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          CCCACGAAAACGATTTGCTTTTT
purE          GCCACGCAACCGTTTTCCCTTG
pyrD          GTTCGGAAAACGTTTGCGTTTT
purT          CACACGCAACCGTTTTCGTTTA
cvpA          CCTACGCAACCGTTTTCTTTTT
purC          GATACGCAACCGTTTGCGTCTG
purM          GTCTCGCAACCGTTTTGCTTTCC
purH          GTTGCGCAACCGTTTTTCGTTAC
purL          TCTACGCAACCGGTTTTCGTCGG
consensus     ACGCAAACGTttTTtCGT
### Positional weight matrix

\[ W(b,j) = \ln(N(b,j) + 0.5) - 0.25 \sum_i \ln(N(i,j) + 0.5) \]
Positional weight matrix

\[
S = 1.1 + 1.9 + 2.2 + 0.5 + 2.2 + 2.2 + 1.9 + 2.2 + 2.2 - 0.1 + 1.9 + 2.2 + 1.2 + 2.2 + 0.6 + 2.2
\]
Phylogenetic footprinting

*rbs* operon in Enterobacteriaceae

**Start codon of** \textit{rbsD}
Phylogenetic footprinting

*rbs* operon in Enterobacteriaceae

... regulated by CRP and RbsR

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>b3748</td>
<td>TAA</td>
<td>TAC</td>
<td>ACC</td>
<td>CAT</td>
<td>GGT</td>
<td>ATT</td>
<td>CAG</td>
<td>GAT</td>
<td>CGG</td>
<td>ACA</td>
</tr>
<tr>
<td>KP_4306</td>
<td>CTC</td>
<td>CCT</td>
<td>GGT</td>
<td>GCT</td>
<td>CAG</td>
<td>ACA</td>
<td>TTA</td>
<td>ATT</td>
<td>ACC</td>
<td>AGT</td>
</tr>
<tr>
<td>SMA4113</td>
<td>TCT</td>
<td>CCT</td>
<td>CAG</td>
<td>ACC</td>
<td>GAA</td>
<td>ACG</td>
<td>AAC</td>
<td>CGA</td>
<td>AGT</td>
<td>TTT</td>
</tr>
<tr>
<td>YE0008</td>
<td>TCC</td>
<td>AAT</td>
<td>TTT</td>
<td>GGA</td>
<td>CAA</td>
<td>AAC</td>
<td>TTT</td>
<td>CCT</td>
<td>CAG</td>
<td>ACA</td>
</tr>
<tr>
<td>YP0007</td>
<td>TCT</td>
<td>TCT</td>
<td>CTA</td>
<td>TCT</td>
<td>AGA</td>
<td>AGT</td>
<td>GCT</td>
<td>GCA</td>
<td>AGT</td>
<td>CAA</td>
</tr>
<tr>
<td>ECA0010</td>
<td>TTA</td>
<td>ACC</td>
<td>TTC</td>
<td>CTT</td>
<td>TTT</td>
<td>TTG</td>
<td>TTA</td>
<td>GCA</td>
<td>AGA</td>
<td>AGT</td>
</tr>
</tbody>
</table>

**CRP binding site**

**RbsR binding site**

**Start codon of rbsD**
Conservation of regulation =>
consistency check

Set of known sites

PWM

Genome 1
Genome 2
Genome N
2. Example. Conserved motif upstream of \textit{nrd} genes

<table>
<thead>
<tr>
<th>α-proteobacteria</th>
<th><img src="image1" alt="DNA motif" /></th>
</tr>
</thead>
<tbody>
<tr>
<td>β-proteobacteria</td>
<td><img src="image2" alt="DNA motif" /></td>
</tr>
<tr>
<td>γ-proteobacteria</td>
<td><img src="image3" alt="DNA motif" /></td>
</tr>
<tr>
<td>δ-proteobacteria</td>
<td><img src="image4" alt="DNA motif" /></td>
</tr>
<tr>
<td>Bacillus/Clostridium</td>
<td><img src="image5" alt="DNA motif" /></td>
</tr>
<tr>
<td>Thermotogales</td>
<td><img src="image6" alt="DNA motif" /></td>
</tr>
<tr>
<td>Thermus/Deinococcus</td>
<td><img src="image7" alt="DNA motif" /></td>
</tr>
<tr>
<td>Chlamydiales</td>
<td><img src="image8" alt="DNA motif" /></td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td><img src="image9" alt="DNA motif" /></td>
</tr>
</tbody>
</table>
**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*, *dnal*, *polA*.
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 \textit{(nrd genes)}: FNR, DnaA, NrdR

