Recent advances in RNA secondary structure prediction with machine learning and deep learning

Tokyo Denki University (until Jul. 31)

Tokyo Institute of Technology (from Aug. 1)

Institute of Science Tokyo (from Oct. 1)

Kengo SATO

satoken@sato-lab.org

https://www.sato-lab.org/

Table of Contents

- Overview of RNA secondary structure prediction
  - Architecture
    - Nussinov algorithm, Nearest neighbor model
  - Inference
    - MFE, MEA
  - Parametrization
    - Machine learning, Deep learning
- Future direction
  - Chemical probing
  - RNA modification
  - Pseudoknots

What is RNA secondary structure prediction?

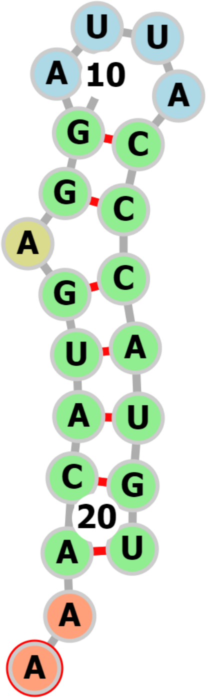
- Given an RNA sequence, predict its secondary structure

RNA sequence

AAACAUGAGGAUUACCCAUGU

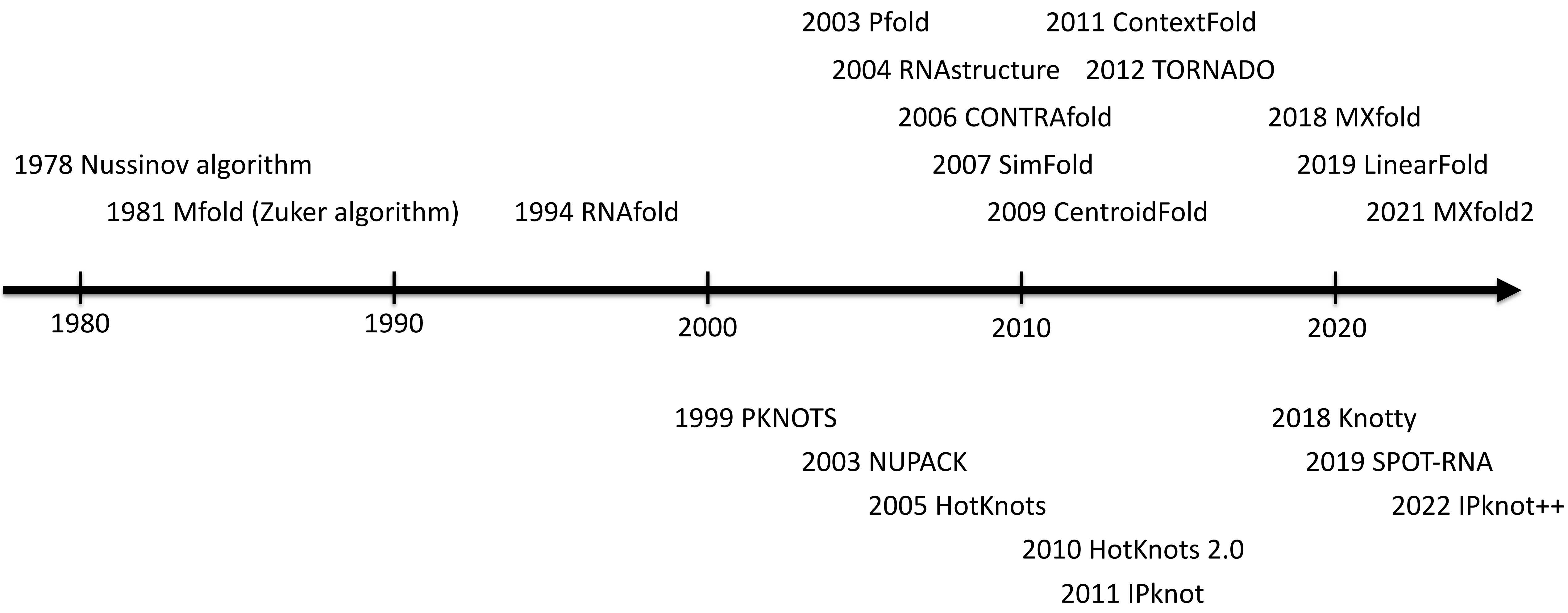
Predict

Secondary Structure



..(((((.((....)))))))

Prediction of RNA secondary structures



2003 Pfold

2004 RNAstructure

2006 CONTRAfold

2007 SimFold

2009 CentroidFold

2011 ContextFold

2012 TORNADO

2018 MXfold

2019 LinearFold

2021 MXfold2

1978 Nussinov algorithm

1981 Mfold (Zuker algorithm)

1994 RNAfold

1999 PKNOTS

2003 NUPACK

2005 HotKnots

2010 HotKnots 2.0

2011 IPknot

2018 Knotty

2019 SPOT-RNA

2022 IPknot++

Nussinov algorithm

- Observation 1:
  The greater the number of base-pairs, the more energetically stable.

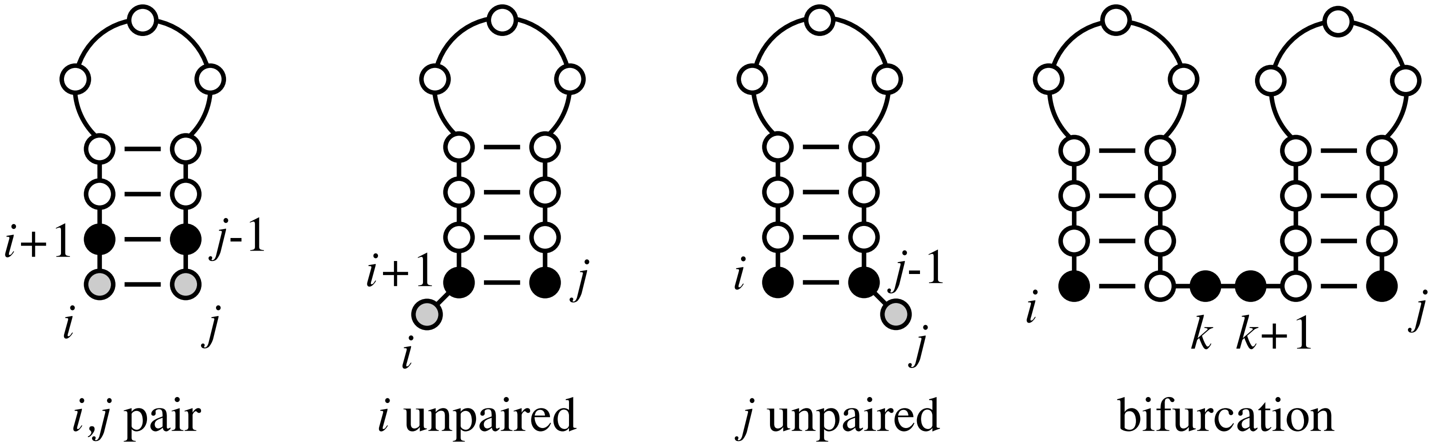
  ⇒ Nussinov algorithm predicts a secondary structure that maximizes the number of base-pairs.

- Observation 2:
  The optimal structure of a given sequence can be constructed from the optimal structures of shorter subsequences.

  ⇒ Dynamic programming

Nussinov algorithm

- The optimal structure of a subsequence $[i, j]$ can be computed from a slightly smaller subsequence.



1. Add a base-pair (i, j) to the optimal structure of the subsequence $[i+1, j$-$1]$.
2. Add an unpaired base i to the optimal structure of the subsequence $[i+1, j]$.
3. Add an unpaired base j to the optimal structure of the subsequence $[i, j$-$1]$.
4. Concatenate the two optimal substructures $[i, k]$ and $[k+1, j]$.

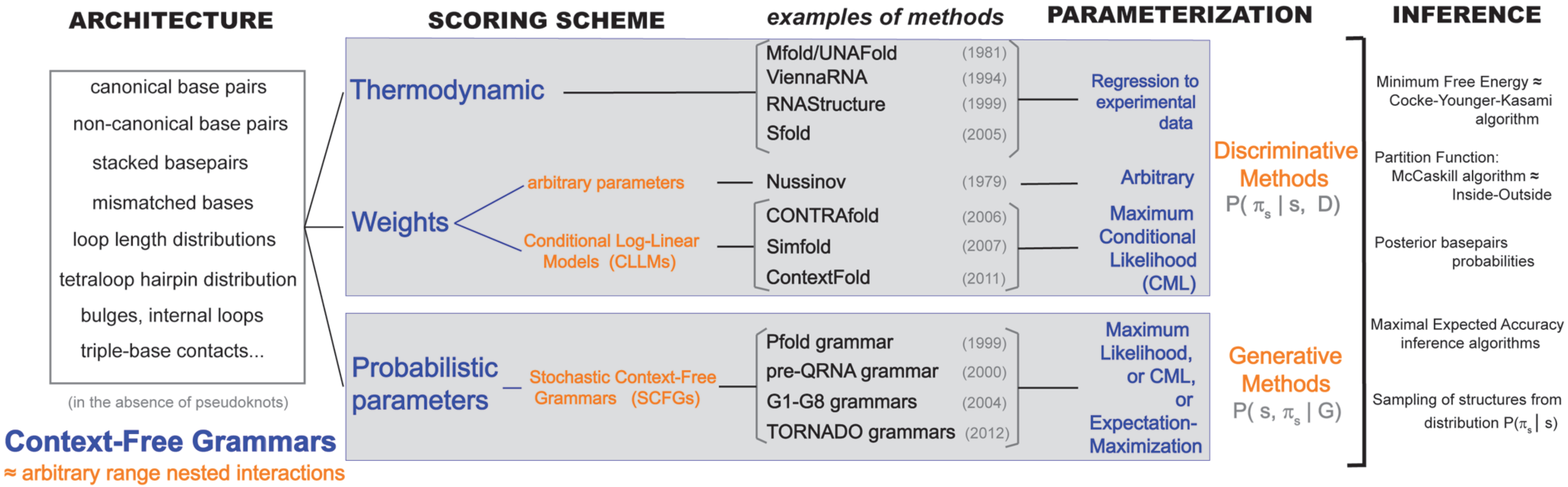
Nussinov algorithm

- Observation:
  The greater the number of base pairs, the more energetically stable.

$$s(i,j) = \max \begin{cases} s(i+1, j-1) + 1 \longleftarrow \boxed{\text{if the i-th base and j-th base are allowed to form base pairs}} \\ s(i+1, j) \\ s(i, j-1) \\ \max_k [s(i,k) + s(k+1, j)] \end{cases}$$

- Computational complexity: $O(L^3)$ time, $O(L^2)$ space

The four ingredients of RNA secondary structure prediction



ARCHITECTURE

canonical base pairs

non-canonical base pairs

stacked basepairs

mismatched bases

loop length distributions

tetraloop hairpin distribution

bulges, internal loops

triple-base contacts...

(in the absence of pseudoknots)

Context-Free Grammars

≈ arbitrary range nested interactions

SCORING SCHEME

Thermodynamic

Weights
- arbitrary parameters
- Conditional Log-Linear Models (CLLMs)

Probabilistic parameters
- Stochastic Context-Free Grammars (SCFGs)

examples of methods

Mfold/UNAFold (1981)
ViennaRNA (1994)
RNAStructure (1999)
Sfold (2005)

Nussinov (1979)

CONTRAfold (2006)
Simfold (2007)
ContextFold (2011)

Pfold grammar (1999)
pre-QRNA grammar (2000)
G1-G8 grammars (2004)
TORNADO grammars (2012)

PARAMETERIZATION

Regression to experimental data

Arbitrary

Maximum Conditional Likelihood (CML)

Maximum Likelihood, or CML, or Expectation-Maximization

Discriminative Methods
$P(\pi_s \mid s, D)$

Generative Methods
$P(s, \pi_s \mid G)$

INFERENCE

Minimum Free Energy ≈ Cocke-Younger-Kasami algorithm

Partition Function: McCaskill algorithm ≈ Inside-Outside

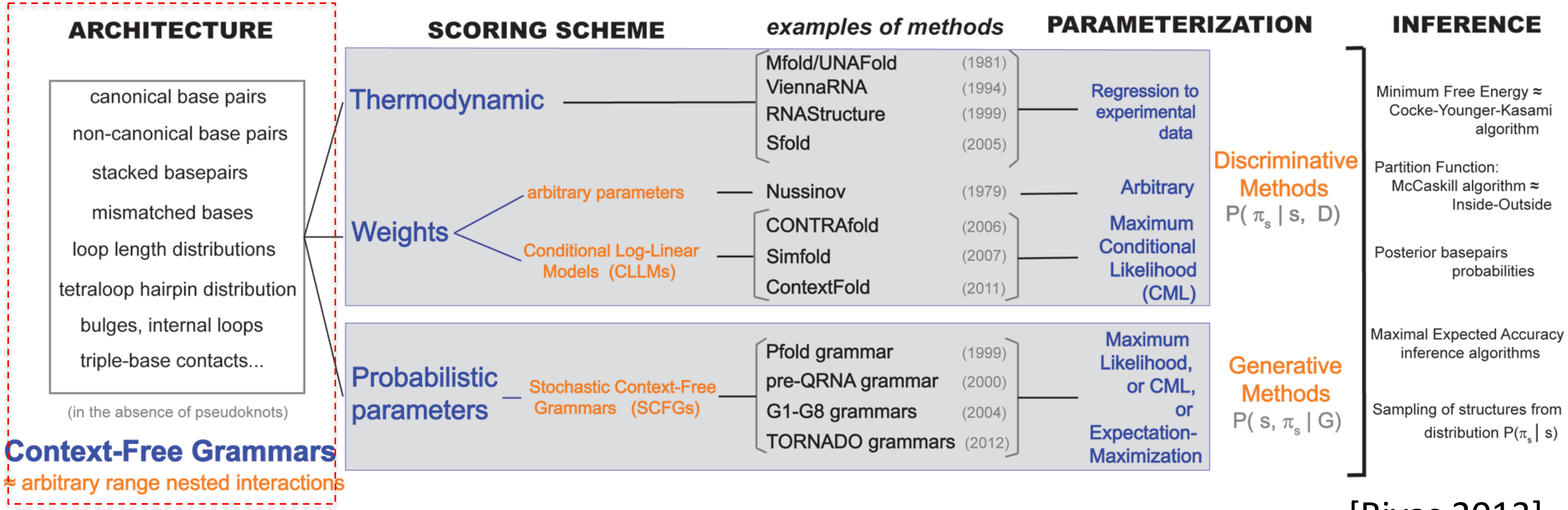
Posterior basepairs probabilities

Maximal Expected Accuracy inference algorithms

Sampling of structures from distribution $P(\pi_s \mid s)$

[Rivas 2013]

The four ingredients of RNA secondary structure prediction



ARCHITECTURE

canonical base pairs
non-canonical base pairs
stacked basepairs
mismatched bases
loop length distributions
tetraloop hairpin distribution
bulges, internal loops
triple-base contacts...

