RNA–RNA interactions

Stefan E Seemann

Center for non-coding RNA in Technology and Health (RTH),
Univ. of Copenhagen, *seemann@rth.dk*

COAT PhD summer school

August 23, 2012

Motivation

OPINION

The edges of understanding

Arthur D Lander*

Abstract

A culture's icons are a window onto its soul. Few would disagree that, in the culture of molecular biology that dominated much of the life sciences for the last third of the 20th century, the dominant icon was the double helix. In the present, post-modern, 'systems biology' era, however, it is, arguably, the hairball.

By hairball I refer here to those stunningly complicated network diagrams that grace the pages (and covers) of major journals with some regularity, in which the vertices or 'nodes' are annotated with symbols representing genes, proteins or metabolites, and the connectors or 'edges' are usually so numerous as to strain the resolution of monitors and printers (Figure 1).

While lacking much of the aesthetic appeal of a double helix, the hairball can be seen as iconic because it succinctly captures the distinctive flavor of systems biology. A molecular biologist and a systems biologist

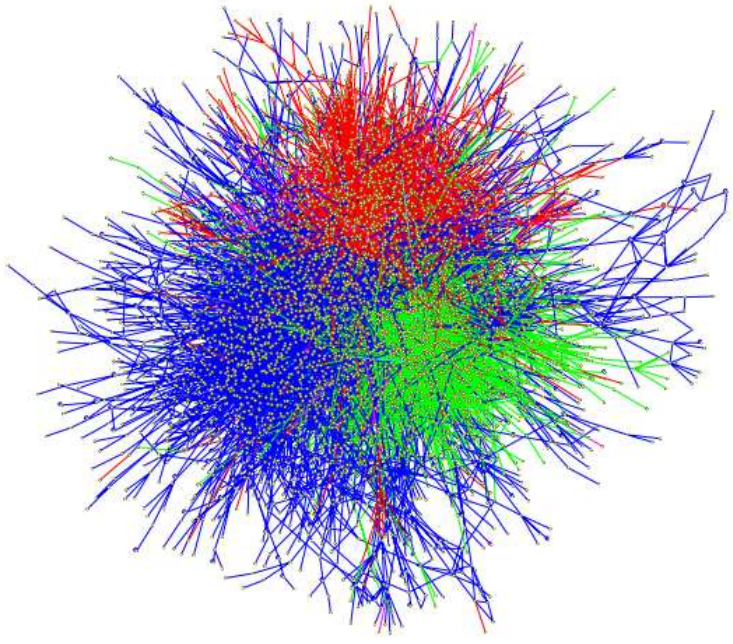
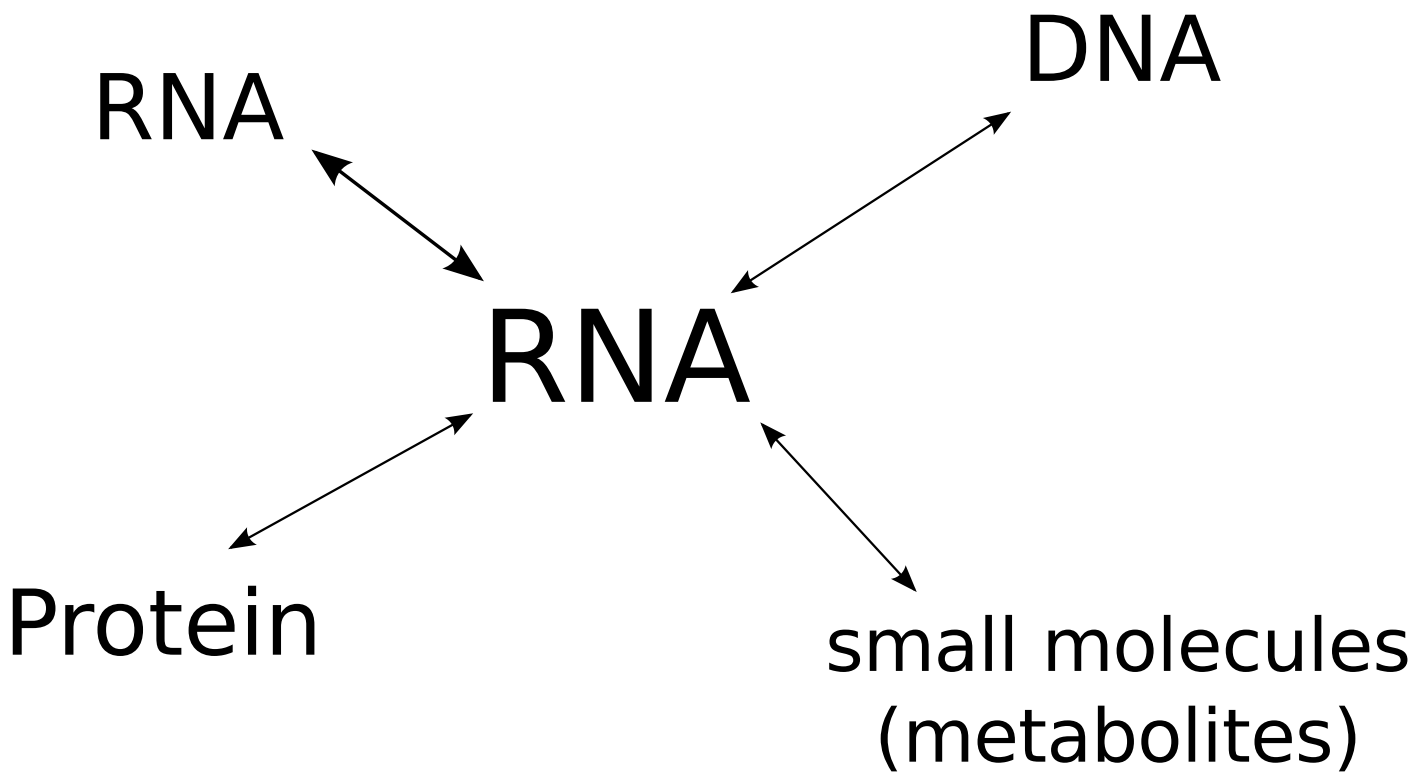
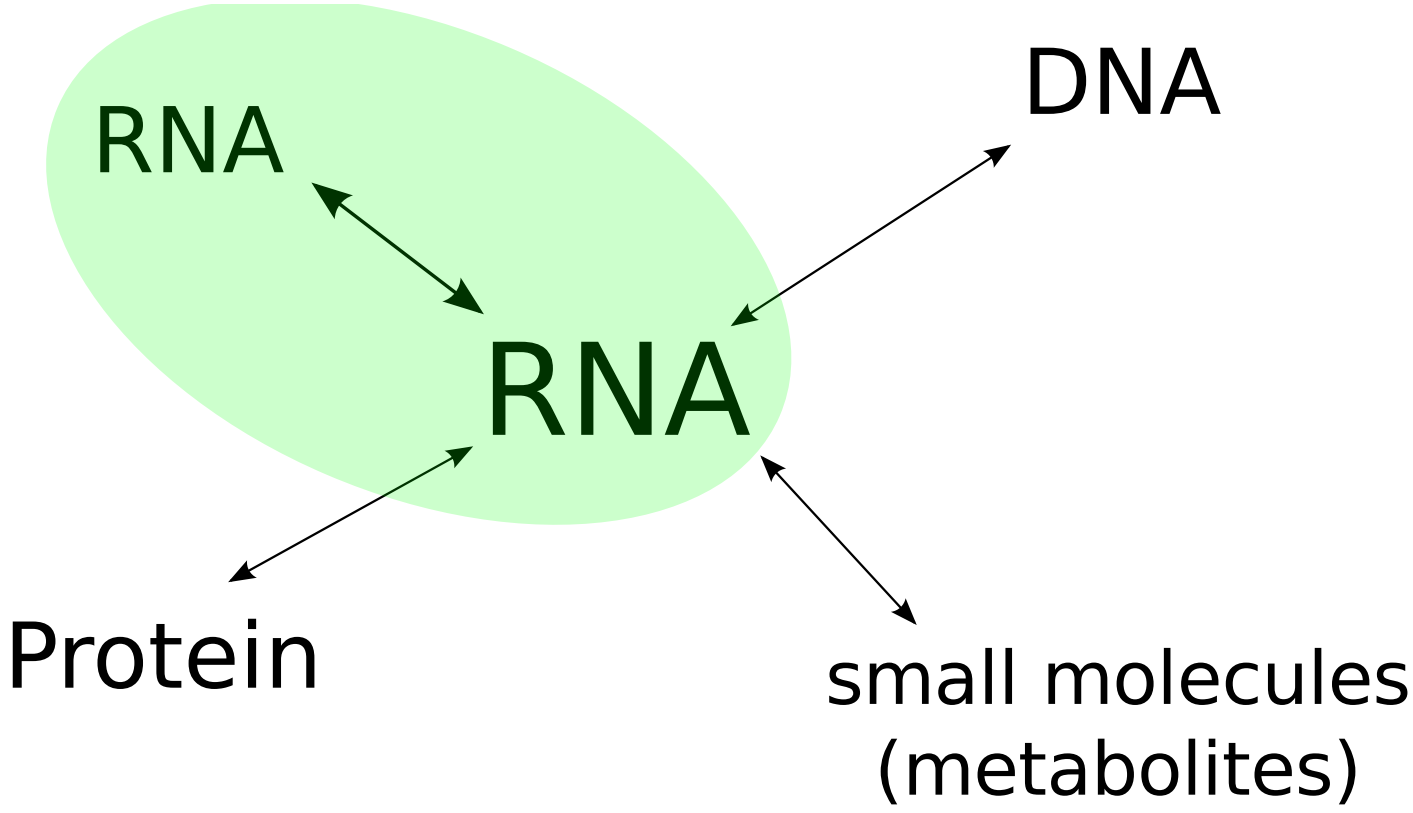


