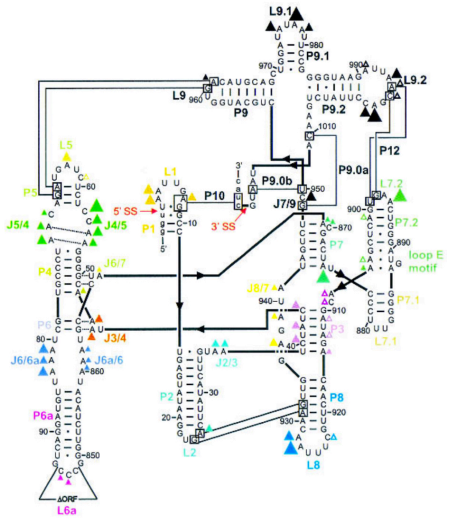
Unspecific binding but specific disruption of the group I intron by the StpA chaperone

Vladimir Reinharz

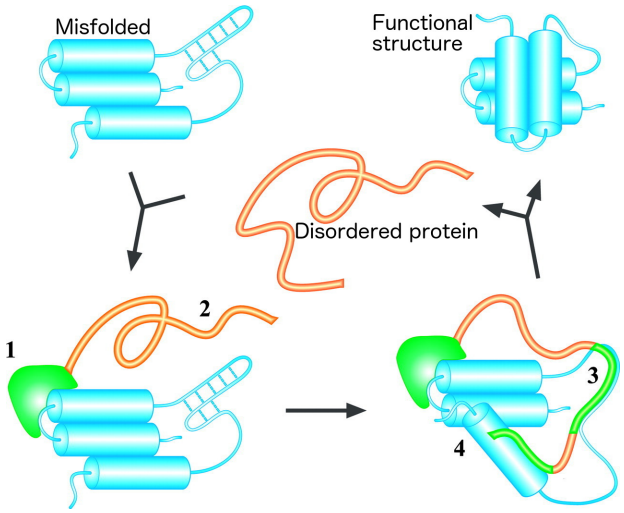
CSLM / IBS

Group I intron SS



Waldish et al., Genes and Development, 2018

StpA entropy transfer (hypothesis)

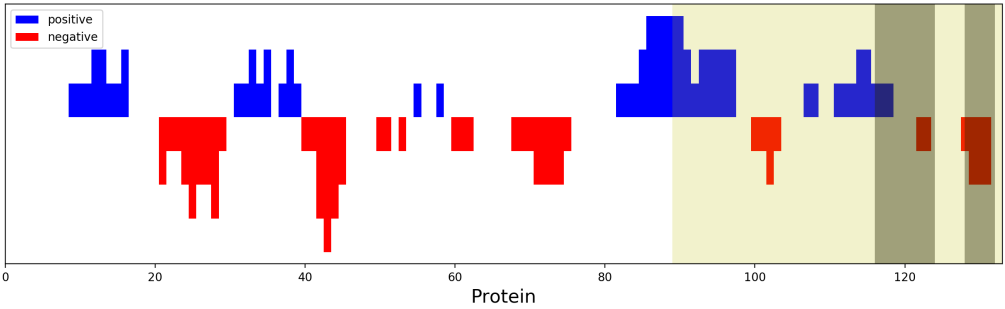


Tompa & Csermely, FASEB 2004

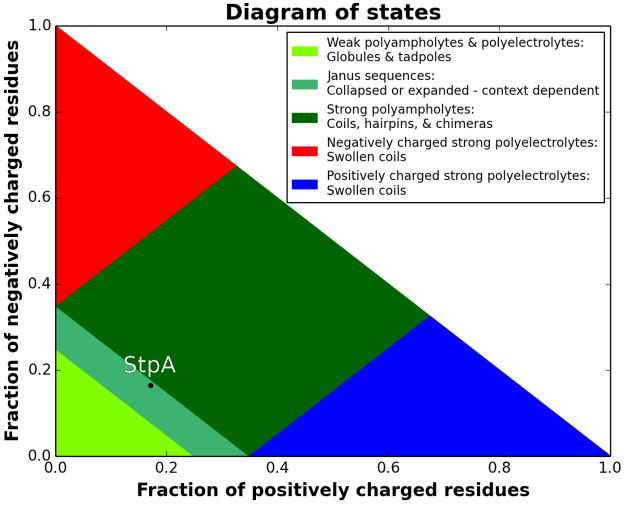
Hypothesis

1. Protein “disorderedness” comes from charge distribution
2. Protein-RNA fast and transient interactions are mediated by the charged regions
3. Mechanism evolutionary conserved

