RNAcontext: A new method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins

Presented by: Ria X. Peschutter

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins

Hilal Kazan[1], Debashish Ray[2], Esther T. Chan[3], Timothy R. Hughes[2,3,4], Quaid Morris[1,2,3,4]*

1 Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, **2** Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada, **3** Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada, **4** Donnelley Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada

Received September 23, 2009; **Accepted** May 25, 2010; **Published** July 1, 2010

Overview

1) Biological Background

2) RNACompete

3) RNAContext

   a) RNA annotation

   b) Motif model

   c) Fitting of the model

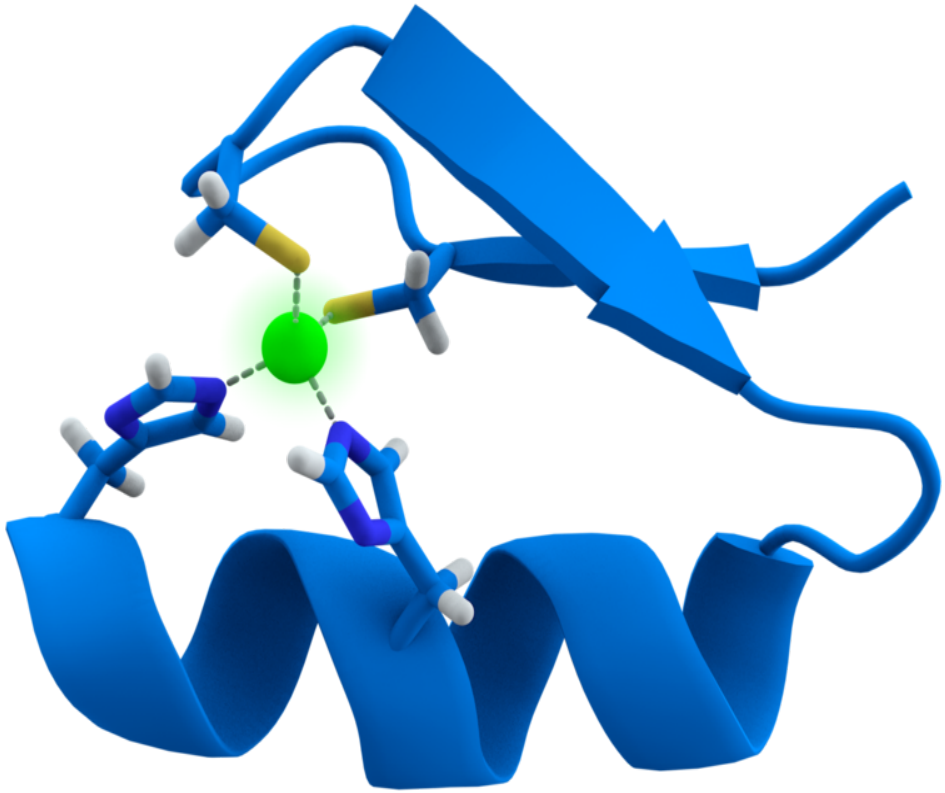
4) Results

RNA binding proteins - RBPs

- Critical roles in numerous cellular processes
- Essential during germline and early embryonic development

Challenges:

- Binding preferences not well characterized
- Multiple RNA targets
- Identification of mechanism behind function