Figure 1. Human proteome, and its binding interactions. Depiction of the data as a hairball, an increasingly familiar image in the biology literature. Figure kindly provided by Nicolas Simonis and Marc Vidal, see [14].

Motivation

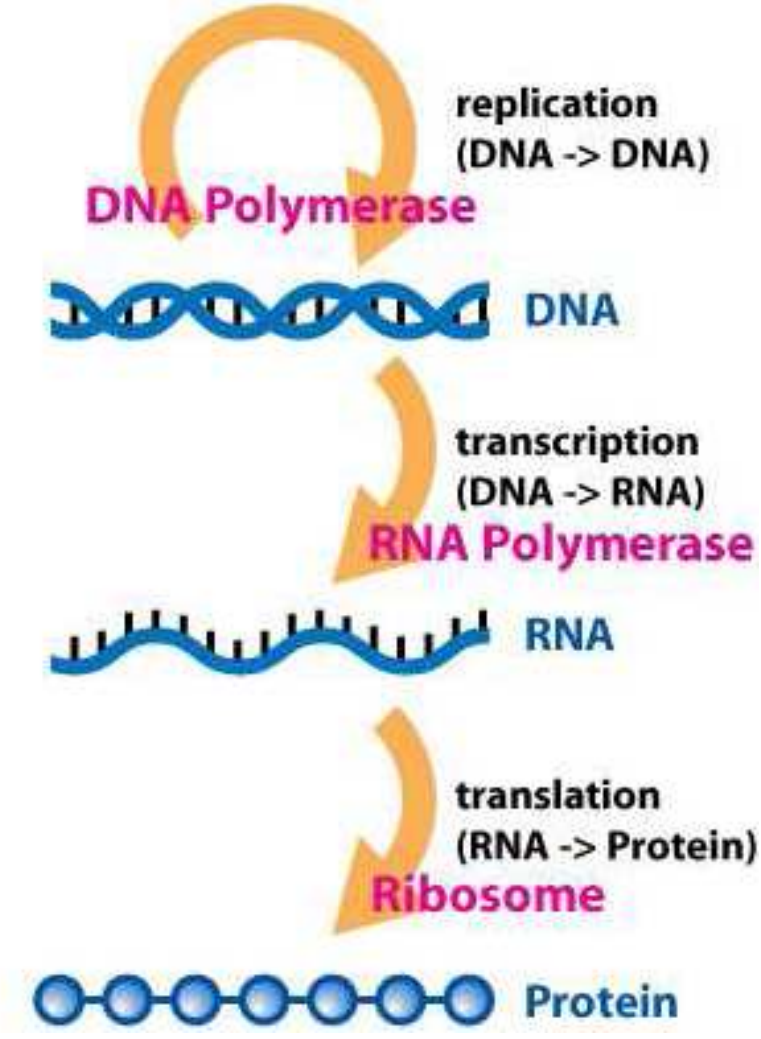


Motivation



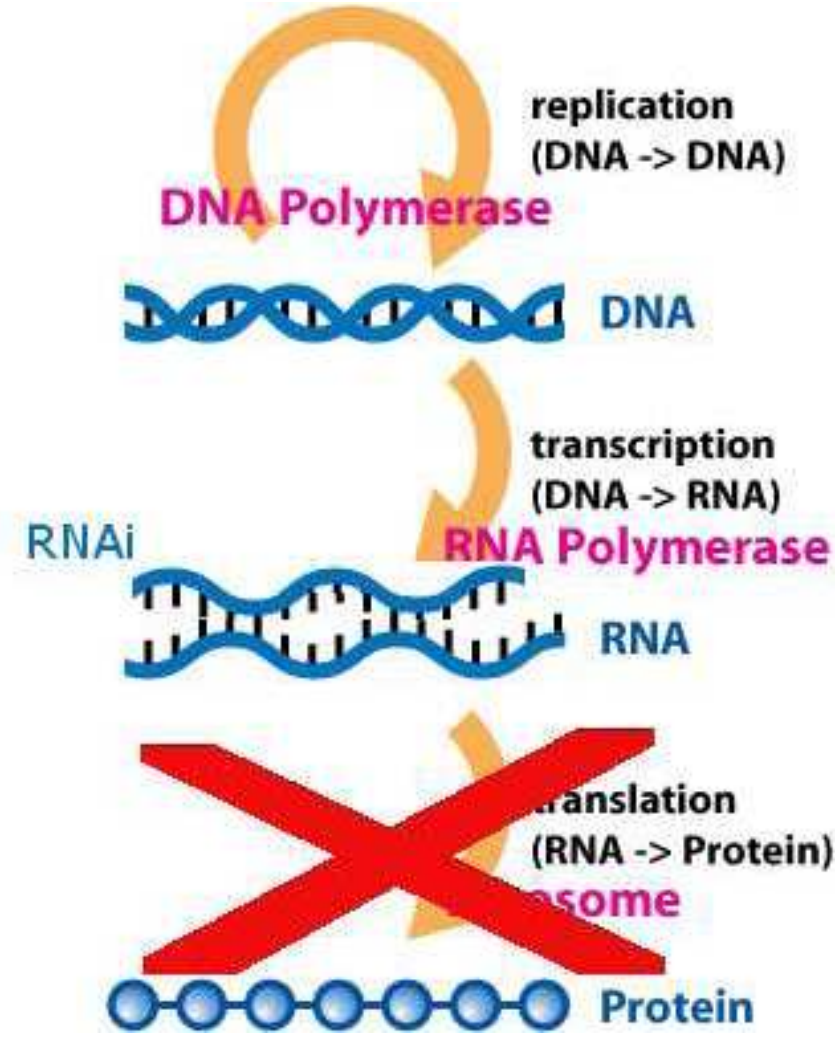
Central dogma

DNA $\rightarrow$ RNA $\rightarrow$ protein



Modified central dogma

