Folding with Fancy Constraints
Secondary Structures from Probing Data

Peter F. Stadler

Bioinformatics Group, Dept. of Computer Science &
Interdisciplinary Center for Bioinformatics,

University of Leipzig

Max Planck Institute for Mathematics in the Sciences

Center for noncoding RNA in Technology and Health

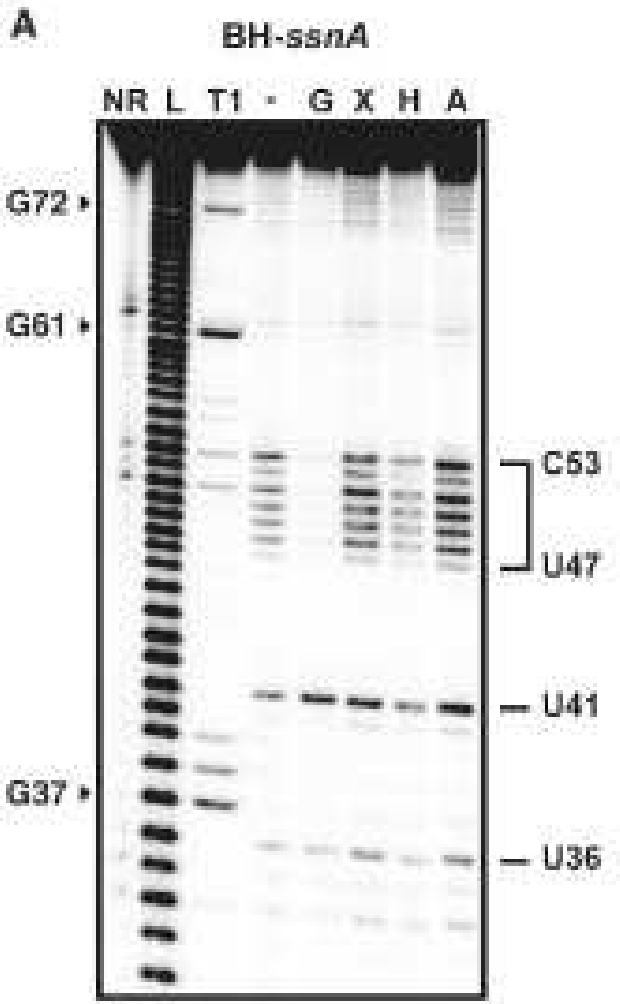
RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology

Institute for Theoretical Chemistry, Univ. of Vienna (external faculty)

The Santa Fe Institute (external faculty)