(in the absence of pseudoknots)

Context-Free Grammars

≈ arbitrary range nested interactions

SCORING SCHEME

Thermodynamic

Weights: arbitrary parameters; Conditional Log-Linear Models (CLLMs)

Probabilistic parameters: Stochastic Context-Free Grammars (SCFGs)

examples of methods

Mfold/UNAFold (1981)
ViennaRNA (1994)
RNAStructure (1999)
Sfold (2005)
Nussinov (1979)
CONTRAfold (2006)
Simfold (2007)
ContextFold (2011)
Pfold grammar (1999)
pre-QRNA grammar (2000)
G1-G8 grammars (2004)
TORNADO grammars (2012)

PARAMETERIZATION

Regression to experimental data

Arbitrary

Maximum Conditional Likelihood (CML)

Discriminative Methods $P(\pi_s \mid s, D)$

Maximum Likelihood, or CML, or Expectation-Maximization

Generative Methods $P(s, \pi_s \mid G)$

INFERENCE

Minimum Free Energy ≈ Cocke-Younger-Kasami algorithm

Partition Function: McCaskill algorithm ≈ Inside-Outside

Posterior basepairs probabilities

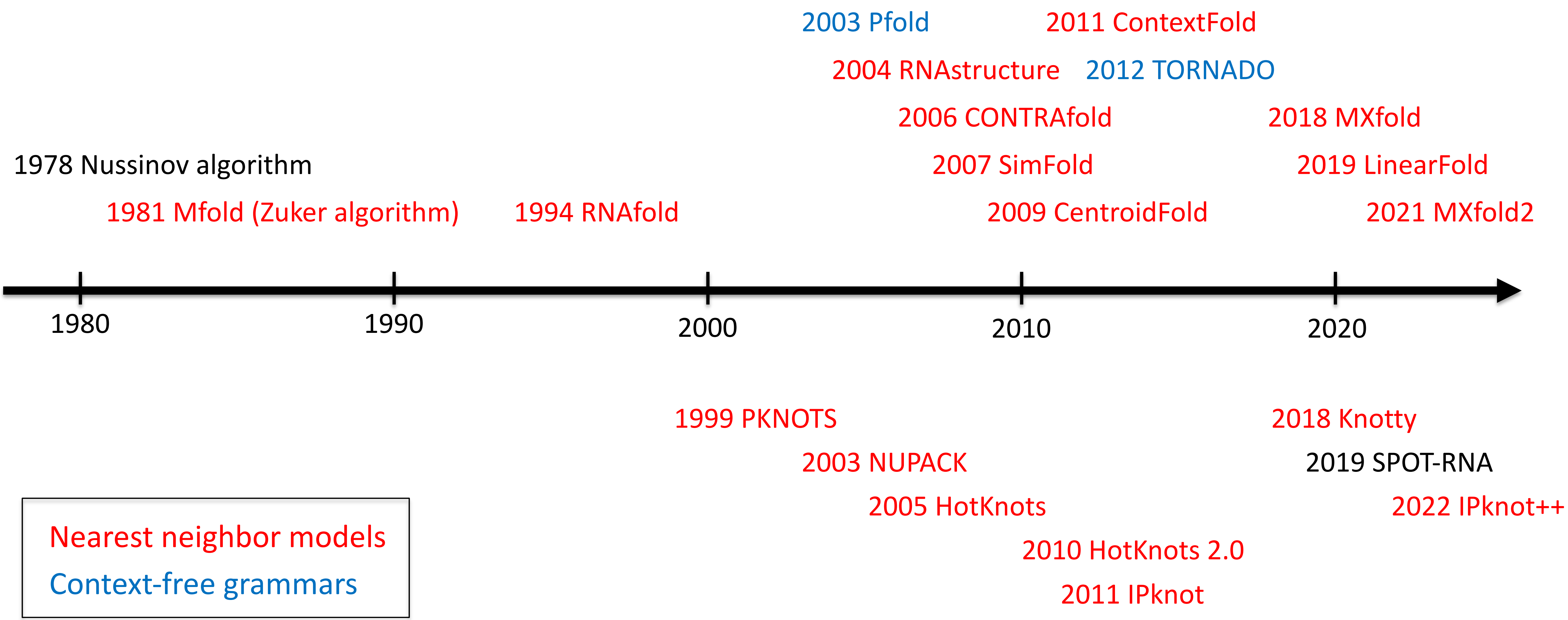
Maximal Expected Accuracy inference algorithms

Sampling of structures from distribution $P(\pi_s \mid s)$

[Rivas 2013]

1. Architecture
   - Nearest neighbor model
   - Context-free grammars

Prediction of RNA secondary structures



2003 Pfold

2011 ContextFold

2004 RNAstructure

2012 TORNADO

2006 CONTRAfold

2018 MXfold

1978 Nussinov algorithm

2007 SimFold

2019 LinearFold

1981 Mfold (Zuker algorithm)

1994 RNAfold

2009 CentroidFold

2021 MXfold2

1999 PKNOTS

2018 Knotty

2003 NUPACK

2019 SPOT-RNA

2005 HotKnots

2022 IPknot++

2010 HotKnots 2.0

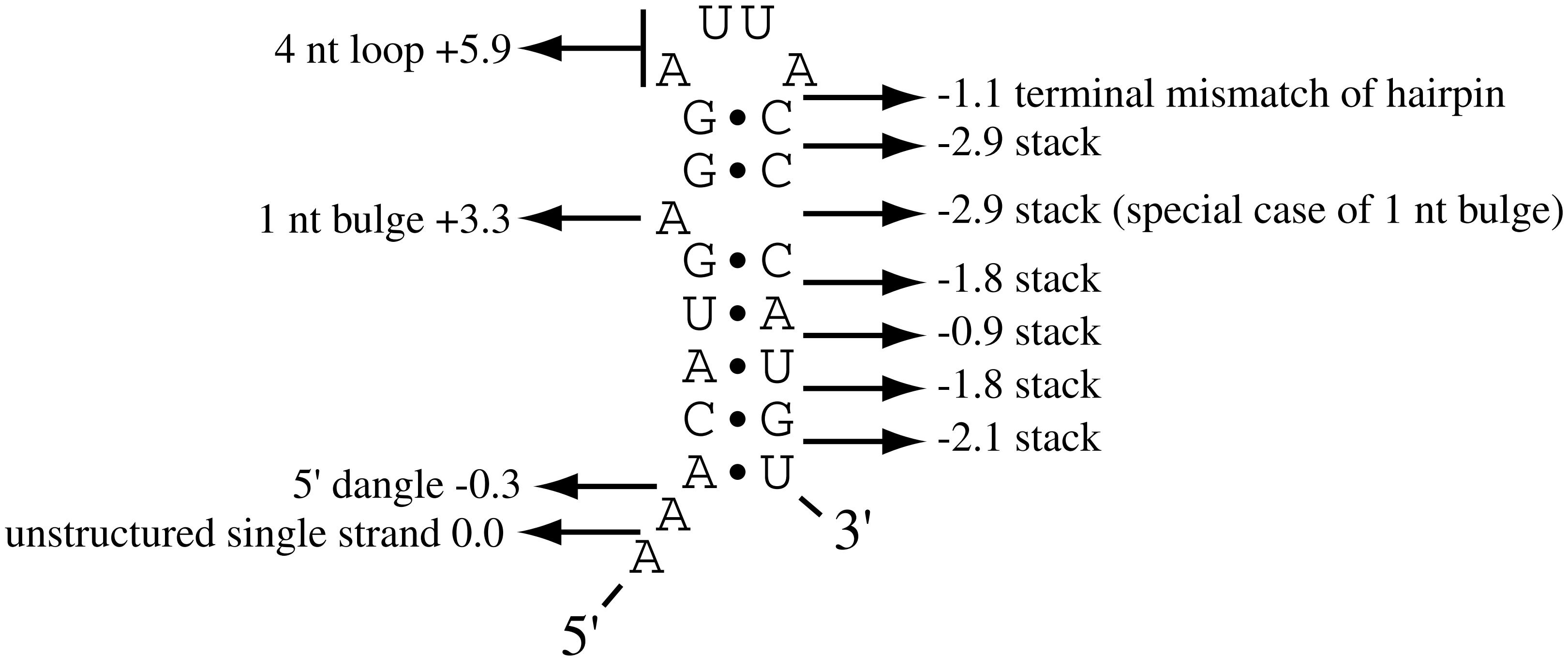
2011 IPknot

Nearest neighbor models

Context-free grammars

Nearest neighbor model

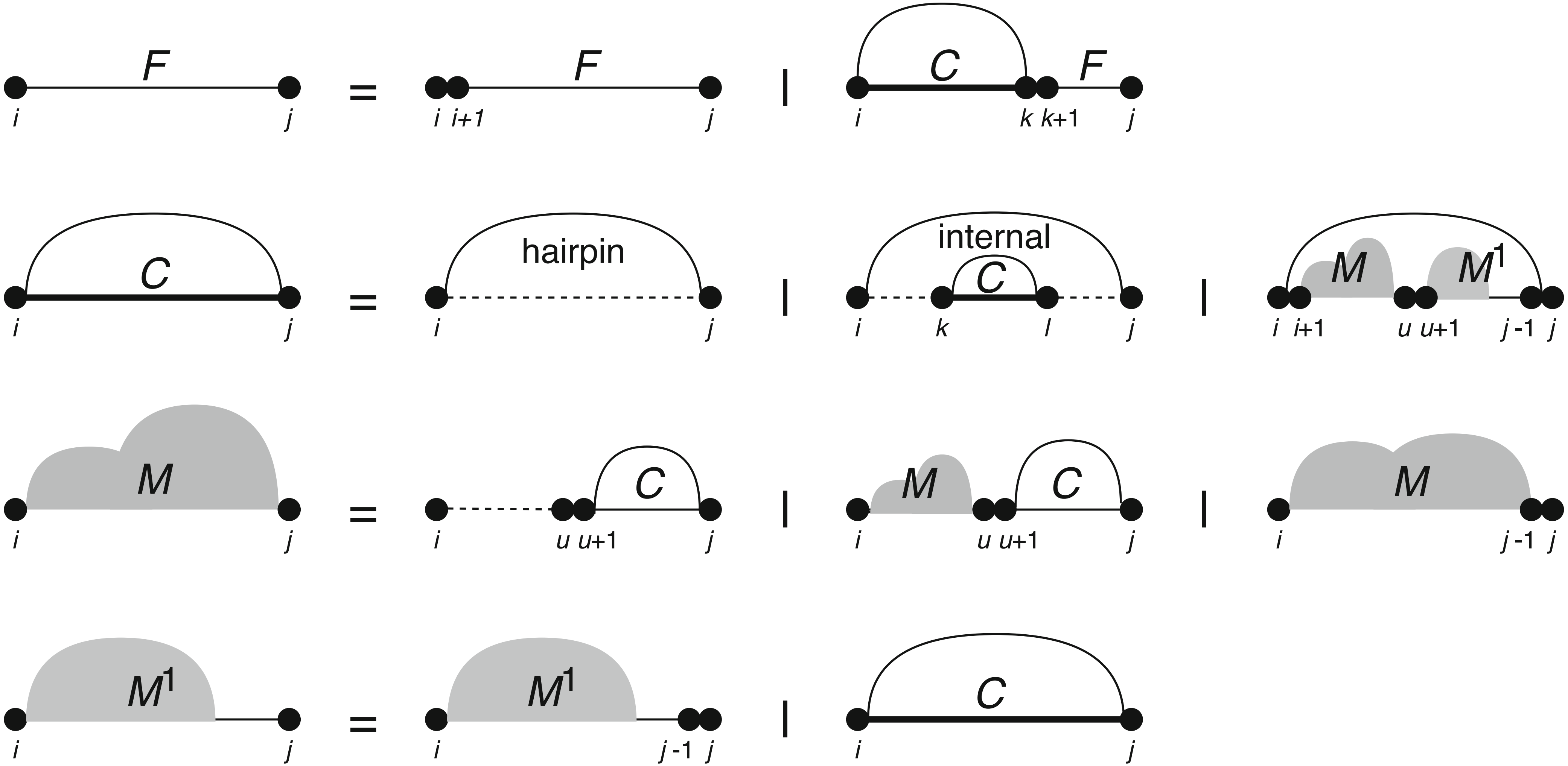
- Nearest neighbor model [Zuker&Stiegler81; Zuker03]
  - The free energy of a secondary structure is the sum of the free energy of its substructures.



overall ΔG = -4.6 kcal/mol

Nearest neighbor model

- Decomposition of RNA secondary structure with the nearest neighbor model



Nearest neighbor model

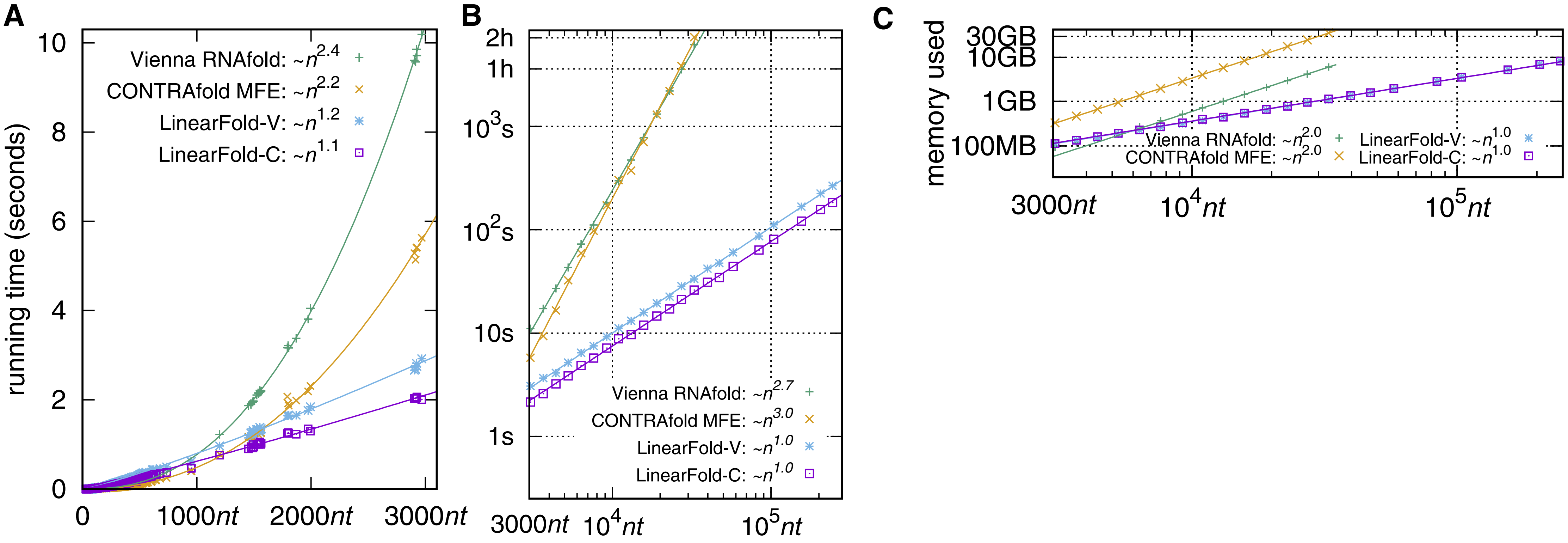
- Recursive equation for Zuker algorithm [1981]

$$
\begin{aligned}
F_{ij} &= \min\left\{F_{i+1,j},\ \min_{i<k\leq j} C_{ik} + F_{k+1,j}\right\} \\
C_{ij} &= \min\left\{\mathcal{H}(i,j),\ \min_{i<k<l<j} C_{kl} + \mathcal{I}(i,j;k,l),\ \min_{i<u<j} M_{i+1,u} + M^{1}_{u+1,j-1} + a\right\} \\
M_{ij} &= \min\left\{\min_{i<u<j}(u-i+1)c + C_{u+1,j} + b,\ \min_{i<u<j} M_{iu} + C_{u+1,j} + b,\ M_{i,j-1} + c\right\} \\
M^{1}_{ij} &= \min\left\{M^{1}_{i,j-1} + c,\ C_{ij} + b\right\} \\
F_{ii} &= 0,\ C_{ii} = M_{ii} = M^{1}_{ii} = \infty,
\end{aligned}
$$

- Computational complexity: $O(L^3)$ time, $O(L^2)$ space

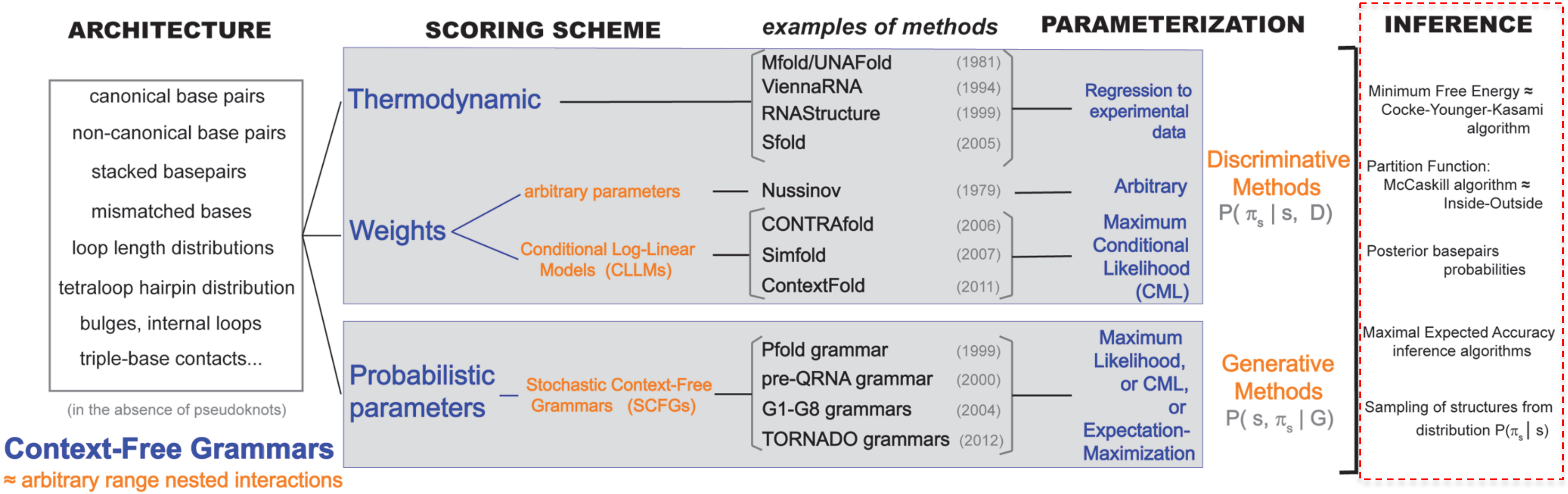
LinearFold algorithm

- [Huang *et al.*, 2019] developed LinearFold algorithm using:
  - left-to-right incremental dynamic programming, and
  - the beam search approximate to reduce search space.



- Computational complexity: $O(L)$ time, $O(L)$ space

The four ingredients of RNA secondary structure prediction



ARCHITECTURE

canonical base pairs
non-canonical base pairs
stacked basepairs
mismatched bases
loop length distributions
tetraloop hairpin distribution
bulges, internal loops
triple-base contacts...