http://en.wikipedia.org/wiki/File:Zinc_finger_rendered.png

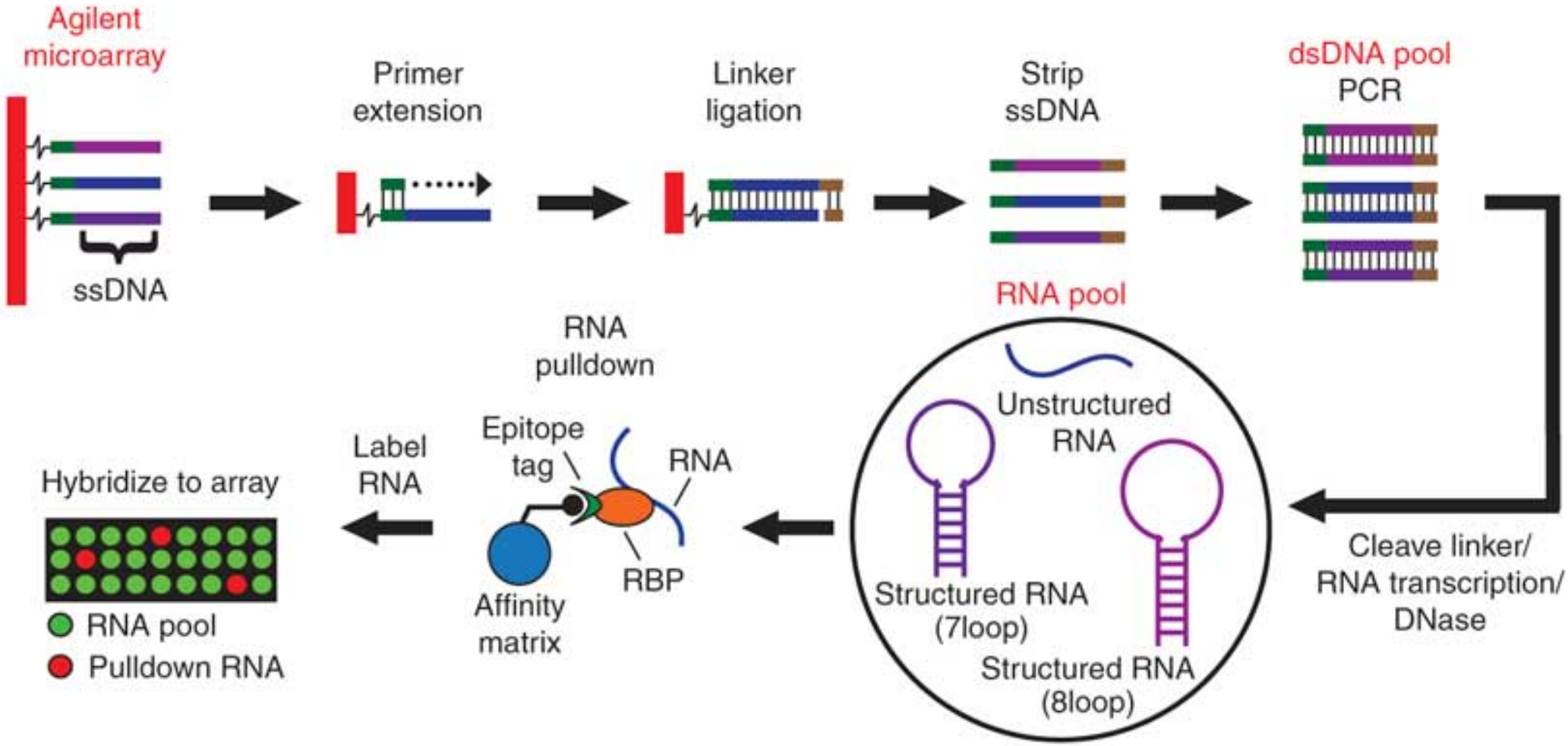
nature biotechnology

Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins

Debashish Ray[1,4], Hilal Kazan[2,4], Esther T Chan[3], Lourdes Peña Castillo[1], Sidharth Chaudhry[3], Shaheynoor Talukder[1], Benjamin J Blencowe[1,3], Quaid Morris[1–3] & Timothy R Hughes[1,3]

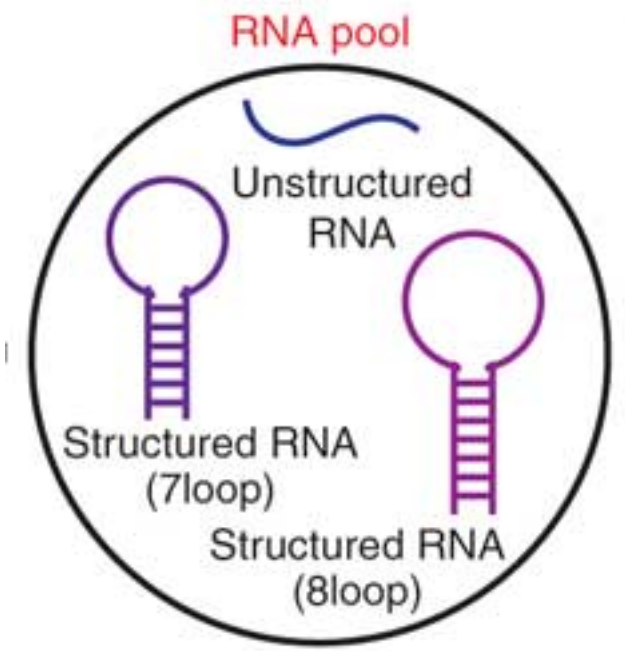
RNACompete

- Estimates binding affinity data
- Pool of unique short RNA sequences (29- to 38-nt)