Bled, Feb 13-20 2010

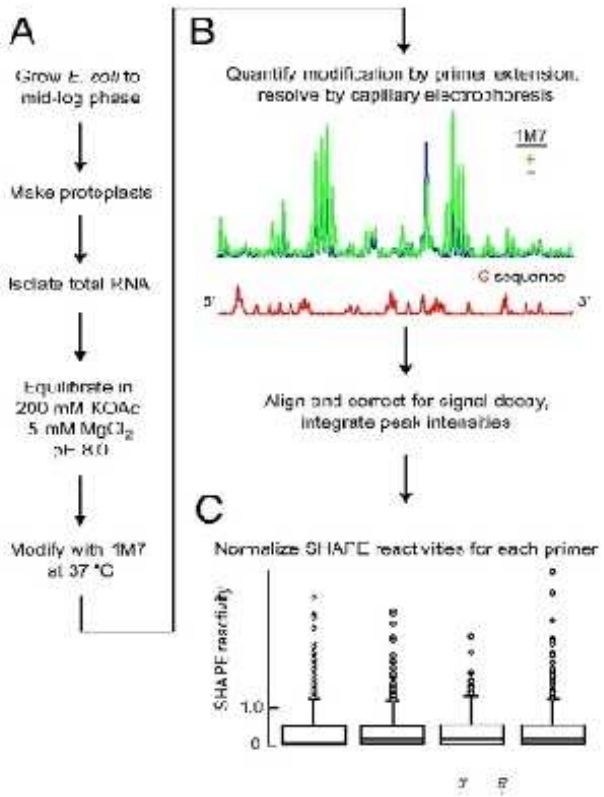
In-line probing

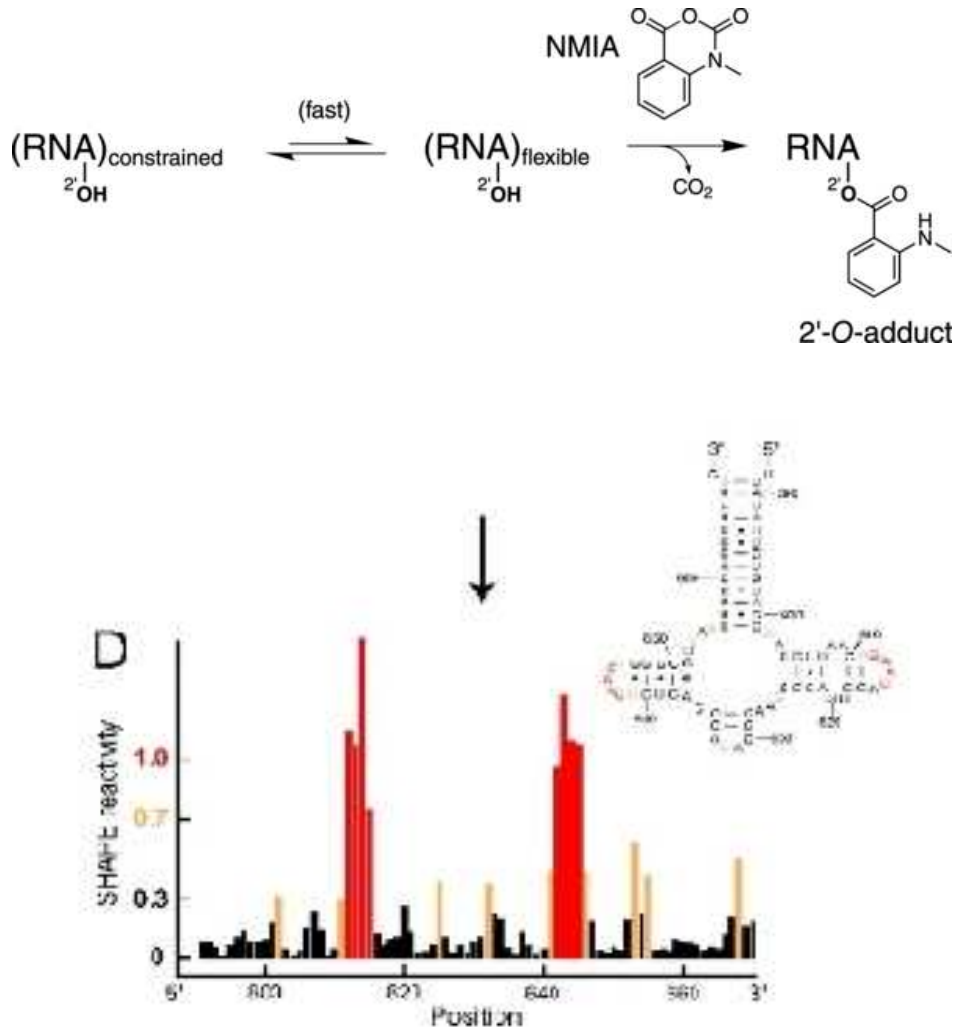


Structure-dependent cleavage of RNA

Guanine riboswitch

SHAPE Analysis

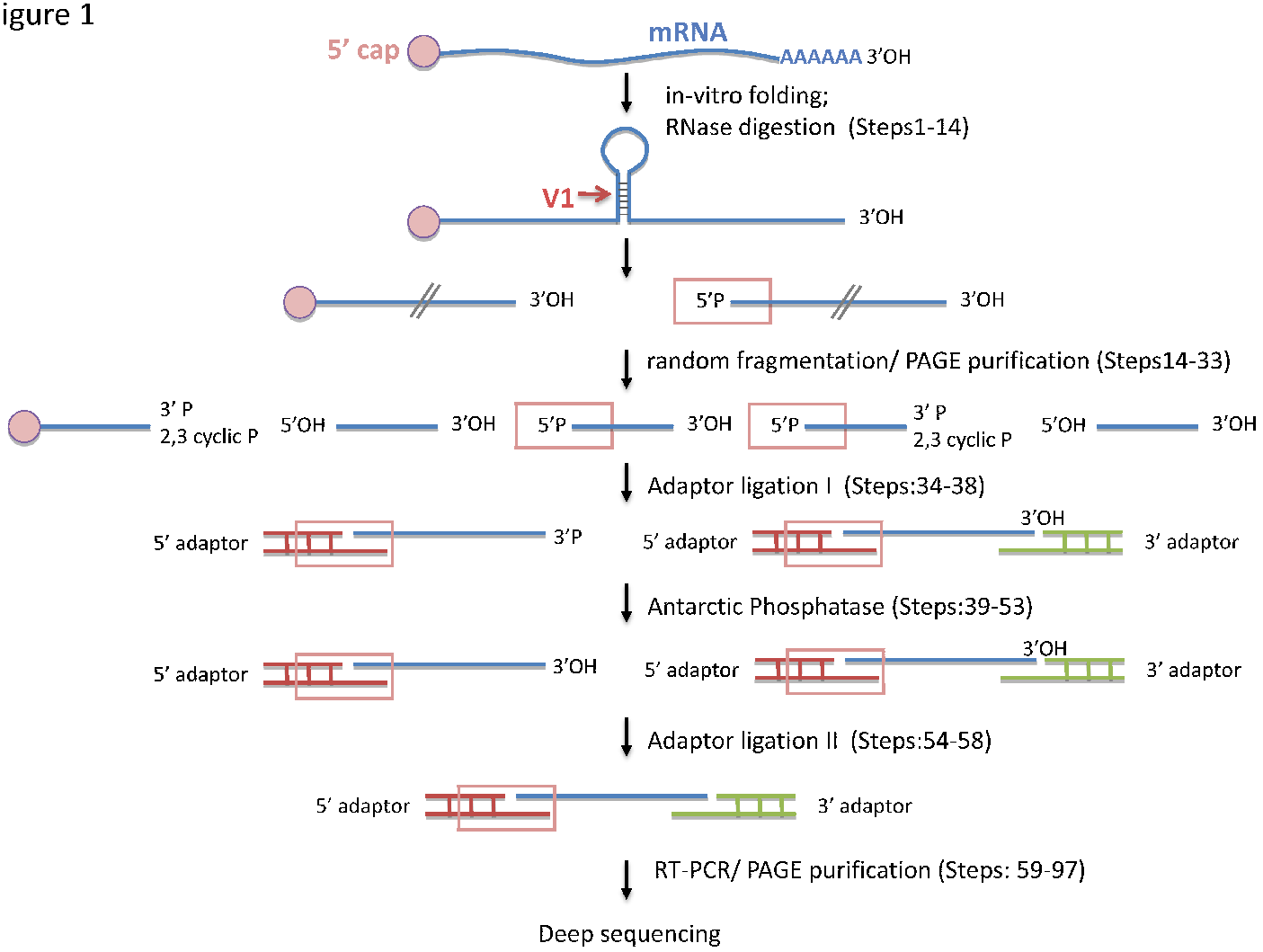




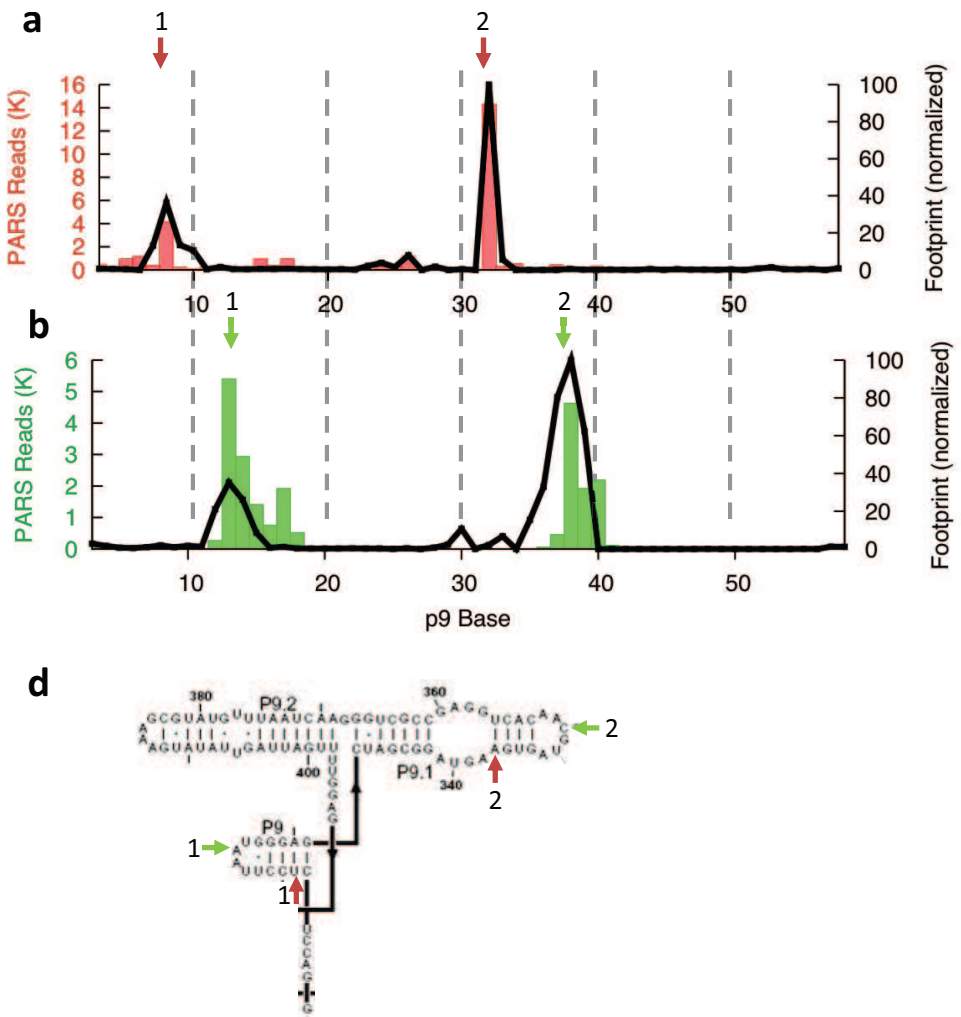
Deigan, Li, Mathews, Weeks 2009

Segal's PARS Protocol

Figure 1



Segal's PARS Protocol



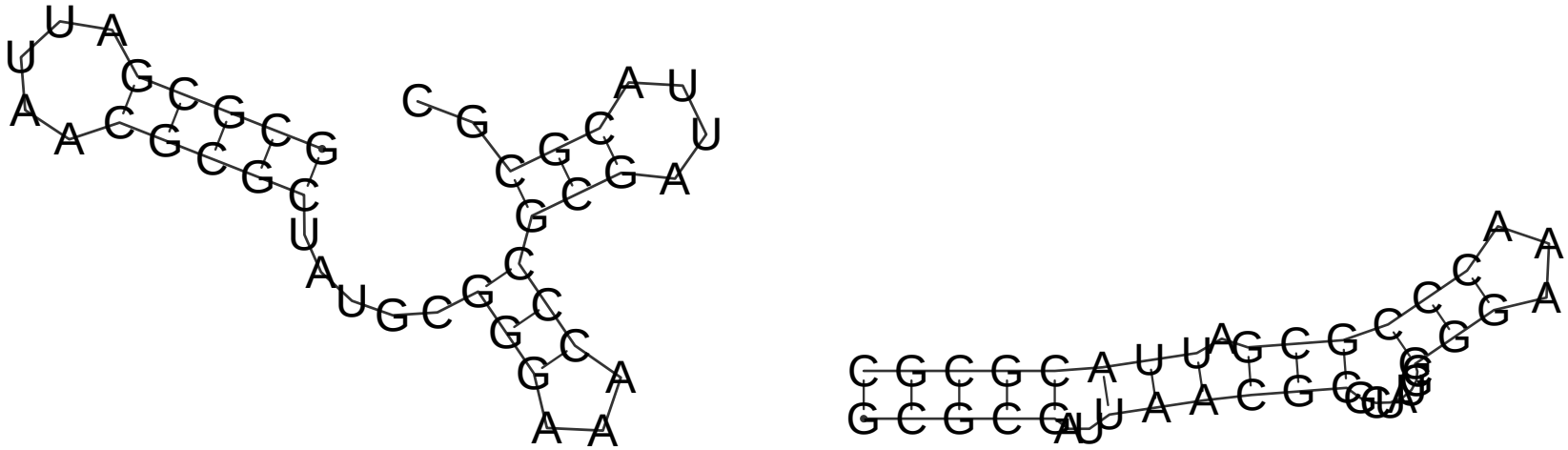
double-stranded cutter RNase V1

single-stranded cutter RNase S1

How to analyze such data?

1. Convert the measures signal (either SHAPE reactivities or paired and unpaired PARS intensities) to a probability $q(k)$ that position k is unpaired.
2. Use $q(k)$ to infer infer the secondary structure structure

Incompleteness



```
