From sequence/structure analysis to sequence design of RNA

Kiyosi Asai

Department of Computational Biology & Medical Sciences

University of Tokyo

&

Computational Biology Research Consortium（CBRC）

Notes

- In this talk, I have no intension to insist that
  - our algorithms/tools are superior to any other tools
  - theoretically “better” means practically/biologically better
- I am very happy if
  - you get a hint to combine/improve(?) your methods
  - and of course, compare our tools with your tools

Probability & free energy of 2D structures

Probability that an RNA sequence x form a structure σ

free energy

$$P(\sigma \mid x) = \frac{1}{Z(x)} \exp \frac{-E(\sigma, x)}{RT}$$

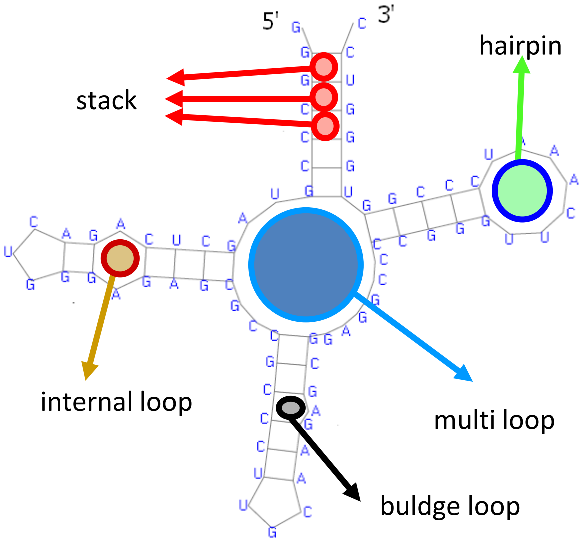
probability

partition function

$$Z(x) = \sum_{\xi \in \Omega} \exp \frac{-E(\xi, x)}{RT}$$

R : constant

T : temperature



The complete information of the 2D structure
Is only represented by distribution, or Z(x).

"hard" prediction of a single structure &
"soft" marginal probabilities (e.g. BPPs)
does not represent the complete information.

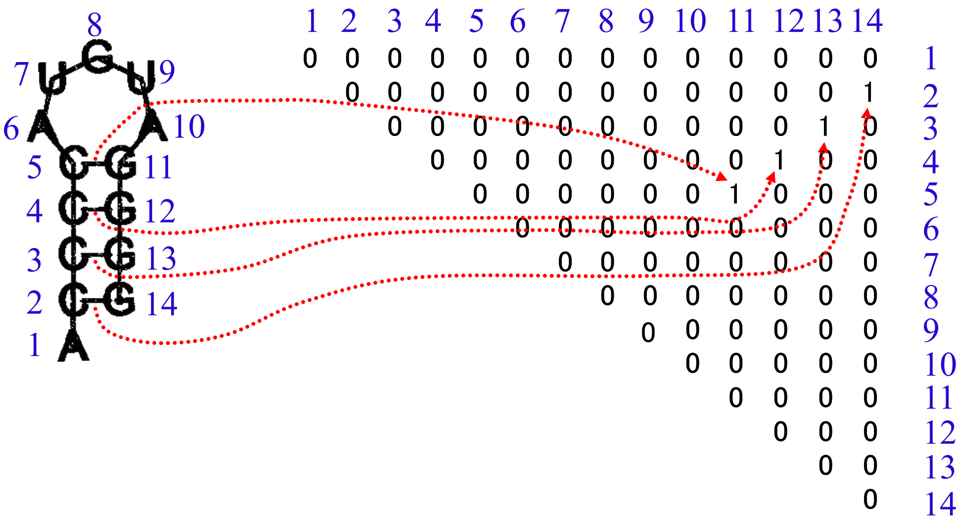
2D structure prediction of an RNA sequence

Given $D = \{x\}$: an RNA sequence

predict the secondary (2D) structure of x

⇒ predict a point in $Y = S(x)$,

the set of all the possible 2D structures of x



1	2	3	4	5	6	7	8	9	10	11	12	13	14	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	0	0	0	0	0	0	0	0	0	0	0	0	1	2
		0	0	0	0	0	0	0	0	0	0	1	0	3
			0	0	0	0	0	0	0	0	1	0	0	4
				0	0	0	0	0	0	1	0	0	0	5
					0	0	0	0	0	0	0	0	0	6
						0	0	0	0	0	0	0	0	7
							0	0	0	0	0	0	0	8
								0	0	0	0	0	0	9
									0	0	0	0	0	10
										0	0	0	0	11
											0	0	0	12
												0	0	13
													0	14

A 2D structure is a point in a subspace of a binary space whose dimension is $|x|^2$

Each cell is not independent

$S(i,j)=1 \Rightarrow S(i,k)=0$ for $k \neq j$

$S(i,j)=1 \Rightarrow S(i,k)=0$ for $k \neq j$

$$\boxed{Y = S(x) \subset \{0,1\}^n}$$

2D structure prediction of RNA

Probability that an RNA sequence x form a structure σ

energy

$$P(\sigma \mid x) = \frac{1}{Z(x)} \exp \frac{-E(\sigma, x)}{RT}$$

probability

Probability Distribution ⟺ Energy Model

Maximum Likelihood
（ML）

=

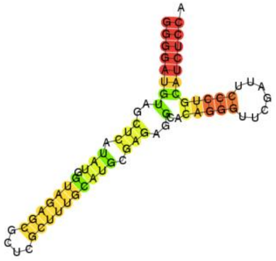
Minimum Free Energy
（MFE）

$$\hat{\sigma}^{ML} = \arg\max_{\sigma} P(\sigma \mid x)$$

$$\hat{\sigma}^{MFE} = \arg\min_{\sigma} E(\sigma, x)$$

Problem of MFE/MLE for RNA 2D structure

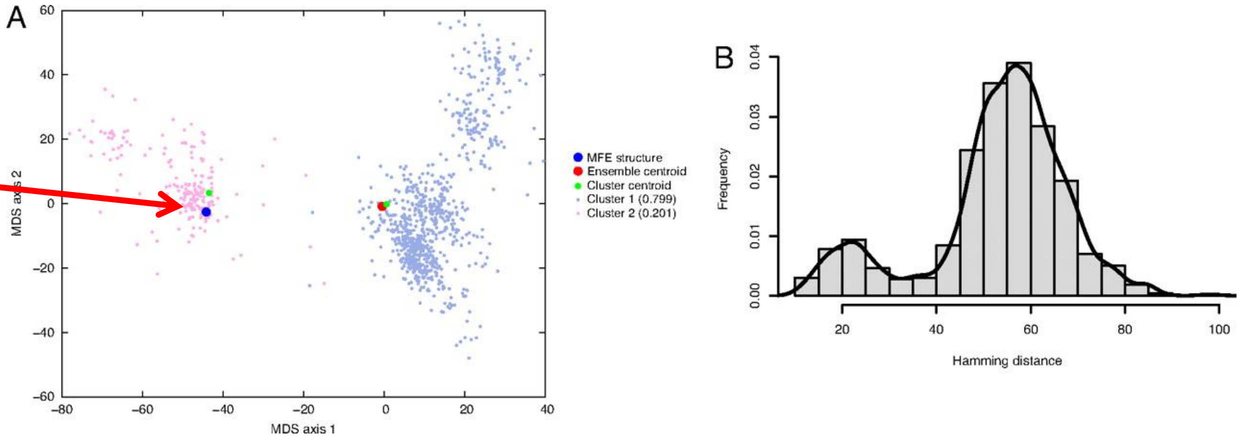
- The probability of MFE/MLE structure is very very small.
  - e.g. tRNA
    8,262,197,946,800,760 patterns



0.66%

- Probability sum of "Clusters" may give different picture
  - e.g. Rnase P

MFE



Multidimensional scaled distribution (A) and histogram of distances to cluster 2 centroid (B) derived from 1,000 representative samples from Sfold for the secondary structure of Dermocarpa sp.

Carvalho & Lawrence PNAS 08

More on MFE/MLE 2D structures

MFE structure = ML estimator
maximizes the probability that the estimator is
exactly same to "correct" structure

$$\hat{\sigma}^{ML} = \arg\max_{\sigma} P(\sigma \mid x) = \arg\max_{\sigma} \sum_{\theta \in Y} \delta(\theta, \sigma) P(\theta \mid x)$$

Drawback of ML estimator:
the probability for the ML estimator is **extremely small**
(10^{-5}～10^{-30})

⇒ General drawback in estimation problem
in high-dimensional binary space

M. Hamada et al. PLoS one 6:2 (2011)

No good solution? But still we try point estimation

MLE maximize the probability

that the estimator is exactly same as “correct” structure

$$\hat{\sigma}^{ML} = \arg\max_{\sigma} P(\sigma \mid x) = \arg\max_{\sigma} \sum_{\theta \in Y} \delta(\theta, \sigma) P(\theta \mid x)$$

MEG（Maximum Expected Gain） estimator is defined as

$$\hat{y}^{(MEG)} = \arg\max_{y \in Y} \sum_{\theta \in Y} G(\theta, y) P(\theta \mid D)$$

Gain Function $G(\theta, y) : Y \times Y \to \mathbb{R}^{+}$ ($\theta \in Y$, $y \in Y$)

ML estimator is the MEG for $G(\theta, y) = \delta(\theta, y)$

M. Hamada et al. PLoS one 6:2 (2011)

Generalized centroid estimator（ γ-centroid ）

γ-centroid estimator is the MEG estimator for the gain function:

$$G(\theta, y) = \sum_{i=1}^{n} \{ I(\theta_i = 0) I(y_i = 0) + \gamma \times I(\theta_i = 1) I(y_i = 1) \}$$

$$= TN + \gamma \times TP$$

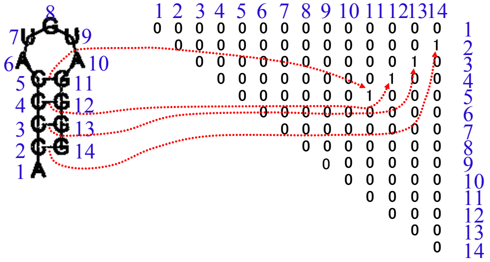
TP: # of true positives

TN: # of true negatives

for $\gamma = \dfrac{\alpha_1 + \alpha_4}{\alpha_2 + \alpha_3}$, γ-centroid estimator

is equivalent to MEG for

$$G(\theta, y) = \alpha_1 TP + \alpha_2 TN - \alpha_3 FP - \alpha_4 FN$$



1	2	3	4	5	6	7	8	9	10	11	12	13	14	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	0	0	0	0	0	0	0	0	0	0	0	0	1	2
		0	0	0	0	0	0	0	0	0	0	1	0	3
			0	0	0	0	0	0	0	0	1	0	0	4
				0	0	0	0	0	0	1	0	0	0	5
					0	0	0	0	0	0	0	0	0	6
						0	0	0	0	0	0	0	0	7
							0	0	0	0	0	0	0	8
								0	0	0	0	0	0	9
									0	0	0	0	0	10
										0	0	0	0	11
											0	0	0	12
												0	0	13
													0	14