A.

\begin{align*}
\text{EC} & \quad \text{TGCTTTTACTTTGGAGCTACACACA} & \text{AAAAAGCTCAAAACATCGTGGATGCAAACACTATATATTAAATGGCGTCCCA} \\
\text{ST} & \quad \text{TAGTTTTTACCTTTGTTCTTACACAA} & \text{TATAATTTGAAACATCGTGGATGCAAACACTATATATTAAATGGCGTCCCA} \\
\text{KP} & \quad \text{ACCTTATTACCTTTGGTCTGCAACAA} & \text{TATAATTTGAAACATCGTGGATGCAAACACTATATATTAAATGGCGTCCCA}
\end{align*}

B.

\begin{align*}
\text{YP} & \quad \text{AACAGGGAATAACCCATAACGCC} & \text{AAAAAGCTCAAAACATCGTGGATGCAAACACTATATATTAAATGGCGTCCCA} \\
\text{YE} & \quad \text{AACAGGGAATAACCCATAACGCC} & \text{TATAATTTGAAACATCGTGGATGCAAACACTATATATTAAATGGCGTCCCA} \\
\text{Eca} & \quad \text{AGTCTTAACAACTTTGTTCTAATGCAC} & \text{AAAAAGCTCAAAACATCGTGGATGCAAACACTATATATTAAATGGCGTCCCA} \\
\text{Ech} & \quad \text{AGTAGCTTAACAACTTTGTTCTAATGCAC} & \text{AAAAAGCTCAAAACATCGTGGATGCAAACACTATATATTAAATGGCGTCCCA}
\end{align*}
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
<table>
<thead>
<tr>
<th>Taxonomic Group</th>
<th>Conserved Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-proteobacteria</td>
<td>CAC AT TG</td>
</tr>
<tr>
<td>β-proteobacteria</td>
<td>CAC TA TG</td>
</tr>
<tr>
<td>γ-proteobacteria</td>
<td>CAT TG</td>
</tr>
<tr>
<td>δ-proteobacteria</td>
<td>CAC AT TG</td>
</tr>
<tr>
<td>Bacillus/Clostridium</td>
<td>CAC AT TG</td>
</tr>
<tr>
<td>Thermotogales</td>
<td>CAC AT TG</td>
</tr>
<tr>
<td>Thermus/Deinococcus</td>
<td>CAC AT TG</td>
</tr>
<tr>
<td>Chlamydiales</td>
<td>CAC AT TG</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>C C TG</td>
</tr>
</tbody>
</table>
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)
Changes: CRP/FNR family

- FNR
- HcpR
- CooA
- Gamma
- Desulfovibrio

TGTCGGCnnGCCGACA
TTGTgAnnnnnnTcACAA
TTGTGAnnnnnnTCACAA
TTGATnnnnATCAA

CRP

Gamma

Desulfovibrio

FNR

Gamma

Desulfovibrio

HcpR

Desulfovibrio
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:
- RE: 1\textsuperscript{st} arginine
- GA: glutamate and 2\textsuperscript{nd} arginine

<table>
<thead>
<tr>
<th>Protein</th>
<th>Amino Acid Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD COOA</td>
<td>ALTTEQLSLHMGATQTVSTLLNLVR</td>
</tr>
<tr>
<td>DV COOA</td>
<td>ELTMEQLAGLGVGTRQTASTLLNDMIR</td>
</tr>
<tr>
<td>EC CRP</td>
<td>KITRQEIGQIVGCSRETGRILKMLED</td>
</tr>
<tr>
<td>YP CRP</td>
<td>KXTRQEIGQIVGCSRETGRILKMLED</td>
</tr>
<tr>
<td>VC CRP</td>
<td>KITRQEIGQIVGCSRETGRILKMLEE</td>
</tr>
<tr>
<td>DD HCP R</td>
<td>DVSKSLLAGVLTARETTLSRALAKLVE</td>
</tr>
<tr>
<td>DV HCP R</td>
<td>DVTLLAGLGLTARETTLSRLSRMVE</td>
</tr>
<tr>
<td>EC FNR</td>
<td>TMTRGDIGNYLGLTVETISRLLGRFQK</td>
</tr>
<tr>
<td>YP FNR</td>
<td>TMTRGDIGNYLGLTVETISRLLGRFQK</td>
</tr>
<tr>
<td>VC FNR</td>
<td>TMTRGDIGNYLGLTVETISRLLGRFQK</td>
</tr>
</tbody>
</table>

