCACGCAACCGTTTTCTTGC
pyrD	GTTCGGAAAACGTTTGCGTTTT
purT	CACACGCAAACGTTTTCGTTTA
cvpA	CCTACGCAAACGTTTTCTTTTT
purC	GATACGCAAACGTGTGCGTCTG
purM	GTCTCGCAAACGTTTGCTTTCC
purH	GTTGCGCAAACGTTTTCGTTAC
purL	TCTACGCAAACGGTTTCGTCGG
consensus	ACGCAAACGTTTTCGT

# Positional weight matrix

<i>j</i>	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

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

# Phylogenetic footprinting

## *rbs* operon in Enterobacteriaceae

```
b3748      -----TAATCACCAT----GTAAACGTTTCGAGGTTGATCACATTTCCGTAACGTCAC
KP_4306    -----CTGTCGCTGCCTC-GCGAAACGTTTCGATGGCGATCACATTTCTCTCTTCTGGT
SMA4113    -----TTTTCCACGCGCGAACGAAACGTTTCGATAGCGATCACACTTCTGCATTGTCCC
YE0008     -----TTTCATTTGTTTCGGCGAAACGTTTCGATGGCGATCACAAATTTCAACCAATTGG
YP0007     TCCTTCTTCTTATATCGCTAGCAAAGTGTTCGGTGGCGATCACAAATTTCACTAAATGAG
ECA0010    TTACCTTTCTTTTTTTGTTAGCGAAACGTTTCGATGGCGATCACATTTTTTTTATTCTT
           *                **   *      *      *      *      *      *
```

```
b3748      GATGGTTTTCCCAACTCAGTCAGGATTAAACTGTGGGTCAGCGAAACGTTTCGCTGATGG
KP_4306    GATGGTTTTCT--GCTCACACATTGATAATAATTATTTTAGCGAAACGTTTCGCTAGTGG
SMA4113    GGTTGCCTTCCCCTGCCGTTTTTTAAACTCCTCCAGAGAGCGAAACGTTTCGCTAGCGG
YE0008     GTTTGCCTTCTGCTGCCATTTTTCTAAACTCAGT--ATCAGCGAAACGTTTCGCTGTTGG
YP0007     GTTTGTCTTCTACTGCCGTTTTTTCTAAACTCCTT--GTTAGCGAAACGTTTCGCTCTTGG
ECA0010    CGTTGCCTTTCCCACTTCTTTTTCTAGAATGTT--GTTAGCGAAACGTTTCGCTGGTGG
           * *  **                *                        *      *
```

```
b3748      AG-----AAAAAAATGAAAAAAGGCACCGTTCTTAATTCTGATATTTTCATCGGTGATC
KP_4306    AGCAAAAGAGAAAAAAATGAAAAAAGGCACCGTACTTAACGCTGATATTTCCGCGGTCATT
SMA4113    AGT-----GAGAAGATGAAAAAAGGCATTAATTCTGATATTTTCATCGGTGATC
YE0008     AGT-----AGAAAATGAAAAAAGGTGCATTACTAAATTCTGATATTTCCGCTGTTATC
YP0007     AGT-----AGATCATGAAAAAAGGTGTATTACTGAACGCTGATATTTCCGCGGTTATC
ECA0010    GGT-----GAGAAGATGAAAAAAGGCAGCATTATTGAATTCAGATATTTCTTCCGTGATT
           *                * *  *      *      *      *      *      *      *
```

Start codon of *rbsD*

# Phylogenetic footprinting

*rbs* operon in Enterobacteriaceae  
... regulated by CRP and RbsR

## CRP binding site

b3748  
KP\_4306  
SMA4113  
YE0008  
YP0007  
ECA0010

```
-----TAATCACCAT----GTAAAACGTTTCGAGGTTGATCACATTTCCGTAACGTCAC
-----CTGTCGCTGCCTC-GCGAAACGTTTCGATGGCGATCACATTTCTCTCTTCTGGT
-----TTTTCCACGCGCGAACGAAACGTTTCGATAGCGATCACACTCTGCATTGTCCC
-----TTTCATTTGTTTCGGCGAAACGTTTCGATGGCGATCACAAATTCACCCAATTGG
TCCTTCTTCTTATATCGCTAGCAAAGTGTTCGTTGGCGATCACAAATTCACTAAATGAG
TTACCTTTCTTTTTTTGTTAGCGAAACGTTTCGATGGCGATCACATTTTTTTATTCTT
          *          ** ***** ***** **          *
```

b3748  
KP\_4306  
SMA4113  
YE0008  
YP0007  
ECA0010

```
GATGGTTTTCCCAACTCAGTCAGGATTAAACTGTGGGTCAGCGAAACGTTTCGCTGATGG
GATGGTTTTCT--GCTCACACATTGATAATAATTATTTTAGCGAAACGTTTCGCTAGTGG
GGTTGCCTTCCCCTGCCGTTTTTTAAACTCCTCCAGAGAGCGAAACGTTTCGCTAGCGG
GTTTGCCTTCTGCTGCCATTTTTCTAAACTCAGT--ATCAGCGAAACGTTTCGCTGTTGG
GTTTGTCTTCTACTGCCGTTTTTTCTAAACTCCTT--GTTAGCGAAACGTTTCGCTCTTGG
CGTTGCCTTCCCCTTCTTTTTCTAGAATGTT--GTTAGCGAAACGTTTCGCTGGTGG
          * * **          *          ***** **
```

## RbsR binding site

b3748  
KP\_4306  
SMA4113  
YE0008  
YP0007  
ECA0010

```
AG-----AAAAAATGAAAAAAGGCACCGTTCTTAATTCTGATATTTTCATCGGTGATC
AGCAAAAGAGAAAAAATGAAAAAAGGCACCGTACTTAACGCTGATATTTCCGCGGTCATT
AGT-----GAGAAGATGAAAAAAGGCGTATTACTGAATTCCGACGTGTCTGCCGTGATT
AGT-----AGAAAATGAAAAAAGGTGCATTACTAAATTCTGATATTTCCGCTGTTATC
AGT-----AGATCATGAAAAAAGGTGTATTACTGAACGCTGATATTTCCGCGGTTATC
GGT-----GAGAAGATGAAAAAAGCAGCATTATTGAATTCAGATATTTCTTCCGTGATT
          *          * * ***** * * ** * ** * ** **
```

Start codon of *rbsD*

# Many sites (*nrd*): FNR, DnaA, NrdR

A.

```

EC  TGCTTTTTACTTTGAGCTACATCAAAAAAAGCTCAAACATCCTTGATGCAAAAGCACTTATATATAGACTTTAAAATGCGTCCCAACCCAATATGTTGTATT
ST  TGATTTTTACCTTGTTCTACATCAATAAAATTGCAAACATCCTTGATGCAAAATCACTTACATATAGACTTTAAAATGCGCAGCCGACCCAATATGTTGTATT
KP  ACCTTTTTACCTTGTTCTGGGTCAATAAAATCGCAAACATCCTTGATGCAAAATCACTTACATATAGAACTTAAAATGCGCCTCGGCCCAACATATTTGTATT
      *****  ***   **   *****  *****  *****  *****  *****  *****  *****  *   *****  **  *****

EC  AATCGACTATAATTGCTACTACAGCTCCCCACG--AAAAGGTGCGGCGTTGTGGATAAGC--GGATGGCGATTGCGGA--AAGCACCGGAAAACGAAACGA
ST  AATTGACTACAATTGCTACAACACCTGTTCACT--CGACACAAGGTGAATTGTGGATAACCTGGGTGAGATTGCGGG--AAGTCATTGGAAAAGAGATGA
KP  AATCGTCTATTAT-GTCACCATATCTTGTGCGATGTCTGGCGGTGATGAGATGTGGATAAAACGGGCCGGATCCGAAGGTAAACAGCACGAGCCGTAGCGT
      *** * ***  ** *  ** *  * **   *   *  *  *****  **   *  *  **   *  *  *

EC  AAAAACCGGAAAACGCCTTTCCCAATTTTGTGGATAACCTGTTCTTAAAAATATGGAGCGATCATGACACCGCATGTGATGAAACGAGA
ST  ATAAACCTGTTA-TGCTTTCCCGGCCTCTGTGGATAACCTGTTCTTACAAATATGGAGTGATCATGACACCGCATGTGATGAAACGAGA
KP  GCAGCGCCTTCG-GGATAACCTCCGCCTCTGTGGATAACCTGTTCT---ATATATGGAGTGATCATGACACCGCATGTGATGAAACGAGA
      *  *  *  *  *  *  *****  *****  *  *****  *****  *****  *****  *****  **
  
```

B.

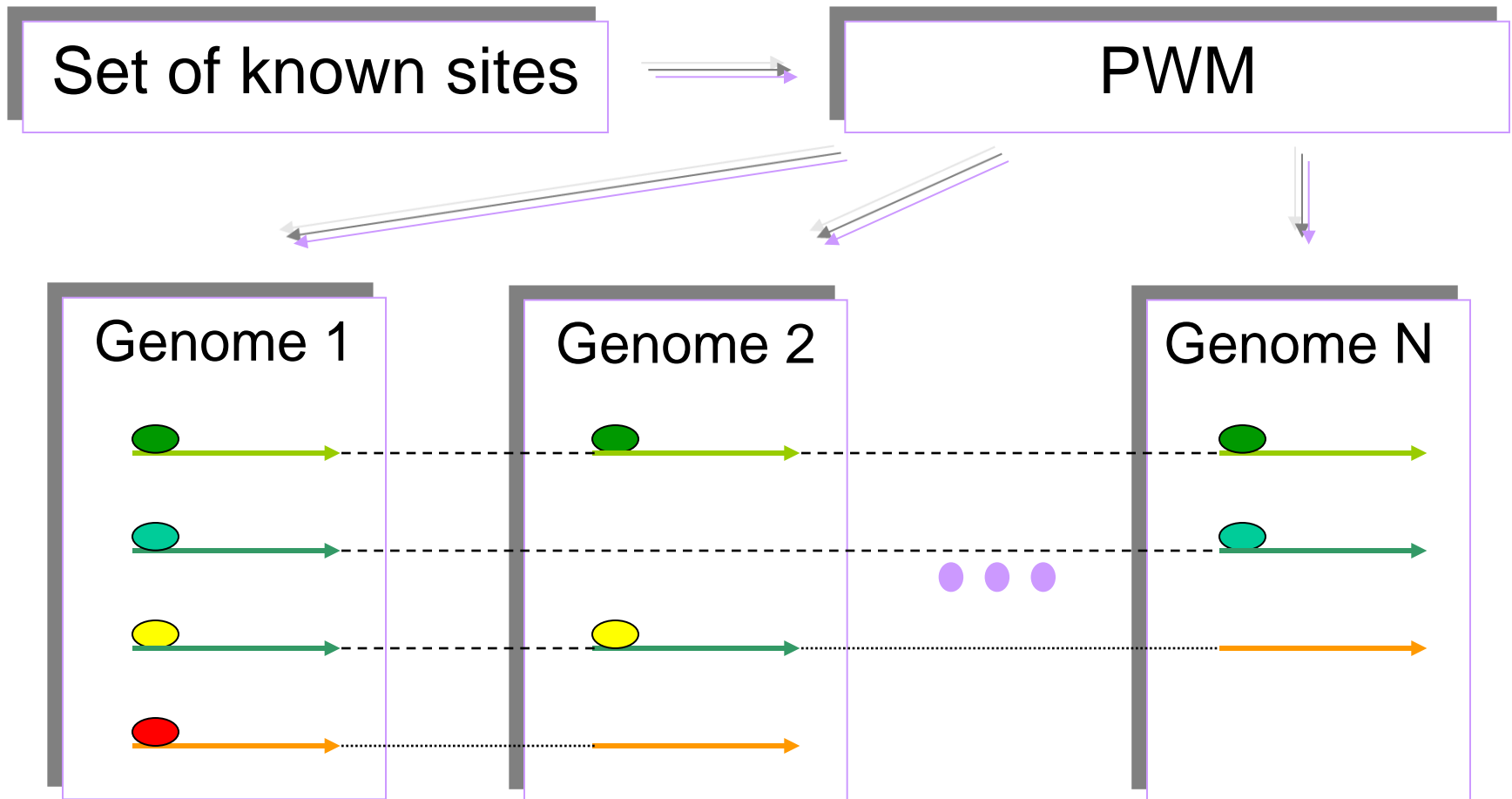
```

YP  AACAGGGAATAACCC-TAACGCC--AATTTCCTTGTTCTAGGTCAACAAATATTGGCTATCAGTTGACTGTCACTCA TCCAGATTACCCATATATAGTGTCT
YE  AACAGGGAATAAACC-TAAAGCT--GATTTCCTTGTTCTAGGTCAATTA-----GTTGACTGTCACTTC TGCCATTACCCATATATAGTGTCT
Eca  AAGTTCGATTTATCTACTAGGGAGGAATTTCCTTGCTCTACATCAATTTTGCAGCGATAAAAGTGC AAACACCCCTACGCAATTTCAATATATAGTGCCT
Ech  AAGACTGATTTCTCTACGATGCCGAATTTCCTTGAGCCAGGTCAATTCTAACGCAATAAAACCGGGTCCCCCTCCAGGCGAATTCAATATATAGTGTCT
      **   ** *  *  *  *  *****  ***  *  *  *****  *   *  *   *  *****  **

YP  ATAATATTTTAAGCATCTATATGTAGTAGTTATCCACAAAAGCATCCACATCCCCCTCGCAGCCCTGATGTGCTGCGGGTTGC-CTTGTGGATAA-----
YE  ATAGTAATTACGATACCTATATGTAGTAGTTATCCACAAAACCATCCACA-CCCCCTCGCAGCCCTGATGTGCTGCGGTTTGC-CTTGTGGATAAGATGG
Eca  ATCCTGTAAACATTACCTACATATAGTGTTTATCCACAAAGTCATCCACA-GCCCTCTGTAACCCTTGCCAGTTACGGTTCTCGCCTGTGGATAAC----
Ech  ATCCTTTTACAAGAACCTACATATTTGTGTTTATCCACAAGAACATCCACA-GCC-TCCGCACCCGTTGTGACCCGCGGCTTCCGTCTGTGGATAAC----
      **  *  *  ***  ** *  **  *****  *****  **   *  *  ** *  *  *  *****

YP  -----CCCTATGCGGCGGTATACAGGAGTGACATTGTGAAAACAGTAGTGATTAAACGGGACGGCTGCCAGGT
YE  TTTTGGGGCTAATCCTACGCGGCAGGATACAGGAGCGACATTGTGAAAACAGTAGTGATTAAACGGGACGGTTGTCAGGT
Eca  -----CTTTCCAG-----AGGAAGAA-AACGTGAAACCAGTAGTGATTAAACGGGACGGTTGCCAGGT
Ech  -----ATCAACAAAGGAAGAACACCGAGGAACAAC---ATGAAACCAGTAGTGATTAAACGGGACGGATGTCAGGT
      *****  *  *****  *****  *****  *****  **  *****
  