(in the absence of pseudoknots)

Context-Free Grammars
≈ arbitrary range nested interactions

SCORING SCHEME

Thermodynamic

Weights
arbitrary parameters
Conditional Log-Linear Models (CLLMs)

Probabilistic parameters
Stochastic Context-Free Grammars (SCFGs)

examples of methods

Mfold/UNAFold	(1981)
ViennaRNA	(1994)
RNAStructure	(1999)
Sfold	(2005)
Nussinov	(1979)
CONTRAfold	(2006)
Simfold	(2007)
ContextFold	(2011)
Pfold grammar	(1999)
pre-QRNA grammar	(2000)
G1-G8 grammars	(2004)
TORNADO grammars	(2012)

PARAMETERIZATION

Regression to experimental data

Arbitrary

Maximum Conditional Likelihood (CML)

Maximum Likelihood, or CML, or Expectation-Maximization

Discriminative Methods
$P(\pi_s \mid s, D)$

Generative Methods
$P(s, \pi_s \mid G)$

INFERENCE

Minimum Free Energy ≈ Cocke-Younger-Kasami algorithm

Partition Function: McCaskill algorithm ≈ Inside-Outside

Posterior basepairs probabilities

Maximal Expected Accuracy inference algorithms

Sampling of structures from distribution $P(\pi_s \mid s)$

[Rivas 2013]

4. Inference

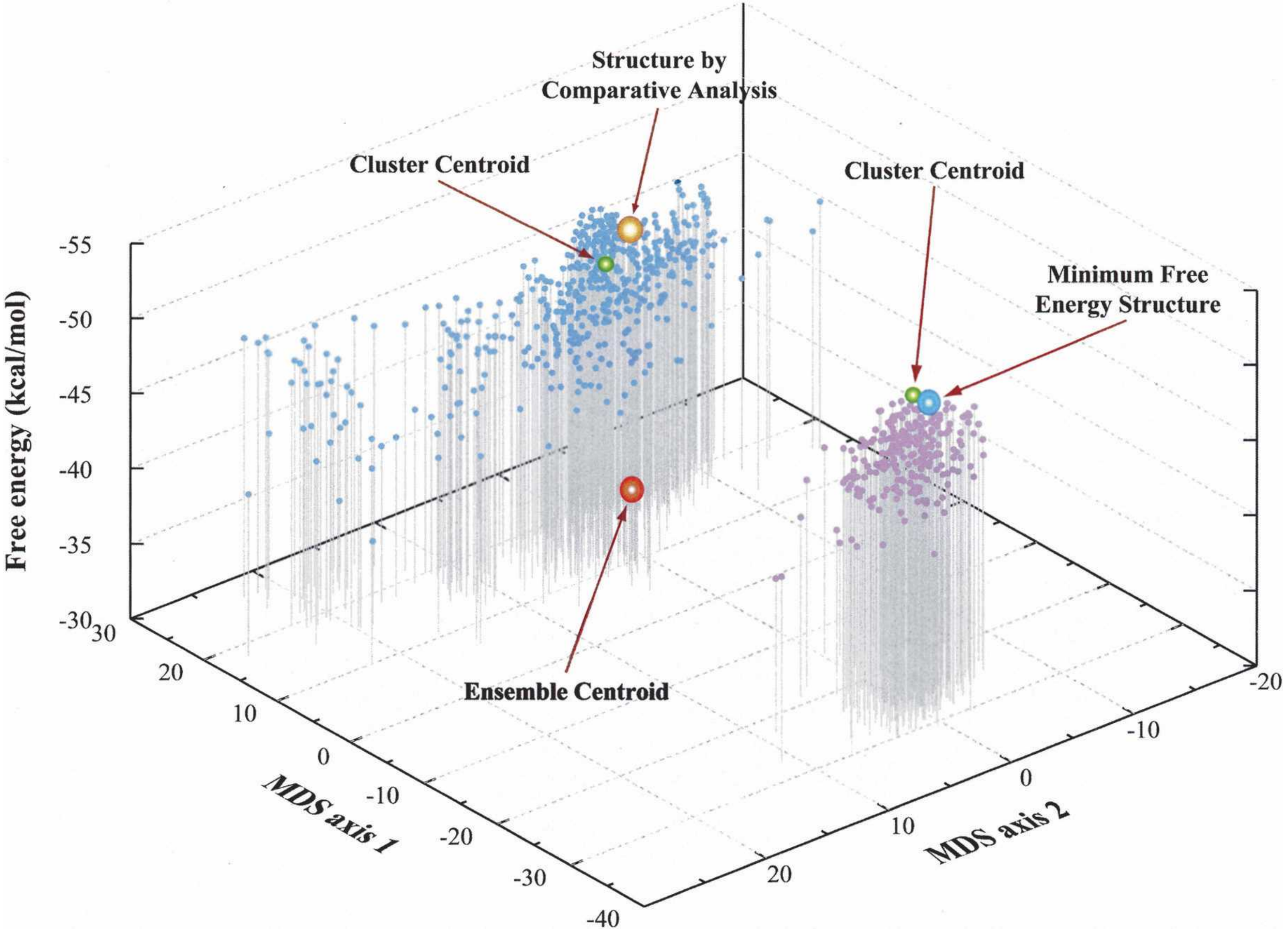
- Minimum free energy (MFE)
- Maximum likelihood estimate (MLE)
- Maximum expected accuracy (MEA)

Inference: MFE/MLE or MEA

- **Inference** focuses on which secondary structure is drawn from the probability distribution of RNA secondary structures.

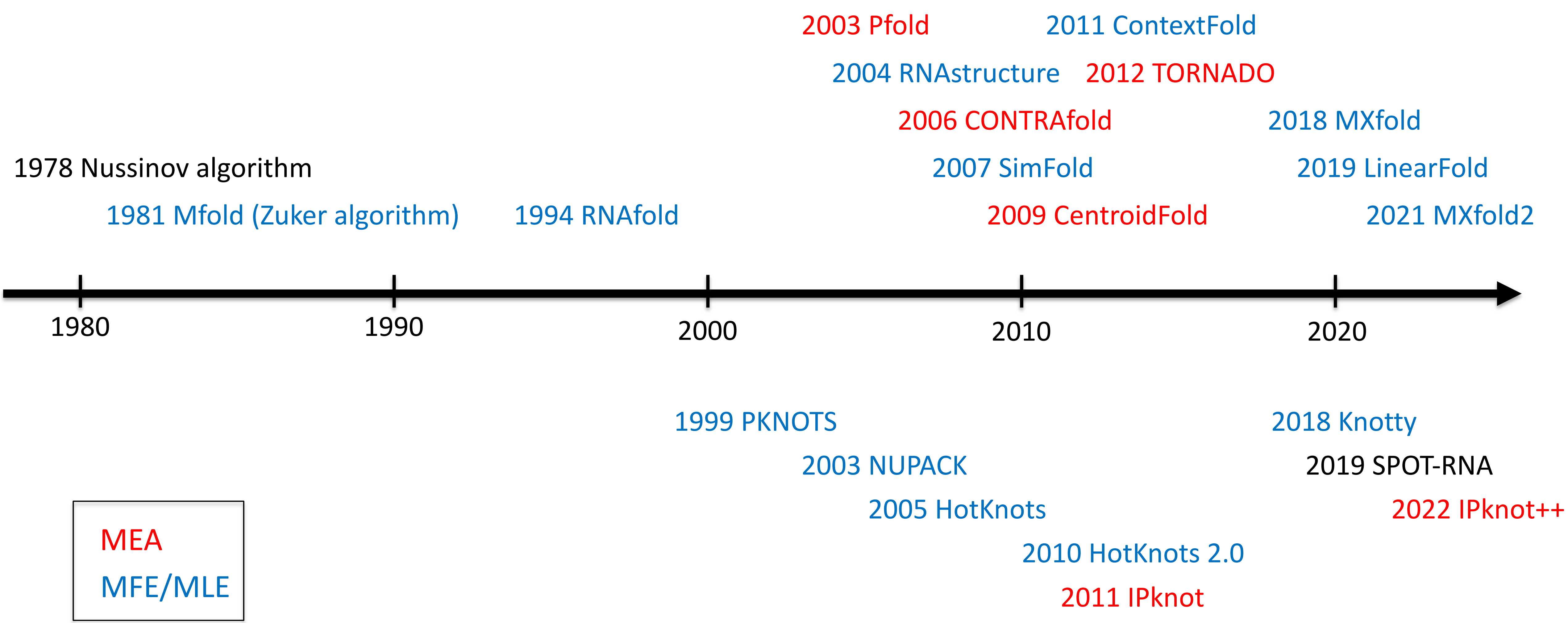
- Predict minimum free energy (MFE) structure
  - Zuker algorithm (Zuker et al., 1981)
  - Software: Mfold / RNAfold
  - Equivalent to maximum likelihood estimate with McCaskill model
- Predict maximum expected accuracy (MEA) structure
  - Prediction by considering "distribution" of secondary structures
  - Software:
    - CONTRAfold (Do et al., 2006)
    - CentroidFold (Hamada et al., 2009, Sato et al. 2009)

MFE structure is not always the best



[Ding et al, 2005]

Prediction of RNA secondary structures



2003 Pfold

2011 ContextFold

2004 RNAstructure

2012 TORNADO

2006 CONTRAfold

2018 MXfold

1978 Nussinov algorithm

2007 SimFold

2019 LinearFold

1981 Mfold (Zuker algorithm)

1994 RNAfold

2009 CentroidFold

2021 MXfold2

1999 PKNOTS

2018 Knotty

2003 NUPACK

2019 SPOT-RNA

2005 HotKnots

2022 IPknot++

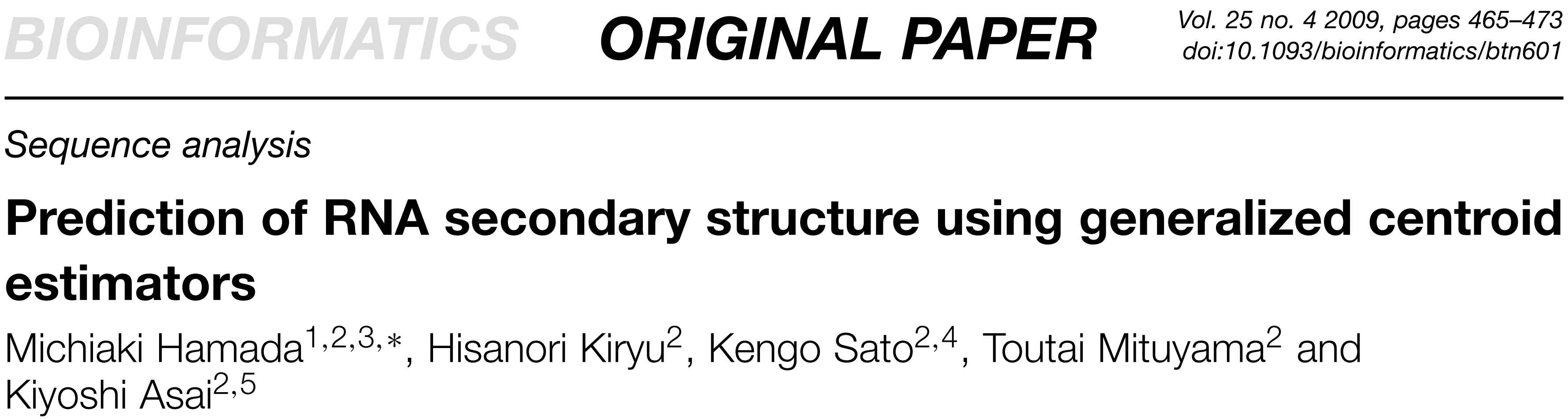
2010 HotKnots 2.0

2011 IPknot

MEA

MFE/MLE

RNA secondary structure prediction using maximizing expected accuracy



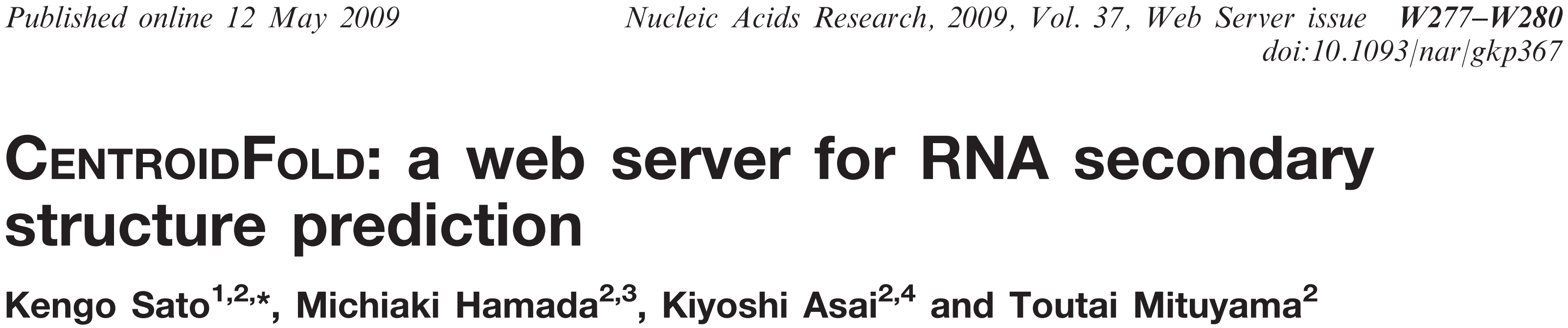
BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 4 2009, pages 465–473
doi:10.1093/bioinformatics/btn601

Sequence analysis

Prediction of RNA secondary structure using generalized centroid estimators

Michiaki Hamada[1,2,3,*], Hisanori Kiryu[2], Kengo Sato[2,4], Toutai Mituyama[2] and Kiyoshi Asai[2,5]



Published online 12 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W277–W280***
doi:10.1093/nar/gkp367

CENTROIDFOLD: a web server for RNA secondary structure prediction

Kengo Sato[1,2,*], Michiaki Hamada[2,3], Kiyoshi Asai[2,4] and Toutai Mituyama[2]

CentroidFold

Maximizing expected accuracy

- Given a space $S(x)$ of secondary structures of RNA sequence x, predict a structure $\hat{y}$ that maximizes an accuracy metric.

y: a reference structure

$\hat{y}$: a predicted structure

◆Gain function for true prediction

$$G(y,\hat{y}) = \gamma \underbrace{TP(y,\hat{y})}_{\text{\# of true positives}} + \underbrace{TN(y,\hat{y})}_{\text{\# of true negatives}} \quad (\gamma > 0)$$

Predict as many correct base pairs as possible

Maximizing expected accuracy

- Given a probability distribution $P(y \mid x)$ over a space $S(x)$ of secondary structures, predict a structure $\hat{y}$ that maximizes expected accuracy

$$\mathop{\arg\max}_{\hat{y} \in \mathcal{S}(x)} \sum_{y \in \mathcal{S}(x)} \boxed{G(y, \hat{y})} P(y \mid x)$$

y: a reference structure

$\hat{y}$: a predicted structure

◆Gain function for true prediction

$$G(y, \hat{y}) = \gamma \underbrace{TP(y, \hat{y})}_{\text{\# of true positives}} + \underbrace{TN(y, \hat{y})}_{\text{\# of true negatives}} \quad (\gamma > 0)$$

Predict as many correct base pairs as possible

Maximizing expected accuracy

- Find $\hat{y}$ that maximizes:

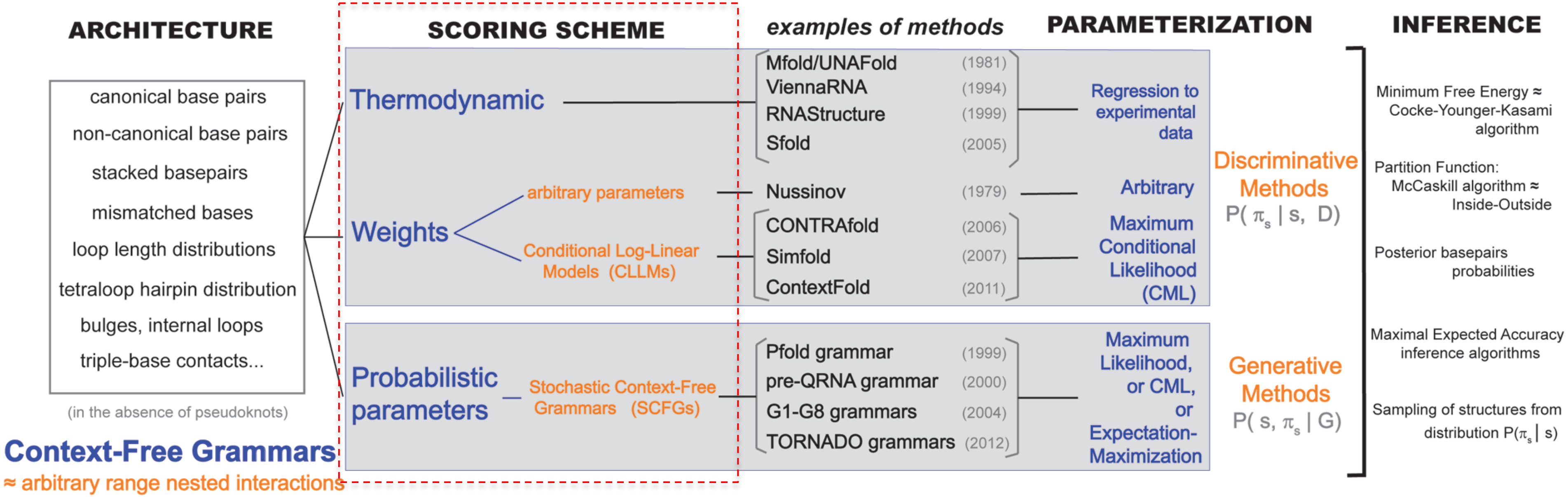
$$\sum_{y \in \mathcal{S}(x)} G(y, \hat{y}) P(y \mid x) = \sum_{i<j} \underline{[(\gamma + 1) p_{ij} - 1]} \hat{y}_{ij} + C$$

base-pairing probability

- Nussinov-style dynamic programming

$$s(i,j) = \max \begin{cases} s(i+1, j-1) + \underline{[(\gamma + 1) p_{ij} - 1]} \\ s(i+1, j) \\ s(i, j-1) \\ \max_k [s(i,k) + s(k+1, j)] \end{cases}$$