RNACompete

- **Eukaryotic RBPs recognize:**
  - ➔ Short unstructured sequences
  - ➔ Loop sequences in RNA stem-loop

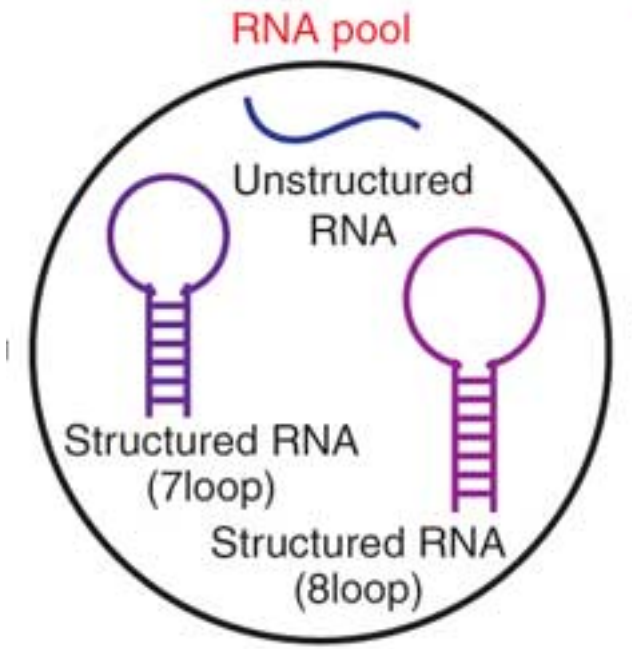


- **Mircoarray design:**
  - ➔ independent duplicate sets of unstructured and stem-loop RNAs
  - ➔ Set A – stem-loop and unstructured sequences
  - ➔ Set B – stem-loop and unstructured sequences

No overlap !

RNACompete

- **Caution:**
  - Unstructured sometimes of stem-loops
  - Stem-loops unexpected structures



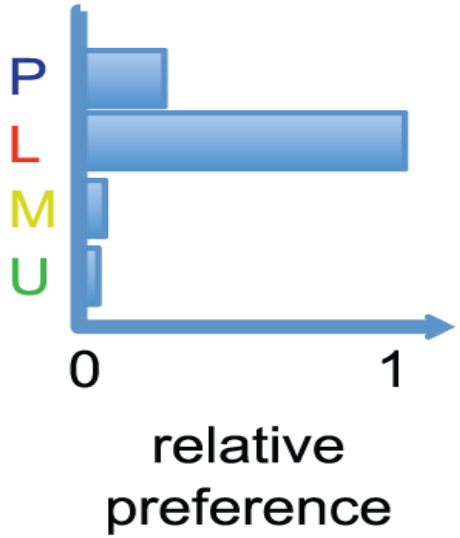
- **Constraints to minimize:**
  - Folding of unstructured RNAs
  - misfolding of structured RNAs
  - Extensive base-pairing among any two RNAs
  - microarray cross-hybridization

RNAcontext

- Motif model
- Predicts sequence and structure preferences of RBPs
- Considers multiple structural preferences simultaneously
- Webserver (restriction on input size)
- Terminal application (no restriction on input size)