γ-centroid represents arbitrary linear combinations of accuracy-related counts, TP, TN, FT, FN

The γ is a parameter to control the valance of sensitivity and PPV（ γ = 1, centroid）

M. Hamada et al. PLoS ONE 6:2 (2011)

DP for γ-centroid estimator of 2D structure
A posterior decoding

$$\hat{y}^{(\gamma)} = \arg\max_{y\in Y} \sum_{\theta\in Y} (TN + \gamma TP \mid \theta, y) P(\theta \mid D)$$

$$M_{i,j} = \max \begin{cases} M_{i+1,j-1} + \boxed{(\gamma+1)P^{(bp)}_{i,j} - 1} \\ M_{i-1,k} \\ M_{i,k-1} \\ \max_k \left[M_{i,k} + M_{k+1,j} \right] \end{cases}$$

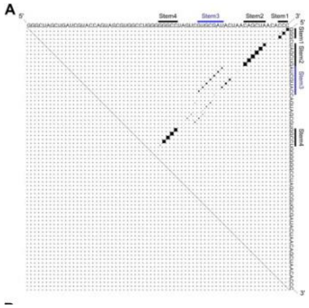
CentroidFold

Hamada et al. *Bioinformatics 25(4), 2009*

Base-pairing probability（BPP）, a posterior probability

We usually need DP for BPP (e.g. McCaskill)

$$P^{(bp)}_{i,j} = P((i,j) \in \sigma | x) = \sum_{\sigma|(i,j)\in\sigma} P(\sigma|x)$$



DP for γ-centroid estimator of 2D structure

A posterior decoding

$$\hat{y}^{(\gamma)} = \arg\max_{y\in Y} \sum_{\theta\in Y} (TN + \gamma TP \mid \theta, y) P(\theta \mid D)$$

$$M_{i,j} = \max \begin{cases} M_{i+1,j-1} + \boxed{(\gamma+1)P^{(bp)}_{i,j} - 1} \\ M_{i-1,k} \\ M_{i,k-1} \\ \max_k \left[M_{i,k} + M_{k+1,j} \right] \end{cases}$$

CentroidFold

Hamada et al. *Bioinformatics 25(4), 2009*

γ-centroid maximizes the expected accuracy of **BASE-PAIR** prediction in terms of

$$TN + \gamma \times TP$$

Can be combined with BPP from any energy model.

DP for γ-centroid estimator of sequence alignment

A posterior decoding

γ-centroid estimator for pairwise alignment

$$M_{i,j} = \max \begin{cases} M_{i-1,j-1} + (\gamma+1)P_{i,j}^{(a)} - 1 \\ M_{i-1,k} \\ M_{i,k-1} \end{cases}$$

Alignment probability is also a marignal probability

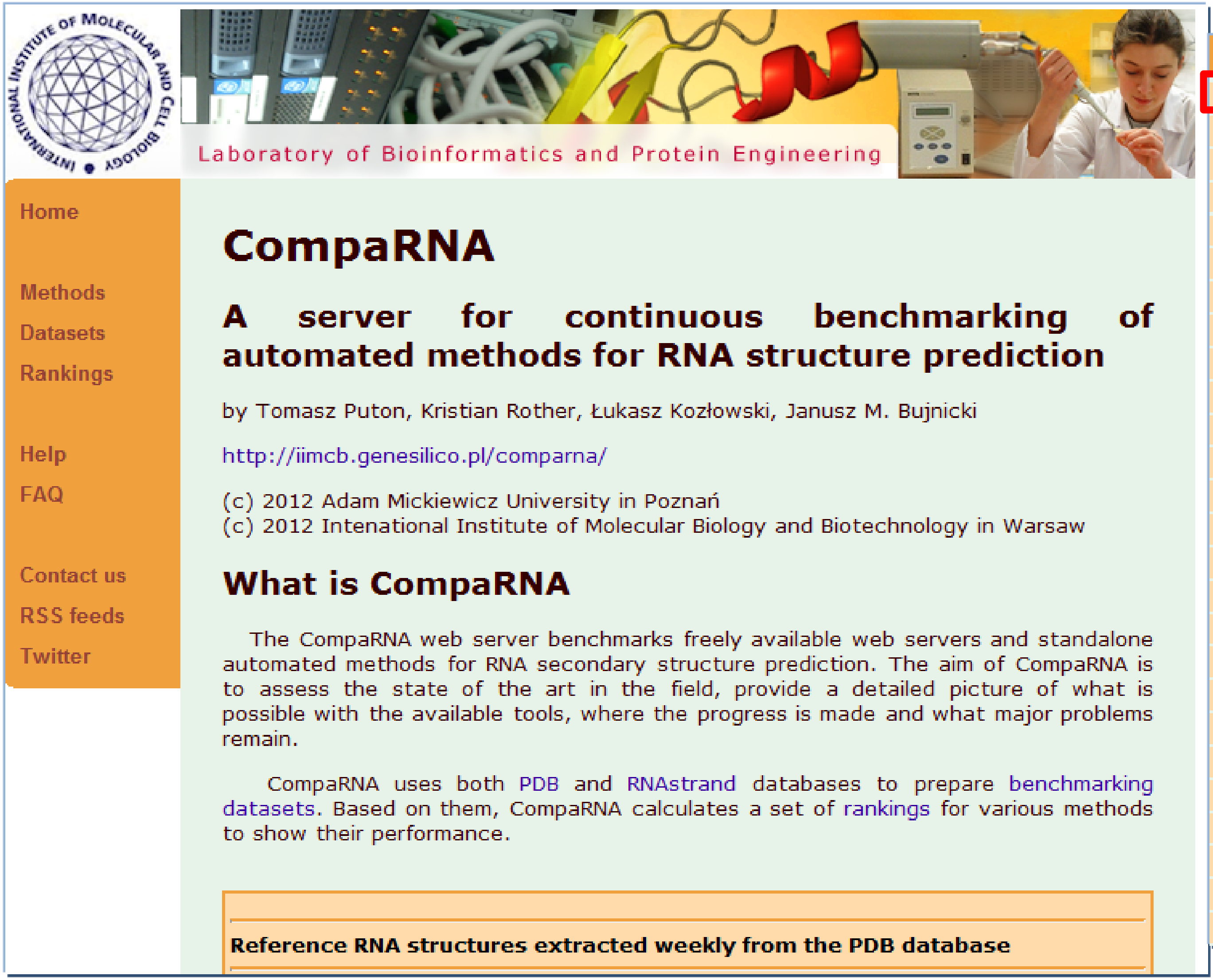
Frith et al. BMC *Bioinformatics 11:80, 2010*

γ-centroid estimator of 2D structure prediction

$$M_{i,j} = \max \begin{cases} M_{i+1,j-1} + (\gamma+1)P_{i,j}^{(bp)} - 1 \\ M_{i-1,k} \\ M_{i,k-1} \\ \max_k \left[M_{i,k} + M_{k+1,j} \right] \end{cases}$$

Hamada et al. *Bioinformatics 25(4), 2009*

CentroidFold in evaluation by 3rd party



INTERNATIONAL INSTITUTE OF MOLECULAR AND CELL BIOLOGY
Laboratory of Bioinformatics and Protein Engineering
Home
Methods
Datasets
Rankings
Help
FAQ
Contact us
RSS feeds
Twitter

CompaRNA

A server for continuous benchmarking of automated methods for RNA structure prediction

by Tomasz Puton, Kristian Rother, Łukasz Kozłowski, Janusz M. Bujnicki

http://iimcb.genesilico.pl/comparna/

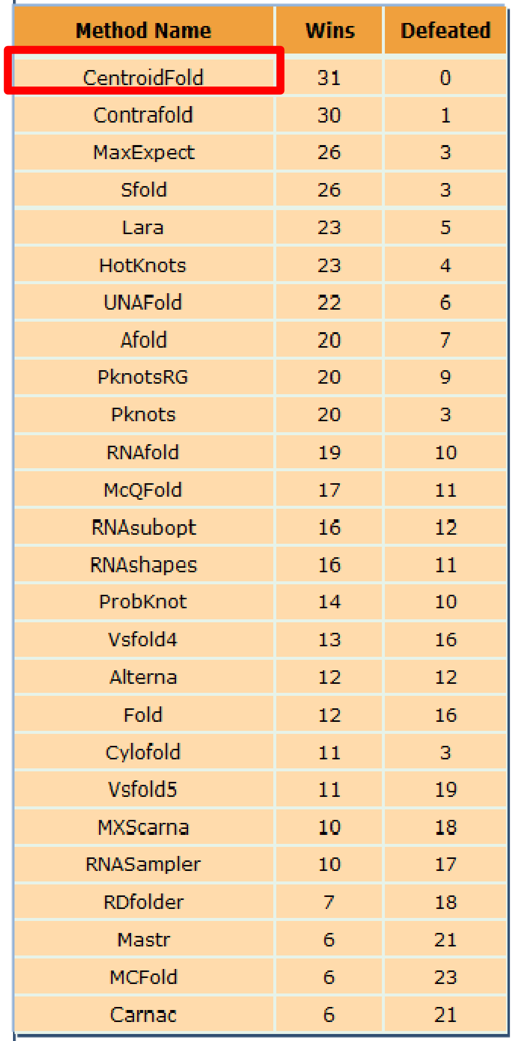
(c) 2012 Adam Mickiewicz University in Poznań
(c) 2012 Intenational Institute of Molecular Biology and Biotechnology in Warsaw

What is CompaRNA

The CompaRNA web server benchmarks freely available web servers and standalone automated methods for RNA secondary structure prediction. The aim of CompaRNA is to assess the state of the art in the field, provide a detailed picture of what is possible with the available tools, where the progress is made and what major problems remain.

CompaRNA uses both PDB and RNAstrand databases to prepare benchmarking datasets. Based on them, CompaRNA calculates a set of rankings for various methods to show their performance.