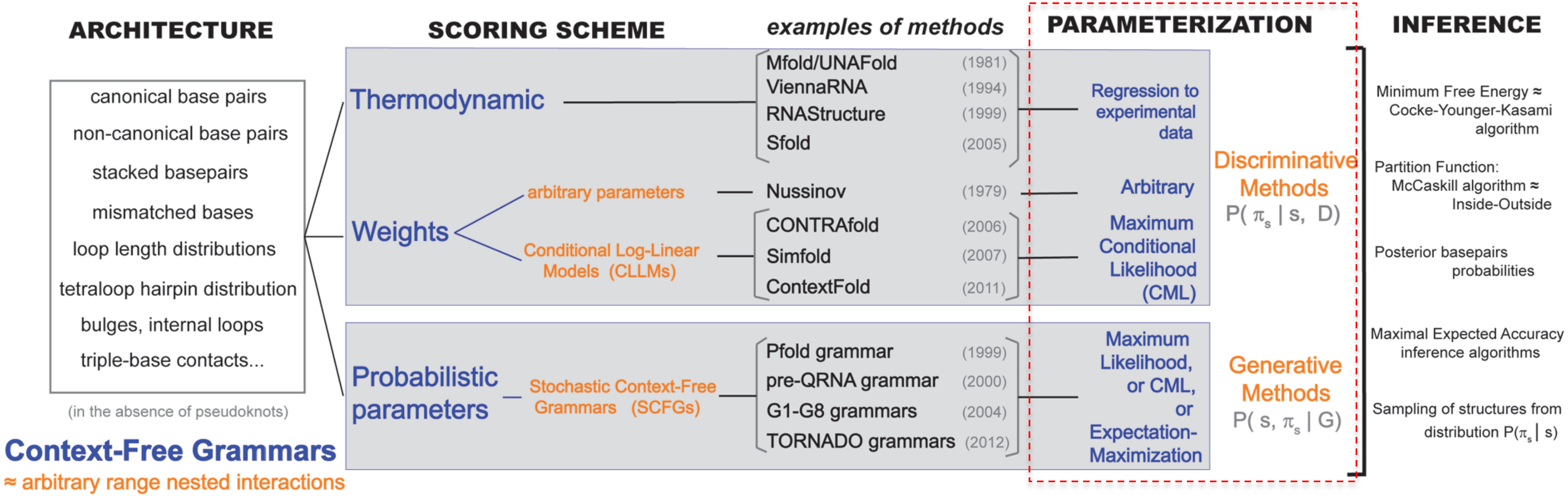
The four ingredients of RNA secondary structure prediction



[Rivas 2013]

2. Scoring scheme
   - Weights
   - Probability distribution

The four ingredients of RNA secondary structure prediction



ARCHITECTURE

canonical base pairs
non-canonical base pairs
stacked basepairs
mismatched bases
loop length distributions
tetraloop hairpin distribution
bulges, internal loops
triple-base contacts...

(in the absence of pseudoknots)

Context-Free Grammars
≈ arbitrary range nested interactions

SCORING SCHEME

Thermodynamic

Weights
arbitrary parameters
Conditional Log-Linear Models (CLLMs)

Probabilistic parameters
Stochastic Context-Free Grammars (SCFGs)

examples of methods

Mfold/UNAFold	(1981)
ViennaRNA	(1994)
RNAStructure	(1999)
Sfold	(2005)
Nussinov	(1979)
CONTRAfold	(2006)
Simfold	(2007)
ContextFold	(2011)
Pfold grammar	(1999)
pre-QRNA grammar	(2000)
G1-G8 grammars	(2004)
TORNADO grammars	(2012)

PARAMETERIZATION

Regression to experimental data

Arbitrary

Maximum Conditional Likelihood (CML)

Discriminative Methods
$P(\pi_s \mid s, D)$

Maximum Likelihood, or CML, or Expectation-Maximization

Generative Methods
$P(s, \pi_s \mid G)$

INFERENCE

Minimum Free Energy ≈ Cocke-Younger-Kasami algorithm

Partition Function: McCaskill algorithm ≈ Inside-Outside

Posterior basepairs probabilities

Maximal Expected Accuracy inference algorithms

Sampling of structures from distribution $P(\pi_s \mid s)$

[Rivas 2013]

3. Parameterization
   - Thermodynamic-based methods
   - Machine learning-based methods
     - discriminative, generative

Prediction of RNA secondary structures

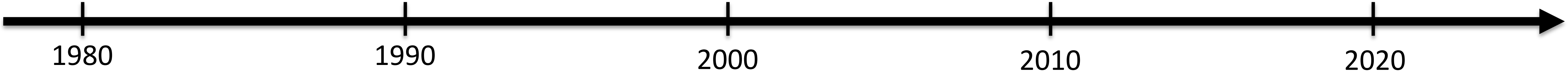
2003 Pfold 2011 ContextFold

2004 RNAstructure 2012 TORNADO

2006 CONTRAfold 2018 MXfold

1978 Nussinov algorithm 2019 LinearFold

1981 Mfold (Zuker algorithm) 1994 RNAfold 2009 CentroidFold 2021 MXfold2



1999 PKNOTS 2018 Knotty

2003 NUPACK 2019 SPOT-RNA

2005 HotKnots 2022 IPknot++

2010 HotKnots 2.0

2011 IPknot

Thermodynamic-based methods

Machine learning-based methods

Hybrid methods

Parameterization

- Thermodynamic-based methods
  - Determine free energy parameters by experiments (e.g., Turner1999, Turner2004)
  - Experimental errors are not negligible.
  - Too simplified models can only be constructed due to the limitations of experimental techniques.

- Machine learning-based methods
  - Rich-parameterized models can be constructed.
  - Potential risk of overfitting due to the inability to provide enough training data.

Potential risk of overfitting

- Fewer parameters can be determined by experiments for the thermodynamic models.
- There is a possibility of overfitting to the training dataset for machine learning-based models.

Comparison of different methods [Rivas *et al.*, 2012]

Method	#parameters	Parameterization	Benchmark: F	
			TestSetA	TestSetB
UNAfold [Markham *et al.*, 2008]	3,500	Thermodynamic	0.510	0.513
RNAfold [Lorenz *et al.*, 2011]	12,700	Thermodynamic	0.537	0.543
CONTRAfold [Do *et al.*, 2006]	300	Machine Learning	0.572	0.579
ContextFold [Zakov *et al.*, 2011]	205,000	Machine Learning	0.644	0.490

Rich parameters

High accuracy

Potential risk of overfitting

- Fewer parameters can be determined by experiments for the thermodynamic models.
- There is a possibility of overfitting to the training dataset for machine learning-based models.

Comparison of different methods [Rivas *et al.*, 2012]

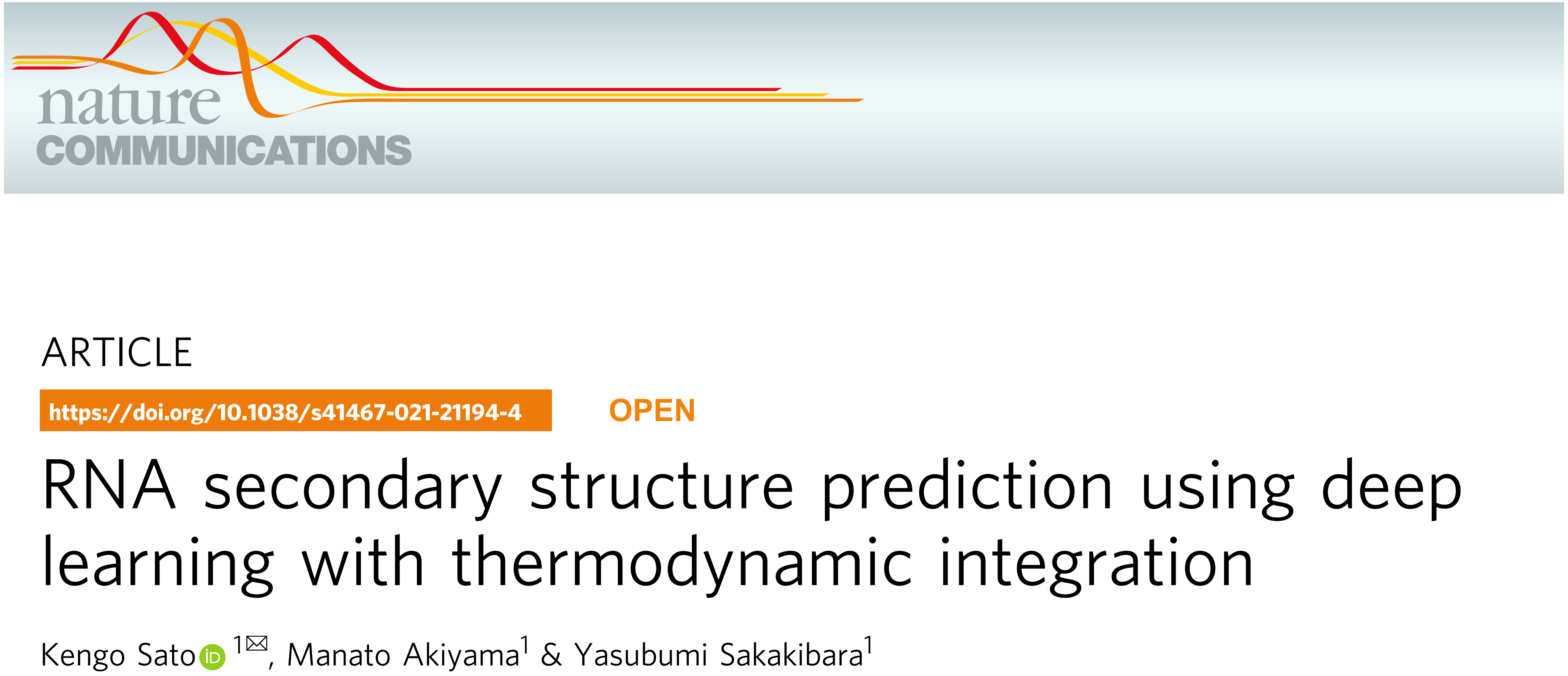
Method	#parameters	Parameterization	Benchmark: F	
			TestSetA	TestSetB
UNAfold [Markham *et al.*, 2008]	3,500	Thermodynamic	0.510	0.513
RNAfold [Lorenz *et al.*, 2011]	12,700	Thermodynamic	0.537	0.543
CONTRAfold [Do *et al.*, 2006]	300	Machine Learning	0.572	0.579
ContextFold [Zakov *et al.*, 2011]	205,000	Machine Learning	0.644	0.490

Rich parameters

High accuracy

overfitting

MXfold2



nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-21194-4 OPEN

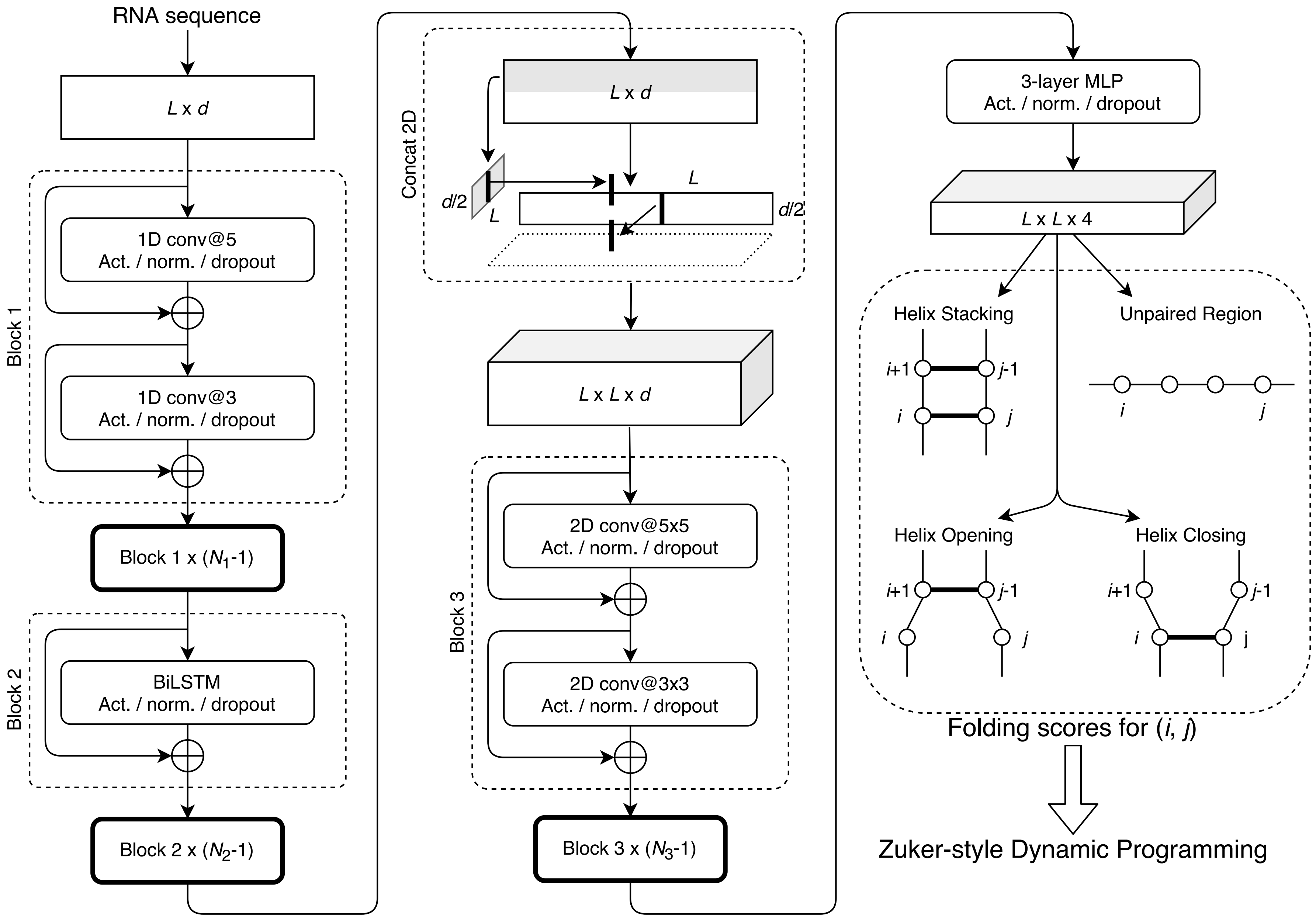
RNA secondary structure prediction using deep learning with thermodynamic integration

Kengo Sato[1✉], Manato Akiyama[1] & Yasubumi Sakakibara[1]

Our Approach

- Develop an algorithm that is robust against the overfitting using…
  - a scoring model that integrates machine learning and thermodynamic approaches,
  - the max-margin based training algorithm a.k.a. structured support vector machines (SSVM), and
  - thermodynamic regularization that ensures that folding scores and the calculated free energy are as close as possible.

Scoring Model

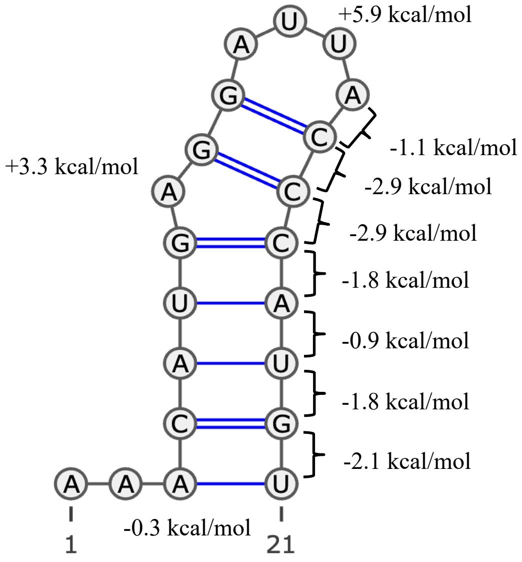


Scoring Model

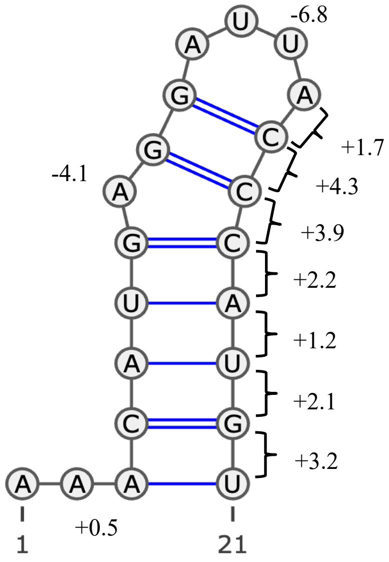
- Integrate the thermodynamic approach and the machine learning approach.

$$f(x, y) = f_T(x, y) + f_W(x, y)$$

thermodynamic



machine learning



x = AAACAUGAGGAUUACCCAUGU

y = ..(((((.((....)))))))