base preference Φ

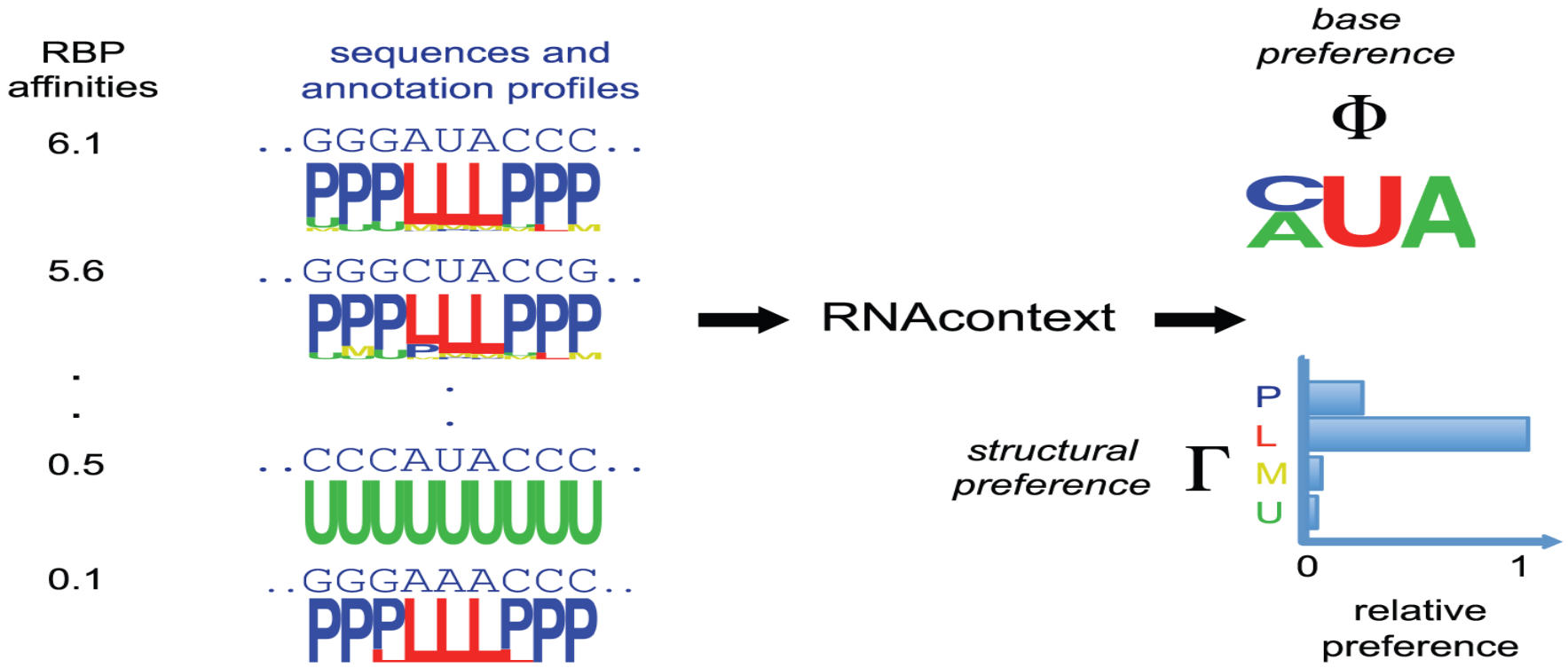
structural preference Γ



Methods

Three steps:

1. How to annotate RNA in terms of structural context
2. Details and mathematical formulation of the motif model
3. How to fit the RNAcontext motif model



Methods - RNA annotation

- Annotation of each base in each structure by context alphabet
- Predicting secondary structure → SFOLD
- RNA multiple distinct stable secondary structures
- Distribution over all its possible context

Alphabet A:

P - paired

L - hairpin loop

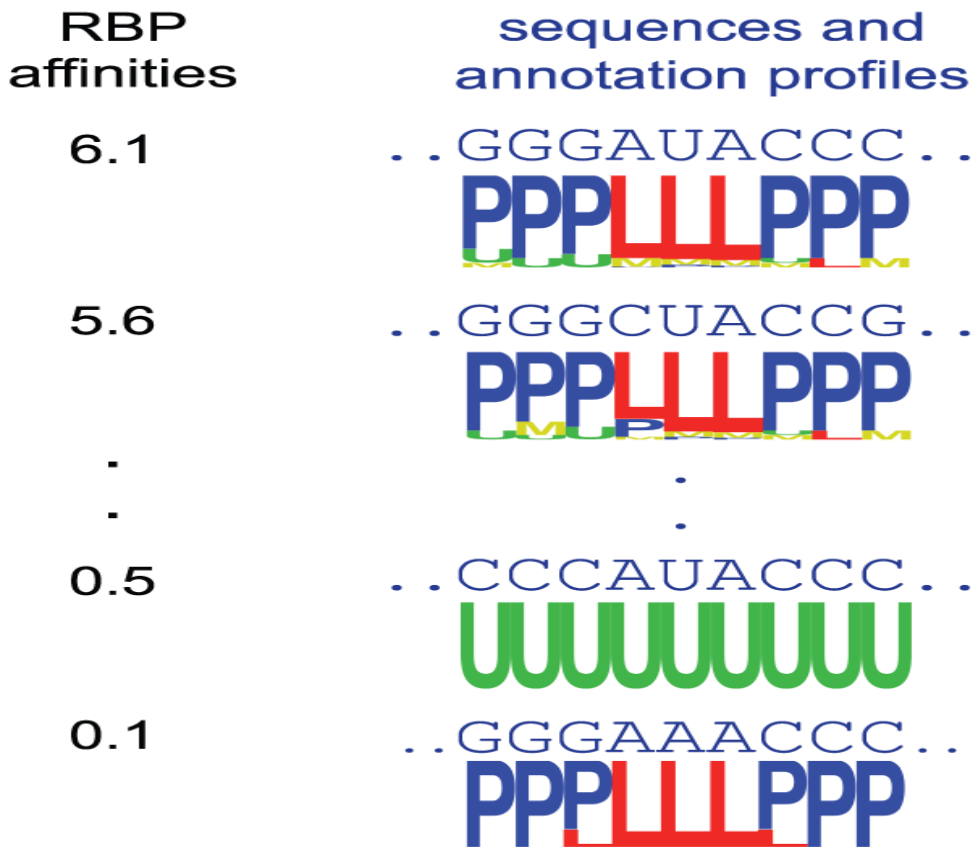
U - unstructured or external region

M - miscellaneous

Methods - Motif model

Input :

- Set of sequences together with their estimated binding affinities (**S**)
- RNA secondary structure annotations of the sequences (**P**)



Methods- Motif model

score
probability that at least one
of the subsequences of length K
is bound by the RBP

$$\underbrace{f(s,p,\Theta)}_{\text{score}} = 1 - \prod_{t=0}^{|s|-K} 1 - \underbrace{N(s_{t+1:t+K}, p_{t+1:t+K}, \Theta)}_{\text{depends on both base content } (s) \text{ and structure } (p)}$$

$$\underbrace{N^{seq}(s_{t+1:t+K}, \Theta)}_{\text{depends only on base content}} * \underbrace{C(p_{t+1:t+K}, \Theta)}_{\text{depends only on structure}}$$

S – sequence set
P – annotation profiles
Θ – model parameter
K – width of binding site

Methods- Motif model

$$\underbrace{N^{seq}(s_{t+1:t+K},\Theta)}_{\text{depends only on base content}} * \underbrace{C(p_{t+1:t+K},\Theta)}_{\text{depends only on structure}}$$

$$\underbrace{N^{seq}(s,\Theta)=\sigma\left(\beta_s+\sum_{k=1}^{K}\Theta_{s_k,k}\right)}_{\text{estimate of probability that RBP binds } s_{t+1:t:k} \text{ in ideal structural context}}$$

$$\underbrace{C(p,\Theta)=\sigma\left(\beta_p+\sum_{\alpha\in A}\Gamma_\alpha * \sum_{k=1}^{K} p_{\alpha,k}\right)}_{\text{estimate of probability that RBP prefers structural context } p_{t+1:t*K}}$$

S – sequence set
P – annotation profiles
K – width of binding site

Θ – model parameter:
βs – sequence affinity bias
βp – structural context bias
Γ – structure annotation

$p_{\alpha,k}$ – probability that base at pos k of s has structural context α

Methods- Motif model

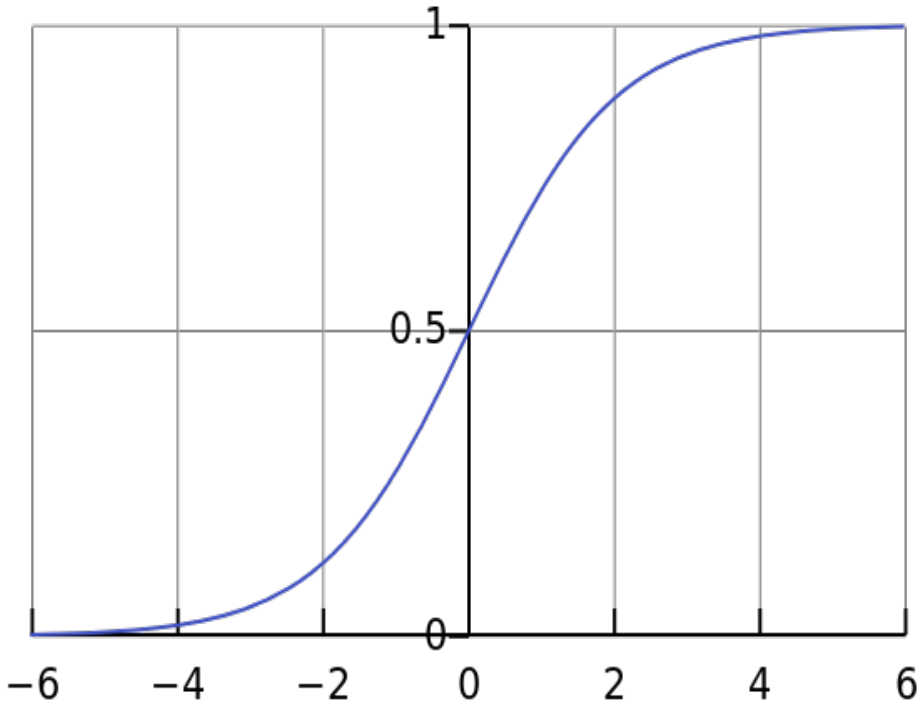
$$N^{seq}(s,\Theta)=\sigma\left(\beta_s+\sum_{k=1}^{K}\Theta_{s_k,k}\right)$$