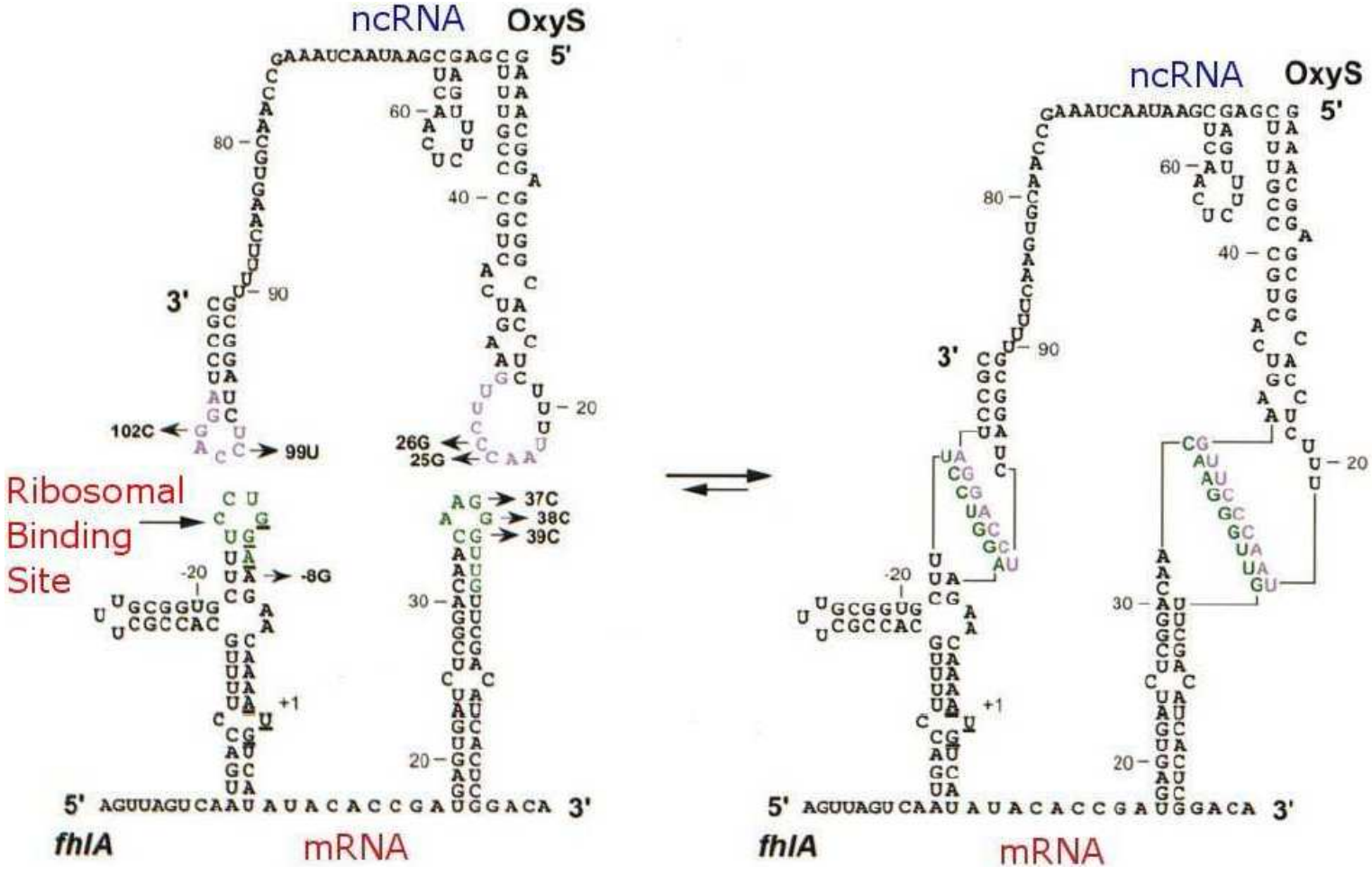
RNAs as active players $\rightarrow$ regulatory functions of ncRNA



Regulatory RNA

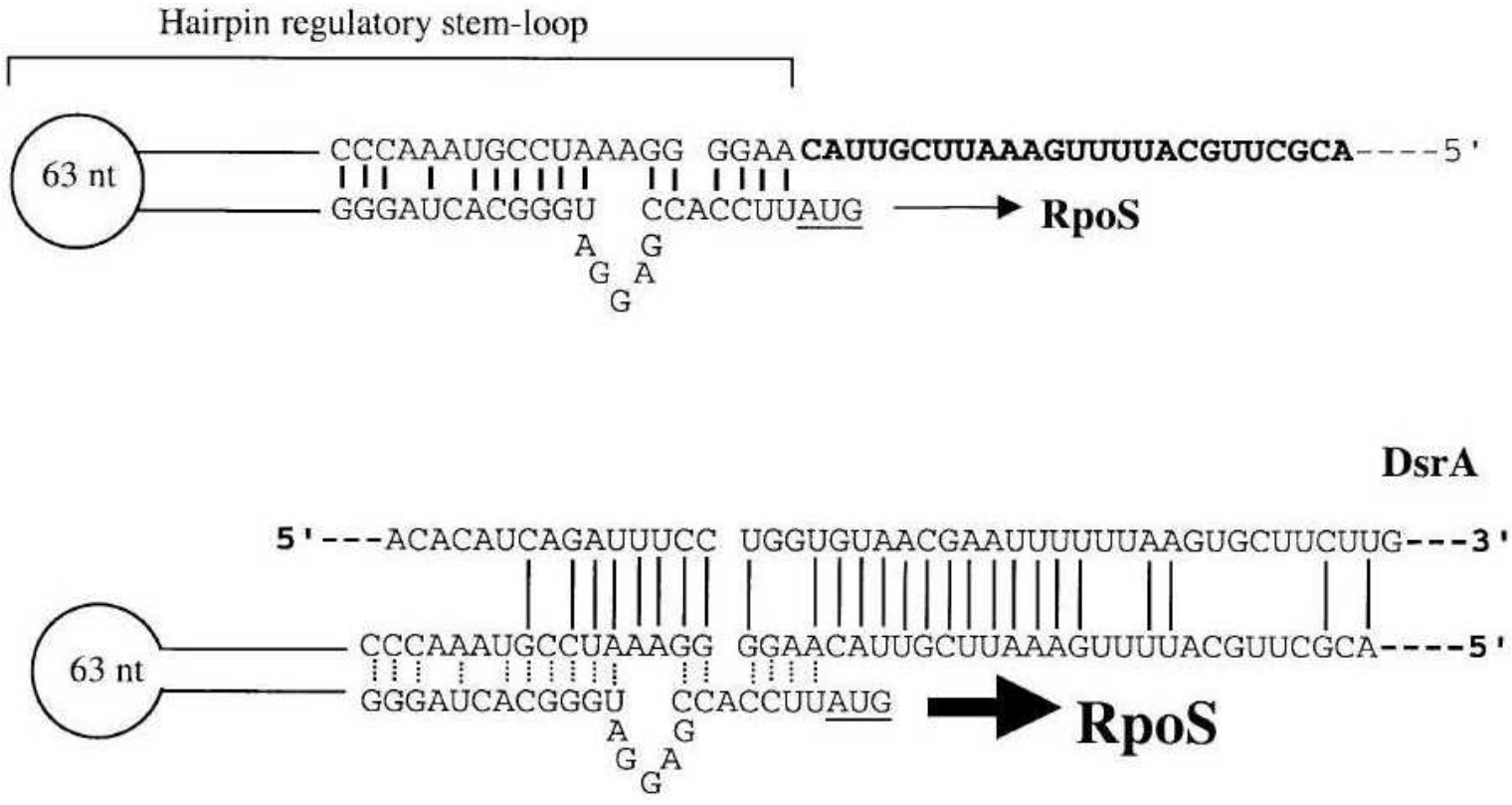
- post-transcriptional regulation of mRNAs
  by miRNAs, siRNAs
- regulation of RNA splicing and transcription factors
  by snRNAs
- guidance of chemical modification of RNAs (e.g. rRNAs)
  by snoRNAs
- ribozymes like RNase P