Training Algorithm

- To optimize the network parameters λ, we employ a max-margin based training algorithm a.k.a structured support vector machines (SSVM) [Tsochantaridis *et al.*, 2005].

Objective function

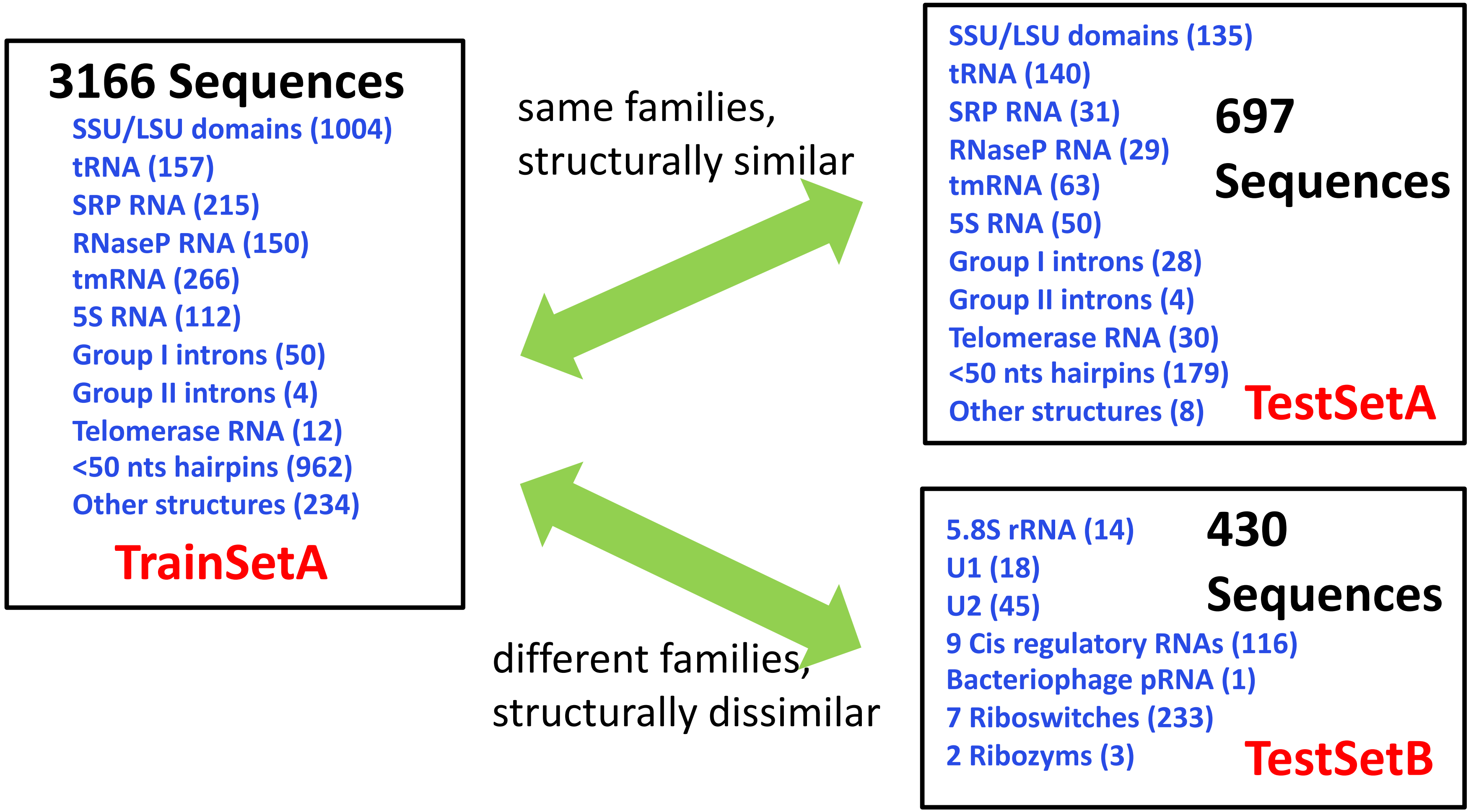
$$\mathcal{L}(\lambda) = \sum_{(x,y)\in\mathcal{D}} \left\{ \underset{\text{Loss term}}{\boxed{\left(\max_{\hat{y}\in\mathcal{S}(x)} [f(x,\hat{y}) + \Delta(y,\hat{y})] - f(x,y)\right)}} + C_1 \underset{\text{Thermodynamic regularization}}{\boxed{[f(x,y) - f_T(x,y)]^2}} + C_2\|\lambda\|_2 \right\},$$

$\Delta(y,\hat{y})$: margin term, $f_T(x,y)$: the free energy of the structure y of the sequence x

- Thermodynamic regularization prevents the folding score of the secondary structure from differing significantly from the free energy of the thermodynamic parameters.

Dataset I

- Assembled by [Rivas *et al*., 2012]



3166 Sequences

- SSU/LSU domains (1004)
- tRNA (157)
- SRP RNA (215)
- RNaseP RNA (150)
- tmRNA (266)
- 5S RNA (112)
- Group I introns (50)
- Group II introns (4)
- Telomerase RNA (12)
- <50 nts hairpins (962)
- Other structures (234)

TrainSetA

same families, structurally similar

- SSU/LSU domains (135)
- tRNA (140)
- SRP RNA (31)
- RNaseP RNA (29)
- tmRNA (63)
- 5S RNA (50)
- Group I introns (28)
- Group II introns (4)
- Telomerase RNA (30)
- <50 nts hairpins (179)
- Other structures (8)

697 Sequences

TestSetA

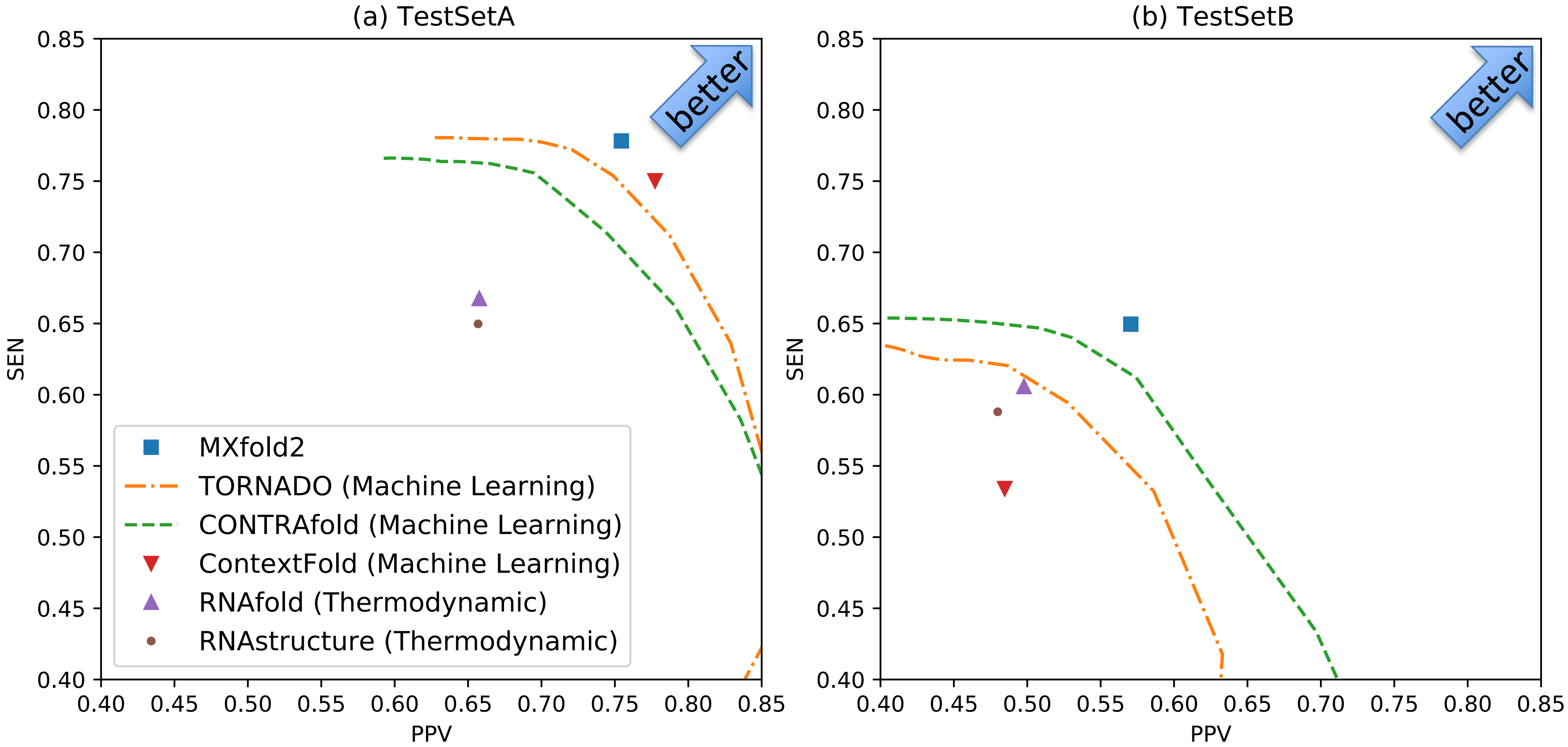
different families, structurally dissimilar

- 5.8S rRNA (14)
- U1 (18)
- U2 (45)
- 9 Cis regulatory RNAs (116)
- Bacteriophage pRNA (1)
- 7 Riboswitches (233)
- 2 Ribozyms (3)

430 Sequences

TestSetB

Comparison with competitive methods



$$PPV = \frac{TP}{TP + FP},\ SEN = \frac{TP}{TP + FN}$$

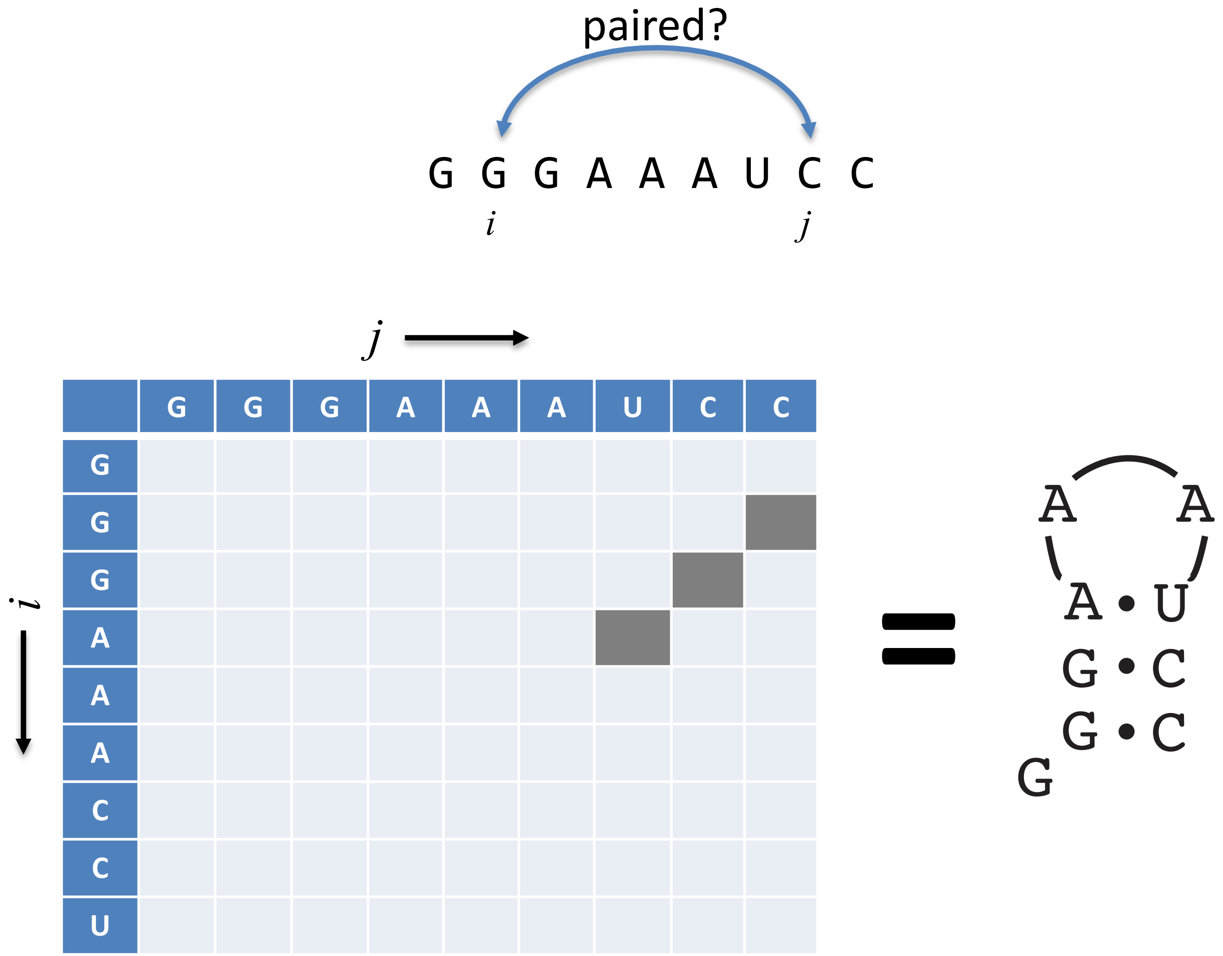
Correlation with free energy

- Dataset
  - T-full dataset [Andronescu *et al*., 2008], which contains sequence-structure-energy triplets

	PPV	SEN	F	RMSE	ρ
MXfold2	0.984	0.978	0.980	3.260	0.833
MXfold2 (w/o thermo. reg.)	0.980	0.972	0.973	3.607	0.538
CONTRAfold	0.963	0.639	0.643	5.781	0.736
RNAfold	0.979	0.964	0.963	2.868	0.909

Other DL-based methods

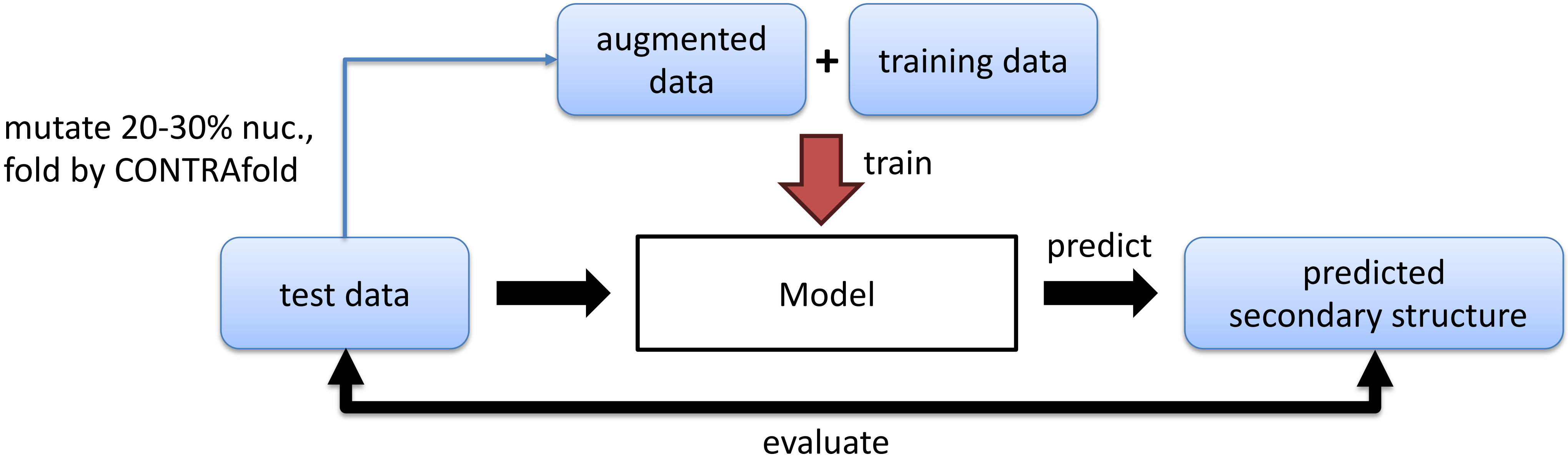
- Multiple binary classifiers for all (i, j) pairs
  - SPOT-RNA [Singh *et al.*, 2019] , E2Efold [Chen *et al.*, 2020], UFold [Fu *et al.*, 2022]