```

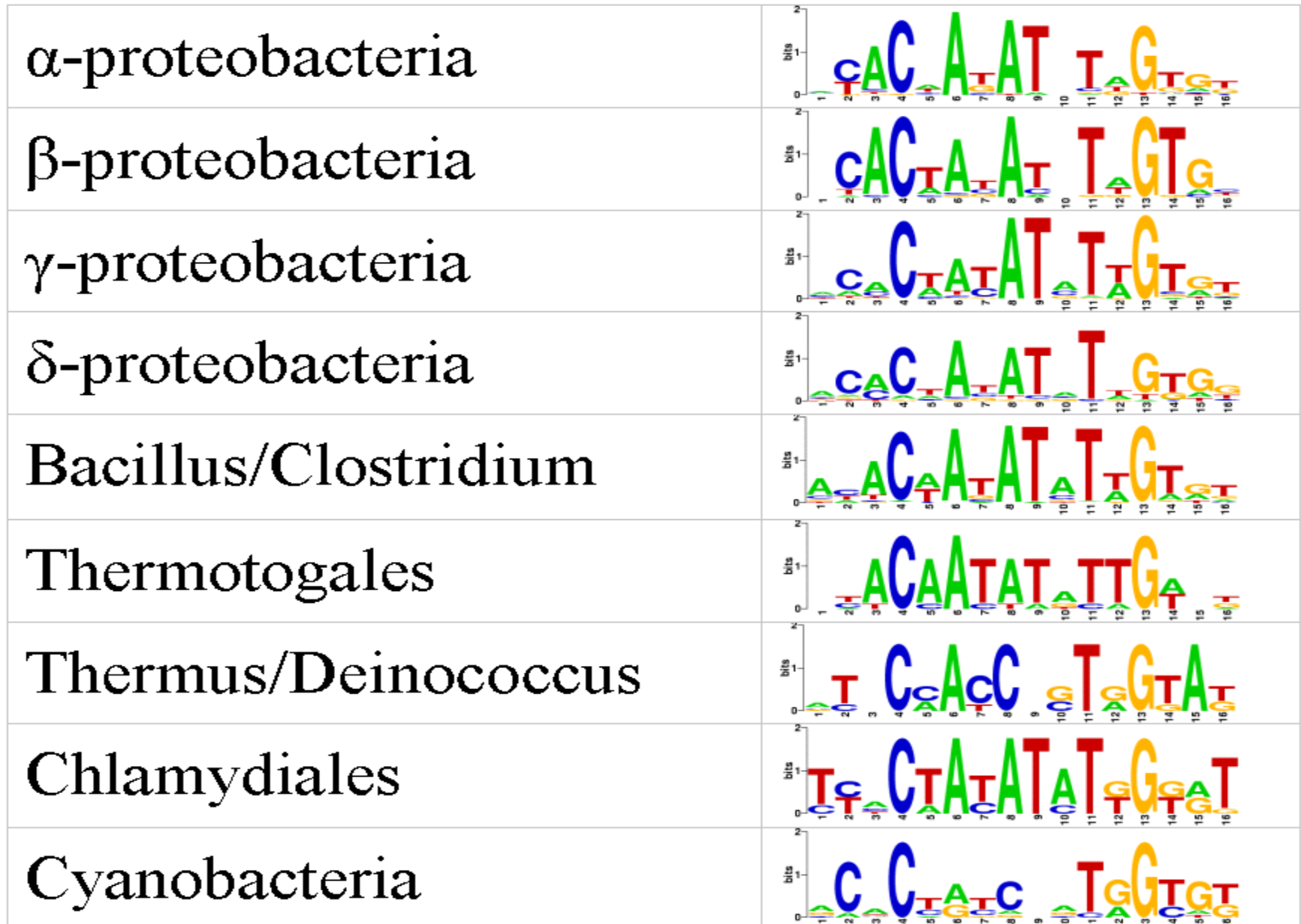
# Conservation of regulation => consistency check





0. Practical use (just one example)

# Conserved motif upstream of *nrd* genes

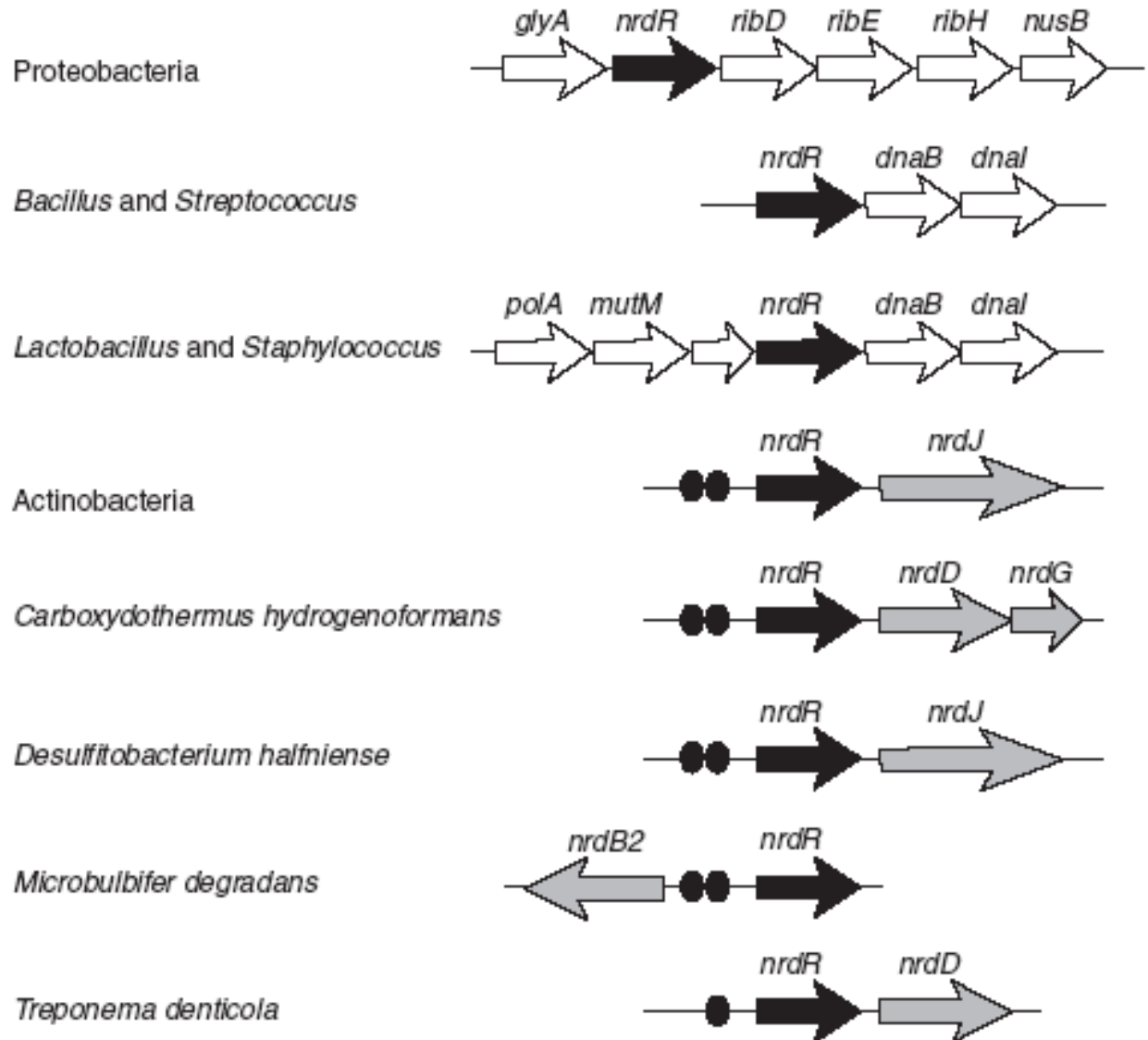


# *ybaD* is regulator of ribonucleotide reductases (*nrdR*)

- 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

# Additional evidence - 1

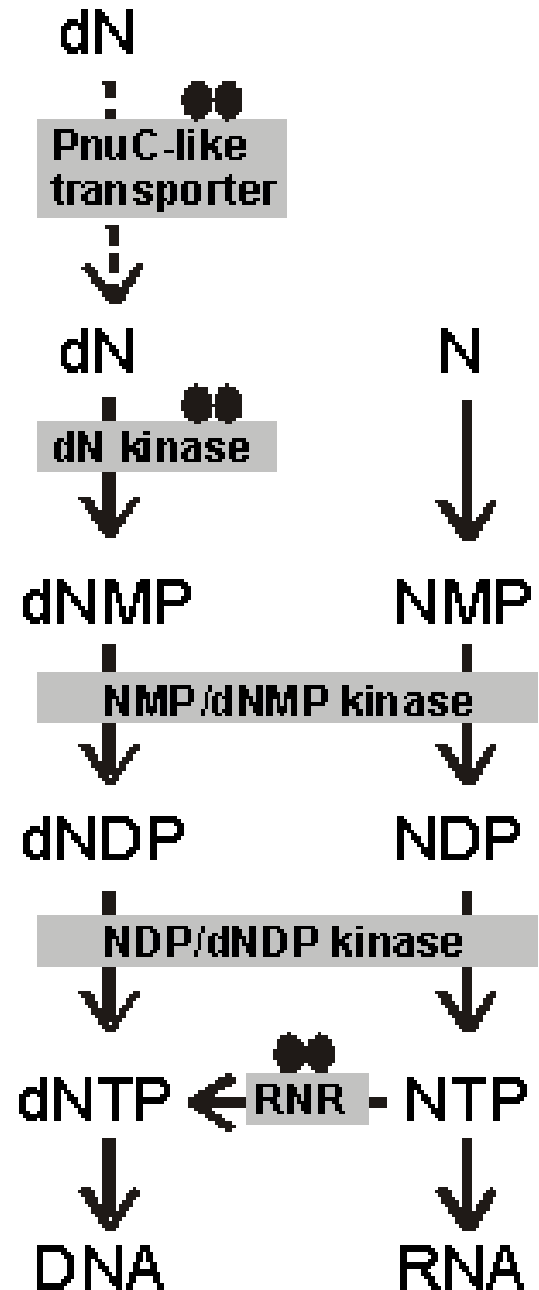
*nrdR* is sometimes clustered with *nrd* genes or with replication genes *dnaB*, *dnaI*, *polA*



## 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

A.

```

EC  TGCTTTTTACTTTGAGCTACATCAAAAAAAGCTCAAACATCCTTGATGCAAAAGCACTTATATATAGACTTTAAAATGCGTCCCAACCCAATATGTTGTATT
ST  TGATTTTTACCTTGTTCTACATCAATAAAATTGCAAACATCCTTGATGCAAAATCACTTACATATAGACTTTAAAATGCGACGCCAACCCAATATGTTGTATT
KP  ACCTTTTTACCTTGTTCTGGGTCAATAAAATCGCAAACATCCTTGATGCAAAATCACTTACATATAGAACTTAAAATGCGCCTCGGCCCAACATATTTGTATT
      *****  ***   **   *****  *****  *****  *****  *****  *****  *****  *   *****  **  *****

EC  AATCGACTATAATTGCTACTACAGCTCCCCACG--AAAAAGGTGCGGCGTTGTGGATAAGC--GGATGGCGATTGCGGA--AAGCACCGGAAAACGAAACGA
ST  AATTGACTACAATTGCTACAACACCTGTTCACT--CGACACAAGGTGAATTGTGGATAACCTGGGTGAGATTGCGGG--AAGTCATTGGAAAAGAGATGA
KP  AATCGTCTATTAT-GTCACCATATCTTGTGCGATGTCTGGCGGTGATGAGATGTGGATAAAACGGGCCGGATCCGAAGGTAAACAGCACGAGCCGTAGCGT
      *** * ***  ** *  ** *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *

EC  AAAAAACCGGAAAACGCCTTTCCCAATTTTGTGGATAACCTGTTCTTAAAAATATGGAGCGATCATGACACCGCATGTGATGAAACGAGA
ST  ATAAACCTGTTA-TGCTTTCCCGGCCTCTGTGGATAACCTGTTCTTACAAATATGGAGTGATCATGACACCGCATGTGATGAAACGAGA
KP  GCAGCGCCTTCG-GGATAACCTCCGCCTCTGTGGATAACCTGTTCT---ATATATGGAGTGATCATGACACCGCATGTGATGAAACGAGA
      *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
  