Reference RNA structures extracted weekly from the PDB database

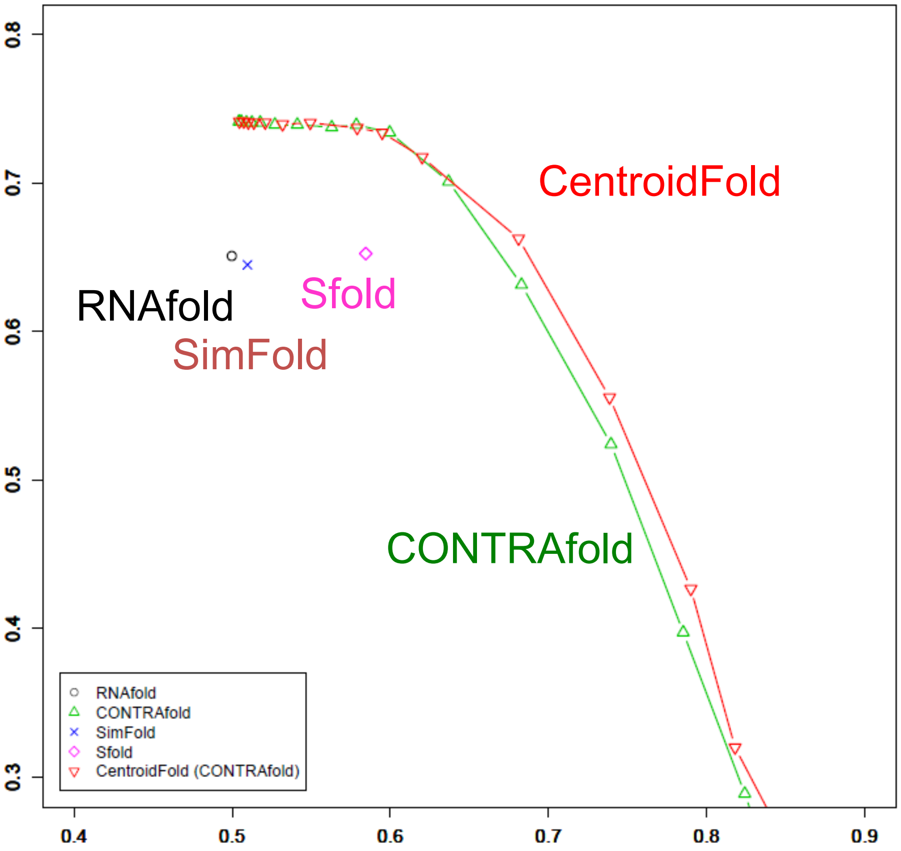


Method Name	Wins	Defeated
CentroidFold	31	0
Contrafold	30	1
MaxExpect	26	3
Sfold	26	3
Lara	23	5
HotKnots	23	4
UNAFold	22	6
Afold	20	7
PknotsRG	20	9
Pknots	20	3
RNAfold	19	10
McQFold	17	11
RNAsubopt	16	12
RNAshapes	16	11
ProbKnot	14	10
Vsfold4	13	16
Alterna	12	12
Fold	12	16
Cylofold	11	3
Vsfold5	11	19
MXScarna	10	18
RNASampler	10	17
RDfolder	7	18
Mastr	6	21
MCFold	6	23
Carnac	6	21

Accuracy in terms of base-pairs prediction

On average, γ-centroid has a very strong position in this evalutation measure if the same energy model is used.

$$\mathrm{SEN} = \frac{TP}{TP + FN}$$



$$\mathrm{PPV} = \frac{TP}{TP + FP}$$

Of course, this does not mean γ-centroid is the “best” method for 2D structure prediction.

Dataset: S-151 [Do+ 2007]

Hamada et al. *Bioinformatics 25(4), 2009*

MEA estimator of 2D structure [Do+2006]

Maximum expected accuracy estimator

Implimented in **CONTRAfold** [Do+2006]

$$\hat{y} = \arg\max_{y\in\mathcal{S}(x)} \sum_{\theta\in\mathcal{S}(x)} \underbrace{G_\gamma^{(mea)}(\theta,y)}\,p(\theta|x)$$

θor yの対称拡張行列

$$G_\gamma^{(mea)}(\theta,y) = \underbrace{\sum_{i=1}^{|x|}}_{\text{Sum for every position } i \text{ in the sequence}} \left[\gamma \underbrace{\sum_{j:j\neq i} I(\theta^*_{ij}=1)I(y^*_{ij}=1)}_{\text{“correct” base-pair in base } i} + \underbrace{\prod_{j:j\neq i} I(\theta^*_{ij}=0)I(y^*_{ij}=0)}_{\text{“correct” loop in base } i} \right]$$

DP for MEA estimator of 2D structure

$$M_{i,j} = \max \begin{cases} M_{i+1,j} \\ M_{i,j-1} \\ M_{i+1,j-1} + 2\gamma p_{ij} - q_i - q_j \\ \max_k \left[M_{i,k} + M_{k+1,j}\right] \end{cases}$$

$$\boxed{q_i = 1 - \sum_{j:j<i} p_{ji} - \sum_{j:j>i} p_{ij}}$$

What’s wrong with MEA estimator of 2D structure?

Relation between MEA estimator and γ-centroid estimator

$$\hat{y}^{(MEG)} = \arg\max_{y \in Y} \sum_{\theta \in Y} G(\theta, y) P(\theta \mid D)$$

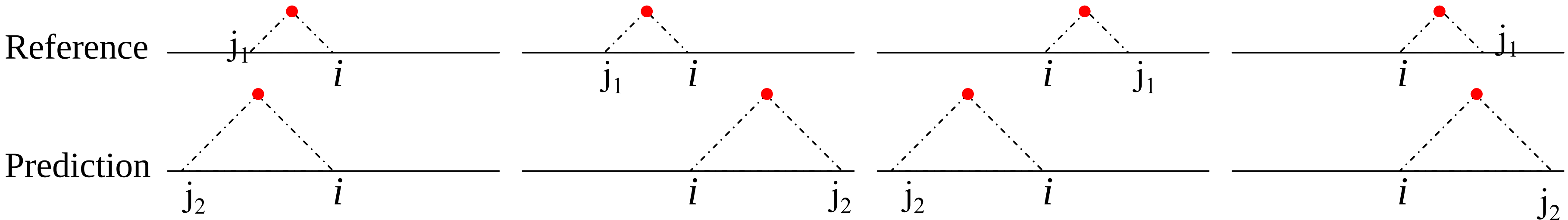
gain function of γ-centroid

Gain function of MEA

$$
\begin{aligned}
G_\gamma^{(mea)}(\theta, y) = 2G_\gamma^{(c)}(\theta, y) + C
&+ \sum_{1 \le i \le |x|} \sum_{\substack{j_1 : j_1 < i \\ j_2 : j_2 < i \\ j_1 \neq j_2}} I(\theta_{j_1 i} = 1) I(y_{j_2 i} = 1) \\
&+ \sum_{1 \le i \le |x|} \sum_{\substack{j_1 : j_1 < i \\ j_2 : j_2 > i}} I(\theta_{j_1 i} = 1) I(y_{i j_2} = 1) \\
&+ \sum_{1 \le i \le |x|} \sum_{\substack{j_1 : j_1 > i \\ j_2 : j_2 < i}} I(\theta_{i j_1} = 1) I(y_{j_2 i} = 1) \\
&+ \sum_{1 \le i \le |x|} \sum_{\substack{j_1 : j_1 > i \\ j_2 : j_2 > i \\ j_1 \neq j_2}} I(\theta_{i j_1} = 1) I(y_{i j_2} = 1)
\end{aligned}
$$

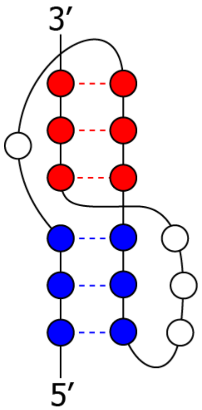
Unfavorable bias for estimating base-pairs

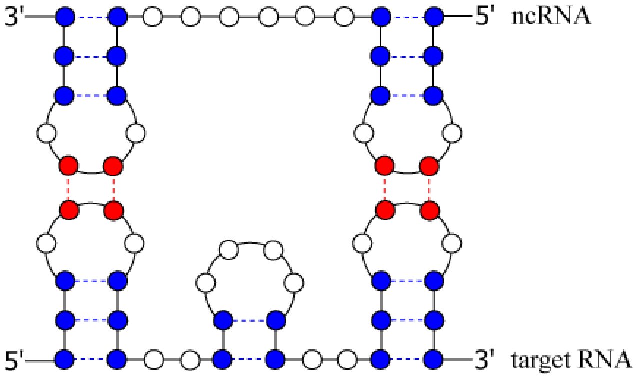
Implying g-centroid is a “better” estimator



When DP is out（more difficult problems）

Kengo SATO
Yuki KATO





RNA secondary structure prediction

RNA−RNA interaction prediction

Model & solve

Integer programming

$$\begin{aligned} \text{maximize} \quad & \boldsymbol{c}^{\mathsf{T}}\boldsymbol{x} \\ \text{subject to} \quad & A\boldsymbol{x} \leq \boldsymbol{b} \\ & \boldsymbol{x} \in \{0,1\}^n \end{aligned}$$

