Current Trends: Non-coding RNAs

“Central Dogma” of molecular biology

DNA → RNA → Protein (mRNA) → Cellular functions

Reverse transcriptase

RNA virus replication

in vitro

(ncRNA)
Non-coding RNAs

- Found in prokaryotes (small RNAs) and eukaryotes (non-coding RNAs).
- Well-characterized examples: tRNA, rRNA

Enzymatic activity
- self-splicing introns
- peptidyl transfer
- viral replication

Regulation of other genes
- eukaryotes: 21-25 nts; micro RNAs
- prokaryotes: 50-550 nts; small RNAs
Eukaryotic vs. Prokaryotic ncRNAs

RNAi
ncRNAs can regulate gene expression at many steps

Gottesman, Trends in Genetics 21:399-404

Targets of RNA Gene Regulation

messenger RNA

UAGCAUGUACGUAGCUAGCUACGAUUGUUAUUACUGUCGUGCUUUCACUUCUCGCAGGAGUCCUCGUAUGGUA

RNA gene

Targets of RNA Gene Regulation
Non-coding RNAs are elusive

- Not annotated in genomes: lack of defined sequence features
- Small, often missed in genetic studies
- Missed in assays for protein function
- None of 70-100 *E. coli* ncRNAs found by mutation

---

**Drosophila bantam gene discovery**

- Overexpression of an intergenic region causes cell and tissue overgrowth
- Deletion of intergenic region surrounding EP element results in slow growth
**bantam** encodes a miRNA that regulates a conserved growth pathway

---

**hippo**  
**salvador**  
**warts**  

↓

**yorkie**  

↓

**bantam**

growth  

suppressors

growth promoter

growth promoter

growth

cell death

---

**Why study ncRNAs in bacteria?**

- We live in a bacterial world
- Bacteria serve as useful model organisms
- Bacteria are diverse
- Understanding bacteria is useful in many important applications

---

- Tuberculosis
- Whooping cough
- Diphtheria
- Typhoid Fever
- Meningitis
- Cholera
- Rocky Mountain Spotted Fever

- Dysentery
- Botulism
- Scarlet Fever
- Peptic Ulcers
- Dental Cavities
- Rheumatic Fever

- Strept Throat
- The Black Plague
- Yaws
- Pneumonia
- Syphilis
- Gonorrhea
- Anthrax
- Tetanus
- Leprosy
- Food Poisoning
- Lyme Disease
- Pneumonia
- Syphilis
- Anthrax
- Tetanus
- Leprosy
• Gram-negative γ-proteobacterium
• Found primarily in deep water anaerobic habitats
• Can use a wide variety of compounds as terminal electron acceptors
• Bioremediation potential: reduces soluble chromium and uranium to insoluble forms

**S. oneidensis genome overview**

- 45.9% G-C content; 85.5% of genome is coding
- 513,141,60 bp total: chromosome is 496,980,3 bp; pMR-1 is 161,613 bp

<table>
<thead>
<tr>
<th>Total genes</th>
<th>5,066</th>
</tr>
</thead>
<tbody>
<tr>
<td>tRNA and rRNA genes</td>
<td>128 (2.5%)</td>
</tr>
<tr>
<td>Protein-coding genes</td>
<td>4,938 (97.5%)</td>
</tr>
<tr>
<td>Genes assigned function</td>
<td>2,915 (59%)</td>
</tr>
<tr>
<td>Conserved hypothetical genes</td>
<td>864 (17.5%)</td>
</tr>
<tr>
<td>Hypothetical genes</td>
<td>1,159 (27%)</td>
</tr>
</tbody>
</table>
Computational ncRNA prediction

- Most ncRNAs are intergenic, function in trans
- Bacterial ncRNAs have promoters, terminators
- Genes have distinct nucleotide composition
- Conserved secondary structures (stem-loop)
- Tiling microarray data
- Data can be integrated into a set of predictions
1) Nucleotide composition

<table>
<thead>
<tr>
<th></th>
<th>Frequency of nucleotides</th>
<th></th>
<th>Frequency of dinucleotides</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.28</td>
<td>A</td>
<td>0.0562</td>
</tr>
<tr>
<td>C</td>
<td>0.22</td>
<td>C</td>
<td>0.0481</td>
</tr>
<tr>
<td>G</td>
<td>0.22</td>
<td>G</td>
<td>0.0732</td>
</tr>
<tr>
<td>T</td>
<td>0.28</td>
<td>T</td>
<td>0.0464</td>
</tr>
</tbody>
</table>

2) Comparative Genomics: Mutation Patterns

Genome (1)

Damage to genome 1 with damage to genome 2.