```

B.

```

YP  AACAGGGAATAACCC-TAACGCC--AATTTCCTTGTTCTAGGTCAACAAATATTGGCTATCAGTTGACTGTCACTCA TCCAGATACCCATATATAGTGTCT
YE  AACAGGGAATAAACC-TAAAGCT--GATTTCCTTGTTCTAGGTCAATTA-----GTTGACTGTCACTTC TGCCATTACCCATATATAGTGTCT
Eca  AAGTTCGATTTATCTACTAGGGAGGAATTTCCTTGTTCTACATCAATTTTGCAGCGATAAAAGTGC AAACACCCCTACGCAATTTCAATATATAGTGCCT
Ech  AAGACTGATTTCTCTACGATGCCGAATTTCCTTGTTCTAGGTCAATTCTAACGCAATAAAACCGGGTCCCCCTCCAGGCGAATTCAATATATAGTGTCT
      **   ** *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *

YP  ATAATATTTTAAGCATCTATATGTAGTAGTTATCCACAAAAGCATCCACATCCCCCTCGCAGCCCTGATGTGCTGCGGGTTGC-CTTGTGGATAA-----
YE  ATAGTAATTACGATACCTATATGTAGTAGTTATCCACAAAACCATCCACA-CCCCCTCGCAGCCCTGATGTGCTGCGGGTTGC-CTTGTGGATAAGATGG
Eca  ATCCTGTAAACATTACCTACATATAGTGTTTATCCACAAAGTCATCCACA-GCCCTCTGTAACCCTTGCCAGTTACGGTTCTCGCCTGTGGATAAC----
Ech  ATCCTTTTACAAGAACCTACATATTTGTGTTTATCCACAAGAACATCCACA-GCC-TCCGCACCCGTTGTGACCCGCGGCTTCCGTCTGTGGATAAC----
      ** *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *

YP  -----CCCTATGCGGCGGTATACAGGAGTGACATTGTGAAAACAGTAGTGATTAAACGGGACGGCTGCCAGGT
YE  TTTTGGGGCTAATCCTACGCGGCAGGATACAGGAGCGACATTGTGAAAACAGTAGTGATTAAACGGGACGGTTGTCAGGT
Eca  -----CTTTCCAG-----AGGAAGAA-AACTGTGAAACCAGTAGTGATTAAACGGGACGGTTGCCAGGT
Ech  -----ATCAACAAAGGAAGAACACCGAGGAACAAC---ATGAAACCAGTAGTGATTAAACGGGACGGATGTCAGGT
      *****  *   *****  *****  *****  *****  *****  *****  *****  *****  *****
  
```

# 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)



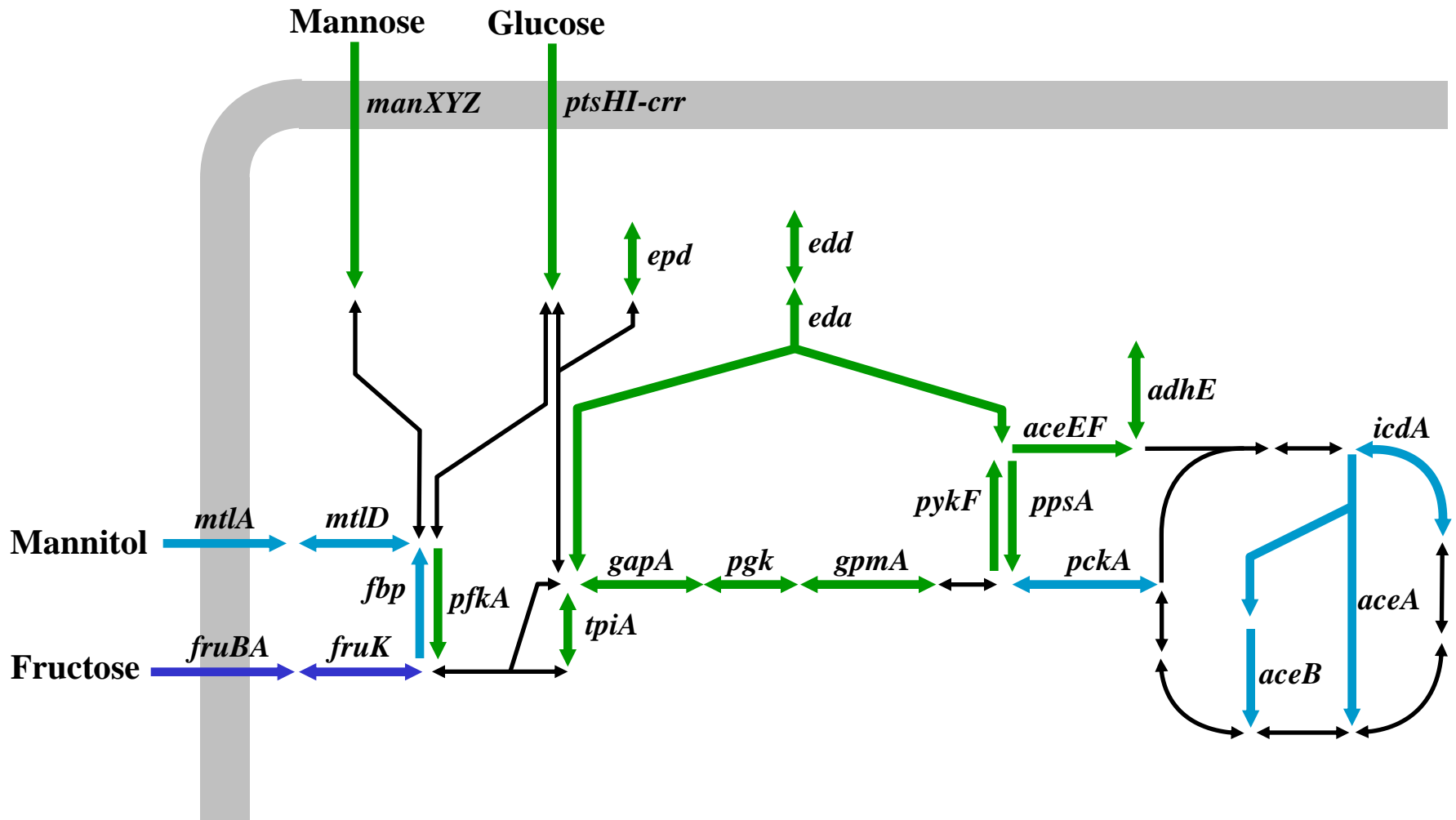
# 1. Evolution of regulatory networks

- 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

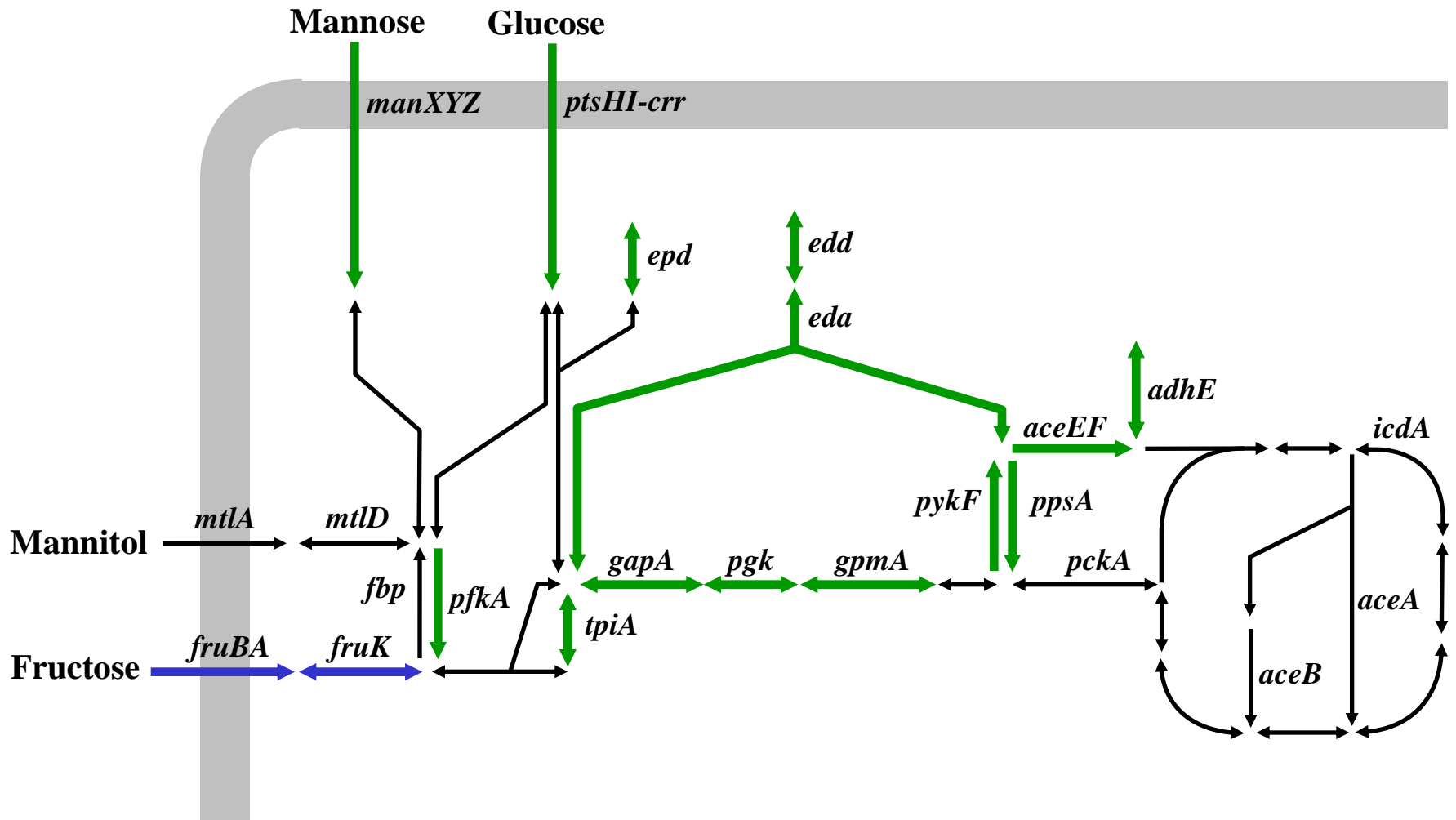
# Regulon expansion, or how FruR has become CRA