Algorithms/software related to γ-centroid

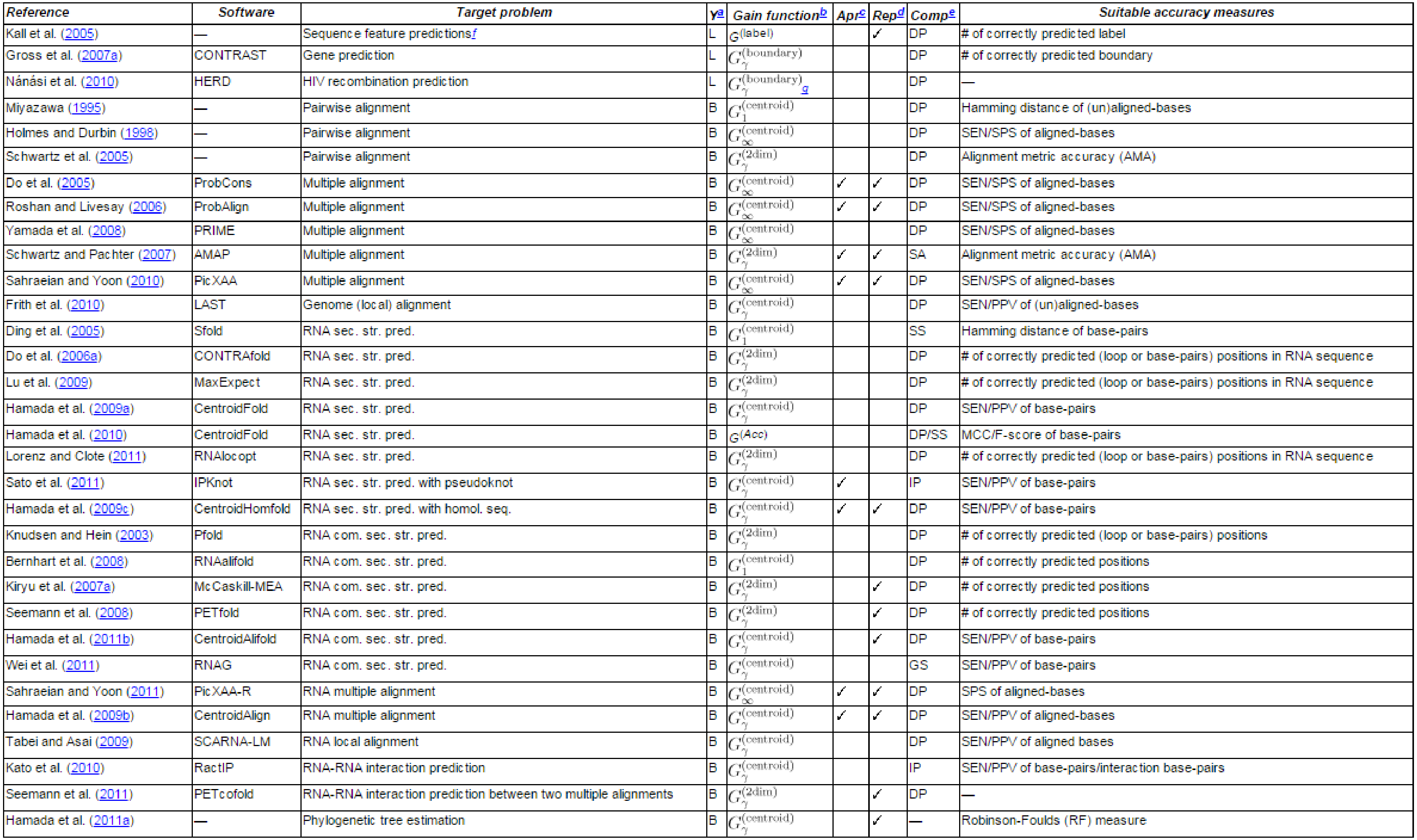
CentroidFold	2D pred.	Hamada+ Bioinformatics 25(4) 2009
CentroidHomfold	2D pred. using similar RNAs	Hamada+ Bioinformatics 25(12) 2009
CentroidAlign	RNA alignment	Hamada+ Bioinformatics 25(24) 2009
RactIP	RNA^2 interaction, integer prog.	Hamada+ Bioinformatics 26(18) 2009
IPknot	2D pred. w. PK integer prog.	Sato+ Bioinformatics 27(13) 2011
McCaskill-MEA	Common 2D pred. MEA	Kiryu+ Bioinformatics 23(4) 2007
CentroidAlifold	Common 2D pred. γ-centroid	Hamada+ Nucleic Acids Res.39(2) 2011
Pseudo-expected Accuracy	2D pred.	Hamada et al. BMC Bioinformatics 2010

For those who want to see more theory

Michiaki Hamada*, Hisanori Kiryu, Wataru Iwasaki, Kiyoshi Asai, Generalized Centroid Estimators in Bioinformatics, **PLoS ONE** 6(2):e16450, 2011. A corrected version is available from arXiv

Michiaki Hamada and Kiyoshi Asai. **A Classification of Bioinformatics Algorithms from the Viewpoint of Maximizing Expected Accuracy (MEA)**
Journal of Computational Biology. May 2012, 19(5): 532-549.

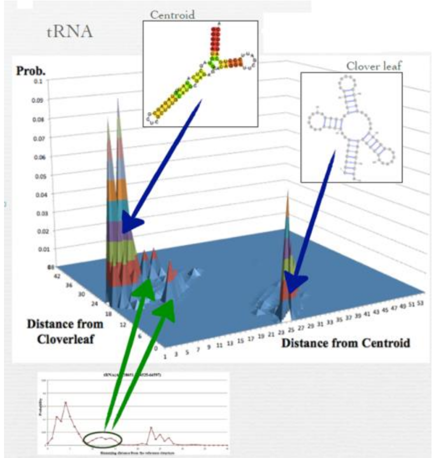
SUMMARY OF MEA ESTIMATIONS IN BIOINFORMATICS

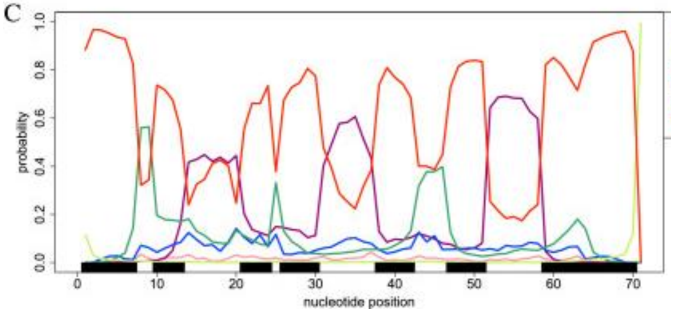


Reference	*Software*	*Target problem*	*Y*[a]	*Gain function*[b]	*Apr*[c]	*Rep*[d]	*Comp*[e]	*Suitable accuracy measures*
Kall et al. (2005)	—	Sequence feature predictions[f]	L	$G^{(\mathrm{label})}$		✓	DP	# of correctly predicted label
Gross et al. (2007a)	CONTRAST	Gene prediction	L	$G_{\gamma}^{(\mathrm{boundary})}$			DP	# of correctly predicted boundary
Nánási et al. (2010)	HERD	HIV recombination prediction	L	$G_{\gamma}^{(\mathrm{boundary})}$[g]			DP	—
Miyazawa (1995)	—	Pairwise alignment	B	$G_{1}^{(\mathrm{centroid})}$			DP	Hamming distance of (un)aligned-bases
Holmes and Durbin (1998)	—	Pairwise alignment	B	$G_{\infty}^{(\mathrm{centroid})}$			DP	SEN/SPS of aligned-bases
Schwartz et al. (2005)	—	Pairwise alignment	B	$G_{\gamma}^{(\mathrm{2dim})}$			DP	Alignment metric accuracy (AMA)
Do et al. (2005)	ProbCons	Multiple alignment	B	$G_{\infty}^{(\mathrm{centroid})}$	✓	✓	DP	SEN/SPS of aligned-bases
Roshan and Livesay (2006)	ProbAlign	Multiple alignment	B	$G_{\infty}^{(\mathrm{centroid})}$	✓	✓	DP	SEN/SPS of aligned-bases
Yamada et al. (2008)	PRIME	Multiple alignment	B	$G_{\infty}^{(\mathrm{centroid})}$			DP	SEN/SPS of aligned-bases
Schwartz and Pachter (2007)	AMAP	Multiple alignment	B	$G_{\gamma}^{(\mathrm{2dim})}$	✓	✓	SA	Alignment metric accuracy (AMA)
Sahraeian and Yoon (2010)	PicXAA	Multiple alignment	B	$G_{\infty}^{(\mathrm{centroid})}$	✓	✓	DP	SEN/SPS of aligned-bases
Frith et al. (2010)	LAST	Genome (local) alignment	B	$G_{\gamma}^{(\mathrm{centroid})}$			DP	SEN/PPV of (un)aligned-bases
Ding et al. (2005)	Sfold	RNA sec. str. pred.	B	$G_{1}^{(\mathrm{centroid})}$			SS	Hamming distance of base-pairs
Do et al. (2006a)	CONTRAfold	RNA sec. str. pred.	B	$G_{\gamma}^{(\mathrm{2dim})}$			DP	# of correctly predicted (loop or base-pairs) positions in RNA sequence
Lu et al. (2009)	MaxExpect	RNA sec. str. pred.	B	$G_{\gamma}^{(\mathrm{2dim})}$			DP	# of correctly predicted (loop or base-pairs) positions in RNA sequence
Hamada et al. (2009a)	CentroidFold	RNA sec. str. pred.	B	$G_{\gamma}^{(\mathrm{centroid})}$			DP	SEN/PPV of base-pairs
Hamada et al. (2010)	CentroidFold	RNA sec. str. pred.	B	$G^{(Acc)}$			DP/SS	MCC/F-score of base-pairs
Lorenz and Clote (2011)	RNAlocopt	RNA sec. str. pred.	B	$G_{\gamma}^{(\mathrm{2dim})}$			DP	# of correctly predicted (loop or base-pairs) positions in RNA sequence
Sato et al. (2011)	IPKnot	RNA sec. str. pred. with pseudoknot	B	$G_{\gamma}^{(\mathrm{centroid})}$	✓		IP	SEN/PPV of base-pairs
Hamada et al. (2009c)	CentroidHomfold	RNA sec. str. pred. with homol. seq.	B	$G_{\gamma}^{(\mathrm{centroid})}$	✓	✓	DP	SEN/PPV of base-pairs
Knudsen and Hein (2003)	Pfold	RNA com. sec. str. pred.	B	$G_{\gamma}^{(\mathrm{2dim})}$			DP	# of correctly predicted (loop or base-pairs) positions
Bernhart et al. (2008)	RNAalifold	RNA com. sec. str. pred.	B	$G_{1}^{(\mathrm{centroid})}$			DP	# of correctly predicted positions
Kiryu et al. (2007a)	McCaskill-MEA	RNA com. sec. str. pred.	B	$G_{\gamma}^{(\mathrm{2dim})}$		✓	DP	# of correctly predicted positions
Seemann et al. (2008)	PETfold	RNA com. sec. str. pred.	B	$G_{\gamma}^{(\mathrm{2dim})}$		✓	DP	# of correctly predicted positions
Hamada et al. (2011b)	CentroidAlifold	RNA com. sec. str. pred.	B	$G_{\gamma}^{(\mathrm{centroid})}$		✓	DP	SEN/PPV of base-pairs
Wei et al. (2011)	RNAG	RNA com. sec. str. pred.	B	$G_{\gamma}^{(\mathrm{centroid})}$			GS	SEN/PPV of base-pairs
Sahraeian and Yoon (2011)	PicXAA-R	RNA multiple alignment	B	$G_{\infty}^{(\mathrm{centroid})}$	✓	✓	DP	SPS of aligned-bases
Hamada et al. (2009b)	CentroidAlign	RNA multiple alignment	B	$G_{\gamma}^{(\mathrm{centroid})}$	✓	✓	DP	SEN/PPV of aligned-bases
Tabei and Asai (2009)	SCARNA-LM	RNA local alignment	B	$G_{\gamma}^{(\mathrm{centroid})}$			DP	SEN/PPV of aligned bases
Kato et al. (2010)	RactIP	RNA-RNA interaction prediction	B	$G_{\gamma}^{(\mathrm{centroid})}$			IP	SEN/PPV of base-pairs/interaction base-pairs
Seemann et al. (2011)	PETcofold	RNA-RNA interaction prediction between two multiple alignments	B	$G_{\gamma}^{(\mathrm{2dim})}$		✓	DP	—
Hamada et al. (2011a)	—	Phylogenetic tree estimation	B	$G_{\gamma}^{(\mathrm{centroid})}$		✓	—	Robinson-Foulds (RF) measure

