Some stories about computing and riboswitches

Walter L. (Larry) Ruzzo

Computer Science & Engineering and Genome Sciences, Univ. of Washington
Fred Hutchinson Cancer Research Center
Seattle, USA

http://www.cs.washington.edu/homes/ruzzo

Central Dogma of Molecular Biology

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

“The central dogma, enunciated by Crick in 1958 and the keystone of molecular biology ever since, is likely to prove a considerable over-simplification.”

Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.
RNA Secondary Structure: RNA makes helices too

Base pairs

A=U
C=G
U=G

Usually single stranded
Gene Regulation: The MET Repressor

Protein

DNA

Alberts, et al. 3e.
Alberts, et al, 3e.

Epshtein, et al., PNAS 2003
Winkler et al., Nat. Struct. Biol. 2003

Not the only way!

Protein way
Riboswitch alternative

SAM
Not the only way!

Protein way
Riboswitch alternatives

SAM-I

SAM-II


Corbino et al., Genome Biol. 2005
Alberts, et al., 3e.

Corbino et al., Genome Biol. 2005

Protein way

Riboswitch alternatives

Not the only way!

Fuchs et al., NSMB 2006

SAM-III


Corbino et al., Genome Biol. 2005

SAM-I

SAM-II

SAM-III
Not the only way!

Protein way

Riboswitch alternatives

Alberts, et al., 3e.

Corbino et al., Genome Biol. 2005

Weinberg et al., RNA 2008


Corbino et al., Genome Biol. 2005

Fuchs et al., NSMB 2006

SAM-I

SAM-II

SAM-III

SAM-IV


Corbino et al., Genome Biol. 2005

Fuchs et al., NSMB 2006

Weinberg et al., RNA 2008
Not the only way!

Protein way

Riboswitch alternatives

Corbino et al., Genome Biol. 2005
Fuchs et al., NSMB 2006
Weinberg et al., RNA 2008
Meyer, et al., BMC Genomics 2009
Translational Control

Transcriptional Control

Detail of Translational Control

The Glycine Riboswitch

(Mandal, Lee, Barrick, Weinberg, Emilsson, Ruzzo, Breaker, Science 2004)
Fig. 3. Cooperative binding of two glycine molecules by the VC I-II RNA. Plot depicts the fraction of VC II (open) and VC I-II (solid) bound to ligand versus the concentration of glycine. The constant, $n$, is the Hill coefficient for the lines as indicated that best fit the aggregate data from four different regions (fig. S3). Shaded boxes demark the dynamic range (DR) of glycine concentrations needed by the RNAs to progress from 10%- to 90%-bound states.
Riboswitches

UTR structure that directly senses/binds small molecules & regulates mRNA
widespread in prokaryotes
some in eukaryotes & archaea
~ 20 ligands known; multiple nonhomologous solutions for some (e.g. SAM)
dozens to hundreds of instances of each
on/off; transcription/translation; splicing; combinatorial control
all found since ~2003; most via bioinformatics
Why is RNA hard to deal with?

A: *Structure* often more important than sequence
RNA Secondary Structure: can be fixed while sequence evolves
Impact of RNA homology search

(Barrick, et al., 2004)

glycine riboswitch operon

B. subtilis

L. innocua

A. tumefaciens

V. cholera

M. tuberculosis

(and 19 more species)
Impact of RNA homology search

(Barrick, *et al.*, 2004)

Using our techniques, we found...

- Glycine riboswitch
- Operon

- *B. subtilis* (and 19 more species)
- *L. innocua* (and 42 more species)
- *A. tumefaciens* (and 42 more species)
- *V. cholera* (and 42 more species)
- *M. tuberculosis* (and 42 more species)
A  mRNA leader

<table>
<thead>
<tr>
<th>mRNAs</th>
<th>Leader sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bsu</td>
<td>TTGCATACTTACG</td>
</tr>
<tr>
<td>Bhs</td>
<td>TTGCATATTTACT</td>
</tr>
<tr>
<td>Oh</td>
<td>UUGGACUGUGUGACC</td>
</tr>
<tr>
<td>Bce</td>
<td>UAAACGAGUGUGAAGC</td>
</tr>
<tr>
<td>Gba</td>
<td>CAUCUGGUACUGUACG</td>
</tr>
<tr>
<td>Soc</td>
<td>AUAAUGCAUGAUGCCG</td>
</tr>
<tr>
<td>Lma</td>
<td>TTTGTAAGTGGGCA</td>
</tr>
<tr>
<td>Ssu</td>
<td>AUACUGAGUGUGUAGC</td>
</tr>
<tr>
<td>Cp</td>
<td>TTTTTCAAGGGCG</td>
</tr>
<tr>
<td>Chy</td>
<td>UACCCGGCGGCAGGC</td>
</tr>
<tr>
<td>Svo</td>
<td>AAAAGGCUGUGUGUAGC</td>
</tr>
<tr>
<td>Ame</td>
<td>UUGGACUGUGUGAAGC</td>
</tr>
<tr>
<td>Dre</td>
<td>UGGUUCUGUGUGAAGC</td>
</tr>
<tr>
<td>Spn</td>
<td>AUAAUGCAUGAUGCCG</td>
</tr>
<tr>
<td>Smu</td>
<td>TTTTTCAAGGGCG</td>
</tr>
<tr>
<td>Lpl</td>
<td>TTTTTCAAGGGCG</td>
</tr>
<tr>
<td>Efa</td>
<td>UACCCGGCGGCAGGC</td>
</tr>
<tr>
<td>Ljo</td>
<td>UUGGACUGUGUGAAGC</td>
</tr>
<tr>
<td>Sth</td>
<td>AAAAGGCUGUGUGUAGC</td>
</tr>
<tr>
<td>Lac</td>
<td>UUGGACUGUGUGAAGC</td>
</tr>
<tr>
<td>Spy</td>
<td>UACCCGGCGGCAGGC</td>
</tr>
<tr>
<td>Lsa</td>
<td>UUGGACUGUGUGAAGC</td>
</tr>
<tr>
<td>Lel</td>
<td>UUGGACUGUGUGAAGC</td>
</tr>
<tr>
<td>Fnu</td>
<td>UUGGACUGUGUGAAGC</td>
</tr>
</tbody>
</table>