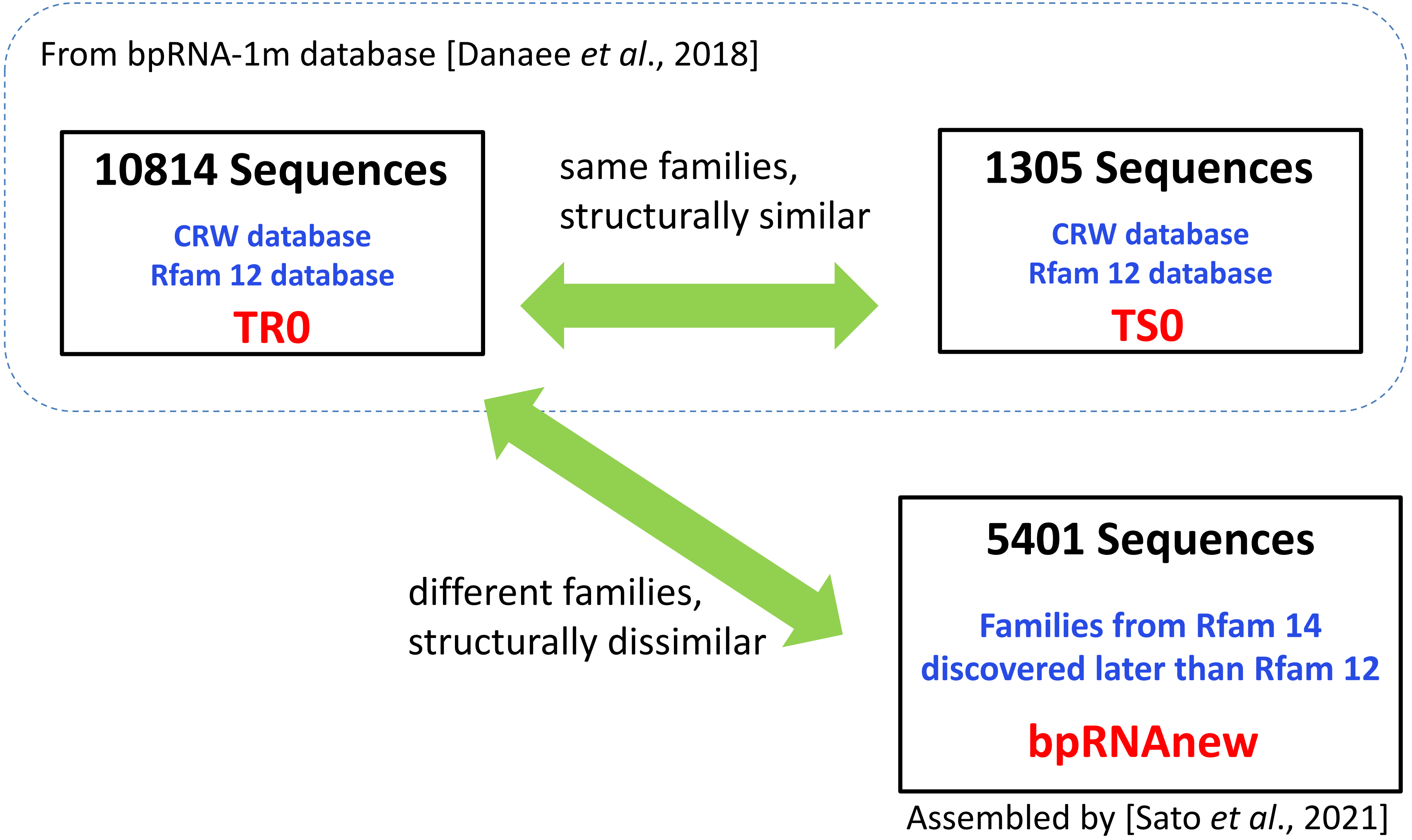
	G	G	G	A	A	A	U	C	C
G									
G									■
G								■	
A							■		
A									
A									
C									
C									
U									

Strategies for overfitting in other DL-based methods

- SPOT-RNA [Singh *et al*., 2019]
  - Ensemble of five different DL models
- E2Efold [Chen *et al*., 2020]
  - None
- UFold [Fu *et al*., 2022]
  - Data augmentation using **test data** mutated



Dataset II



From bpRNA-1m database [Danaee *et al*., 2018]

10814 Sequences

CRW database
Rfam 12 database

TR0

same families,
structurally similar

1305 Sequences

CRW database
Rfam 12 database

TS0

different families,
structurally dissimilar

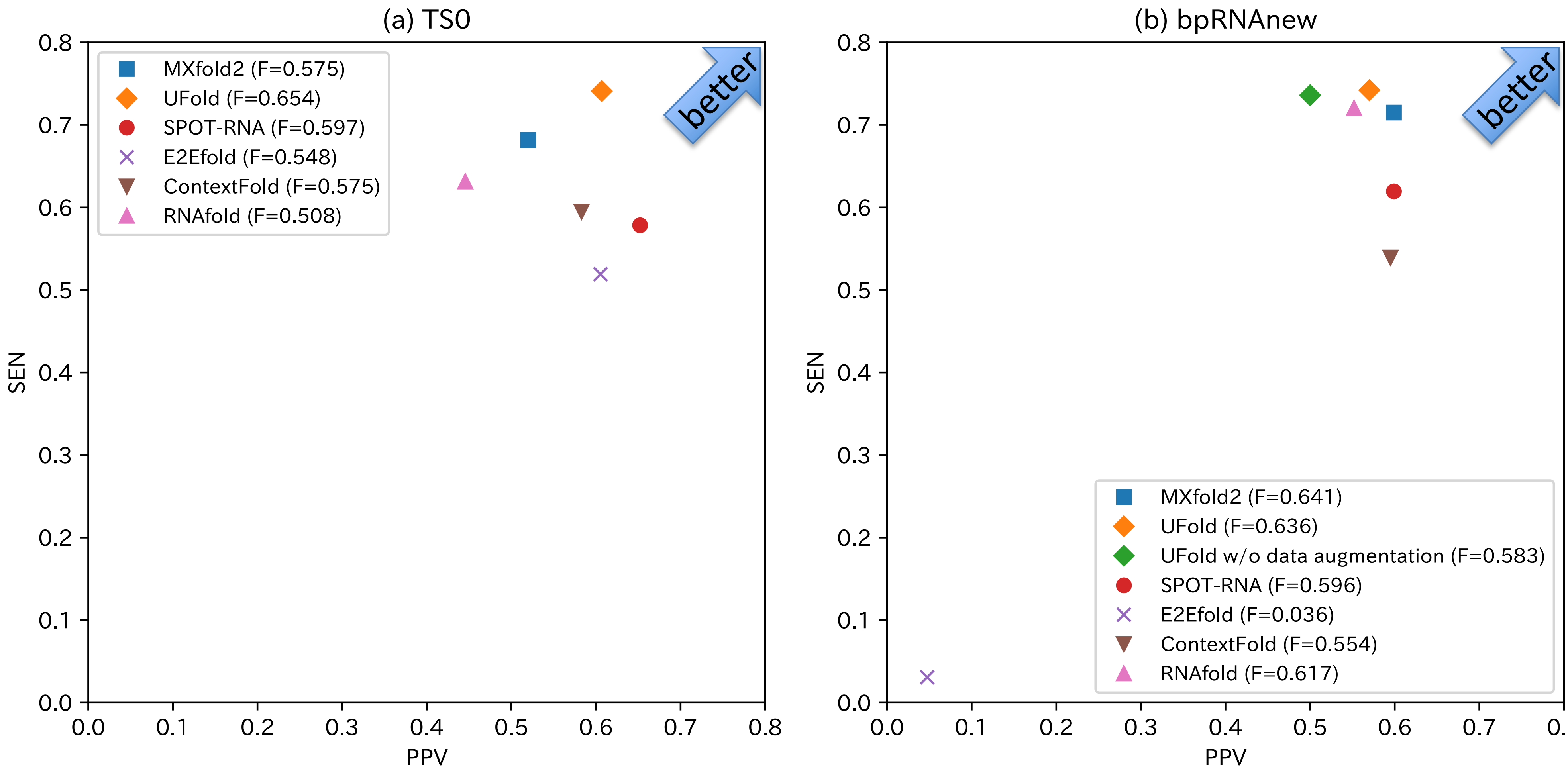
5401 Sequences

Families from Rfam 14
discovered later than Rfam 12

bpRNAnew

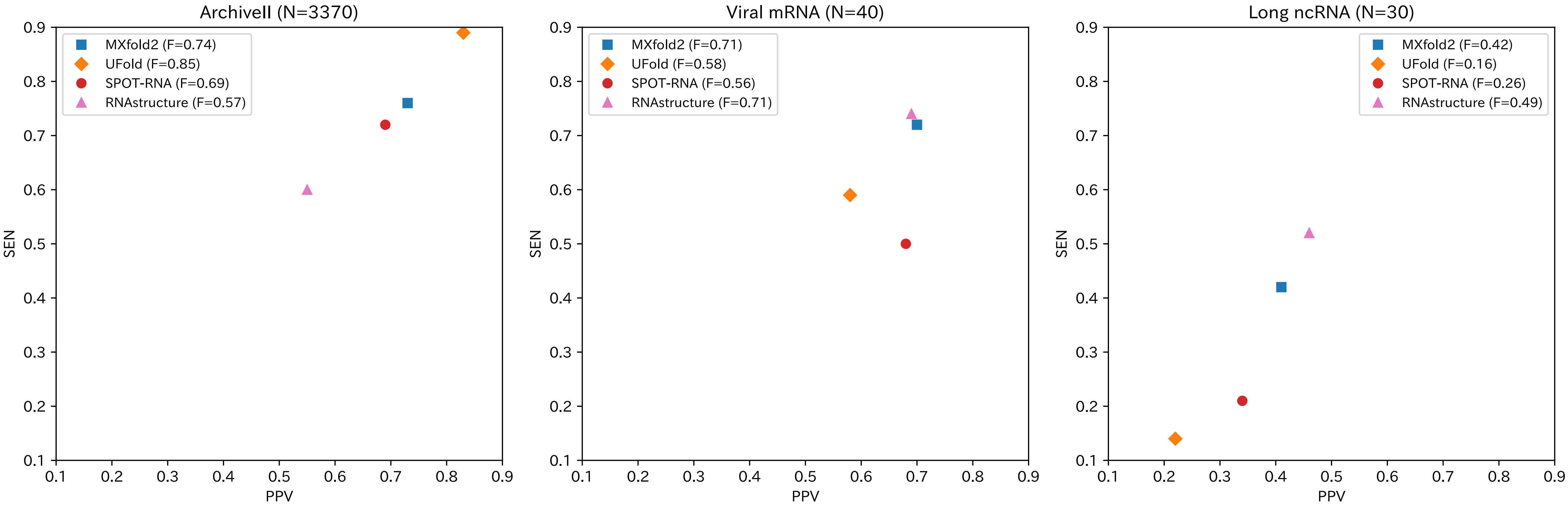
Assembled by [Sato *et al*., 2021]

Comparison with other DL-based methods



$$PPV = \frac{TP}{TP+FP}, \; SEN = \frac{TP}{TP+FN}, \; F = \frac{2 \times PPV \times SEN}{PPV + SEN}$$

Comparison with other DL-based methods on another study

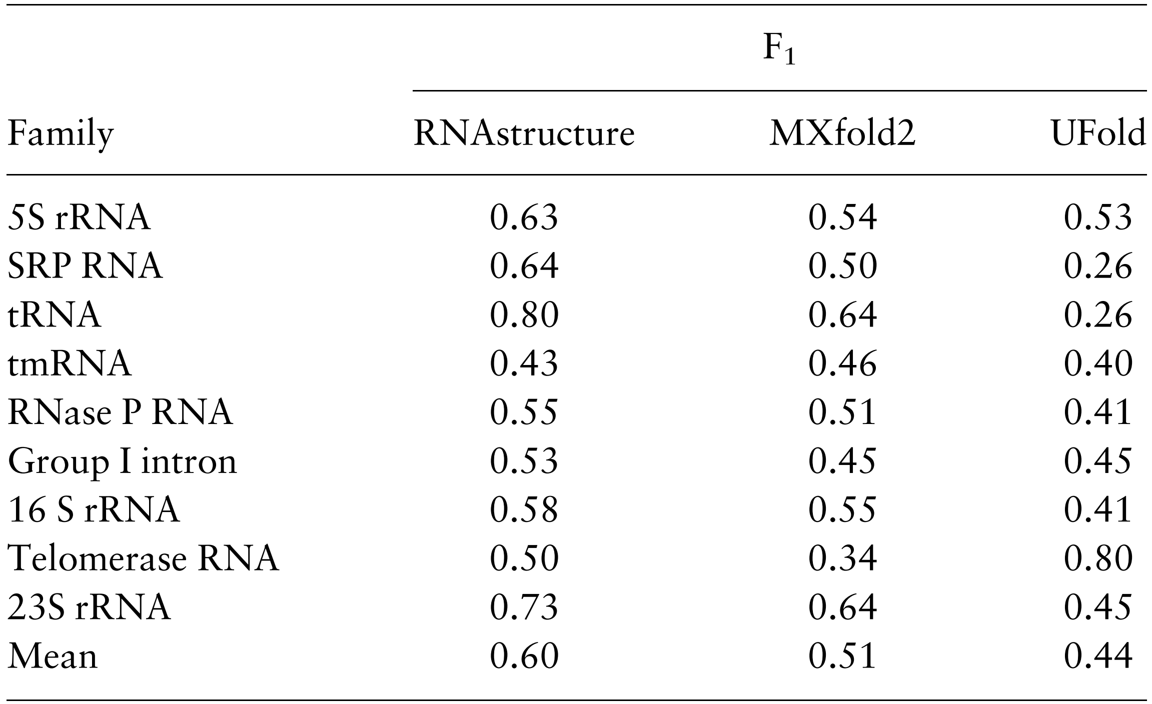


- UFold’s data augmentation is not likely to be helpful.

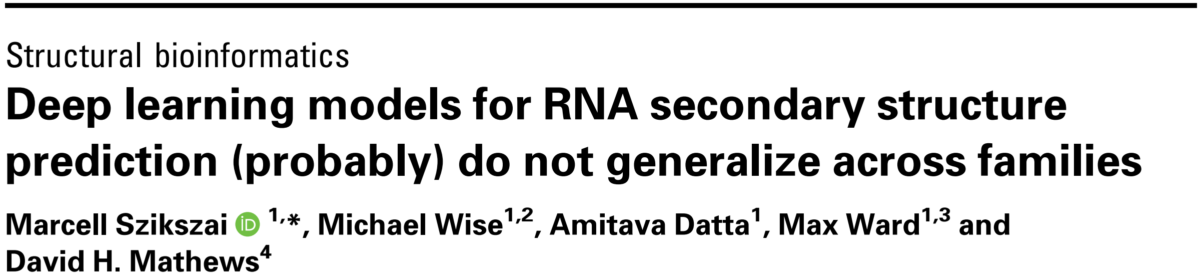
Values are taken from [de Lajarte *et al.*, 2024]

But not perfect

- Family-wise cross validation on Archive II dataset [Szikszai et al., 2022]



	F_1		
Family	RNAstructure	MXfold2	UFold
5S rRNA	0.63	0.54	0.53
SRP RNA	0.64	0.50	0.26
tRNA	0.80	0.64	0.26
tmRNA	0.43	0.46	0.40
RNase P RNA	0.55	0.51	0.41
Group I intron	0.53	0.45	0.45
16 S rRNA	0.58	0.55	0.41
Telomerase RNA	0.50	0.34	0.80
23S rRNA	0.73	0.64	0.45
Mean	0.60	0.51	0.44



Structural bioinformatics

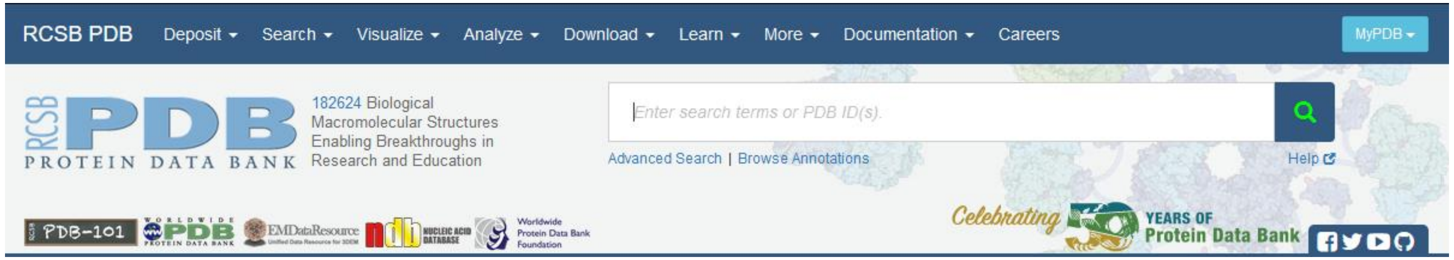
Deep learning models for RNA secondary structure prediction (probably) do not generalize across families

Marcell Szikszai [1,*], Michael Wise[1,2], Amitava Datta[1], Max Ward[1,3] and David H. Mathews[4]