Hamada+ J. Comp. Biol. 19(5) 2012

Algorithms & tools for 2D structure analysis





Risa Kawaguchi
Hisanori Kiryu

Ryota Mori
Kiyoshi Asai

Importance of detailed analysis of 2D structures

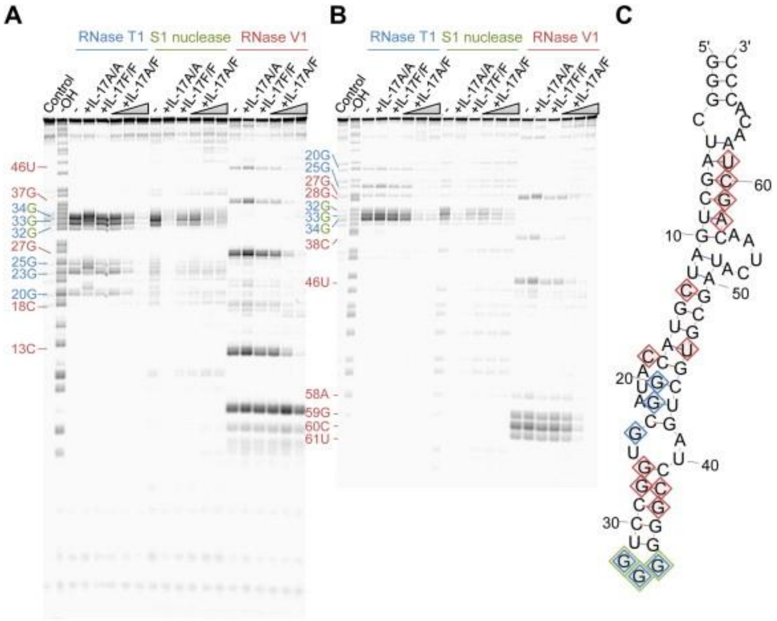


Fig. 5 RNase protection analysis of AptAF42dope1. (A and B) RNase footprinting of 5′- (A) and 3′- (B) FAM labeled AptAF42dope1 (5 pmol) in the presence of IL-17 proteins (IL-17A/F, 16.6, 33.3, 66.5 pmol; IL-17A/A and IL-17F/F, 66.5 pmol). Experimental conditions and procedures are as described in Materials and methods. (C) Mapping of nucleotides in AptAF42dope1 protected from RNase cleavage in the presence of IL-17A/F. Symbols: blue diamonds represent protection from RNase T1 cleavage; red diamonds represent protection from RNase V1 cleavage; and green diamonds represent protection from S1 nuclease cleavage, respectively.

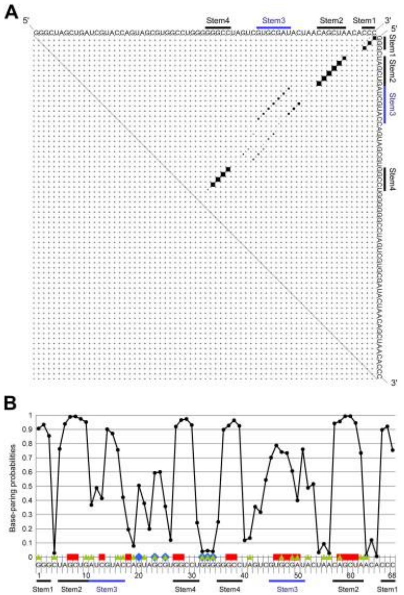


Fig. 6 Base-pairing probabilities. (A) The estimated probabilities indicated by dot plot for base-pairs in the AptAF42dope1 sequence. The dots in the (*i*, *j*)-cell, with *i* < *j*, indicates the base-pairing probability of the base-pair between *i*-th and *j*-th nucleotides in the sequence, where larger dots represent higher probabilities. In the calculation, the McCaskill model with Boltzmann Likelihood (BL) parameters were adopted as the probability distribution of the secondary structures. (B) Base-pairing probabilities of each position of the AptAF42dope1 sequence. The horizontal axis indicates positions of AptAF42dope1 and the vertical axis indicates the sum of base-pairing probabilities for the position. Cleavage sites obtained from ribonuclease digestion assay are also shown in the figure. Blue diamonds represent RNase T1 cleavage sites; red squares represent RNase V1 cleavage sites; green triangles represent S1 nuclease cleavage sites, respectively.

Hironori Adachi , Akira Ishiguro , Michiaki Hamada , Eri Sakota , Kiyoshi Asai , Yoshikazu Nakamura

Antagonistic RNA aptamer specific to a heterodimeric form of human interleukin-17A/F

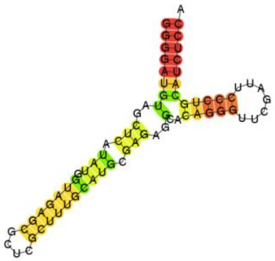
Biochimie, Volume 93, Issue 7, 2011, 1081 - 1088

Nothing has been solved on those problems weakness of point estimation

- The probability of MFE/MLE structure is very very small.
  - e.g. tRNA

γ-centroid?

8,262,197,946,800,760 patterns

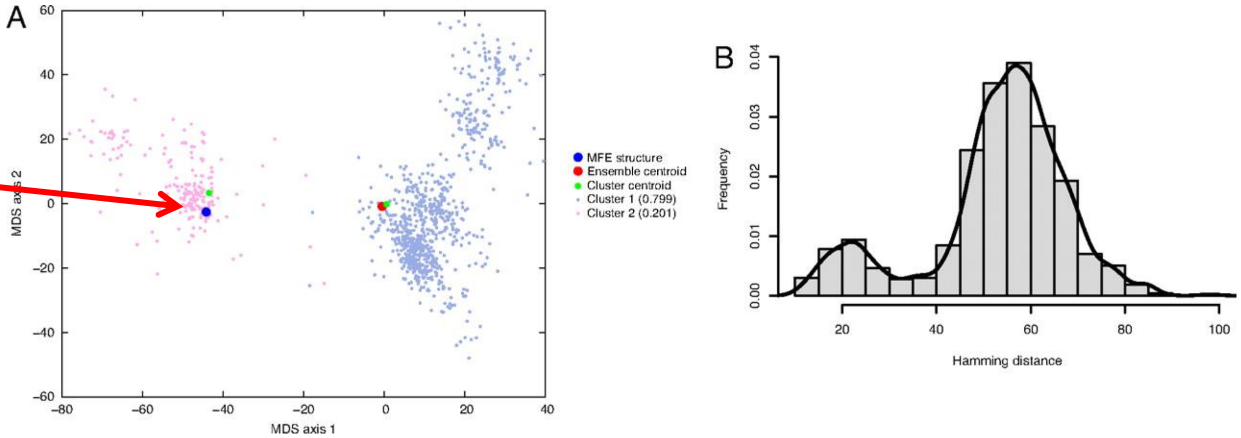


0.66%

- Probability sum of "Clusters" may give different picture
  - e.g. Rnase P

MFE

γ-centroid?



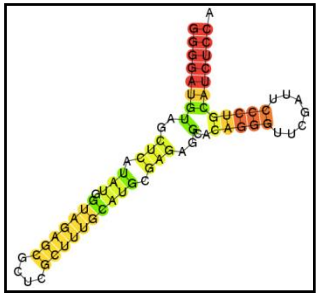
Multidimensional scaled distribution (A) and histogram of distances to cluster 2 centroid (B) derived from 1,000 representative samples from Sfold for the secondary structure of Dermocarpa sp.

Carvalho & Lawrence PNAS 08

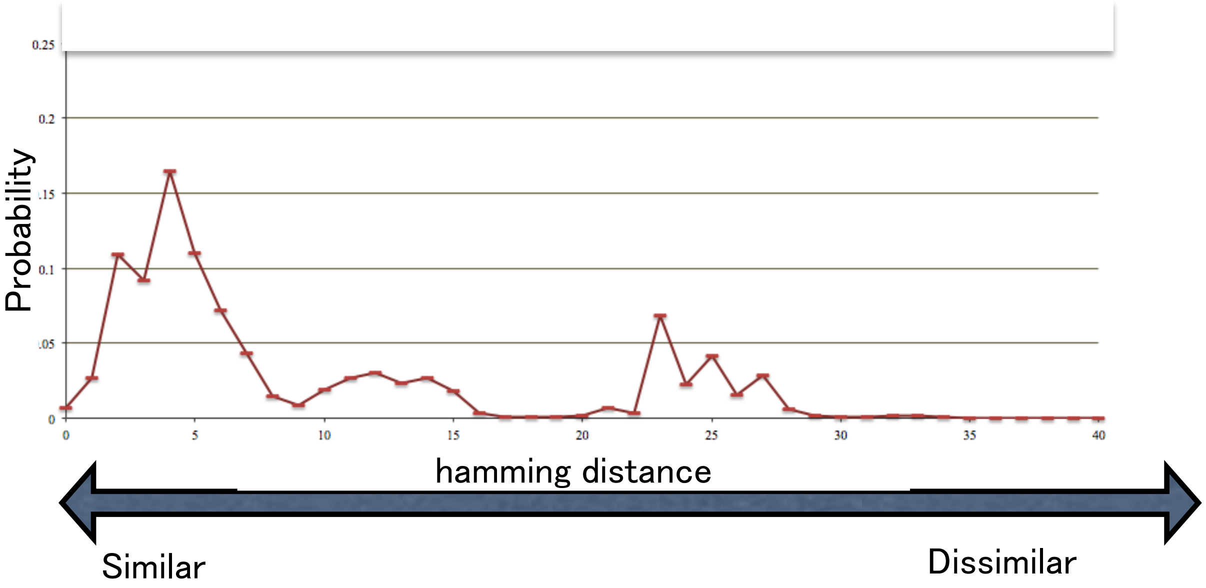
Efficient calculation of exact probability distributions of integer features on RNA secondary structures

Calculating the complete distributions of integer score S which is assigned to each RNA structure considering the whole RNA structure ensemble.