- 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*

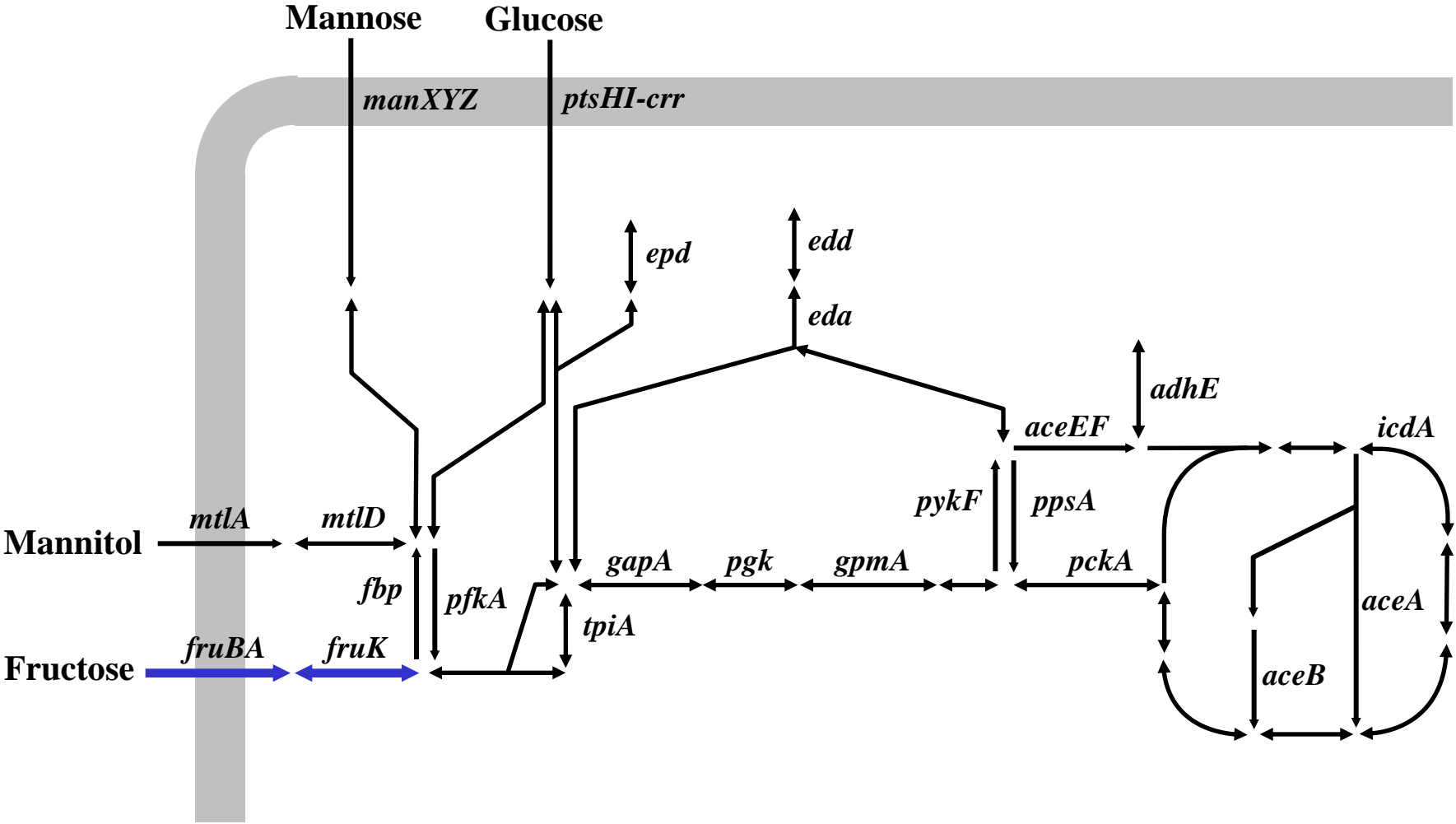
# *E. coli* - experimental data



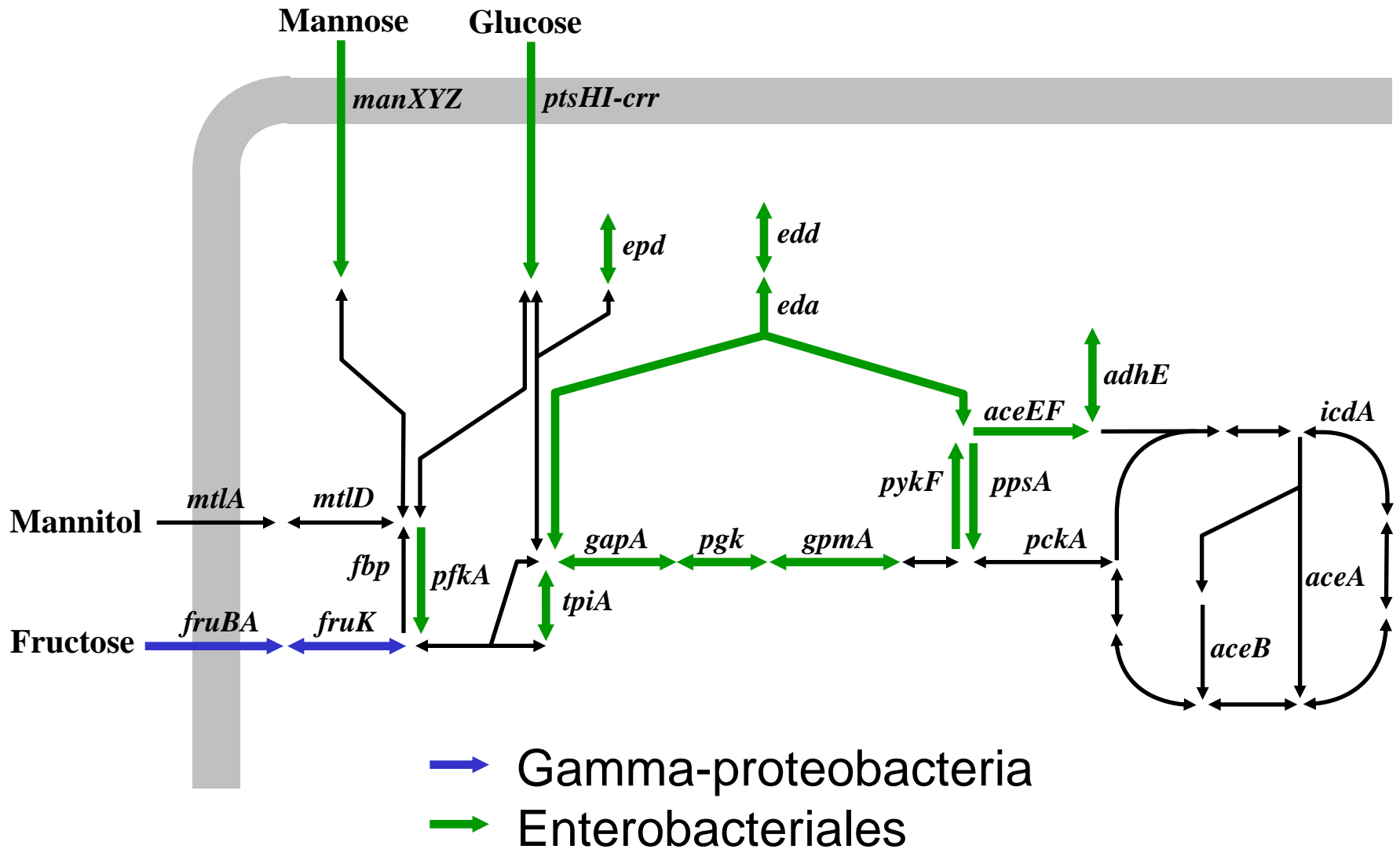
# Sites conserved in the Enterobacteriales



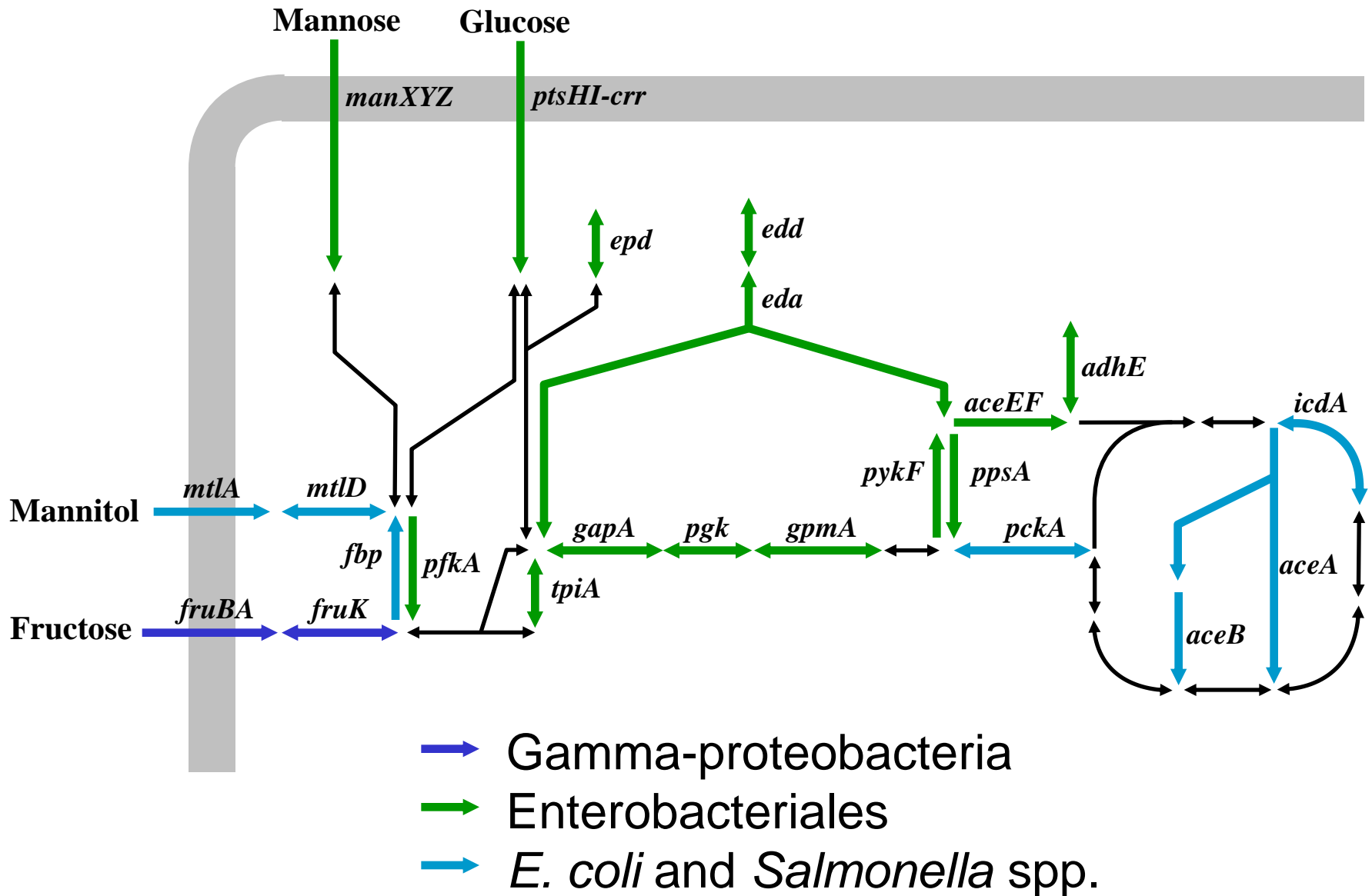
# Common ancestor of the Enterobacteriales and Vibrionales



# Common ancestor of the Enterobacteriales



# Common ancestor of *Escherichia* and *Salmonella*



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

Iron:

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

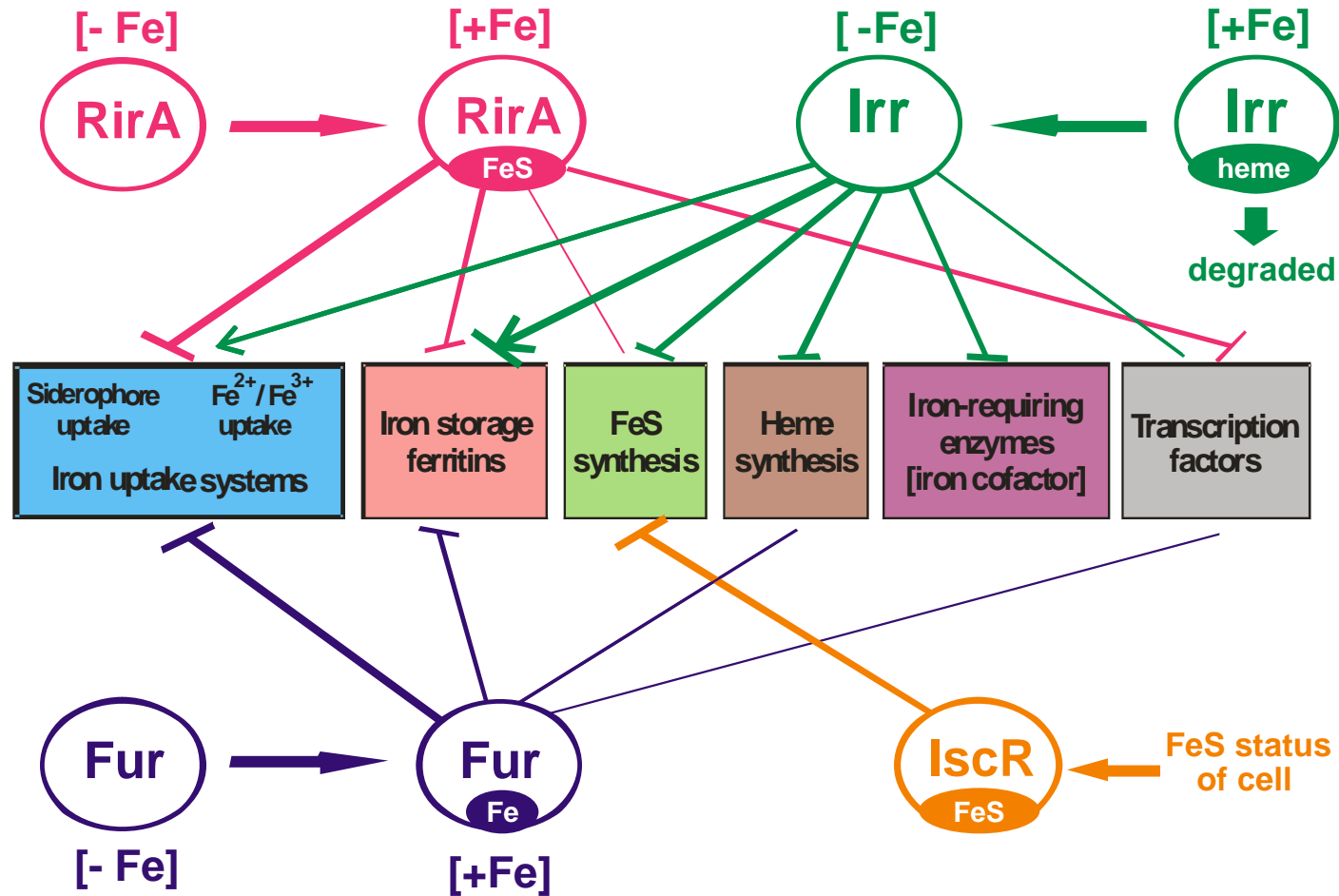
FUR (responds to iron):

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

Similar in *Bacillus subtilis*



# Regulation of iron homeostasis in $\alpha$ -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	
Rhizobiales	Rhizobiaceae	<i>Sinorhizobium meliloti</i>	SM	+	+	-	+	-	
		<i>Rhizobium leguminosarum</i>	RL	++	+	-	+	-	
		<i>Rhizobium etli</i>	RHE	+	+	-	+	-	
		<i>Agrobacterium tumefaciens</i>	AGR	+	+	-	+	-	
		<i>Mesorhizobium loti</i>	ML	+	-	+	+	-	
		<i>Mesorhizobium sp. BNC1</i>	MBNC	+	++	-	+	-	
		<i>Brucella melitensis</i>	BME	++	+	-	+	-	
		<i>Bartonella quintana</i> and spp.	BQ	+	+	-	+	-	
	Bradyrhizobiaceae	<i>Bradyrhizobium japonicum</i>	BJ	++	+	-	-	-	
		<i>Bradyrhizobium sp. BTA1</i>	Brad	++	+	+	-	-	
		<i>Rhodopseudomonas palustris</i>	RPA	++	+	-	-	-	
		<i>Nitrobacter hamburgensis</i>	Nham	+	+	-	-	-	
		<i>Nitrobacter winogradskyi</i>	Nwi	+	+	-	-	-	
	Rhodobacteriales	Rhodobacteraceae	<i>Rhodobacter capsulatus</i>	RC	+	-	+	-	+
			<i>Rhodobacter sphaeroides</i>	RSP	+	+	-	-	+
<i>Silicibacter sp. TM1040</i>			TM1040	+	+	-	-	+	
<i>Silicibacter pomeroyi</i>			SPO	+	+	-	-	+	
<i>Jannaschia sp. CC51</i>			Jann	+	+	-	-	+	
<i>Rhodobacteriales bacterium HTCC2654</i>			RB2654	+	+	-	-	+	
<i>Roseobacter sp. MED193</i>			MED193	+	+	-	-	+	
<i>Roseovarius nubinhibens</i> ISM			ISM	+	+	-	-	+	
<i>Roseovarius sp. 217</i>			ROS217	+	+	-	-	+	
<i>Loktanella vestfoldensis</i> SKA53			SKA53	+	+	-	-	+	
<i>Sulfitobacter sp. EE-36</i>			EE36	+	+	-	-	+	
<i>Oceanicola batsensis</i> HTCC2597			OB2597	+	+	-	-	+	
<i>Oceanicaulis alexandrii</i> HTCC2633			OA2633	-	+	-	-	+	
<i>Caulobacter crescentus</i>			CC	-	+	-	-	+	
<i>Parvularcula bermudensis</i> HTCC2503			PB2503	-	+	-	-	+	
Sphingomonadales	<i>Erythrobacter litoralis</i>	ELI	-	+	-	-	+		
	<i>Novosphingobium aromaticivorans</i>	Saro	-	+	-	-	+		
	<i>Sphingopyxis alaskensis</i> RB2256	Sala	-	+	-	-	+		
	<i>Zymomonas mobilis</i>	ZM	-	+	-	-	+		
Rhodospirillales	<i>Gluconobacter oxydans</i>	GOX	-	+	+	-	+		
	<i>Rhodospirillum rubrum</i>	Rru	+	+	-	-	++		
	<i>Magnetospirillum magneticum</i> AMB1	Amb	++	++	-	-	+		
	<i>Magnetospirillum magnetotacticum</i> MS-1	Magn	++	++	+	-	+		
Rickettsiales	SAR11 cluster	<i>Pelagibacter ubique</i> HTCC1002	PU1002	+	+	-	-	+	
		<i>Rickettsia</i> and <i>Ehrlichia</i> species		-	-	-	-	+	