Contacting residues:
- TGTCGGCnnGCGACA
- TTTGGAnnnnnnTCACAA
- TTTGGAnnnnnnTCACAA
- TTTGATnnnnnATCAA
The correlation holds for other factors in the family.
Engineering transcription factors with novel DNA-binding specificity using comparative genomics

Tasha A. Desai¹, Dmitry A. Rodionov²,³, Mikhail S. Gelfand³,⁴, Eric J. Alm⁵,* and Christopher V. Rao¹,*
LacI family: systematic analysis

- 1369 DNA-binding domains in 200 orthologous rows
  \(<\text{Id}>=35\%, \ <\text{L}>=71\ \text{a.o.}\)
- 4484 binding sites, \(L=20\ \text{nt.}, \ <\text{Id}>=45\%\)

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

**Protein alignment**

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

- \( \tilde{I}_{i,j} \)

**Sites**

- tTAaTGgCTTTATcGcACATAT
- TTAaaGTAaTTACCATAA
- AaAtTGTcTTATGcACATAT
- TTATGGTAaTTcTACCATAA
- TTATGGTAaTTcTACCATAA
- TTATGGTAaTTcTACCATAA
- TTATGGTAaTTcTACCATAA
- tTAaTGgCTTTATcGcACATAT

**Mutual information**

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

**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 contacts
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*

- Fructose: fruK, fruBA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
- Mannose: manXYZ
- Glucose: ptsHI-crr, epd, edd, eda, gapA, pgk, gpmA, pykF, ppsA, pckA
- Mannitol: mtlA, mtlD, fbp, pfkA
- Fructose: fruK, fruBA
Sites conserved in Enterobacteriales

- Mannose
  - manXYZ
- Glucose
  - ptsHI-crr

Metabolites:
- Mannitol
  - mtlA
  - mtlD
- Fructose
  - fruBA
  - fruK
- Fructose
  - epd
- Fructose
  - edd
- Glucose
  - eda
- Glucose
  - pykF
- Glucose
  - pckA
- Glucose
  - ppsA
- Glucose
  - aceEF
- Glucose
  - adhE
- Fructose
  - gapA
- Fructose
  - pgk
- Fructose
  - gpmA
- Fructose
  - pckA
- Fructose
  - pfkA
- Fructose
  - tpiA

Additional enzymes:
- aceB
- aceA
- icdA
Common ancestor of Enterobacteriales and Vibrionales

- **Mannose**: manXYZ
- **Glucose**: ptsHI-crr
- **Fructose**: fruK, fruBA
- **Mannitol**: mtlA, mtlD, fbp
- **Galactose**: not shown
- **Glucose**: ptsHI-crr
- **Pyruvate**: not shown
- **Fumarate**: not shown

**Gene Expressions**
- **fruK**: Fructose kinase
- **fruBA**: Fructose bisphosphate aldolase
- **fbp**: Fructose-1,6-bisphosphatase
- **mtlA, mtlD**: Mannitol transporters
- **mtlA, mtlD**: Methylotrophic transporters
- **pfkA**: Fructose bisphosphate aldolase
- **gapA, pgk**: Glycolytic enzymes
- **eda, edd**: Entner-Doudoroff pathway enzymes
- **tpiA**: Triose phosphate isomerase
- **gapA, pgk, gpmA**: Gluconeogenesis enzymes
- **pckA**: Phosphoenolpyruvate carboxykinase
- **pykF**: Pyruvate kinase
- **adhE**: Alcohol dehydrogenase
- **aceEF, aceA, aceB, icdA**: Acetate metabolism enzymes

**Other Metabolites**
- **Fructose**: fruK, fruBA
- **Mannose**: manXYZ
- **Mannitol**: mtlA, mtlD, fbp
- **Glucose**: ptsHI-crr

**Phenotype**
- **Mannitol-utilizing strain**: can grow on mannitol
- **Non-mannitol-utilizing strain**: cannot grow on mannitol
Common ancestor of Enterobacteriales