StpA charge distribution (average over 5 AA)



StpA disordered state



Direct Information

$$\mathrm{DI}_{ij} = \sum_{A,B} P_{ij}(A,B) ln \frac{\overbrace{P_{ij}(A,B)}^{\text{"}\mathbb{P}\text{" of observing } A \text{ at pos } i,\ B \text{ at pos } j}}{\underbrace{f_i(A)}_{\text{Frequency of } A \text{ at position } i}\ f_j(B)}$$

where

$$P_{i,j}(A,B) = \frac{1}{Z_{ij}} \mathrm{e}^{e_{ij}(A,B)+\tilde{h}_i(A)+\tilde{h}_j(B)}$$

for amino acids (nucleotides) A and B at positions i and j.
Matrix e is inverse of correlation matrix

Different methods

The goal is to estimated the correlation matrix without the cross links
What's new:

- Different alphabets (amino acids and nucleotides)
- Different sequence diversity (normalization factor)

StpA protein sequences

- From *Escherichia coli*
- 134 AA
- 5749 unique homologues identified with `Jackhmmer`
- 7539 different taxons

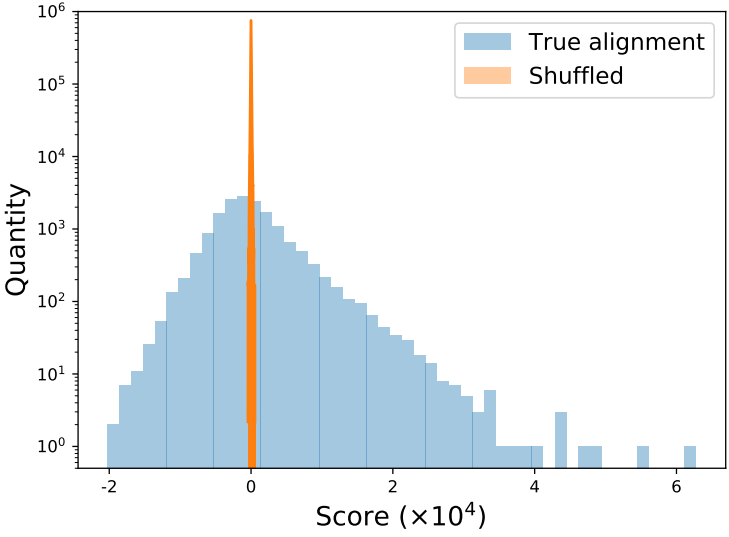
group I intron sequences

- 14 sub-families
- Best match `IA2` (e-val 1.7×10^{-36})
- Download all 633 GB of sequences (whole genome sequences / single read)
- `InfeRNAl` identifies 7542 matches (471 unique)