Repression example



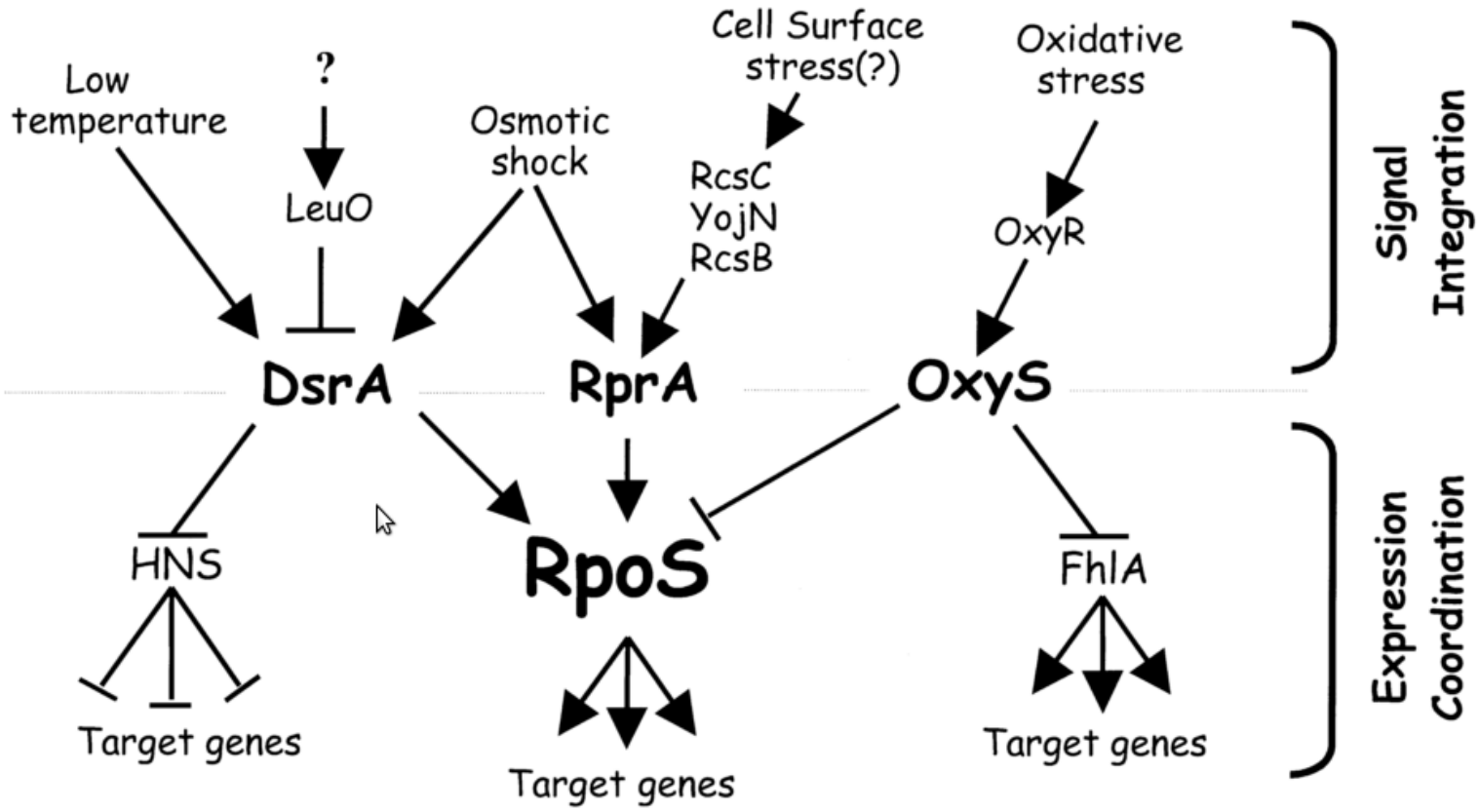
(Argaman et al, JMB 2000)

Activation example



(Repoila, Majdalani, and Gottesman, Mol. Microbiol. 2003)

Put it together



(Repoila, Majdalani, and Gottesman, Mol. Microbiol. 2003)

RNA folding algorithm

Each nested RNA (sub-)structure can be of only two different forms:

$\rightarrow$ dynamic programming

Minimum free energy (MFE):

$$E_{ij} = \min\left\{ E_{i+1,j}, \min_{i+3<k\leq j} \left\{ C_{ik} + E_{k+1,j} \right\} \right\}$$

Ensemble of all possible structures (Z . . . partition function):

$$Z_{ij} = Z_{i+1,j} + \sum_{k=i+4,j} Z^{C}_{ik} Z_{k+1,j}$$

How does comp. RNA target recognition work?

idea 0: find reverse complementary sequences
`BLAST` or better `GUUGle` (Gerlach & Giegerich 2006) (suffix trees + GU bps)

How does comp. RNA target recognition work?

idea 0: find reverse complementary sequences

`BLAST` or better `GUUGle` (Gerlach & Giegerich 2006) (suffix trees + GU bps)

```
5' C     A   C          3'
    CGUGG CCU AUUG GCA
    GCAUU GGA UAAC CGU
3'       A        U   A 5'
```

BUT: no internal loops, no energies

How does comp. RNA target recognition work?

idea 0: find reverse complementary sequences

`BLAST` or better `GUUGle` (Gerlach & Giegerich 2006) (suffix trees + GU bps)

```
5' C     A   C          3'
    CGUGG CCU AUUG GCA
    GCAUU GGA UAAC CGU
3'       A        U   A 5'
```