estimate of probability that RBP binds $s_{t+1:t:k}$ *in ideal structural context*

$$C(p,\Theta)=\sigma\left(\beta_p+\sum_{\alpha\in A}\Gamma_\alpha * \sum_{k=1}^{K} p_{\alpha,k}\right)$$

estimate of probability that RBP prefers structural context $p_{t+1:t*K}$

Logistic function: $\sigma(x)=(1+\exp(-x))^{-1}$

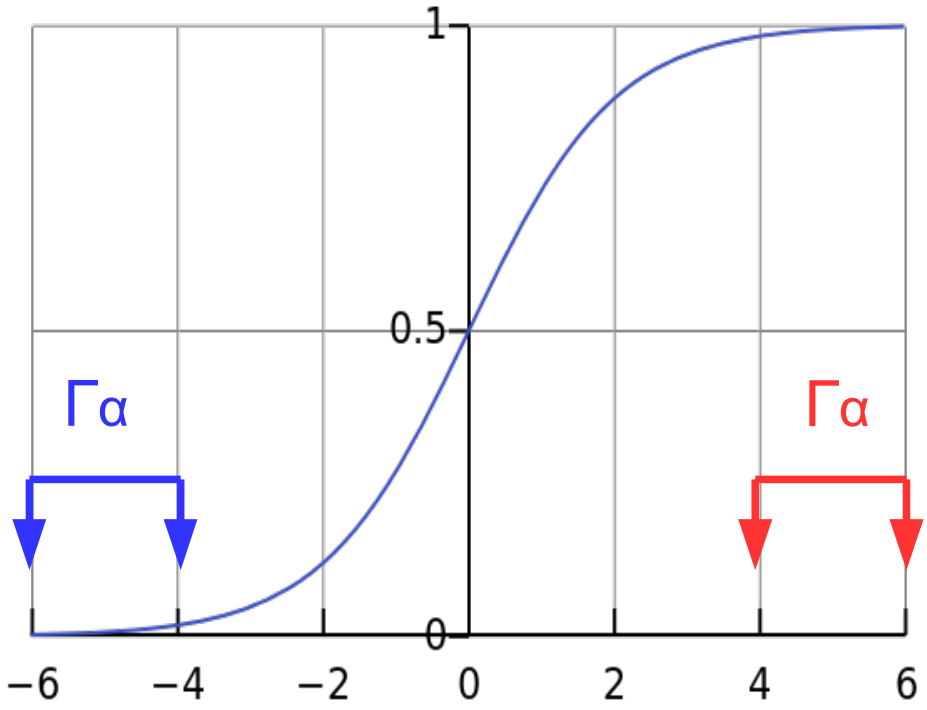


http://en.wikipedia.org/wiki/Logistic_function

Methods- Motif model

$$\overbrace{f(s,p,\Theta)}^{\text{score}} = 1 - \prod_{t=0}^{|s|-K} 1 - N(s,p,\Theta) = 1 - \prod_{t=0}^{|s|-K} 1 - \left(N^{seq}(s,\Theta) * C(p,\Theta)\right)$$

- Γα → favored structure
  - → N(s,p,Θ) ≈ N(s,Θ)
  - → determined by s
- Γα → unfavored structure
  - → C(p, Θ) ≈ 0
  - → N(s,p, Θ) ≈ 0
  - → determined by p