B  mRNA leader switch

- **P1**: Watson-Crick base pair G-C
- **P2**: Other base interaction

C  mRNA leader switch?

- **L19**: compensatory mutations
  - 97% nucleotide identity
  - 97% present
  - 90% nucleotide identity
  - 90% present
  - 75% nucleotide identity
  - 75% present
  - 50% nucleotide identity
  - 50% present

- **3' stem loop**: always present
Covariation is strong evidence for base pairing
Alignment Matters

Structural conservation ≠ Sequence conservation

Alignment without structure information is unreliable

CLUSTALW alignment of SECIS elements with flanking regions

```plaintext
GCGATCAATTGGAAGAAGCAGCTT - ACCTGATTAATGCTGATGACGAG - AATGAGAATGGTGCCGTAACGAA - TCTGCAGAGCTGGATCGCACTTATCACTTATACAATTGTTGCTCGCTGAATCTTGTACTTGACGACTTATCAGAGCTTATGACGGTGTCCTGAA
```

same-colored boxes should be aligned
How to model an RNA “Motif”? 

Conceptually, start with a profile HMM:

- from a multiple alignment, estimate nucleotide/insert/delete preferences for each position
- given a new seq, estimate likelihood that it could be generated by the model, & align it to the model
How to model an RNA “Motif”? 

Add “column pairs” and pair emission probabilities for base-paired regions
Covariance Models
(specialized stochastic CFGs)

Sequences

\[ S_1 \rightarrow cS_2g \text{ or } aS_2u \]

\[ S_2 \rightarrow a \]

Example parse of CAG

\[ S_1 \rightarrow cS_2g \rightarrow \text{cag} \]
**Stochastic context-free grammar**

\[
S_1 \rightarrow cS_2g \mid aS_2u
\]

\[
S_2 \rightarrow a
\]

Example parse of CAG

\[
S_1 \rightarrow cS_2g \rightarrow cag
\]

Classification

Is \( \Pr(\text{parse of CAG}) \geq \text{threshold} \)

(e.g., vs \( \Pr(\text{CAG in null model}) \))
Application: *cis*-regulatory ncRNA discovery in prokaryotes

Key issue is exploiting prior knowledge to focus on promising data
CMFinder

Simultaneous alignment, folding & motif description
Yao, Weinberg & Ruzzo, Bioinformatics, 2006

Folding predictions
Smart heuristics
Mutual Information
Candidate alignment
Realign
CM

Combines folding & mutual information in a principled way.
CMfinder Accuracy
(on Rfam families with flanking sequence)
Use the Right Data; Do Genome Scale Search

Right Data:
• 5-10 examples amidst 20 extraneous ones OK; (but not 5 in 200 or 2000)
• length 1k (not 100k)

How:
• Regulators near regulatees
• Get UTRs of homologs

Genome Scale Search:
• Many riboswitches are present in ~5 copies per genome
• More examples = better model + clues to function
Processing Times

Input from ~70 complete Firmicute genomes available in late 2005-early 2006, totaling ~200 megabases

- Identify CDD group members
  - 2946 CDD groups
  - < 10 CPU days

- Retrieve upstream sequences

- Footprinter ranking
  - < 10 CPU days

- CMfinder
  - 35975 motifs
  - 1 ~ 2 CPU months

- Motif postprocessing
  - 1740 motifs

- RaveNnA
  - 10 CPU months

- CMfinder refinement
  - < 1 CPU month

- Motif postprocessing
  - 1466 motifs
S-adenosyl homocystein
cyclic di-GMP
boxed = confirmed riboswitch
(+2 more:
• S-adenosyl methionine
• molybdenum cofactor)
queuosine precursor

Riboswitch Summary

RNA elements that control (“switch”) gene expression, *without* involvement of (transcription factor) proteins

Varied mechanism: Transcriptional, translational, on, off, combinatorial... Aptamer/expression platform.

Large diversity: Dozens of ligands, multiple aptamers for some, many operons, hundreds of species

Computationally challenging search/discovery

Many open problems!