GCGCGATTAACGCGCTATGCGGGAAACCCGCGATTACGCGC
((((((.....))))))...((((((...))))))((((.....))))))    -9.30
((((((.....((((((...))))))((((...))))))....))))))    -8.50
XXXXXX.....XXXXXX...XXXXXX...XXXXXXX....XXXXXX
```

Secondary structure is not uniquely determined by accessibility, i.e., the probability that individual base pairs are unpaired. The left (upper) structure is the most stable alternative.

SHAPE analysis

Use a position-dependent “pseudoenergy” contribution

$$\Delta G = m \ln[1 + q(k)] + b$$

to give a bonus for unpaired bases with high SHAPE reactivity

- used successfully e.g. for HIV RNA, implemented in `RNAstructure`
- this works (surprisingly) well, but it is not a very *elegant* solution:
  1. why give a bonus to positions that are already predicted correctly?
  2. there is no good interpretation of the folding energies with the pseudoenergies
  3. it seems hard to get a thermodynamic prediction out of the combined model

A more fancy approach

- **Observation:** both the measure of exposure, $\vec{q}$, and the standard Turner energy model contains errors and inaccuracies.
- **Given** and energy model, we can compute the probability p_i that nucleotide i remains unpaired.
  Use `RNAfold -p` and sum rows/columns of the base pair probability matrix.
- We can try to compute explicit corrections to the energy model to fit the data by adding a extra correction terms ϵ_μ to the energy model. These are obtained by minimizing the error functional

$$F(\vec{\epsilon}) = \sum_\mu \frac{\epsilon_\mu^2}{\tau_\mu^2} + \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(p_i(\vec{\epsilon}) - q_i\right)^2$$

Solving the Minimization Problem

The minimum of the error term satisfies $\partial F / \partial \epsilon_\mu = 0$ for all parameters.
Note that computing $p_i(\vec{\epsilon})$ needs the evaluation of the partition function for a perturbed energy model.
We use a gradient optimizer

$$\epsilon'_\mu = \epsilon_\mu - a_t \frac{\partial F}{\partial \epsilon_\mu} = \left(1 - \frac{2a_t}{\tau_\mu^2}\right) \epsilon_\mu - 2a_t \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(p_i(\vec{\epsilon}) - q_i\right) \frac{\partial p_i}{\partial \epsilon_\mu}(\vec{\epsilon})$$

The parameter a_t for the stepsize adjustments can be estimated.