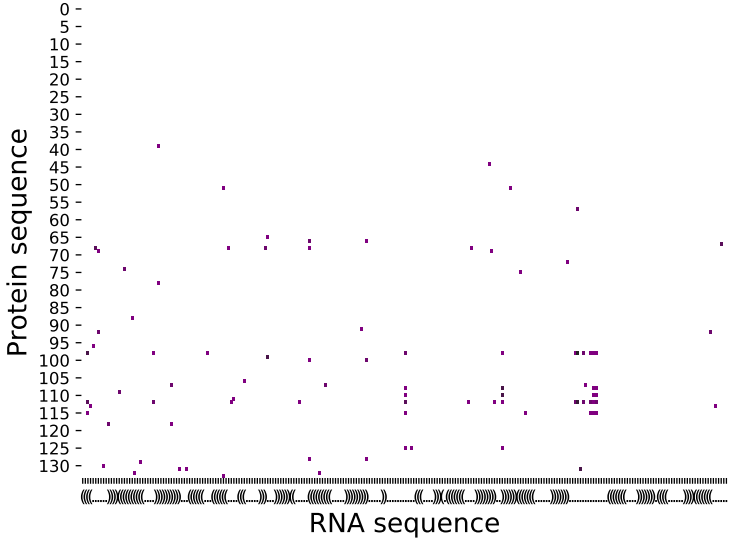
StpA + group I introns

- Concatenated every combinations from same taxon
- 10 013 unique pairs
- 95aa and 184nt per row after $> 50\%$ gap removal