http://en.wikipedia.org/wiki/Logistic_function

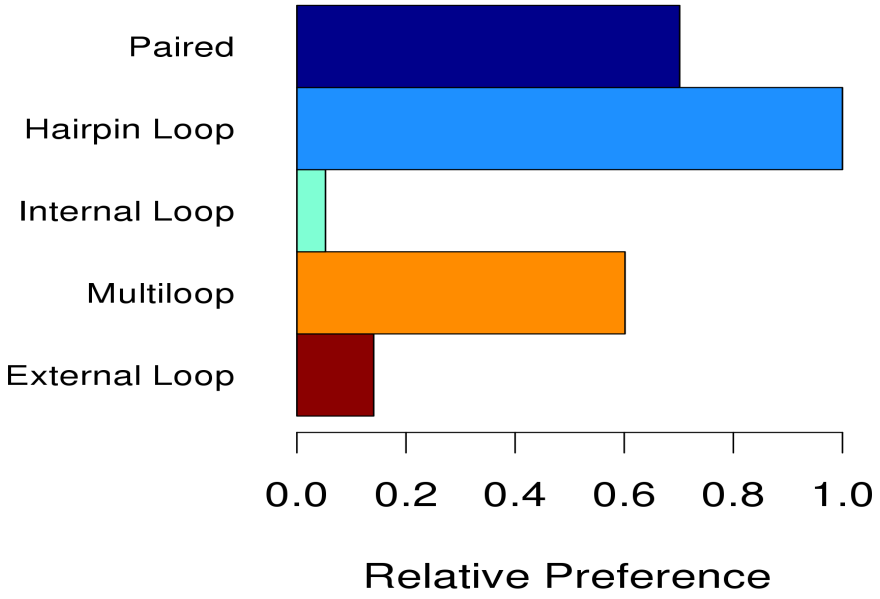
Methods - Motif model fitting

- Model affinity of sequence as linear model
- Minimizing least squares cost function by L-BFGS-B
  - ➔ Constrain slope of linear model to only take positive values
  - ➔ Making use of position weight matrix (PWM)
  - ➔ Small constant to ensure unique global minimum

- Cost function multi-modal
  - ➔ different initializations → different outputs
  - ➔ Take the best

RNAcontex-Output

- Structure parameters:
  - ➔ Scaled → most preferred context =1



- Sequence parameters:
  - ➔ Position Weight Matrix (PWM)
  - ➔ Displayed with EnoLOGOS