Constrained partition functions

Since ϵ_μ denotes the energy contribution that is added to all secondary structures that contain a particular "structural feature" μ, we can subdivide the structure ensemble into those structure that "have μ", and those that do not.
$Z[i](\epsilon_\mu)$ . . . partition function with i unpaired and energy model ϵ_μ.

$$
\begin{aligned}
Z[i](\epsilon_\mu) &= Z[i](0) - Z[i,\mu](0) + Z[i,\mu](\epsilon_\mu) \\
Z(\epsilon_\mu) &= Z(0) - Z[\mu](0) + Z[\mu](\epsilon_\mu) \\
Z[\mu](\epsilon_\mu) &= Z[\mu](0)\exp(-\epsilon_\mu/RT) \\
Z[i,\mu](\epsilon_\mu) &= Z[i,\mu](0)\exp(-\epsilon_\mu/RT)
\end{aligned}
$$

$p_i(\,.\,) = Z[i](\,.\,)/Z(\,.\,)$

Constrained partition functions

$$\left.\frac{\partial p_i}{\partial \epsilon_\mu}\right|_{\epsilon_\mu \to 0} = \left.\frac{\partial}{\partial \epsilon_\mu} \frac{Z[i](0) - Z[i,\mu](0)\,(1-\exp(-\epsilon_\mu/RT))}{Z(0) - Z[\mu](0)\,(1-\exp(-\epsilon_\mu/RT))}\right|_{\epsilon_\mu \to 0}$$

$$= \frac{1}{RT}\left[\frac{Z[i](0)}{Z(0)}\frac{Z[\mu](0)}{Z(0)} - \frac{Z[i,\mu](0)}{Z[i](0)}\frac{Z[i](0)}{Z(0)}\right]$$

$$= \frac{1}{RT} p_i(0)\,[p[\mu](0) - p[\mu|i](0)]$$

This simplifies for position-specific corrections ϵ_j only:

$$\left.\frac{\partial p_i}{\partial \epsilon_j}\right|_{\epsilon_j \to 0} = \frac{1}{RT} p_i(0)\left[p_j(0) - p[j|i](0)\right]$$