BUT: no internal loops, no energies

idea 1: maximize hybridization energy

`RNAhybrid` (Rehmsmeier 2004), `RNAduplex` (Bernhart 2006), `UNAFold` (Markham 2008), `TargetRNA` (Tjaden 2006), `RNAplex` (Tafer 2011), `RIsearch` (Wenzel 2012)

How does comp. RNA target recognition work?

idea 0: find reverse complementary sequences

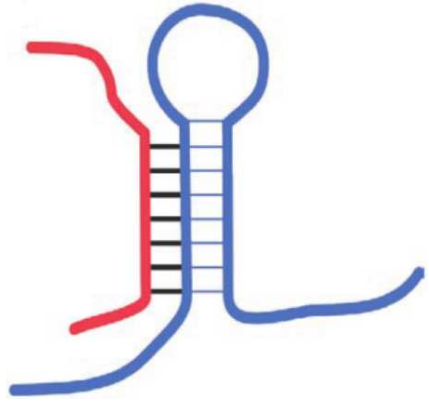
`BLAST` or better `GUUGle` (Gerlach & Giegerich 2006) (suffix trees + GU bps)

```
5' C     A   C          3'
    CGUGG CCU AUUG GCA
    GCAUU GGA UAAC CGU
3'       A        U   A 5'
```

BUT: no internal loops, no energies

idea 1: maximize hybridization energy

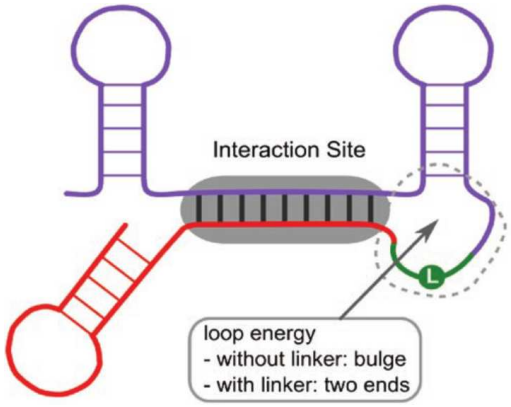
`RNAhybrid` (Rehmsmeier 2004), `RNAduplex` (Bernhart 2006), `UNAFold` (Markham 2008), `TargetRNA` (Tjaden 2006), `RNAplex` (Tafer 2011), `RIsearch` (Wenzel 2012)



BUT: no internal structure $\rightarrow$ **accessibility**

How does comp. RNA target recognition work?

Concatenation approaches:

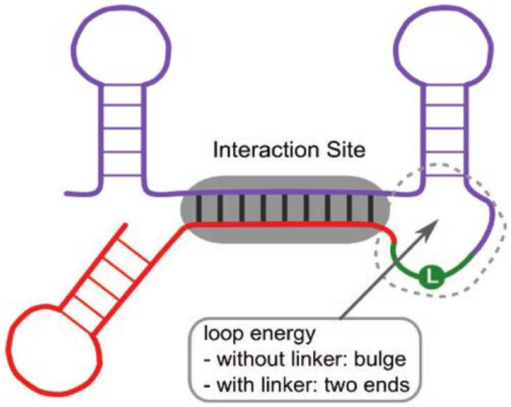


idea 2: common structure of concatenated sequences

`RNAcofold` (Bernhart 2006), `Pairfold` (Andronescu 2005)

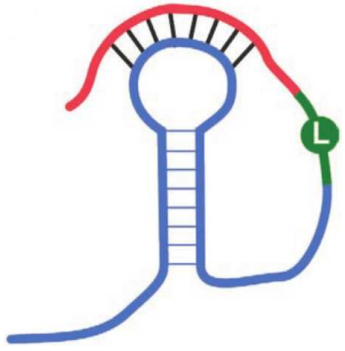
How does comp. RNA target recognition work?

Concatenation approaches:



idea 2: common structure of concatenated sequences

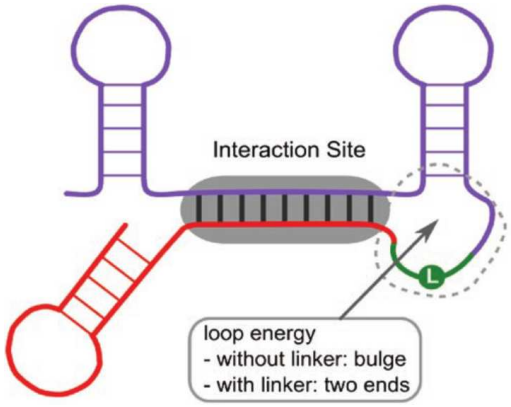
`RNAcofold` (Bernhart 2006), `Pairfold` (Andronescu 2005)



BUT: no pseudo-knots, no kissing hairpins

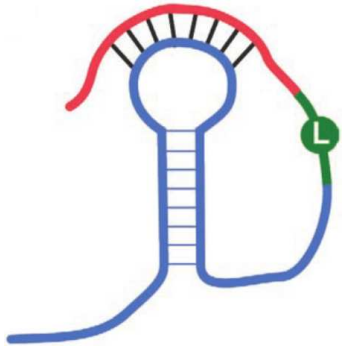
How does comp. RNA target recognition work?