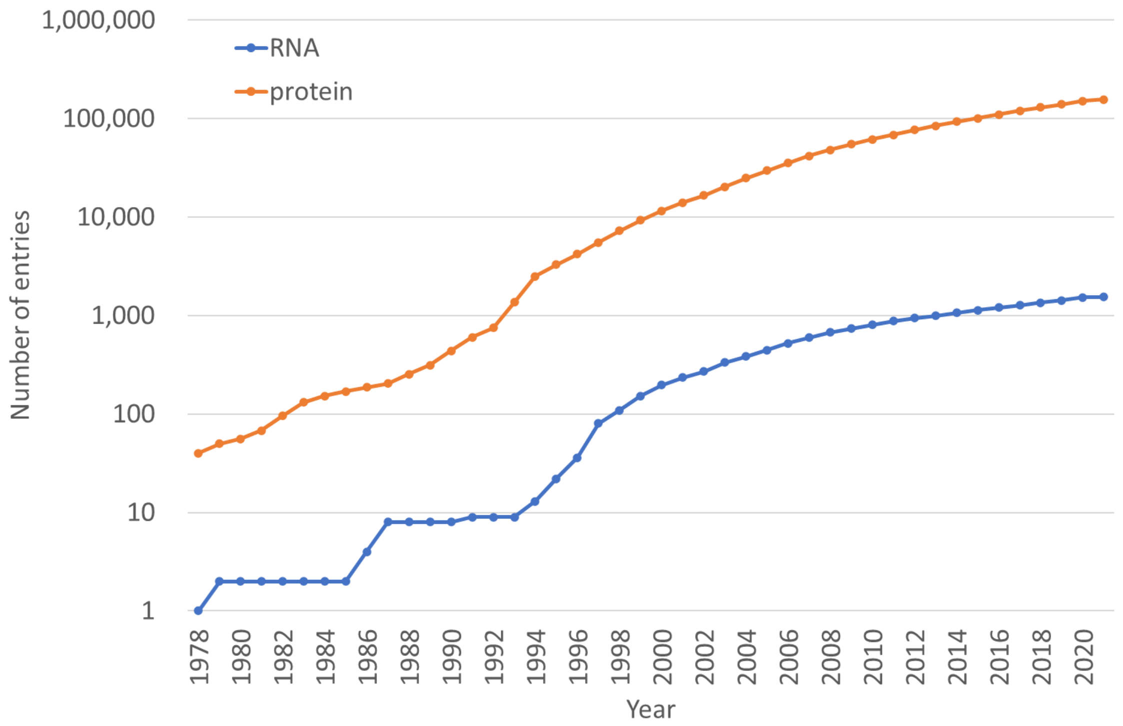
Table of Contents

- Overview of RNA secondary structure prediction
  - Architecture
    - Nussinov algorithm, Nearest neighbor model
  - Inference
    - MFE, MEA
  - Parametrization
    - Machine learning, Deep learning
- Future direction
  - Chemical probing
  - RNA modification
  - Pseudoknots

The number of known structures of proteins and RNA


RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
Documentation
Careers
MyPDB
RCSB PDB
PROTEIN DATA BANK
Enter search terms or PDB ID(s).
Advanced Search | Browse Annotations
Help
PDB-101
WORLDWIDE PDB PROTEIN DATA BANK
EMDataResource
ndb NUCLEIC ACID DATABASE
Worldwide Protein Data Bank Foundation
Celebrating 50 YEARS OF Protein Data Bank

182624 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education



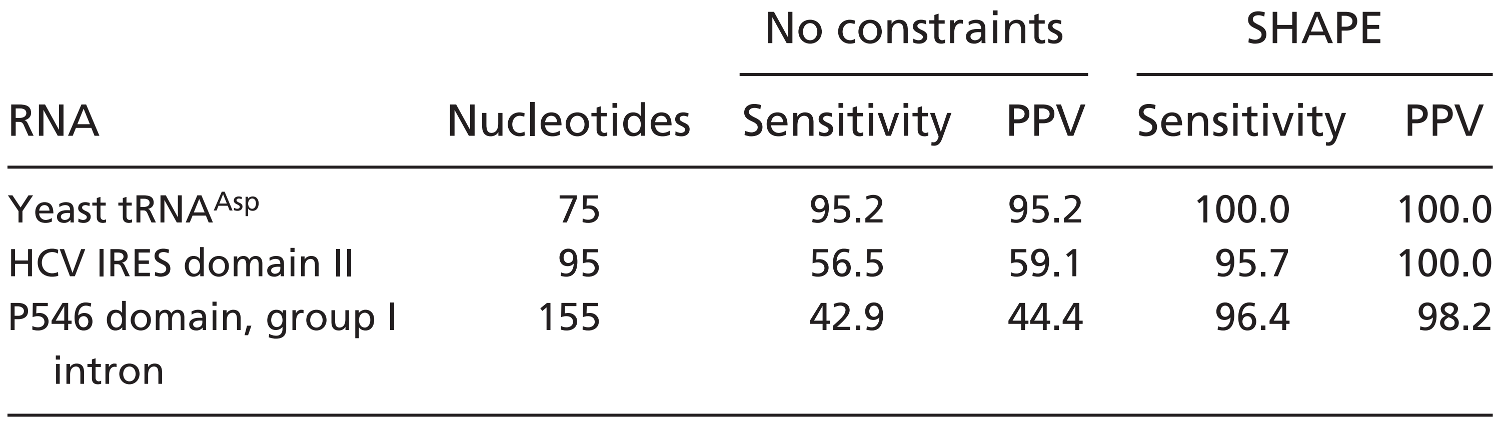
- The number of known RNA structures is 100 times less than that of proteins.

SHAPE-directed folding

- RNAstructure [Deigan et al. *PNAS.* 2009]
  - adds pseudo-energy for i-th base for base-pairing:

$$\Delta G_{\mathrm{SHAPE}}(i) = m\ \ln[\mathrm{SHAPE\ reactivity}(i) + 1] + b$$

  - shows significant improvement in prediction accuracy

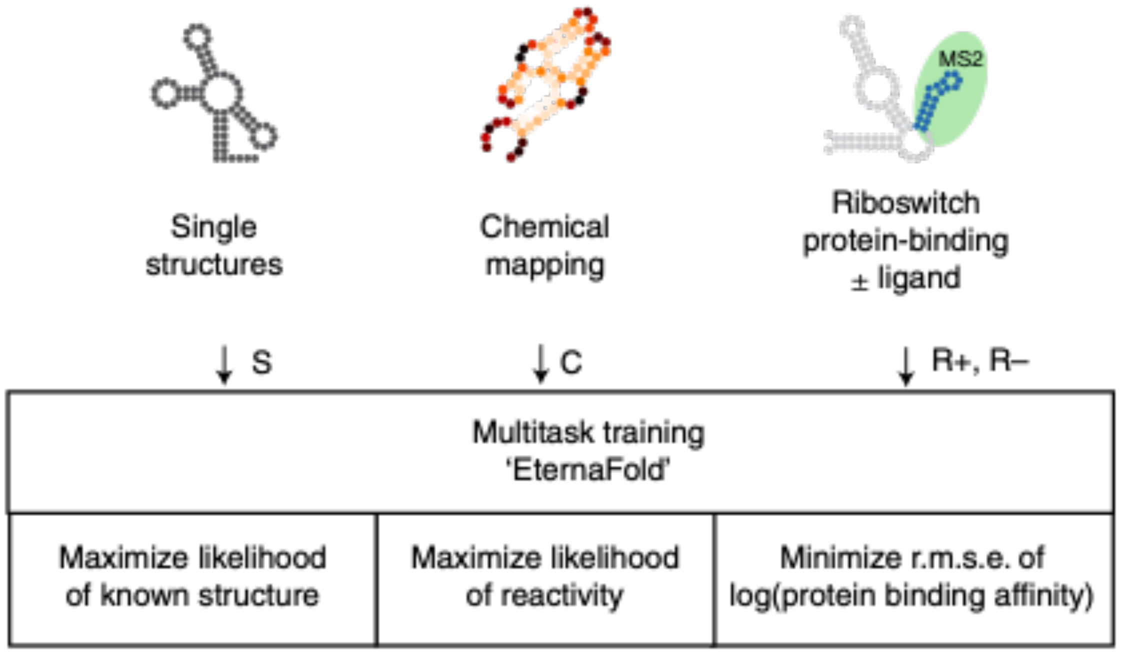


		No constraints		SHAPE	
RNA	Nucleotides	Sensitivity	PPV	Sensitivity	PPV
Yeast tRNA^{Asp}	75	95.2	95.2	100.0	100.0
HCV IRES domain II	95	56.5	59.1	95.7	100.0
P546 domain, group I intron	155	42.9	44.4	96.4	98.2

- We implemented SHAPE-directed folding in MXfold2 following this same approach.

Training from chemical probing data

- EternaFold [Wayment-Steele et al. *Nat. Methods.* 2022]
  - Multi-task learning based on the CONTRAfold model



Multitask training 'EternaFold'

Maximize likelihood of known structure	Maximize likelihood of reactivity	Minimize r.m.s.e. of log(protein binding affinity)

- We implemented SHAPE-directed "training" in MXfold2 while avoiding the computation of the partition function.

Training from chemical probing data

- Key idea:
  - SHAPE-directed folding make a perfect prediction, so use it as the reference structure.

- Update the model parameter θ for a sequence x with chemical probing data
  1. predict secondary structure y of x using SHAPE-directed folding with parameter θ
  2. predict secondary structure $\hat{y}$ of x using normal folding with parameter θ
  3. update parameter: $\theta \leftarrow \theta - \eta \nabla_{\theta} loss(y, \hat{y})$

Training from chemical probing data

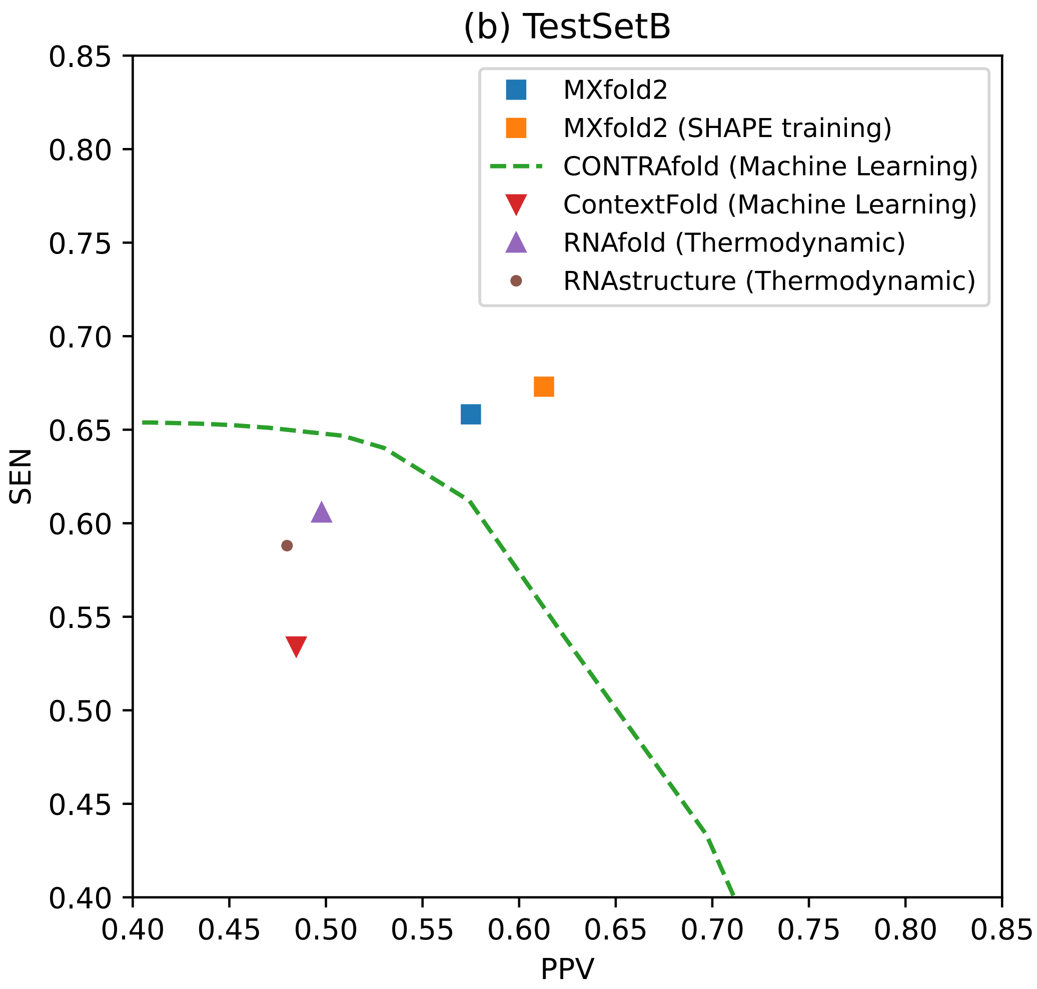
- Training data
  - MXfold2

    TrainSetA only

  - MXfold2 (SHAPE-directed training)

    TrainSetA
    +
    simulated SHAPE reactivity of TrainSetB

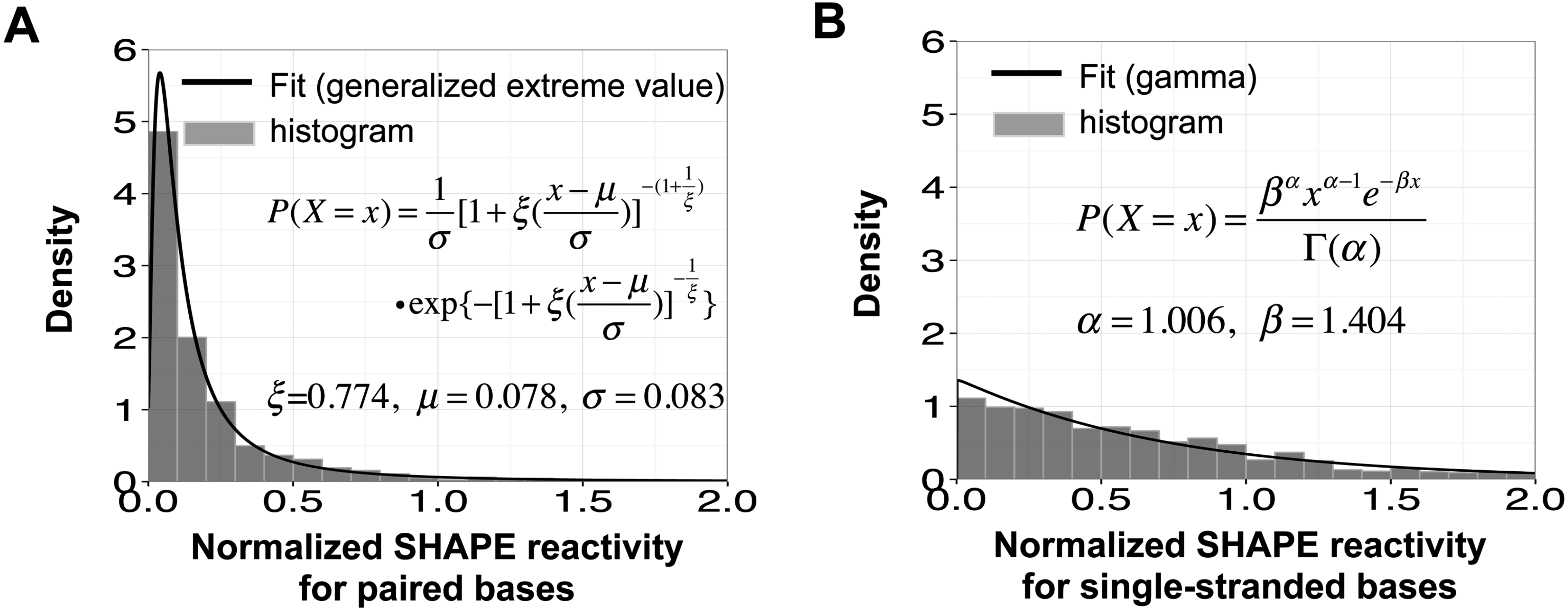
- This enables training on sequences for which few training data have been available so far (e.g., lncRNA, mRNA).



[Sato et al., in prep]

Distribution of SHAPE reactivity

- [Wu et al. *NAR*. 2015]



Large scale training dataset including chemical probing

- Performance will be improved by significantly scaling up both the quality and quantity of training data.