# Regulation of genes in functional subsystems

Rhizobiales

Bradyrhizobiaceae

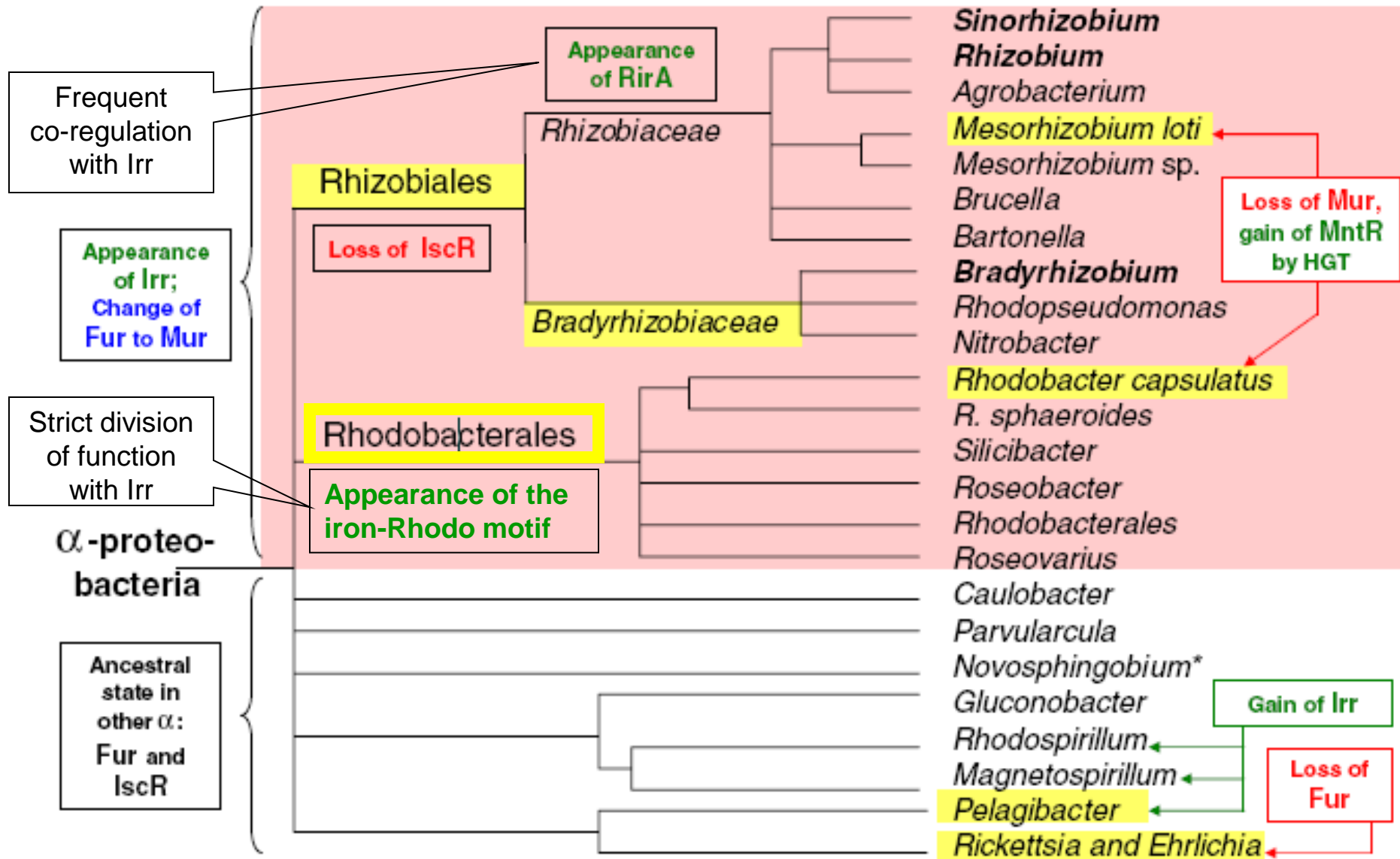
Rhodobacteriales

The Zoo (likely ancestral state)

Genome of $\alpha$ -proteobacteria	Components of iron/manganese regulatory networks										Iron Uptake (heme, siderophores, Fe <sup>2+</sup> , Fe <sup>3+</sup> )	Iron Storage	Fe-S synth.	Fe-dependent enzymes	Heme synth.	Transcript. regulators	Mn <sup>2+</sup> uptake													
	hmuRSTUV	exbBD-tonB	fhuBCD	fepBCDG	fetBCDE	fecBCDE	irp6ABC	OM receptors	pluB, pluC	mxcB	fbpABC	feoAB	FTR-chpA	mbfA	bfd-bfr	irpA	dps	surS/BCD	fssA	katG	fumA	ccpA	rubrerythrin	hema	ccm-ABCDG	rfa	fecIR	araX	si/ABCD	mntH
<i>Sinorhizobium meliloti</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Rhizobium leguminosarum</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Rhizobium etli</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Agrobacterium tumefaciens</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Mesorhizobium loti</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Mesorhizobium</i> sp. BNC1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Brucella melitensis</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Bartonella quintana</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Bradyrhizobium japonicum</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Bradyrhizobium</i> sp. BTAI1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Rhodospseudom. palustris</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Nitrobacter hamburgensis</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Nitrobacter winogradskyi</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Rhodobacter capsulatus</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Rhodobacter sphaeroides</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Silicibacter</i> sp. TM1040	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Silicibacter pomeroyi</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Jannaschia</i> sp. CC51	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Rhodobacteriales</i> HTCC2654	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Roseobacter</i> MED193	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Roseovarius</i> ISM	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Roseovarius</i> 217	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Loktanelle vestfold.</i> SKA53	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Sulfitobacter</i> EE-36	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Oceanicola</i> bat. HTCC2597	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Oceanicaulis</i> al. HTCC2633	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Caulobacter crescentus</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Parvularcula</i> ber.HTCC2503	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Erythrobacter litoralis</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Sphingopyxis</i> alas. RB2256	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Novosphingobium</i> aromat.	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Zymomonas mobilis</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Gluconobacter oxydans</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Rhodospirillum rubrum</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Magnetospirillum</i> AMB1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Magnetospirillum</i> MS-1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Pelagibacter</i> HTCC1002	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Rickettsia</i> and <i>Ehrlichia</i>	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

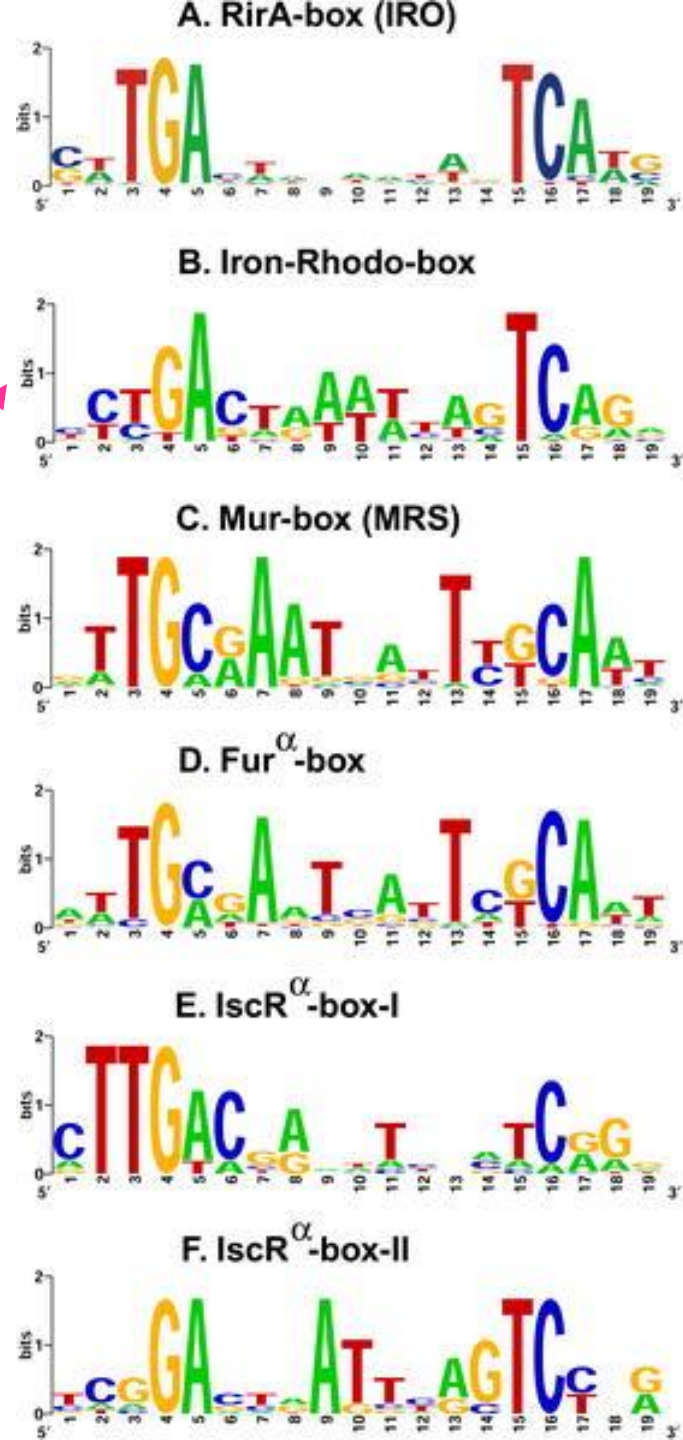
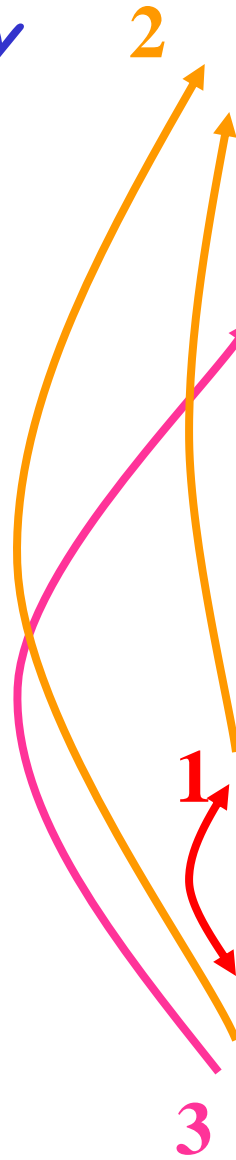
■ RirA-box, 
 ■ Iron-Rhodo-box, 
 ■ Irr-box (ICE), 
 ■ Fur-box, 
 ■ Mur-box, 
 ■ MntR-box, 
 ■ IscR-box, 
 ■ uncertain regulation

# Reconstruction of history



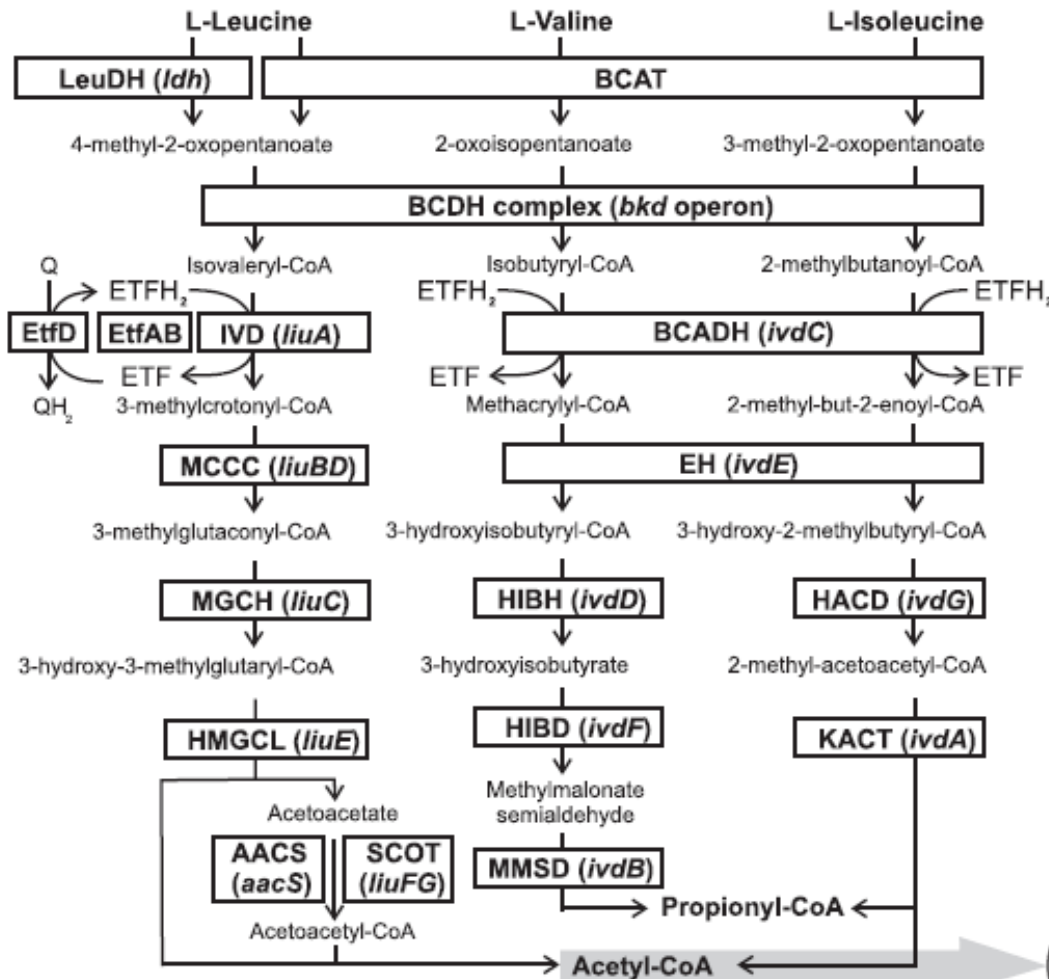
# All logos and *Some Very Tempting Hypotheses:*