For example, S can be the hamming distance from the specific reference structure.



reference structure



R. Mori et al. BMC Genomics 2014, 15(Suppl 10):S6

References on this topic

Newberg LA, Lawrence CE: Exact calculation of distributions on integers, with application to sequence alignment.

J Comput Biol 2009, 16(1):1-18.

Freyhult E, Moulton V, Clote P: RNAbor: a web server for RNA structural neighbors.

Nucleic Acids Res 2007, 35(Web Server):305-309.

Senter E, Sheikh S, Dotu I, Ponty Y, Clote P: Using the fast fourier transform to accelerate the computational search for RNA conformational switches.

PLoS ONE 2012, 7(12):50506

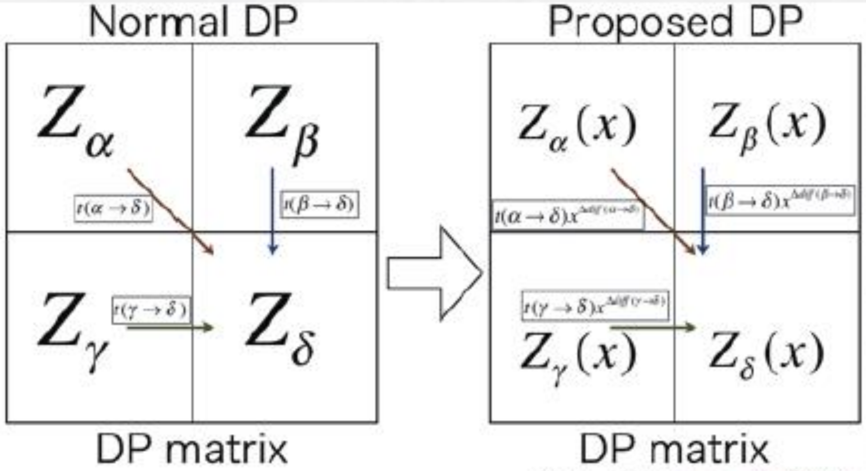
Lorenz R, Flamm C, Hofacker IL: 2D Projections of RNA Folding Landscapes.

Senter E, Dotu I, Clote P: Efficiently computing the 2D energy landscape of RNA.

Math Biol 2014. OpenURL

How to calculate the sum of the probabilities

- The basic idea is adopting polynomials which includes information on score when we calculate partition function.
- Simplified concept is illustrated in the following:

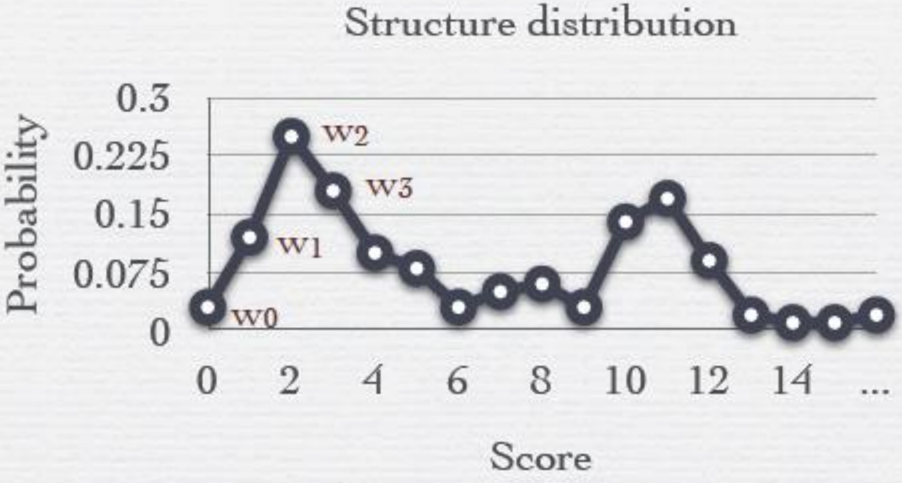


$$Z_\delta = Z_\alpha \cdot t(\alpha \to \delta) + Z_\beta \cdot t(\beta \to \delta) + Z_\gamma \cdot t(\gamma \to \delta)$$

$Z_\bullet$: *scalar*

$$\begin{aligned} Z_\delta(x) = {} & Z_\alpha(x) \cdot t(\alpha \to \delta) x^{\Delta diff(\alpha \to \delta)} \\ & + Z_\beta(x) \cdot t(\beta \to \delta) x^{\Delta diff(\beta \to \delta)} \\ & + Z_\gamma(x) \cdot t(\gamma \to \delta) x^{\Delta diff(\gamma \to \delta)} \end{aligned}$$

$Z_\bullet(x)$: *polynomial*



$$Z_n = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots$$

Efficient calculation of exact probability distributions of integer features on RNA secondary structures

･Algorithm construction

We modify the McCaskill model, which is the standard DP procedure for the partition function of RNA secondary structure ensemble.

$$Z^b_{ij} = e^{-[F_1(i,j)/kT]} + \sum_{h=i+1}^{j-2}\sum_{l=h+1}^{j-1} Z^b_{hl}e^{-[F_2(i,j,h,l)/kT]} + \sum_{h=i+1}^{j-1} Z^m_{i+1,h-1}Z^{m1}_{h,j-1}e^{-[(a+b)/kT]}$$

–

$$Z^b_{ij}(x) = e^{-[F_1(i,j)/kT]}x^{a_{ij}} + \sum_{h=i+1}^{j-2}\sum_{l=h+1}^{j-1} Z^b_{hl}e^{-[F_2(i,j,h,l)/kT]}x^{b_{ijhl}} + \sum_{h=i+1}^{j-1} Z^m_{i+1,h-1}Z^{m1}_{h,j-1}e^{-[(a+b)/kT]}x^{g_{ijh}}$$

Discrete Fourier Transform reduces time complexity of computations in order-level, and decentralizes the procedure.

$$x_l \equiv \exp\left(2\pi i \frac{l}{d_{max}+1}\right) \quad (l = 0, 1, \cdots, d_{max})$$

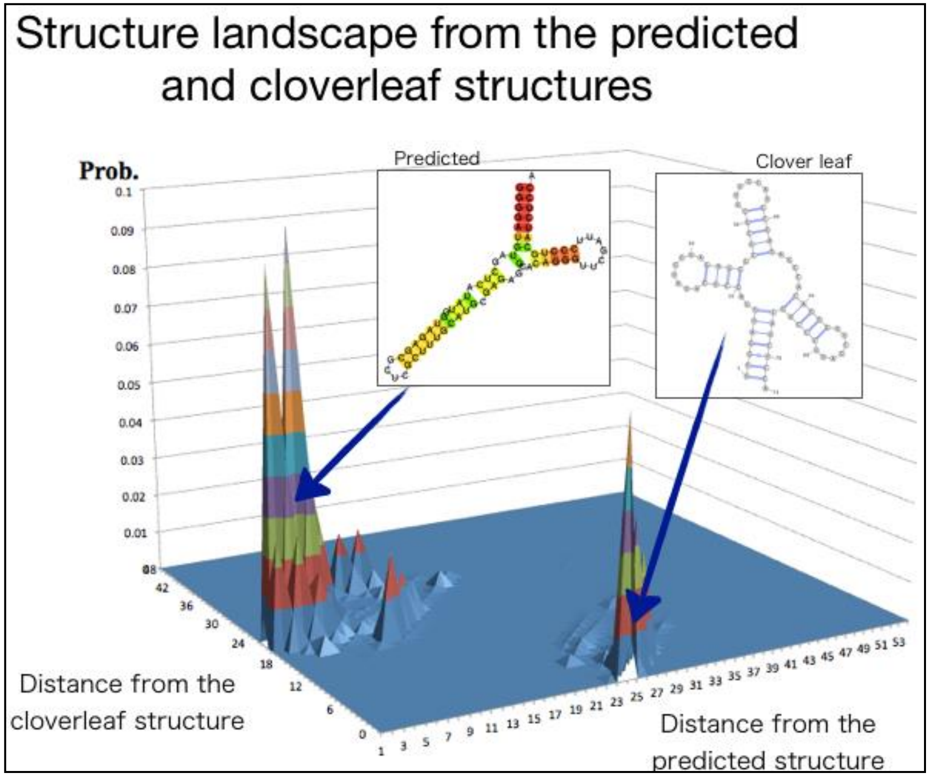
$$F_l = \frac{1}{d_{max}+1} Z_l(x) = \frac{1}{d_{max}+1} \sum_{d=0}^{d_{max}} w_d x_l^d \quad (l = 0, 1, \cdots, d_{max})$$

$$DFT(F_0, F_1, \cdots, F_{d_{max}}) \rightarrow w_0, w_1, \cdots, w_{d_{max}}$$

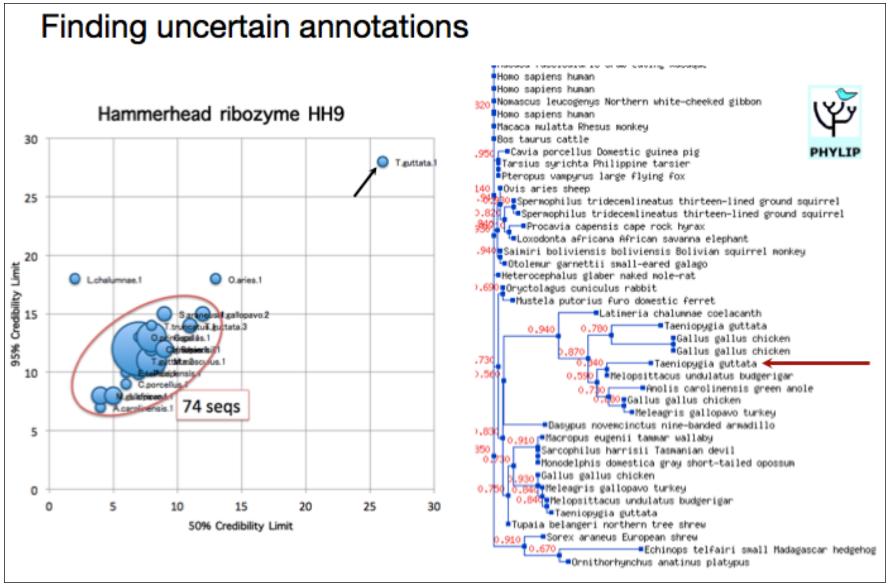
R. Mori et al. BMC Genomics 2014, 15(Suppl 10):S6