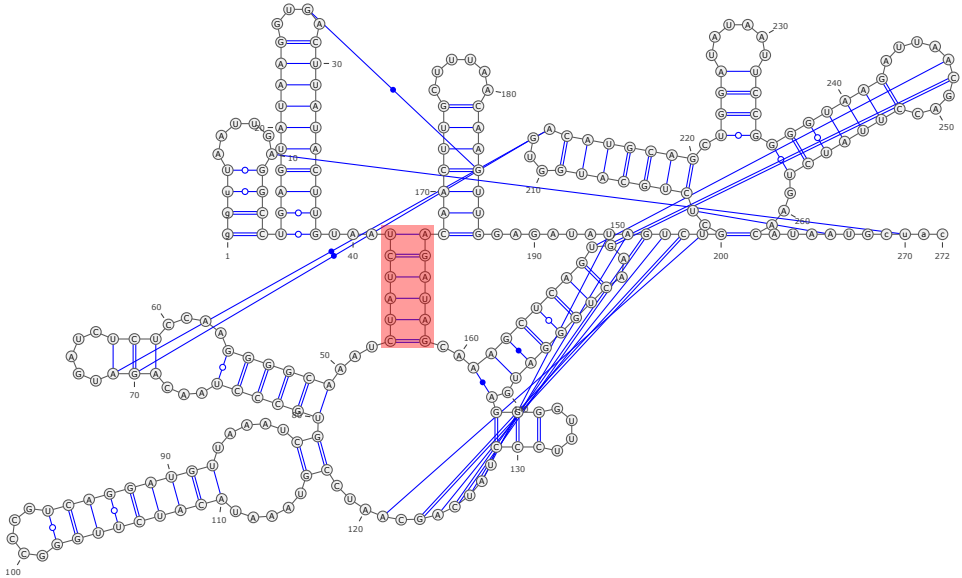
DCA scores distribution



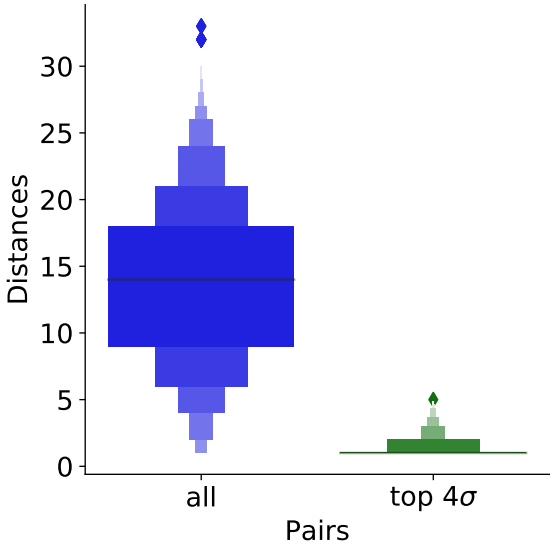
DCA scores heatmap



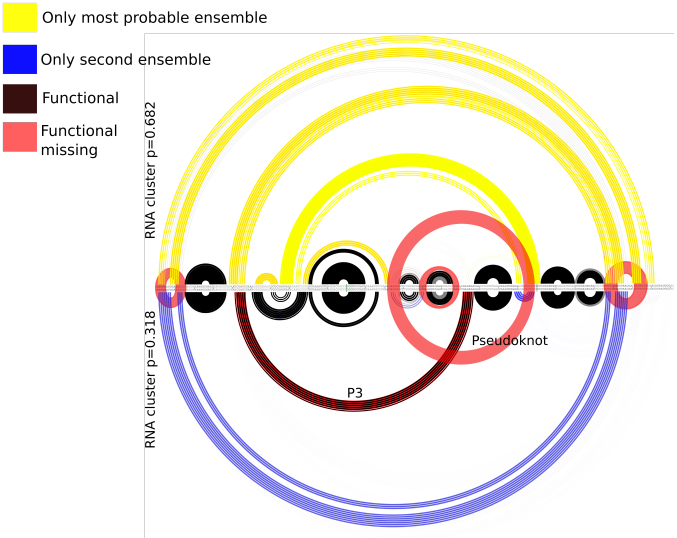
RNA secondary structure



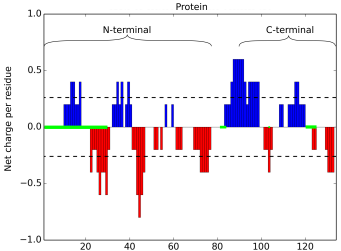
Top scores distances in RNA

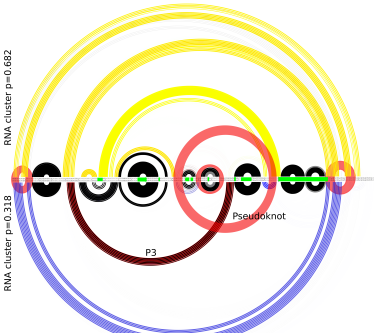


RNA 2 main structures ensembles

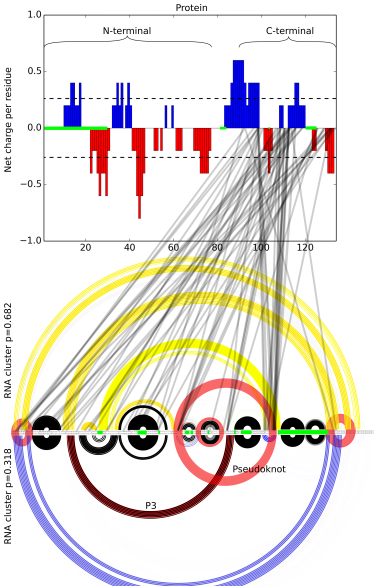


RNA–protein

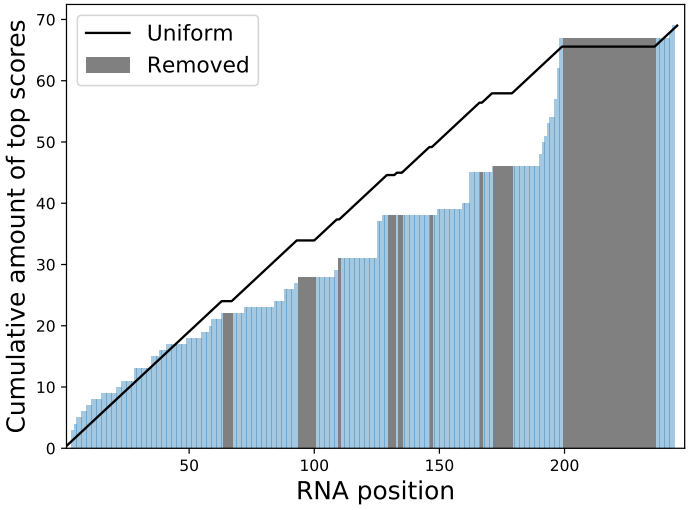

Protein
N-terminal
C-terminal
Net charge per residue
1.0
0.5
0.0
−0.5
−1.0
20
40
60
80
100
120