Numerical Evaluation

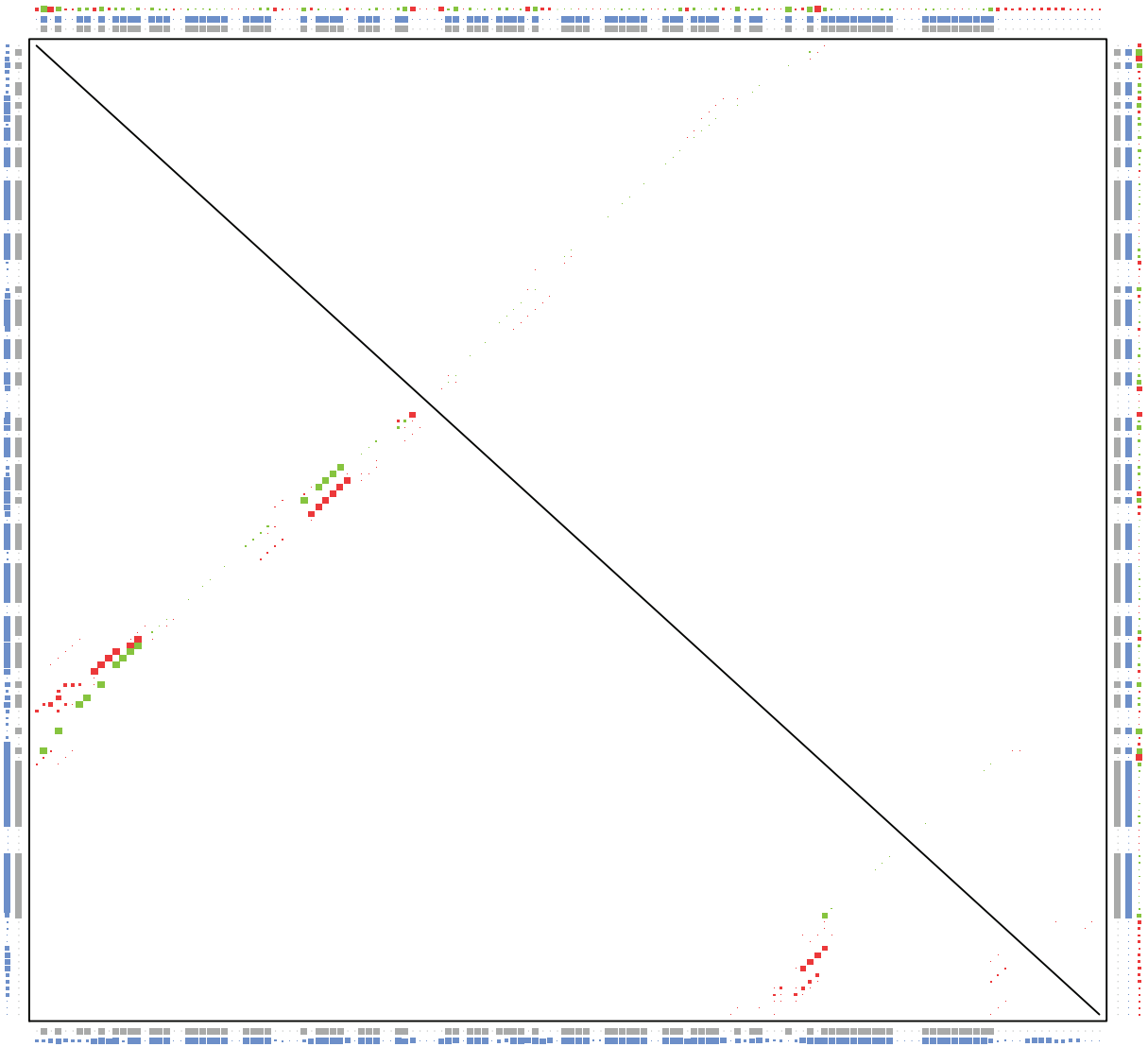
- Computing constrained partition functions with position i unpaired need n partition function computations, thus $O(n^4)$
- Suboptimal folding: sample $p_i(0)$ and $p[j|i](0)$ approximate, but runs in $O(n^3)$

Does it work?

Construct a test system where we know the outcome:

- Use `RNAfold` as the ground truth.
  Compute $\vec{q}$ as the vector of base pairing probabilities
- Use the Nussinov (maximum matching) algorithm with $\beta = -3$, -2, or -1 for GC, AU, or GU pairs
- compute the correction energies by minimizing $F(\vec{\epsilon})$.
- Compare the base pairing probability matrix computed with "Nussinov $+\vec{\epsilon}$" with the `RNAfold` ground truth.

It works!



lower left: difference between Nussinov and `RNAfold` for a domain from a 16S rRNA

upper right: difference between "Nussinov $+\vec{\epsilon}$" and `RNAfold`

in prediction but not in reference

in reference but not in prediction

What is this good for?

1. Infer secondary structures from (large scale) probing data
   this can of course also be done by Matthews method
2. Detect discrepancies between observations and folding prediction:
   Are there **localized** regions where energy corrections are necessary?
   This could be used to detect possible binding regions of ligands (small molecular or protein) or as locations of un-usual RNA motifs (such as G quartetts)
3. Possible applications also better understanding refolding in which ligands are involved

Ongoing work

1. Beautify implementation of `RNApbfold` (according to Wash we might rename it `RNAsegfault`)
2. Understand why Matthews simple bonus energies work as well or even better when affine parameters are optimized (possibly because that optimizes predictive power on a relatively small sample?)
3. deal with missing data:
   Often, probing does not “touch” certain parts of an RNA, leading to a large fraction of missing data.
4. **Open Problem**: Convert different types of measurements to $\vec{q}$ for different types of experiments

Joint Work

Stefan Washietl (MIT) & Ivo L. Hofacker (TBI)