Efficient calculation of exact probability distributions of integer features on RNA secondary structures

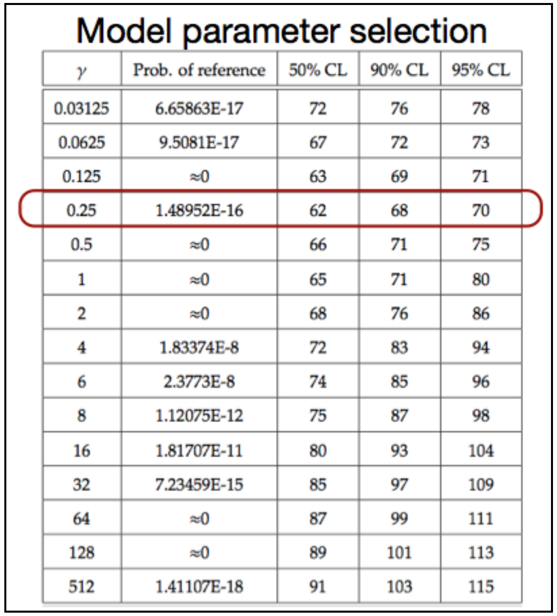
- Analyses by the proposed method



Structure landscape from the predicted and cloverleaf structures



Finding uncertain annotations



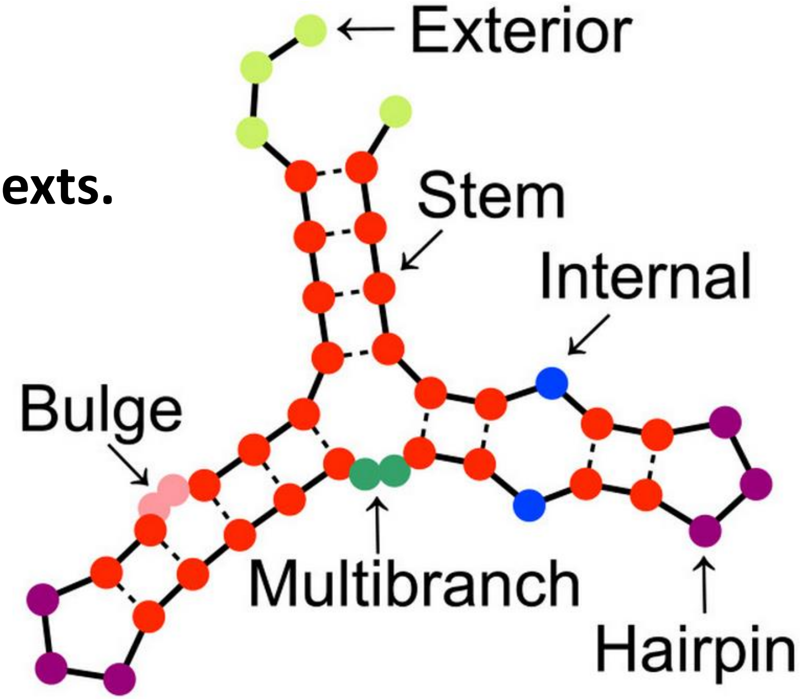
Model parameter selection

γ	Prob. of reference	50% CL	90% CL	95% CL
0.03125	6.65863E-17	72	76	78
0.0625	9.5081E-17	67	72	73
0.125	≈0	63	69	71
0.25	1.48952E-16	62	68	70
0.5	≈0	66	71	75
1	≈0	65	71	80
2	≈0	68	76	86
4	1.83374E-8	72	83	94
6	2.3773E-8	74	85	96
8	1.12075E-12	75	87	98
16	1.81707E-11	80	93	104
32	7.23459E-15	85	97	109
64	≈0	87	99	111
128	≈0	89	101	113
512	1.41107E-18	91	103	115

CapR

computes marginal probability for each structural context

The six structural contexts.

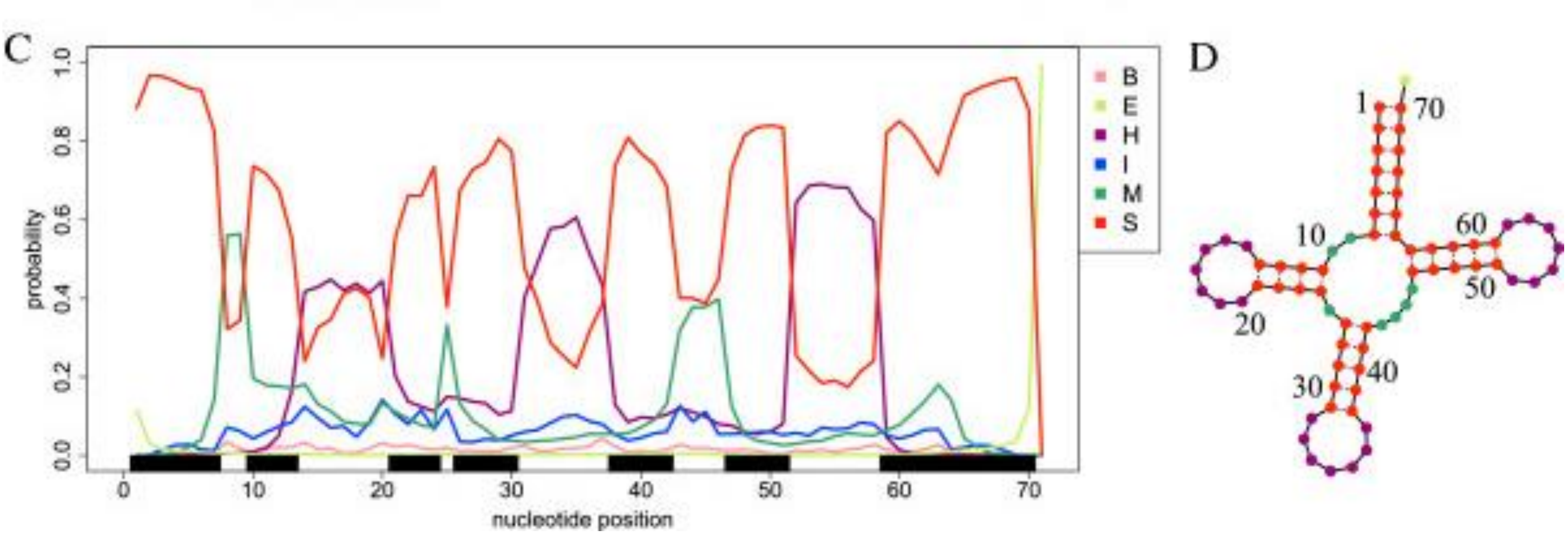


The six structural contexts are represented by six colors: stems (red), exterior loops (light green), hairpin loops (purple), bulge loops (pink), internal loops (blue) and multibranch loops (green). The unstructured context is the union of the exterior and multibranch loops. These colors are used throughout the paper.

Fukunaga *et al. Genome Biology* 2014 **15**:R16

CapR

Probabilistic structural profile of RNA

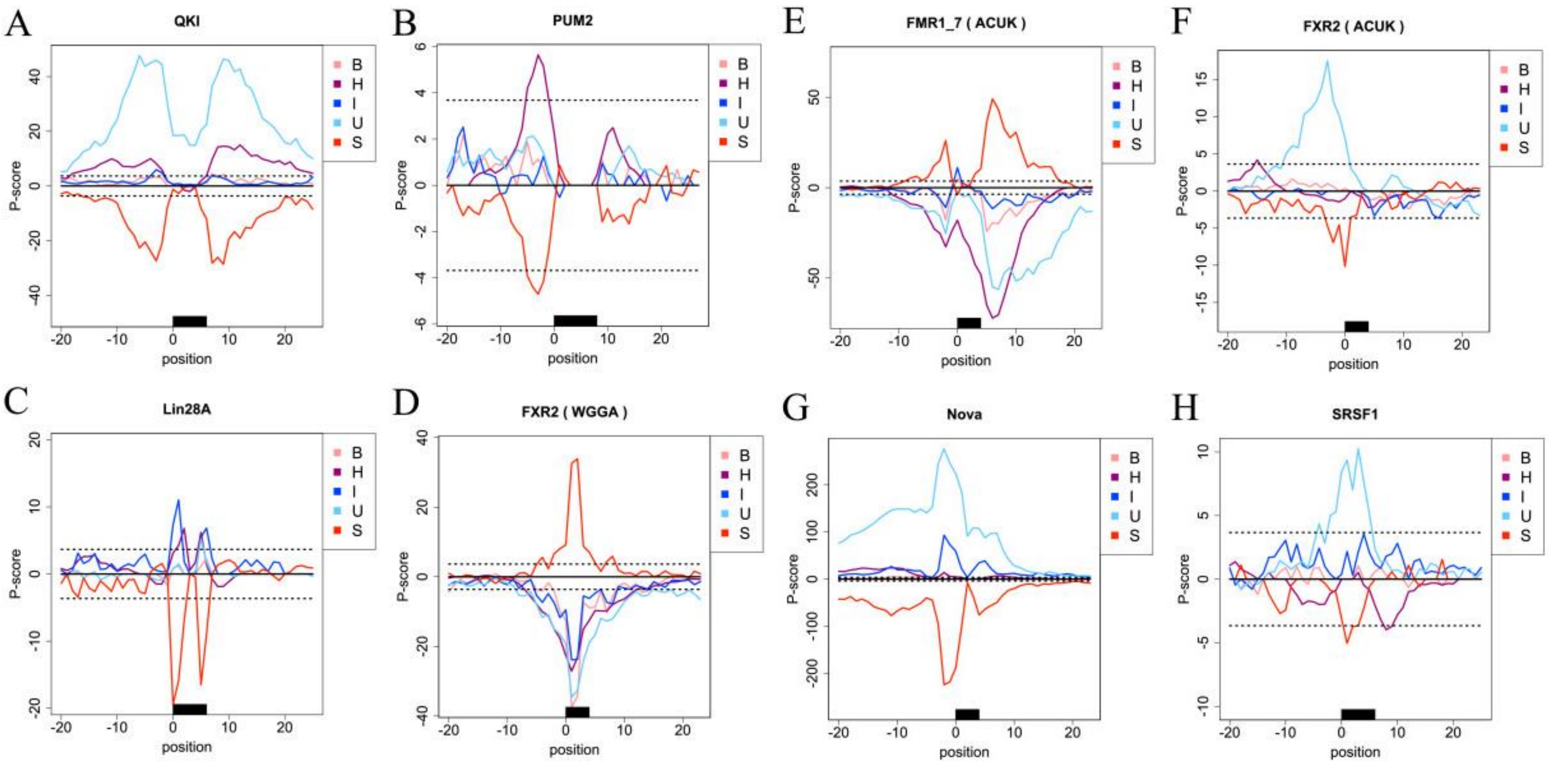


Performance of CapR. (C) The structural profiles of tRNAs. The *x*-axis represents the nucleotide positions from 5′ to 3′. The *y*-axis represents averaged probabilities that each base belongs to each structural context across all tRNA genes in the Rfam dataset [22]. The black boxes represent the nucleotides annotated as stem in Rfam. **(D)** tRNA cloverleaf structure annotated in Rfam. B, bulge loop; E, exterior loop; H, hairpin loop; I, internal loop; M, multibranch loop; S, stem.