http://rnamotif.org/help.html

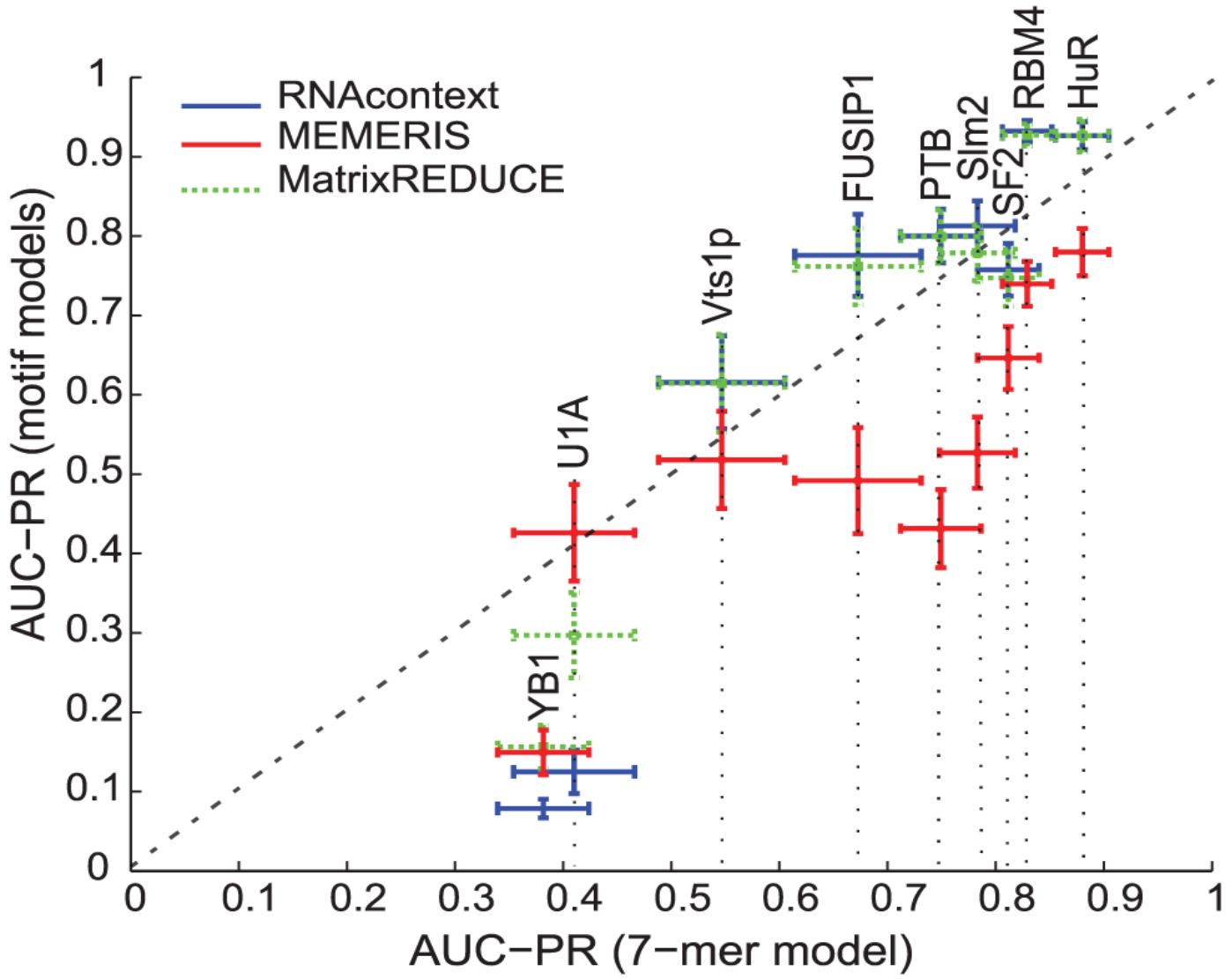
http://rnamotif.org/help.html

Results

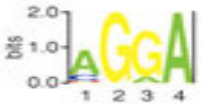
- Compared MEMERIS(MEME), RNAcontext and MatrixReduce
- Each method trained individually
- making use of each possible parameter setting
- best model

Proteins	RNAcontext	MEMERIS	MatrixREDUCE
RBM4	**0.91**	0.43	0.63
FUSIP1	**0.53**	0.31	0.32
Vts1p	**0.65**	0.58	0.56
YB1	**0.17**	0.07	0.11
SLM2	**0.81**	0.49	0.77
SF2	**0.70**	0.50	0.66
U1A	**0.30**	0.27	0.21
HuR	**0.96**	0.74	0.94
PTB	**0.69**	0.26	0.67

Results


RNAcontext
MEMERIS
MatrixREDUCE
AUC−PR (motif models)
AUC−PR (7−mer model)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
YB1
U1A
Vts1p
FUSIP1
PTB
Slm2
SF2
RBM4
HuR

Results - RNAcompete

Protein	Domain(s)	Known motif/ binding site	Our highest-correlating motif
Vts1	One SAM domain	 in stem-loop	
SLM2	One KH domain	NA	
YB1	Full-length; one cold-shock domain		
RBM4	Full-length; two RRM domains	NA	
SF2/ASF	Two RRM domains		
FUSIP1	One RRM domain		
HuR	Full-length; three RRM domains		
U1A	One of two RRM domains		
PTB	Full-length; four RRM domains		

Results - RNAContext

	Domain(s)	Previously reported binding site	RNAcontext predicted motifs
Vts1p	one SAM domain	 in stem-loop source: [17]	 context: P L L L L L L
SLM2	one KH domain	 source: [5]	 context: U U U U U U U U
RBM4	full-length two RRM domains	 source: [5]	 context: U U U U U U U
SF2/ASF	two RRM domains	 source: [43]	 context: M M M M P
FUSIP1	one RRM domain	 source:[5]	 context: U U U U U U U U U U U
HuR	full-length; three RRM domains	 source: [12]	 context: U U U U U U U U P
PTB	full-length; four RRM domains	 source: [44]	 context: U U U U U

Summary

- New motif model of RBP binding preferences
  - ➔ Sequence and structure
  - ➔ Corresponding algorithm for fitting model
- Recovers previously reported sequences and binding preferences
- Predicts new structure binding preferences
- Interprets RNAcompete data better than correlation analysis
- Initially designed for short RNA sequences can be used for long:
  - RNAplfold

Thank you for your attention!

Questions?

Methods – L-BFGS-B

BFGS :

- iterative method
- solves unconstrained non-linear optimization problems
- Approximates Newton's method by Hessian matrix
- Hessian

  ➔ square matrix of second-order partial derivatives of a function
- stores and computes whole matrix

Methods – L-BFGS-B

L-BFGS:

- Vectors implicitly model Hessian
- Optimization algorithm for parameter estimation in machine learning

L-BFGS-B:

- extends L-BFGS
- handle simple box constraints (aka bound constraints) on variables
- identifies fixed and free variables at every step
- L-BFGS method only free variables

  → higher accuracy