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*

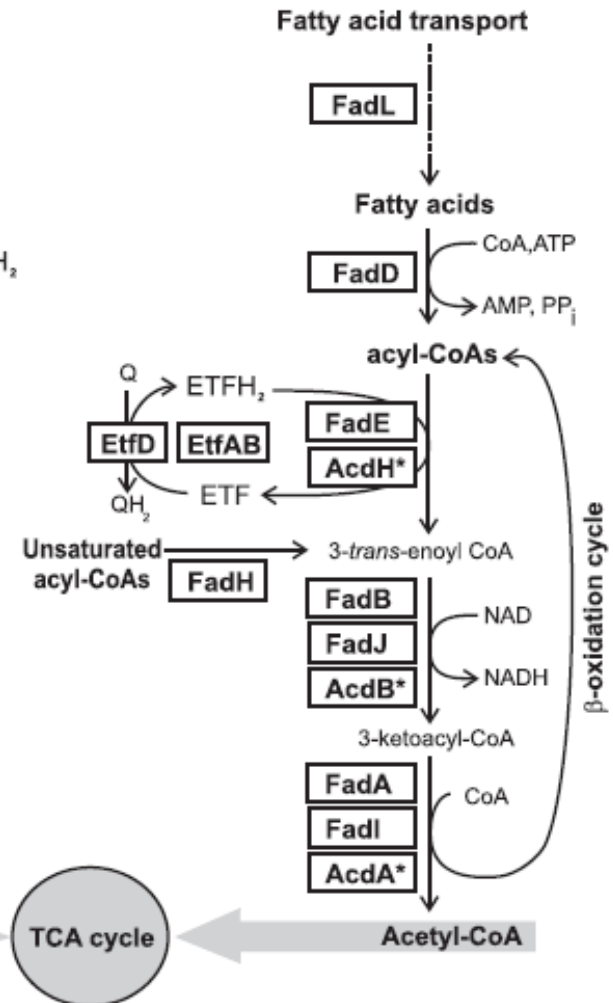


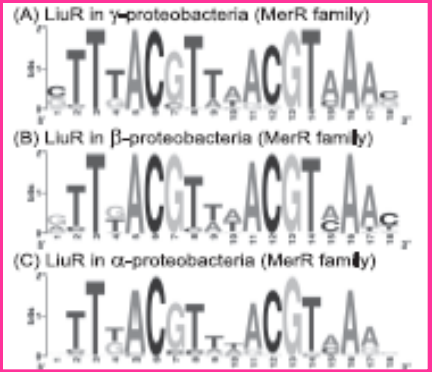
# Large-scale restructuring: Catabolism of branched chain amino acids and fatty acids, gamma- and beta-proteobacteria

(A) Branched-chain amino acid (ILV) degradation

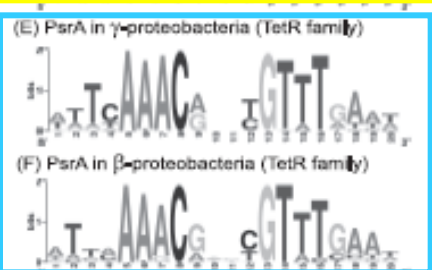


(B) Fatty acid (FA) degradation





MerR



TetR



GntR

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

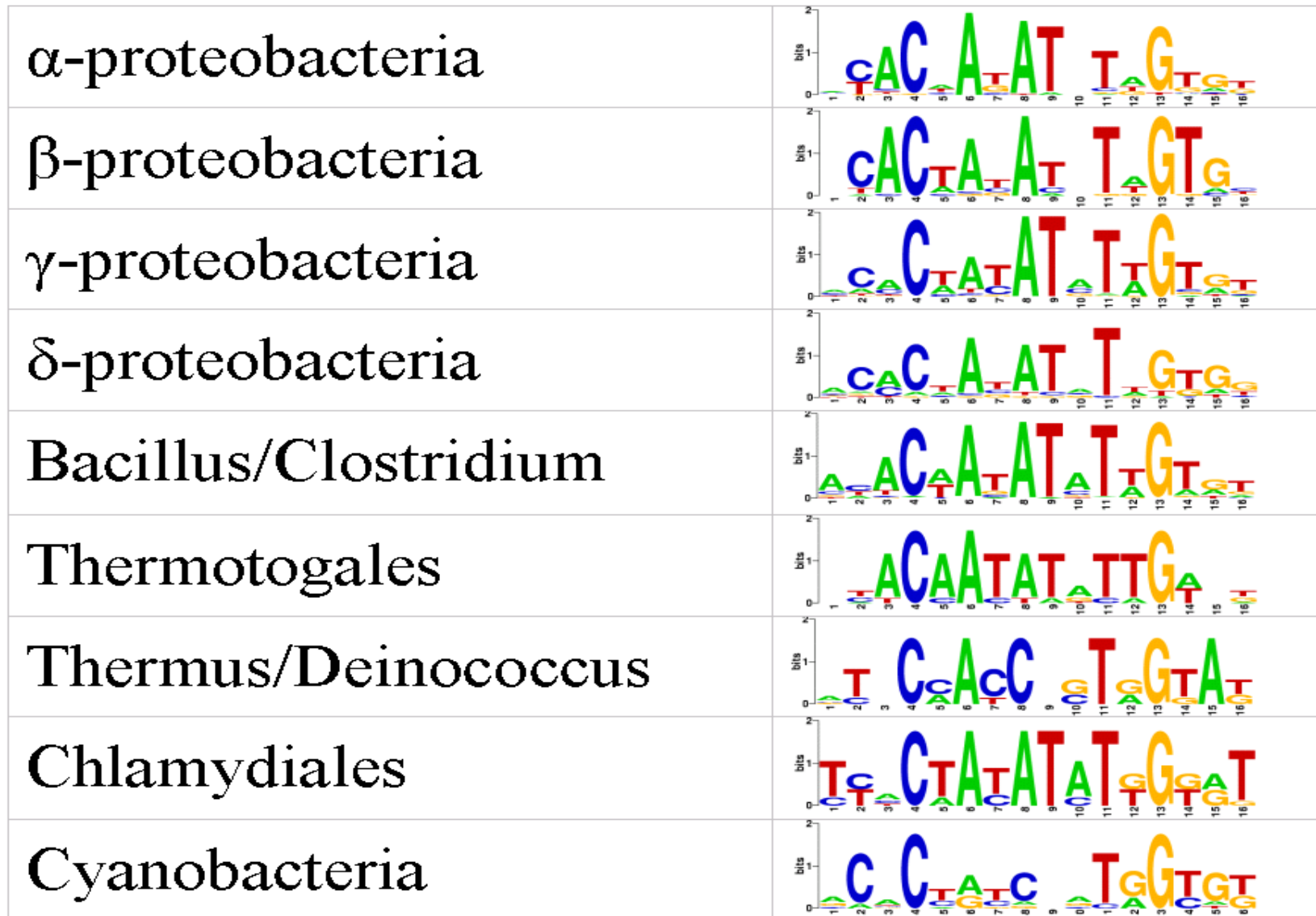
■ LiuR 
 ■ LiuQ 
 ■ FadR 
 ■ PsrA 
 ■ FadP 
 ■ unknown regulation

## 2. 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



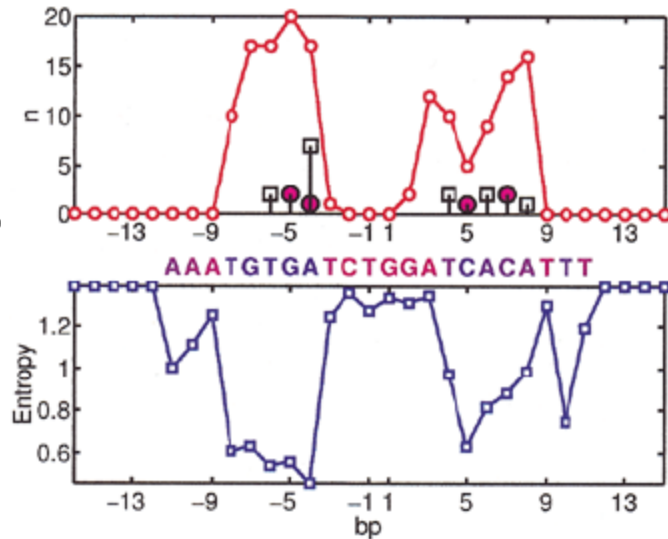
# NrdR (regulator of ribonucleotide reductases and some other replication-related genes): conservation at large distances



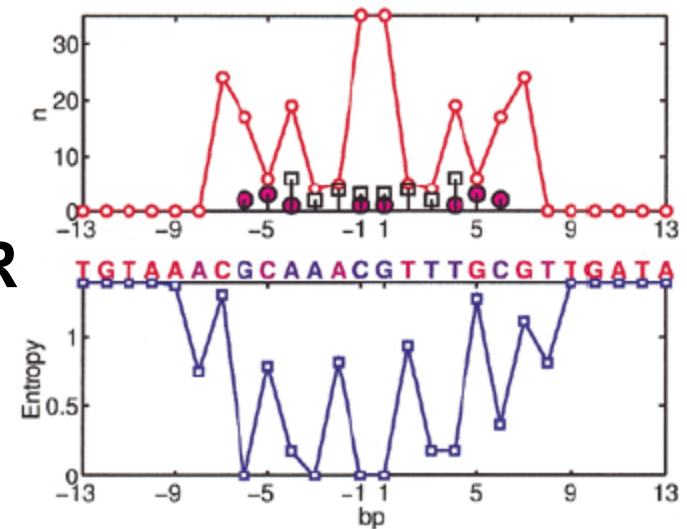
# DNA motifs and protein-DNA interactions