Fukunaga *et al. Genome Biology* 2014 **15**:R16

CapR

Specific patterns of probabilistic structural profile near the binding site of RNA-biding proteins

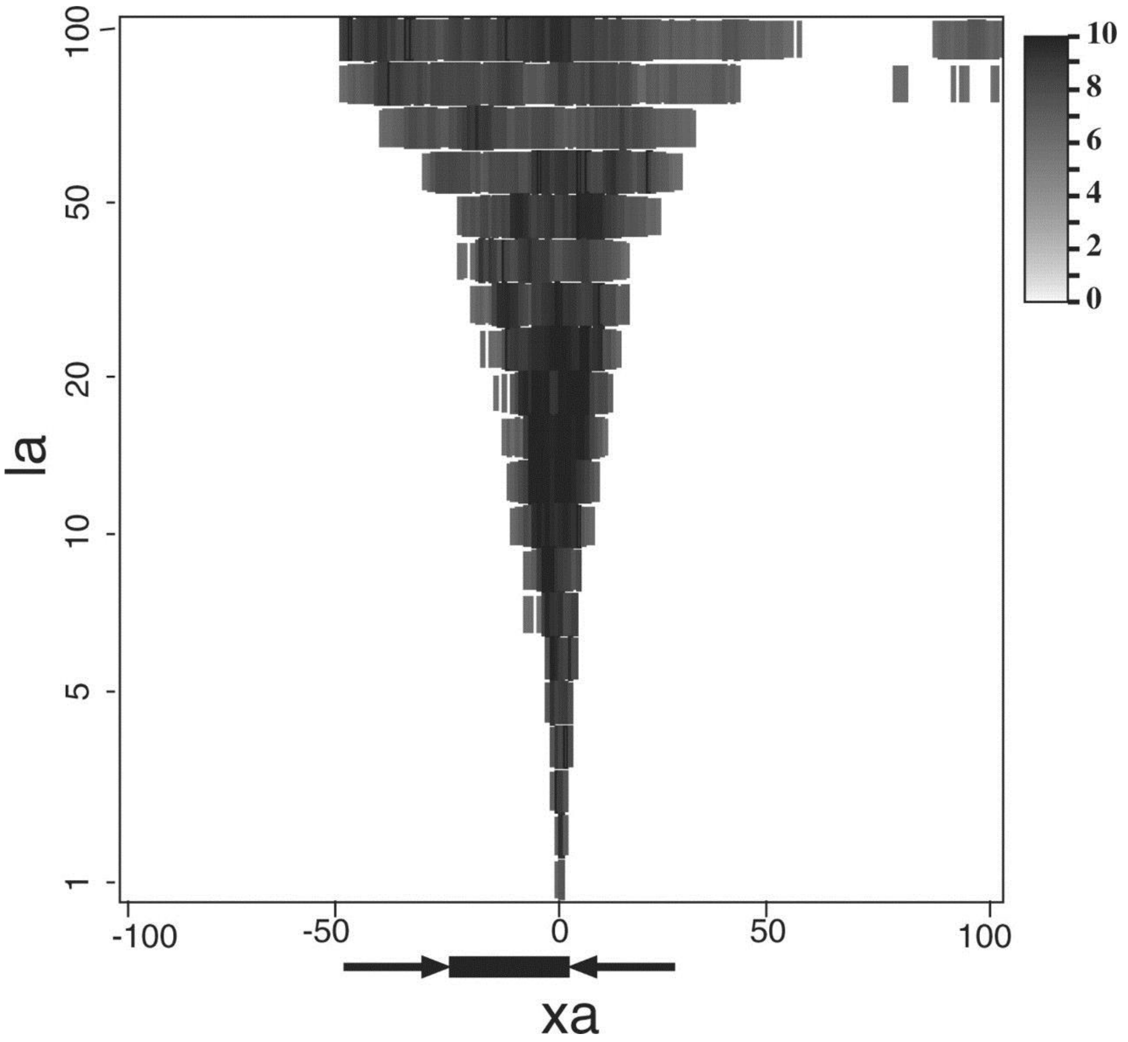


The distribution of theP scores for each RNA-binding protein. The *x*-axis represents the nucleotide positions and the *y*-axis represents the *P* score of ±20 bases around the sequential motif site. The position 0 denotes the start position of the sequential motif. Positive *P* scores for each structural context indicate that the positions tend to prefer the structural context. The black box represents the sequential motif site. The dotted lines show the corrected significance levels of the Bonferroni correction (*α*=0.05). The panels represent the distribution of *P* scores for **(A)**QKI, **(B)** Pum2, **(C)** Lin28A, **(D)** FXR2(WGGA), **(E)** FMR1_7(ACUK), **(F)** FXR2(ACUK), **(G)** Nova and**(H)** SRSF1. B, bulge loop; H, hairpin loop; I, internal loop; S, stem; U, unstructured.

Fukunaga *et al. Genome Biology* 2014 **15**:R16

Raccess

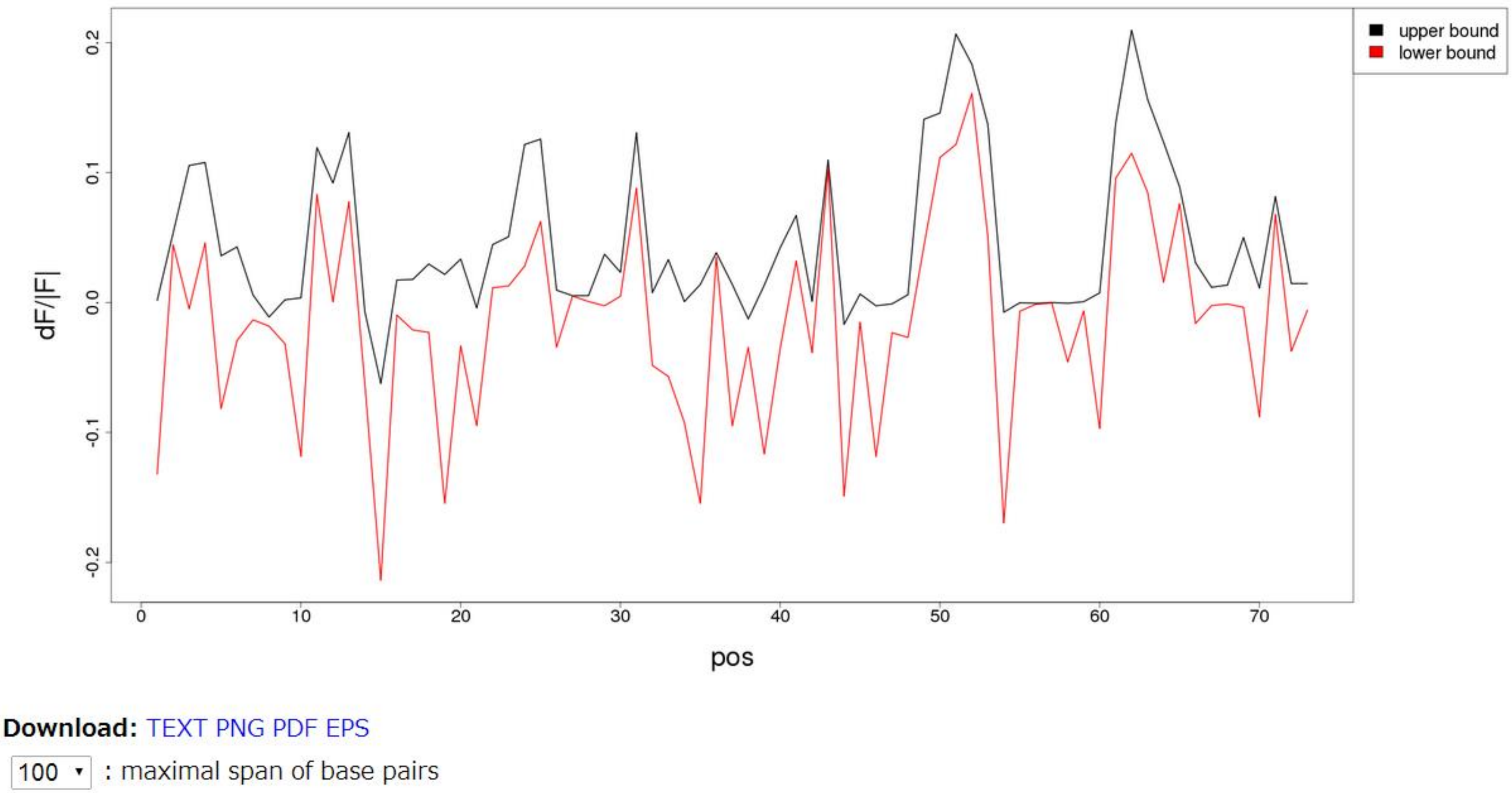
Density plot of siRNA efficacy–accessibility correlations.



Kiryu H et al. Bioinformatics 2011;27:1788-1797

Bioinformatics

Rchange



Download: TEXT PNG PDF EPS

100 ▾ : maximal span of base pairs

Kiryu H , Asai K Bioinformatics 2012;28:1093-1101

Acknowledgments

RNA Algorithms & Software

Michiaki Hamada	Waseda-U
Yukiteru Ono	IMSBIO
Hisanori Kiryu	U-Tokyo
Kengo Sato	Keio-U
Yuki Kato	Kyoto-U
Tsukasa Fukunaga	U-Tokyo
Ryota Mori	Asai Lab.
Toutai Mituyama	AIST

Collaborators in silico

Tomoshi. Kameda	AIST
Junichi Iwakiri	Asai Lab.
Shun Sakuraba	Asai Lab.
Zeng Chao	Asai Lab.
Goro Terai	Intec Inc.

Funding Development of technology to design genes for innovative biomaterial Project of METI