RNA cluster p=0.682
RNA cluster p=0.318
Pseudoknot
P3

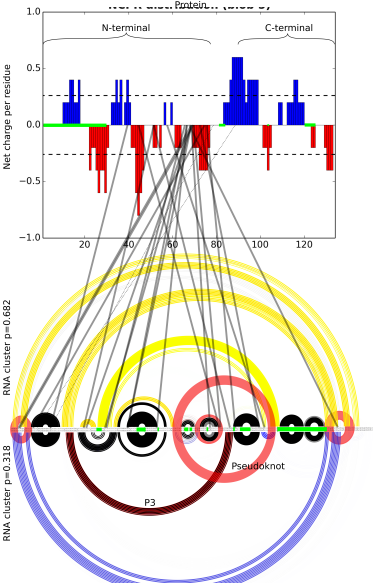
C-terminal binding interactions are spread



C-terminal binding interactions are spread

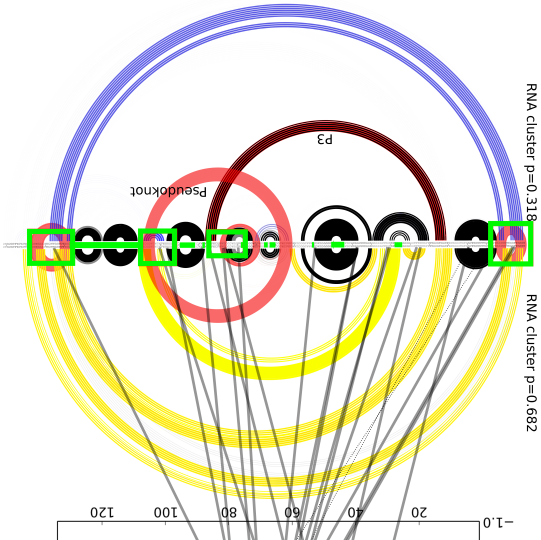


N-terminal disruptive interactions are specific

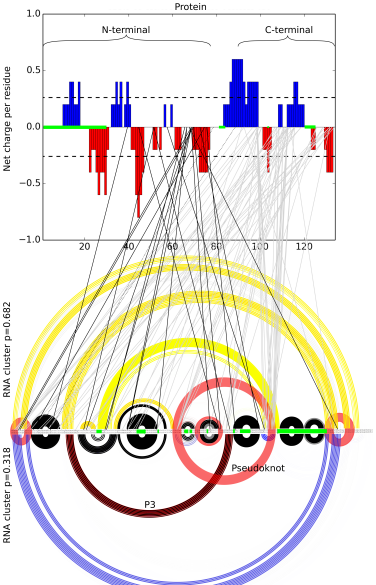


N-terminal disruptive interactions are specific

N-terminal disruptive interactions are specific



Global view


Protein
N-terminal
C-terminal
Net charge per residue
1.0
0.5
0.0
−0.5
−1.0
20
40
60
80
100
120
RNA cluster p=0.682
RNA cluster p=0.318
Pseudoknot
P3

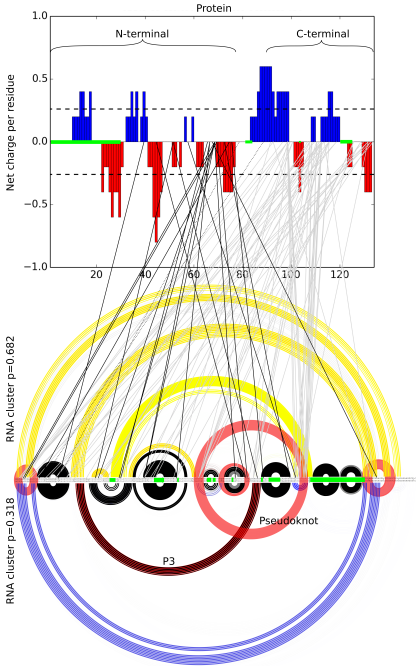
To do

- Compare with Gremlin-H2
  - Must implement variable alphabet / frequency
- Benchmark with *Cyt-18* (1.6TB downloaded and going on)
- Find other examples
- Identify other disordered proteins with that charge pattern

Acknowledgments

Tsvi Tlusty (IBS/CSLM)

Sergey Ovchinnikov (Harvard)


Protein
N-terminal
C-terminal
Net charge per residue
1.0
0.5
0.0
−0.5
−1.0
20
40
60
80
100
120
RNA cluster p=0.682
RNA cluster p=0.318
Pseudoknot
P3