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

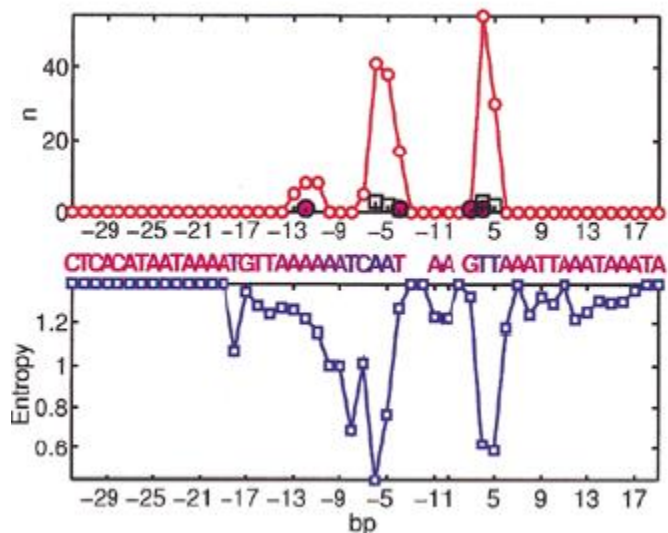
CRP



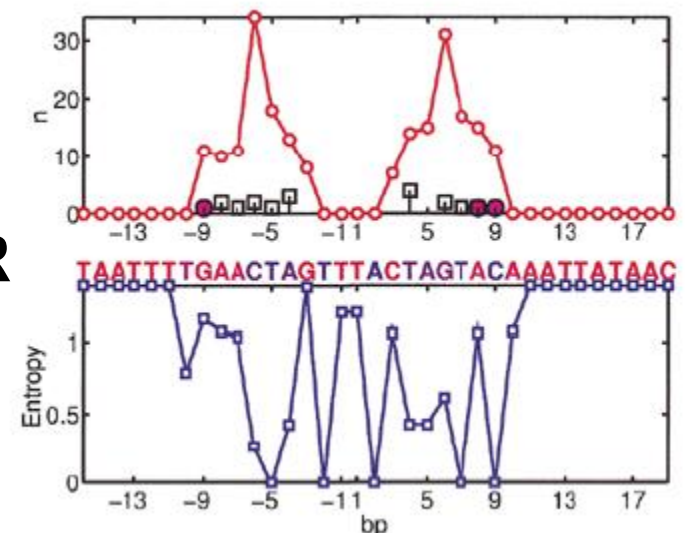
PurR



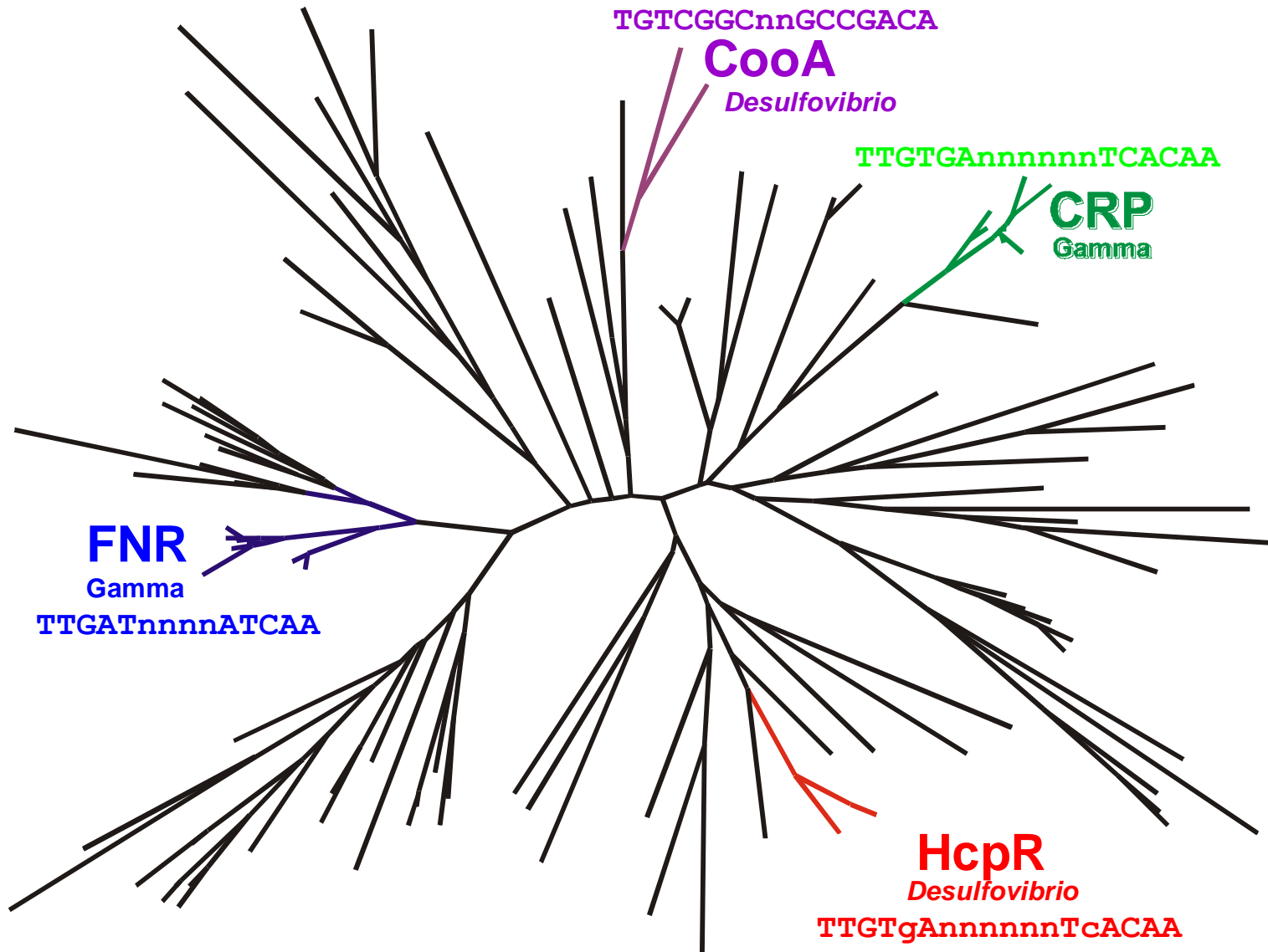
IHF



TrpR



# The CRP/FNR family of regulators



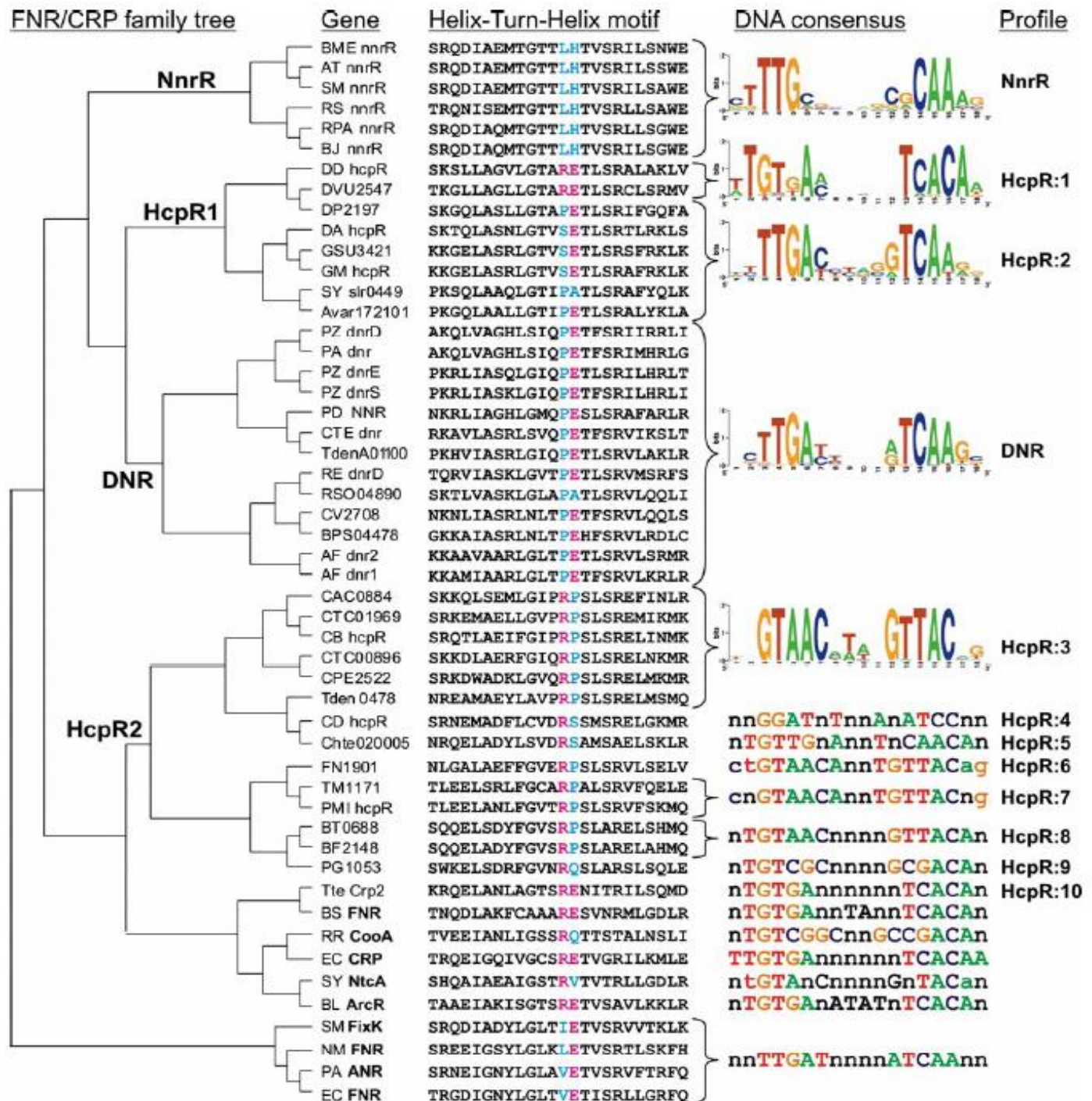
# Correlation between contacting nucleotides and amino acid residues

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

Contacting residues: **REnnnR**  
**TG**: 1<sup>st</sup> arginine  
**GA**: glutamate and 2<sup>nd</sup> arginine

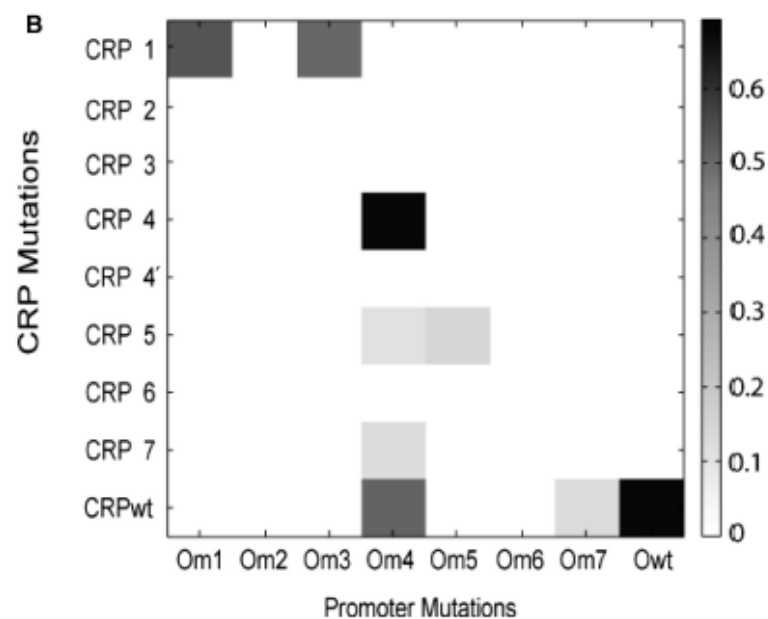
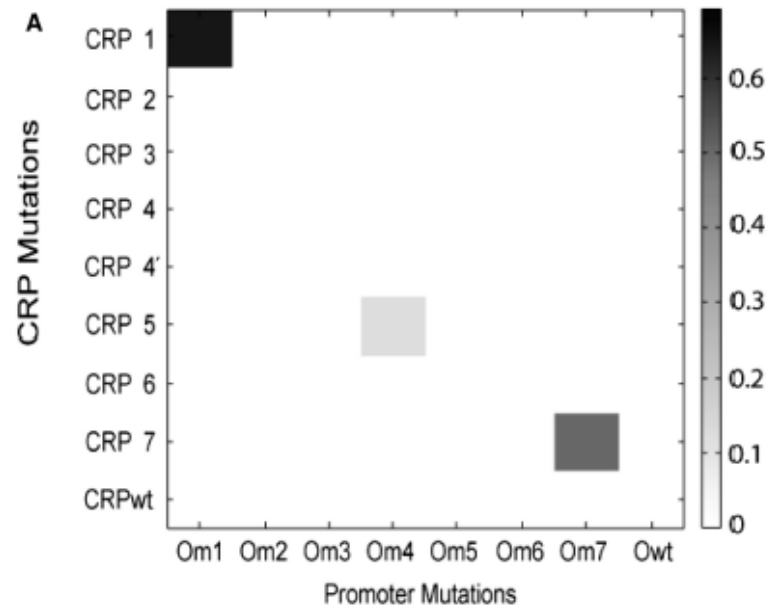
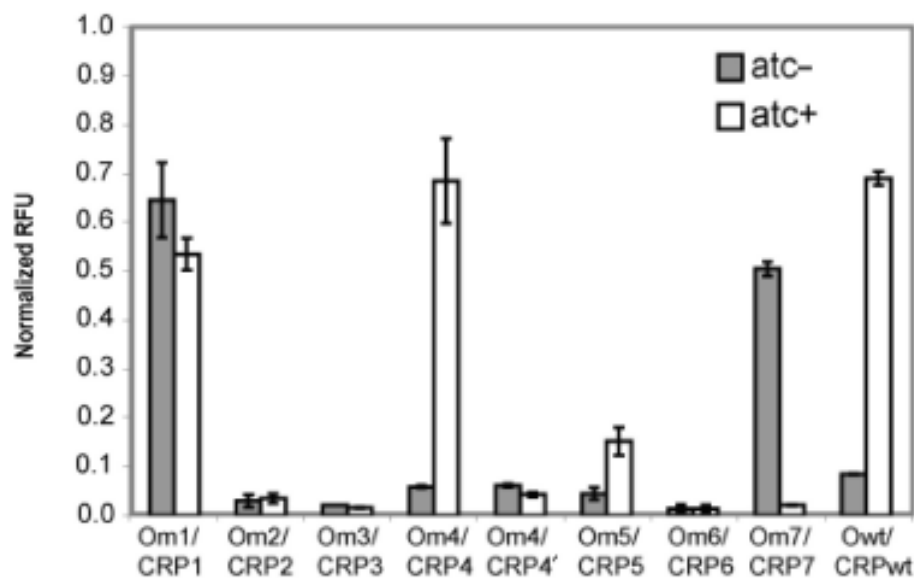
DD	COOA	ALTTEQLSLHMGAT <b>RQ</b> TVSTLLNNLVR	}	<b>TG</b> TCGGCnnGCCG <b>CA</b>
DV	COOA	ELTMEQLAGLVGTT <b>RQ</b> TASTLLNDMIR		
EC	CRP	KITRQEIGQIVGCS <b>RE</b> TVGRILKMLED	}	TT <b>TG</b> <b>GA</b> nnnnnnn <b>TC</b> <b>CA</b> <b>AA</b>
YP	CRP	KXTRQEIGQIVGCS <b>RE</b> TVGRILKMLED		
VC	CRP	KITRQEIGQIVGCS <b>RE</b> TVGRILKMLEE	}	TT <b>TG</b> <b>Tg</b> Annnnnnn <b>Tc</b> <b>CA</b> <b>AA</b>
DD	HCPR	DVSKSLLAGVLGT <b>ARE</b> TL <b>S</b> RALAKLVE		
DV	HCPR	DVTKGLLAGLLGT <b>ARE</b> TL <b>S</b> RCLSRMVE	}	TT <b>GA</b> Tnnnnn <b>AT</b> <b>CA</b> <b>AA</b>
EC	FNR	TMTRGDIGNYLGLT <b>VE</b> TIS <b>R</b> LLGRFQK		
YP	FNR	TMTRGDIGNYLGLT <b>VE</b> TIS <b>R</b> LLGRFQK	}	
VC	FNR	TMTRGDIGNYLGLT <b>VE</b> TIS <b>R</b> LLGRFQK		

The correlation holds for other factors in the family



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

Tasha A. Desai<sup>1</sup>, Dmitry A. Rodionov<sup>2,3</sup>, Mikhail S. Gelfand<sup>3,4</sup>, Eric J. Alm<sup>5,\*</sup> and Christopher V. Rao<sup>1,\*</sup>



# The LacI family: systematic analysis

- 1369 DNA-binding domains in 200 orthologous rows  $\langle Id \rangle = 35\%$ ,  $\langle L \rangle = 71$  a.o.
- 4484 binding sites,  $L = 20$ h.,  $\langle Id \rangle = 45\%$
- Calculate mutual information between columns of TF and site alignments
- Set threshold on mutual information of correlated pairs