- **Fructose**: fruK, fruBA
- **Mannitol**: mtlD, mtlA
- **Mannose**: manXYZ
- **Glucose**: ptsHI-crr
- **Pyruvate Conversion Pathways**:
  - **Gamma-proteobacteria**:
    - fruK, fruBA
    - mtlD, mtlA
  - **Enterobacteriales**:
    - manXYZ
    - ptsHI-crr
- **Central Metabolism**:
  - **fruK**: fructose-1-phosphate aldolase
  - **fruBA**: fructose-bisphosphate aldolase
  - **mtlD, mtlA**: mannitol 1-phosphate dehydrogenase
  - **manXYZ**: mannose 1-phosphate isomerase
  - **ptsHI-crr**: phosphotransferase system
  - **gapA**: glyceraldehyde 3-phosphate dehydrogenase
  - **pgk**: phosphoglycerate kinase
  - **gpmA**: glyceraldehyde 3-phosphate dehydrogenase
  - **pckA**: phosphoenolpyruvate carboxykinase
  - **icdA**: isocitrate dehydrogenase
  - **aceA**: acetate kinase
  - **aceB**: aceA, aceB
  - **aceEF**: aceA, aceEF
  - **adhE**: alcohol dehydrogenase
Common ancestor of *Escherichia* and *Salmonella*

**Mannose**
- manXYZ
- ptsHI-crr

**Glucose**
- eda
- edd
- epd

**Fructose**
- fruK
- fruBA

**Mannitol**
- mtlD
- mtlA

**Fructose**
- fbp
- pfkA

**GapA**
- gapA

**PGK**
- pgk

**GPM**
- gpmA

**PykF**
- pykF

**PPS**
- ppsA

**PCK**
- pckA

**AdhE**
- adhE

**AceEF**
- aceEF

**IcdA**
- icdA

**AceA**
- aceA

**AceB**
- aceB

**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*
Search for candidate motifs and binding sites using standard comparative genomic techniques.
## Regulation of genes in functional subsystems

### Rhizobiales
- Sinorhizobium meliloti
- Rhizobium leguminosarum
- Rhizobium etli
- Agrobacterium tumefaciens
- Mesorhizobium loti
- Mesorhizobium sp. BNC1
- Brucella melitensis
- Bartonella quintana
- Bradyrhizobium japonicum
- Bradyrhizobium sp. BTAI1
- Rhodopseudomonas palustris
- Nitrobacter hamburgensis
- Nitrobacter winogradskyi
- Rhodobacter capsulatus
- Rhodobacter sphaeroides
- Silicibacter sp. TM1040
- Silicibacter pomeroyi
- Jannaschia sp. CC51
- Rhodobacterales HTCC2654
- Roseobacter MED193
- Roseovarius ISM
- Roseovarius 217
- Loktanella vestfold. SKA53
- Sulfitobacter EE-36
- Oceanicola bat. HTCC2597
- Oceanicola as. HTCC2633
- Caulobacter crescentus
- Parvularcula ber. HTCC2503
- Erythrobacter litoralis
- Sphingopyxis alas. RB2256
- Novosphingobium aromat.
- Zymomonas mobilis
- Gluconobacter oxydans
- Rhodospirillum rubrum
- Magnetospirillum AMB1
- Magnetospirillum MS-1
- Pelagibacter HTCC1002
- Rickettsia and Ehrlichia

### Bradyrhizobiaceae

### Rhodobacteriales

### The Zoo (likely ancestral state)
Reconstruction of history

**Appearance of the iron-Rhodo motif**

- **Frequent co-regulation with Irr**
  - Appearance of Irr; Change of Fur to Mur
  - Loss of IscR

- **Strict division of function with Irr**
  - Appearance of the iron-Rhodo motif

**α-proteobacteria**

- Ancestral state in other α: Fur and IscR
- Gain of Irr
- Loss of Fur

**Rhizobiales**

- Appearance of RirA
  - Rhizobiaceae
    - Mesorhizobium
      - Mesorhizobium loti
      - Mesorhizobium sp.
      - Brucella
      - Bartonella
    - Bradyrhizobium
      - Rhodopseudomonas
      - Nitrobacter
      - Rhodobacter capsulatus
        - R. sphaeroides
        - Silicibacter
        - Roseobacter
        - Rhodobacterales
        - Roseokervivus
        - Caulobacter
        - Parvularcula
        - Novosphingobium
          - Gluconobacter
          - Rhodospirillum
          - Magnetospirillum
          - Pelagibacter
          - Rickettsia and Ehrlichia
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
RbsR and PurR: duplication and subsequent change of specificity
PurR regulon

- core
- taxon-specific
- E.coli-specific (experimental)
Evolution of DNA motifs

PurR

RbsR
Pseudomonadales

Different
Conserved
Different
PurR: new ligand, changed default state (bound to DNA), new regulon

RbsR: change of motif

Ancestral state (retained in Pseudomonadales): RbsR regulates the rbs operon by binding in the absence of ribose

Enterobacteriales
Pasteurellales
Vibrionales
Pseudomonadales
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
• 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
В любой непонятной ситуации — эволюционируй!

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

- text