Genome (2)

Damage to genome 1 with damage to genome 2.

Shewanella amazonensis, Shewanella baltica, Shewanella denitrificans, Shewanella frigidimarina, Shewanella loihica, Vibrio cholerae, Yersinia pestis, Photorhabdus luminescens, Photobacterium profundum
2) Mutation Patterns that Conserve RNA Structure

Derive score based on:
- # of compensatory mutations
- Length of sequence
- Sequence structure

2) Score of Conserved RNA Structure

[Graph showing frequency distribution of scores]

Rivas and Eddy, *BMC Bioinformatics* 2:8
3) DNA Microarray Data

Examine correlation of expression for probes near one another in the genome:

1) intergenic regions likely to produce RNA
2) for those much less likely to produce RNA

3) Correlation of Transcript Expression

![Correlation Graph](image)

- Correlation in non-operon IG regions
- Correlation in operon IG regions

# of IG Probes vs Correlation Coefficient
Integrating Heterogeneous Data

1) Sequence Data

ATGCATGCTAGTCATC
GATCGATC

A  0.28  A  0.23
C  0.22  C  0.27
G  0.22  G  0.27
T  0.28  T  0.23

2) Conserved Structure Data

3) DNA Microarray Data

general Markov model

Integrating Heterogeneous Data

1) Sequence Data

ATGCATGCTAGTCATC
GATCGATC

A  0.28  A  0.23
C  0.22  C  0.27
G  0.22  G  0.27
T  0.28  T  0.23

2) Conserved Structure Data

3) DNA Microarray Data

general Markov model
Performance on known *E. coli* ncRNAs

![Performance on known E. coli ncRNAs](image)

- **Primary sequence**
- **Conserved structure**
- **Expression data**

Predicting ncRNAs in *Shewanella*

- Have robust tiling microarray data set: 144 experiments, wide variety of growth conditions
- Generated predictions of ~160 ncRNAs
- Some may be orthologous to characterized ncRNAs
- Some may be novel ncRNAs
- Some may not be ncRNAs at all
Regulation of *RyhB*

**Fur repressor is active when bound to Fe$^{2+}$**

**Fe$^{2+}$ limitation induces *RyhB* expression**

**Validating *Shewanella* ncRNA predictions**

*Putative *ryhB* northern blotting experiments*
Regulation of \textit{spot42} in \textit{E. coli}

\textit{spot42} negatively regulates translation of \textit{galK} but does not affect \textit{galE} translation.

\textit{spot42} expression increases the \textit{GalE:GalK} ratio.

Thus, \textit{glucose} induces \textit{spot42} expression.

Validating \textit{Shewanella} ncRNA predictions

- Putative \textit{spot42} northern blotting experiments

Blot probed with \textit{spot42} probe 1

Lac20, Lac30, Lac20 Glc20, Lac20 Gal20

\text{~110 nt}
Predicting ncRNA targets

- Target prediction is an inexact science
- ncRNA sequences not exact matches to targets
- May be multiple targets
- Validate targets using northern blots, western blots, exogenous expression

Putative *Shewanella* ncRNA targets

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO0006</td>
<td>Outer membrane protein, putative</td>
</tr>
<tr>
<td>SO0076</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0104</td>
<td>P-glycoprotein ABC transporter</td>
</tr>
<tr>
<td>SO0104</td>
<td>Peptide ABC transporter, P-glycoprotein</td>
</tr>
<tr>
<td>SO0366</td>
<td>Outer membrane protein, putative</td>
</tr>
<tr>
<td>SO0456</td>
<td>Transcription regulator, ChtC family</td>
</tr>
<tr>
<td>SO0456</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO2193</td>
<td>OmpA family protein</td>
</tr>
<tr>
<td>lgsA</td>
<td>DNA ligase, NAD-dependent</td>
</tr>
<tr>
<td>msxT</td>
<td>MusM protein</td>
</tr>
<tr>
<td>nmpF</td>
<td>Cytochrome e-type protein, NmpF</td>
</tr>
<tr>
<td>SO0456.2</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0293</td>
<td>Transcription regulator, LyhD family</td>
</tr>
<tr>
<td>SO0410</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0449</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0410</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0449</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0410</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0449</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0410</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0449</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0410</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0449</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0410</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0449</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0410</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0449</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0410</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0449</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0410</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>SO0449</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>soxA</td>
<td>Cytochrome e</td>
</tr>
<tr>
<td>SO0479</td>
<td>Cytochrome e, putative</td>
</tr>
</tbody>
</table>
Questions

• Is the interaction between a ncRNA and its target RNA positive or negative?
• What conditions regulate ncRNA expression?
• What can we learn that will improve ncRNA prediction and understanding of function?