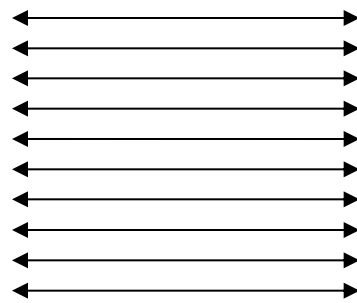
# Definitions

## Protein alignment

## Sites

```

LAFDHDQILQMAQERLQGKVRYP-IGFELLPEKFSRLRQLQRMVETVLGRS---LDKRN
LAFDHNQILDYGYQRLRNKLEYS-IAFEVLPPELFTLNDLFQLYTTVLGED--FADYS
LSFDHNEILAYGHRRLRNKLEYS-VAFEVLPPEMFTLNDLYQLYTTVLGEN--FSDYS
LSFDHNEILAYGHRRLRNKLEYS-VAFEVLPPEMFTLNDLYQLYTTVLGEN--FSDYS
LAFDHSKILAYGHRRLCNKLEYS-VAFDVLPPEYFTLNDLYQFYSTVLGAN--FSDYS
LAFDHSKILAYGHRRLCNKLEYS-VAFDVLPPEYFTLNDLYQFYSTVLGAN--FSDYS
LAFDHNQILDYGYQRLRNKLEYS-IAFEVLPPELFTLNDLFQLYTTVLGED--FADYS
LSFDHNEILAYGHRRLRNKLEYS-VAFEVLPPEMFTLNDLYQLYTTVLGEN--FSDYS
LSFDHNEILAYGHRRLRNKLEYS-VAFEVLPPEMFTLNDLYQLYTTVLGEN--FSDYS
    
```



```

tTAaTGgCtTTAtGcCACTAT
TTAaaGTAAtAaTTACCATAA
AaAtTGTCtTTAtGcCACTAT
TTATGGTAAAttcTACCATAA
TTATGGTAAAttcTACCATAA
TTATgGTCAgTTTcACcAaAA
TTaGTCgAAATAaccaACTAA
TTATCGTCAtCtcGACGACAA
TttAGGTAAGTTATACTTTTA
tTAaTGgCtTTAtGcCACTAT
    
```

$i$

$j$

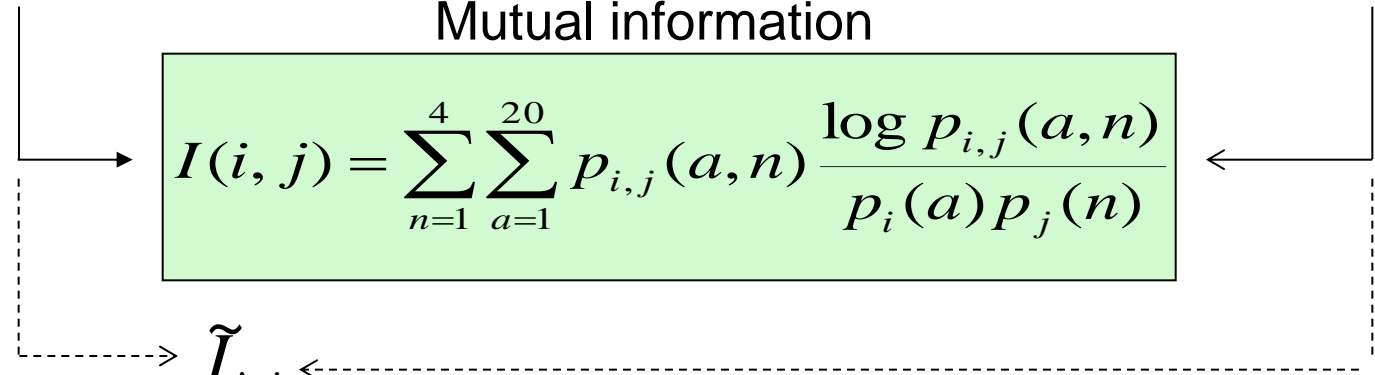
## Mutual information

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

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

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

Z-score

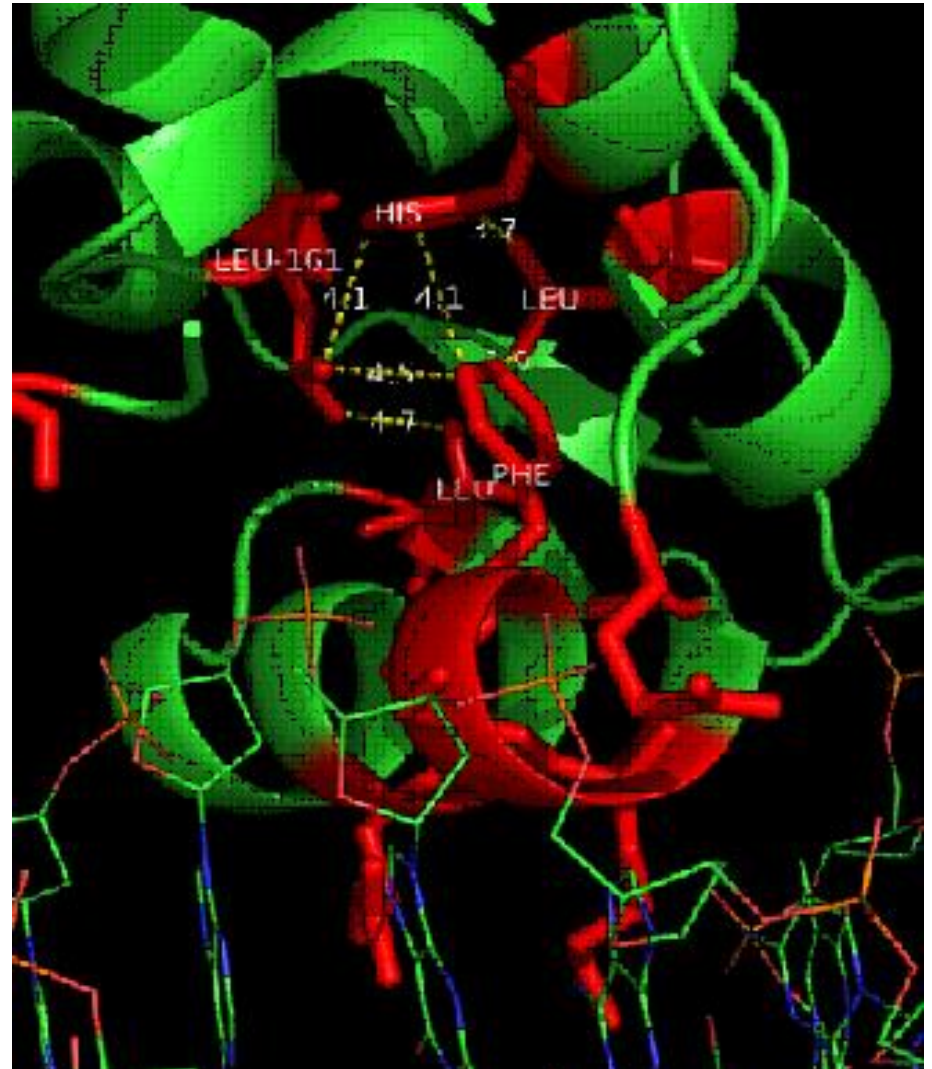
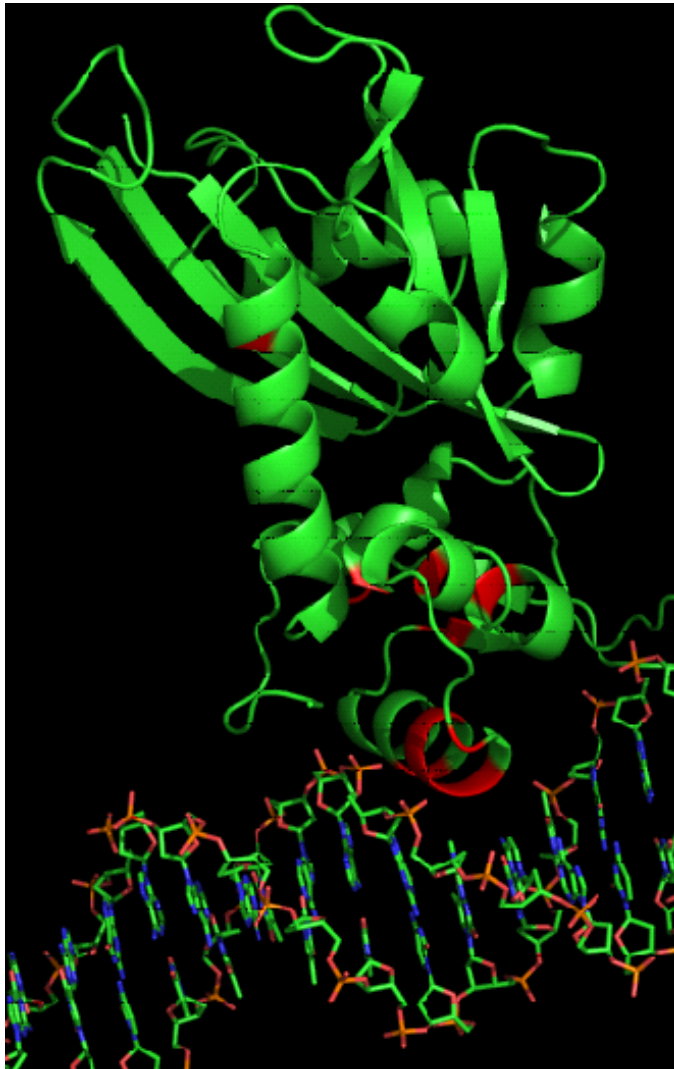






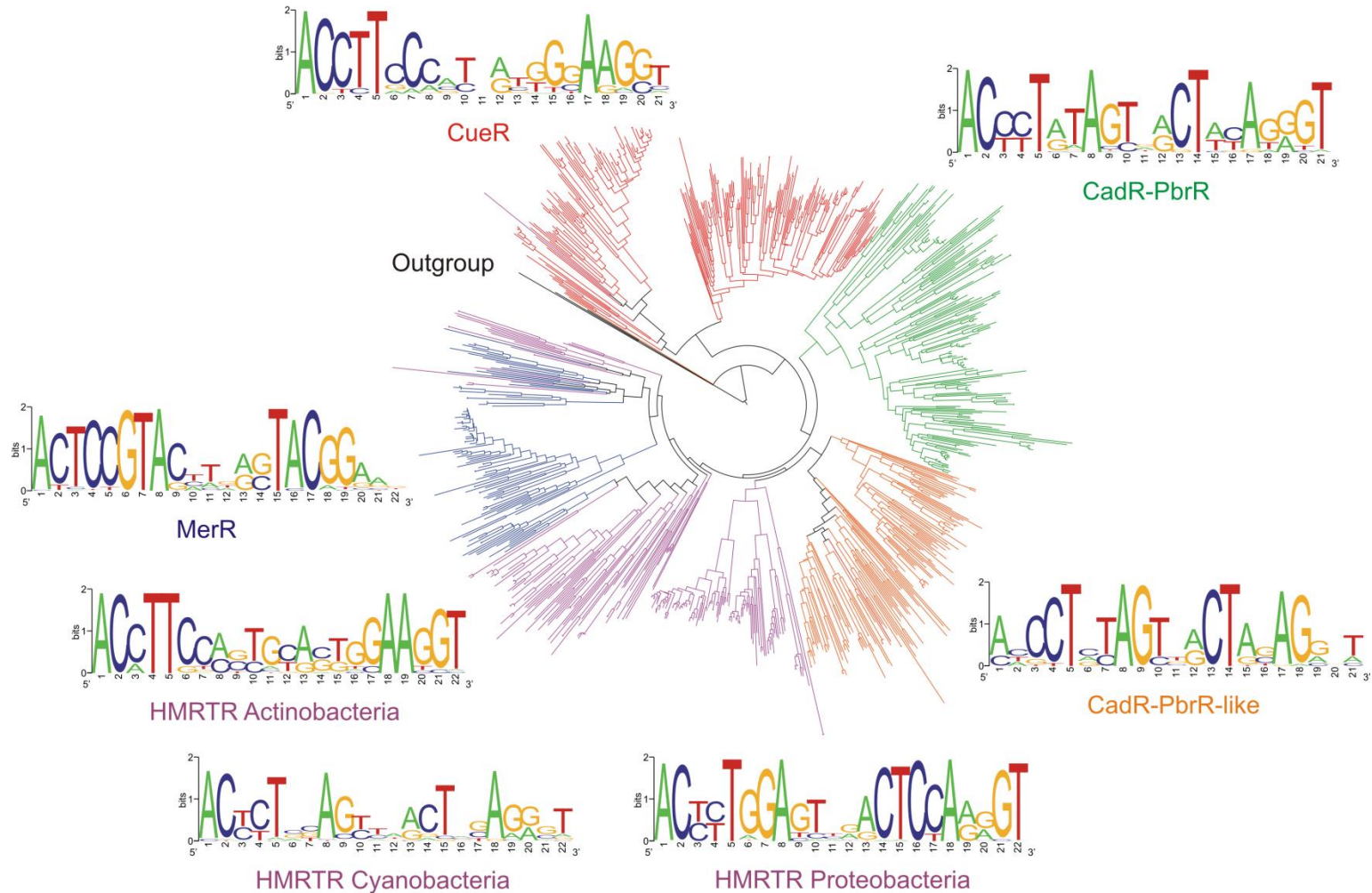


Comparison with the recently solved structure:  
correlated positions indeed bind the DNA  
(more exactly, form a hydrophobic cluster)



# MerR family

Phylogenetic tree of HMR transcriptional regulators from MerR family

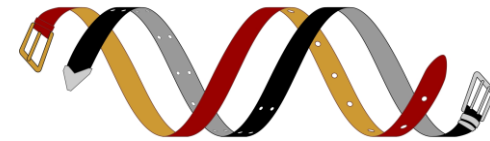


First 3 positions in sequence logos are the 3' end of 10 promoter boxes.



# 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 periphery) genes in relation to metabolic and functional subsystems/roles
  - (TF family-specific) protein-DNA recognition code



- **Andrei A. Mironov** - software, algorithms
- Alexei Kazakov (IITP, LANL)- branched chain amino acids and fatty acids
- \* Olga Kalinina (Saarbrucken Univ.) - SDP
- Yuri Korostelev - protein-DNA correlations
- \* Olga Laikova - LacI
- \* Alexandra Rakhmaninova - SDP, protein-DNA correlations
- Dmitry Ravcheev (IITP, Burnham Institute) - CRA/FruR, PurR/RbsR
- Dmitry Rodionov (IITP, Burnham Institute) - NrdR, iron, fatty acids etc.
- Olga Tsoy - CRA/FruR
- Andy Jonson (U. of East Anglia) - experimental validation (iron)
- Leonid Mirny (MIT) - protein-DNA, SDP
- Russian Ministry of Education and Science
- Russian Foundation of Basic Research
- Russian Academy of Sciences, program "Molecular and Cellular Biology"







- text

template