Concatenation approaches:



idea 2: common structure of concatenated sequences

`RNAcofold` (Bernhart 2006), `Pairfold` (Andronescu 2005)



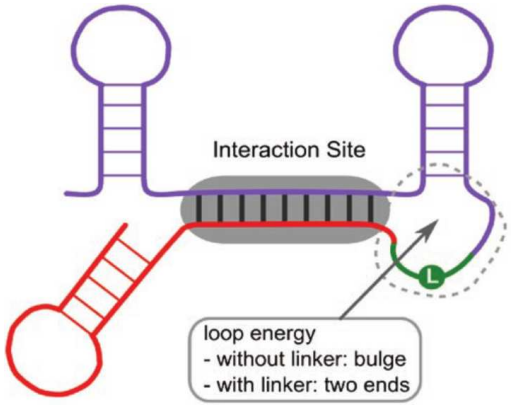
BUT: no pseudo-knots, no kissing hairpins

idea 3: special pseudoknotes of concatenated RNA sequences

`NUPACK` (Dirks 2003,2007), `PETcofold` (Seemann 2010,2011)

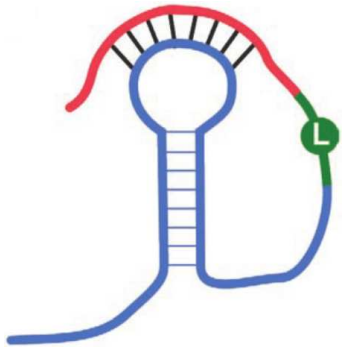
How does comp. RNA target recognition work?

Concatenation approaches:



idea 2: common structure of concatenated sequences

`RNAcofold` (Bernhart 2006), `Pairfold` (Andronescu 2005)



BUT: no pseudo-knots, no kissing hairpins

idea 3: special pseudoknotes of concatenated RNA sequences

`NUPACK` (Dirks 2003,2007), `PETcofold` (Seemann 2010,2011)

BUT: still not all structures

How does comp. RNA target recognition work?

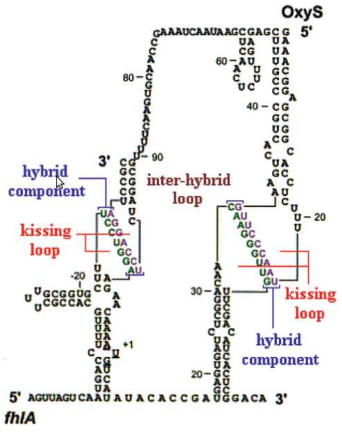
idea 4: combine interaction search with accessibility calculation

`RNAup` (Mückstein 2006), `IntaRNA` (Busch 2008), `RNApredator` (Eggenhofer 2011)

How does comp. RNA target recognition work?

idea 4: combine interaction search with accessibility calculation

`RNAup` (Mückstein 2006), `IntaRNA` (Busch 2008), `RNApredator` (Eggenhofer 2011)

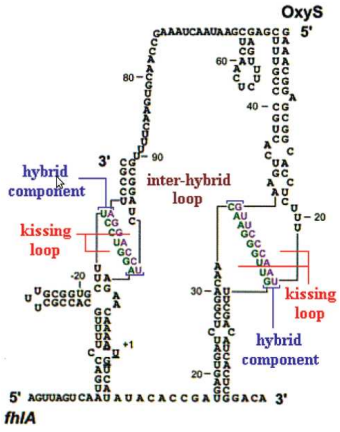


BUT: just one binding site

How does comp. RNA target recognition work?

idea 4: combine interaction search with accessibility calculation

`RNAup` (Mückstein 2006), `IntaRNA` (Busch 2008), `RNApredator` (Eggenhofer 2011)



BUT: just one binding site

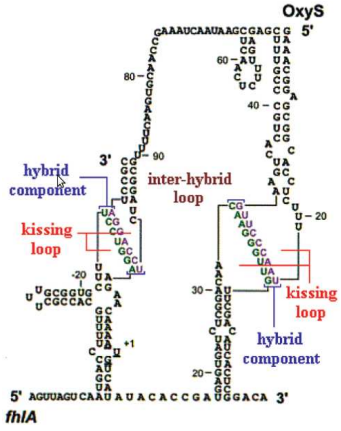
idea 5: consider both inter- and intramolecular base pairing

`IRIS` (Pervouchine 2004), `inteRNA` (Alkan 2005)

How does comp. RNA target recognition work?

idea 4: combine interaction search with accessibility calculation

`RNAup` (Mückstein 2006), `IntaRNA` (Busch 2008), `RNApredator` (Eggenhofer 2011)



BUT: just one binding site

idea 5: consider both inter- and intramolecular base pairing

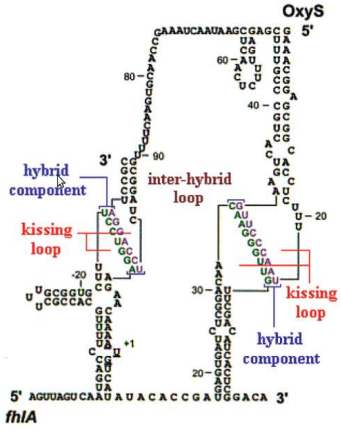
`IRIS` (Pervouchine 2004), `inteRNA` (Alkan 2005)

BUT: computationally expensive; no zigzags and crossing interactions as the general problem is NP-Hard

How does comp. RNA target recognition work?

idea 4: combine interaction search with accessibility calculation

`RNAup` (Mückstein 2006), `IntaRNA` (Busch 2008), `RNApredator` (Eggenhofer 2011)



BUT: just one binding site

idea 5: consider both inter- and intramolecular base pairing

`IRIS` (Pervouchine 2004), `inteRNA` (Alkan 2005)

BUT: computationally expensive; no zigzags and crossing interactions as the general problem is NP-Hard

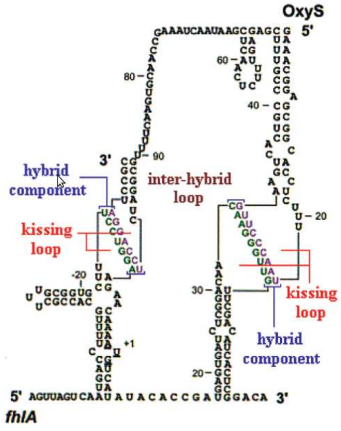
idea 6: complex joint secondary structures

`piRNA` (Chitsaz 2009), `RNArip` (Huang 2010)

How does comp. RNA target recognition work?

idea 4: combine interaction search with accessibility calculation

`RNAup` (Mückstein 2006), `IntaRNA` (Busch 2008), `RNApredator` (Eggenhofer 2011)



BUT: just one binding site

idea 5: consider both inter- and intramolecular base pairing

`IRIS` (Pervouchine 2004), `inteRNA` (Alkan 2005)

BUT: computationally expensive; no zigzags and crossing interactions as the general problem is NP-Hard

idea 6: complex joint secondary structures

`piRNA` (Chitsaz 2009), `RNArip` (Huang 2010)

BUT: very slow

RNA concentration

Bi-molecular reactions are concentration dependent.

Equilibrium constant of RNA–RNA interactions can be calculated as:

$$\frac{[AB]}{[A][B]} = K_{AB} = exp(-(G_{AB} - G_B - G_B)/RT)$$

$[AB]$ . . . concentration of hetero-dimer
$[A]$, $[B]$ . . . concentrations of monomers A
G_{AB} . . . free energy of structure ensemble of hetero-dimer
G_A, G_B . . . free energies of structure ensemble of monomers

Which method to choose?

Why aren't we using the most accurrate method all the time?

Which method to choose?

Why aren't we using the most accurrate method all the time?

`RNAhybrid` $\rightarrow O(mn)$ time

vs

`RNAcofold` $\rightarrow O(n^3)$ time: takes less than a second for two 110nt RNAs (OxyS-*fhlA*)

vs

`piRNA` $\rightarrow O(n^4m^2 + n^2m^4)$ time: takes about 1 day on 64 CPUs with 150GB RAM for two 110nt RNAs (OxyS-*fhlA*)

Which method to choose?

Why aren't we using the most accurrate method all the time?

`RNAhybrid` $\rightarrow O(mn)$ time

vs

`RNAcofold` $\rightarrow O(n^3)$ time: takes less than a second for two 110nt RNAs (OxyS-*fhlA*)

vs

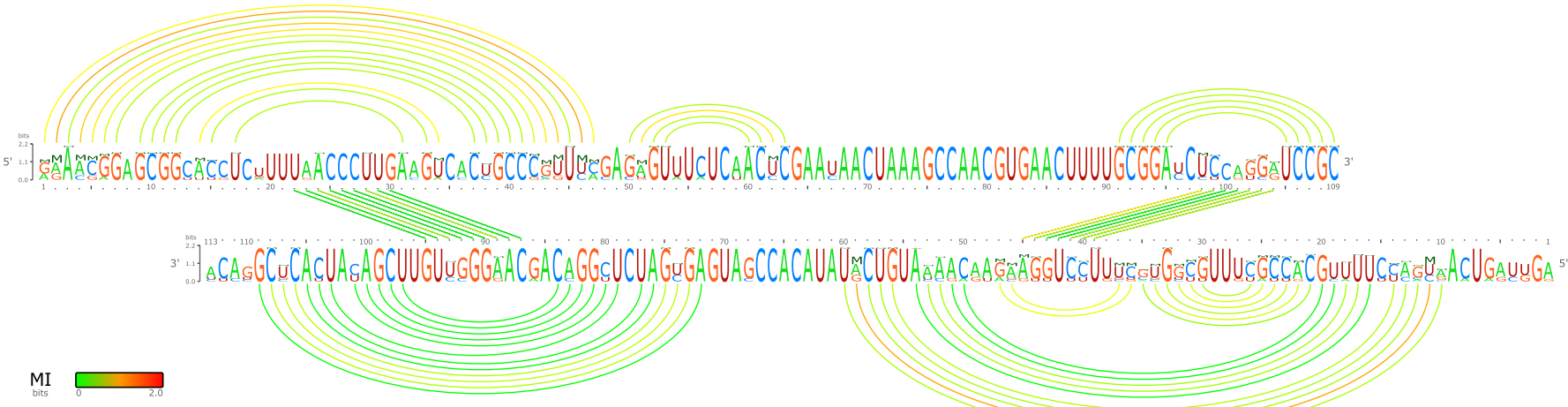
`piRNA` $\rightarrow O(n^4m^2 + n^2m^4)$ time: takes about 1 day on 64 CPUs with 150GB RAM for two 110nt RNAs (OxyS-*fhlA*)

$\Rightarrow$ Emerging need for computational methods that allow efficient detection of RNA–RNA interaction sites on transcriptome-wide scale.

$\Rightarrow$ Conservation is a powerful filter to narrow down the search to conserved interactions.

Interaction site conservation

Conservation of OxyS – *fhlA* interaction:



drawn by `RILogo` (Menzel 2012)

only method that considers conservation is `PETcofold` (Seemann 2011)

Genome-wide target screen

”Scanning variants“ of RNA folding algorithms:

- sRNA–mRNA interactions
  - `TargetRNA` (Tjaden 2006)
  - `RIsearch` (Wenzel 2012)
  - `RNAplex` (Tafer 2011)
  - `RNApredator` (Eggenhofer 2011)
- known binding motives
  - `biRNA` (Chitsaz 2009)
- microRNA specific
  - `miRanda` (Betel 2008)
- H/ACA snoRNA specific
  - `RNAsnoop` (Tafer 2010)

In-silico design of regulators

Small–interfering RNAs (siRNAs):

- assemble into RISC (RNA–induced silencing complex) which cleaves complementary mRNAs
- used for gene silencing
- designed exactly complementary to target site
- off-target effects have to be considered

Predictors:

- `DSIR` (Vert 2006) – linear model to map sequence features of siRNA to its expression efficiency
- `RNAxs` (Tafer 2008) – sequence features + accessibility

Summary

- RNA molecules function through interactions
- Computational problems are:
  1. search genomes or transcriptomes for targets
  2. characterize the joint structure of interacting RNAs
  3. design regulatory RNAs (siRNAs)
- Interaction formation often initiated at well-accessible *intra*-molecular structures:
  1. complementarity $\rightarrow$ interaction energy
  2. accessibility $\rightarrow$ low internal base pair probability
- Large number of methods for different applications