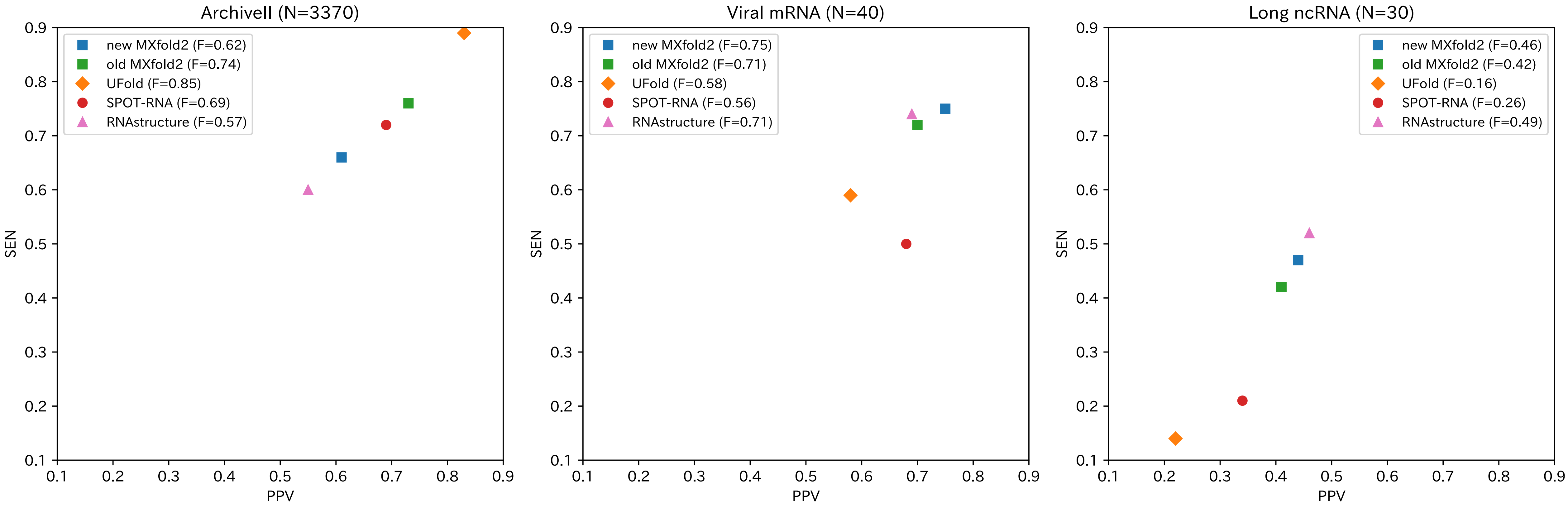
<u>Current</u>

- 4,270 sequences with complete SS
  - TrainSetA+B, TestSetA+B [Rivas et al., 2012]

<u>New</u>

- 19,266 sequences with complete SS
  - TrainSetA+B, TestSetA+B [Rivas et al., 2012]
  - bpRNA-1m [Danaee et al., 2018]
  - bpRNAnew [Sato et al., 2021]
- 48,614 sequences with chemical reactivity
  - 1,456 human mRNA 3’ end, 1,098 human pri-miRNA [de Lajarte et al., 2024]
  - 46,060 Ribonanza data [He et al., 2024]

Large scale training dataset including chemical probing



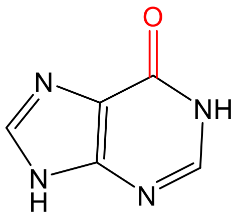
Values are taken from [de Lajarte *et al.*, 2024] [Sato *et al.*, in prep]

Table of Contents

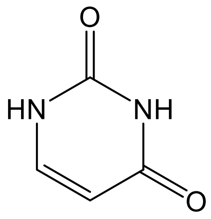
- Overview of RNA secondary structure prediction
  - Architecture
    - Nussinov algorithm, Nearest neighbor model
  - Inference
    - MFE, MEA
  - Parametrization
    - Machine learning, Deep learning
- Future direction
  - Chemical probing
  - RNA modification
  - Pseudoknots

RNA modifications

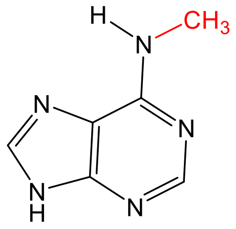
- play important roles in biological processes such as gene regulation [Camper *et al.*, 1984],
- are known to exist >170 types [Nombela *et al.*, 2021], and



Inosine (I)



Pseudouridine (ψ)



N6-methyladenosine (m^{6}A)

- alter RNA secondary structures [Alseth *et al.*, 2014].

However, few methods are available for predicting RNA secondary structures that consider RNA modifications.

Methods: Representation for RNA modifications

One-hot encoding

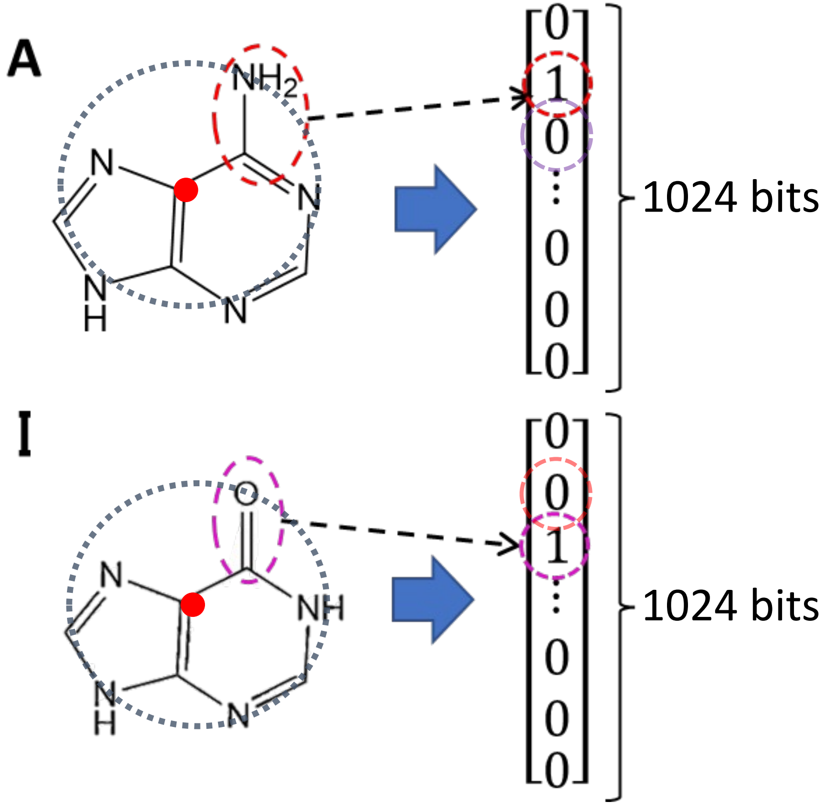
- Takes characters (RNAs) as inputs, and
- Identifies input characters by a set bit.

	4 bits
A →	(1 0 0 0)
U →	(0 1 0 0)
G →	(0 0 1 0)
C →	(0 0 0 1)

	7 bits
A →	(1 0 0 0 0 0 0)
U →	(0 1 0 0 0 0 0)
G →	(0 0 1 0 0 0 0)
C →	(0 0 0 1 0 0 0)
I →	(0 0 0 0 1 0 0)
Ψ →	(0 0 0 0 0 1 0)
m^6A →	(0 0 0 0 0 0 1)

Fingerprint encoding
ECFP (Extended-Connectivity Fingerprint)

- Takes chemical structures as inputs, and
- Represents the presence or absence of substructures in 1024 bits.



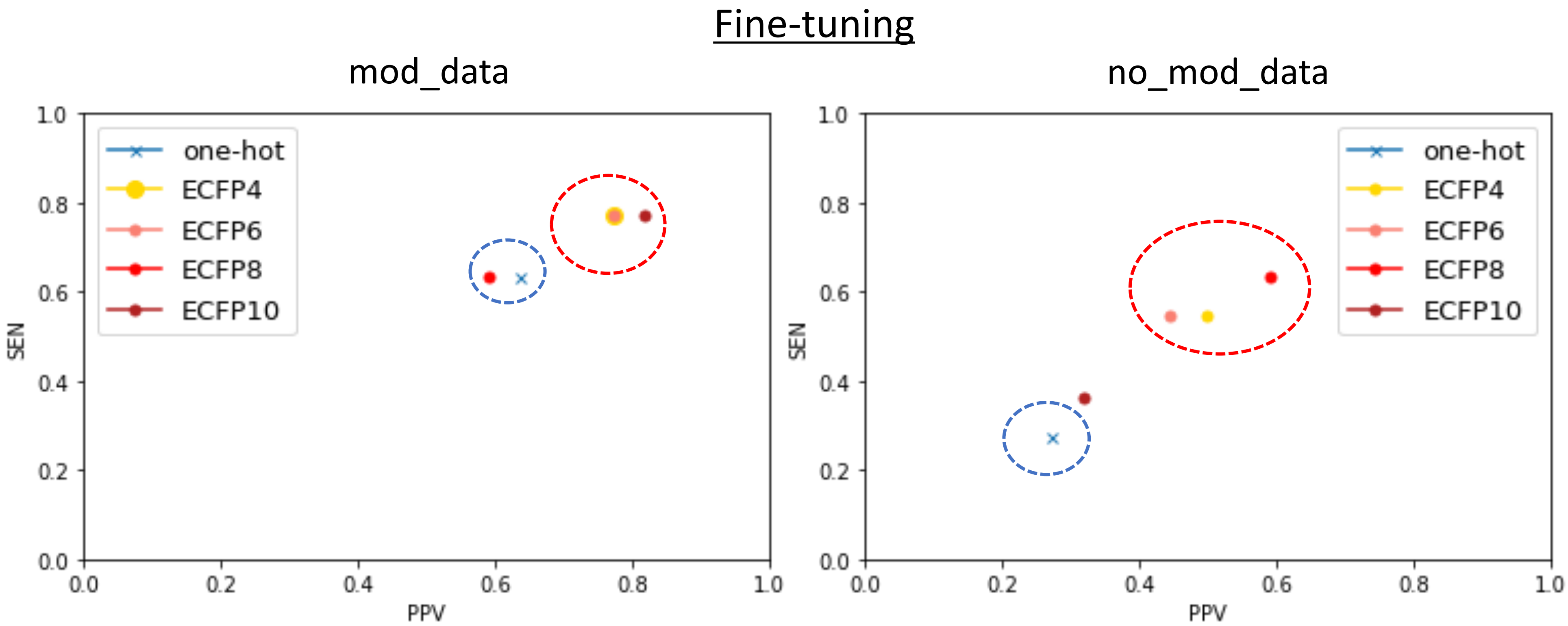
Datasets

	# of seqs.	Modified bases (%)				
		I	ψ	m^6A		
TrainSetA[※1]	3166	0.0	0.0	0.0	RNA seqs without modifications	Pre-training
mod_data[※2]	218	0.16	8.8	0.26	tRNA seqs with modifications	Fine-tuning
no_mod_data[※2]	218	0.0	0.0	0.0	same seqs as mod_data, but no mods	Fine-tuning
pdb_data[※3]	11	1.7	18.0	0.0	tRNA seqs with modifications	Evaluation

[※1] Rivas *et al,* 2012, [※2] Boccaletto *et al.,* 2018, [※3] Lorenz *et al*., 2017, Helm *et al*., 2006 , Guy *et al*., 2014 , Bilbille *et al*., 2011, Swinehart *et al.,* 2020, Keller *et al*., 1999, Jank *et al*., 1977, Kulinska *et al*., 1974, Hayase *et al*., 1974, Tinse *et al*., 2000

Results

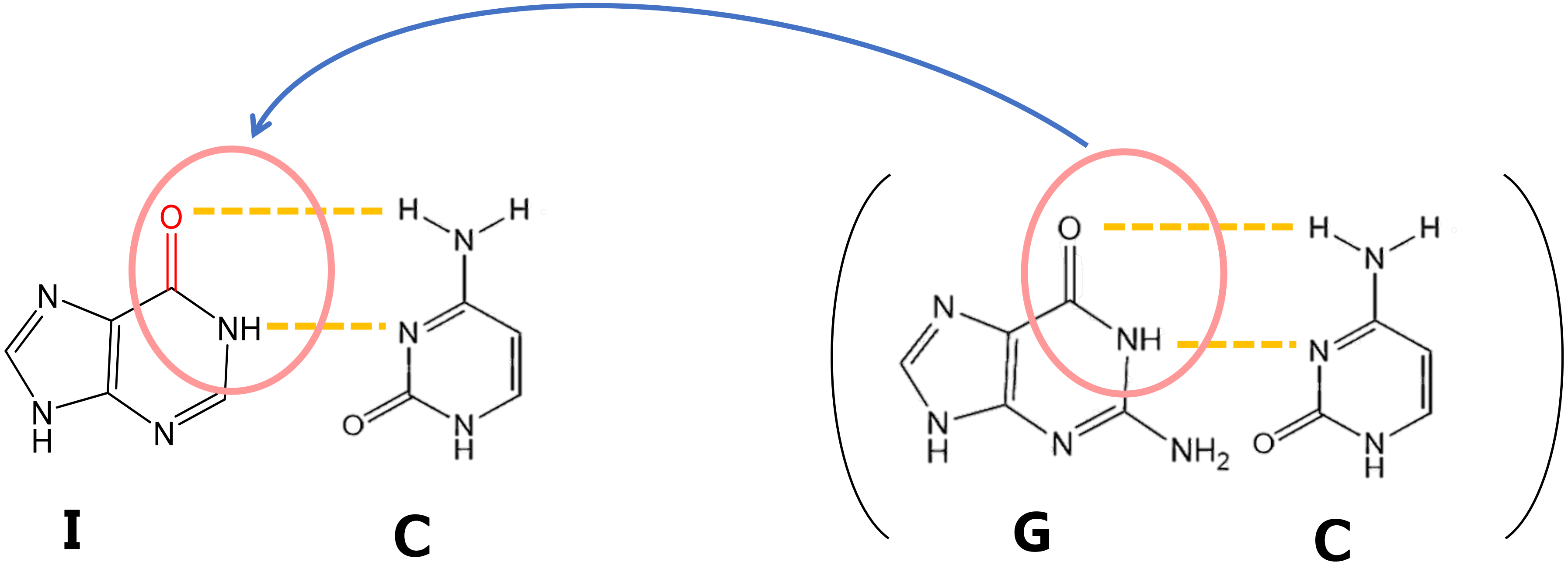
- Pre-train with **TrainSetA,** evaluate positions at modified bases on **pdb_data**



- Fingerprint encoding tends to be more accurate for modified bases than one-hot encoding.

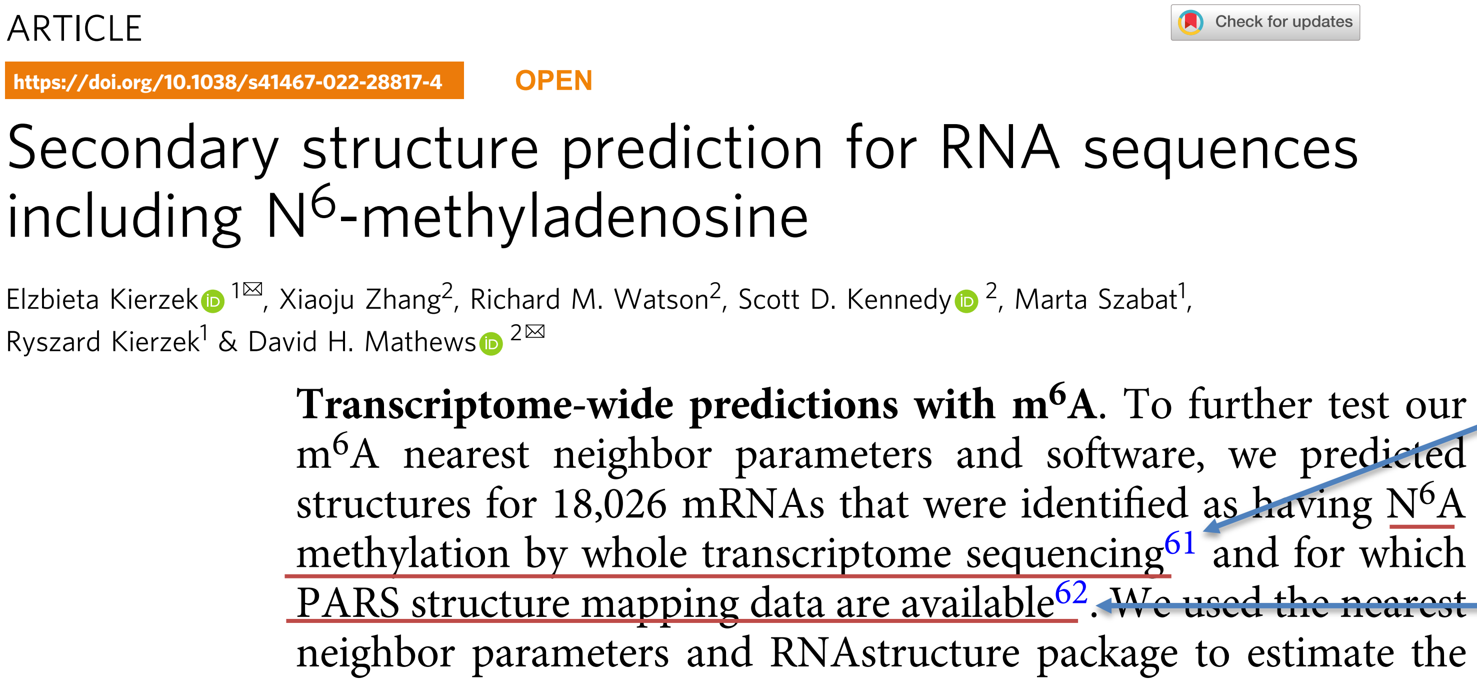
Interpretation of fingerprint encoding

- The fingerprint encoding can share the results of the training for common bits.



How to prepare sequence data with structures including modified bases?

- Few sequence data are available that contain modified bases with complete secondary structures.
- Chemical probing data with modified bases has also never been available.
- We plan to combine experimental data in different experiments as in:



ARTICLE

Check for updates

https://doi.org/10.1038/s41467-022-28817-4 OPEN

Secondary structure prediction for RNA sequences including N^6-methyladenosine

Elzbieta Kierzek[1✉], Xiaoju Zhang[2], Richard M. Watson[2], Scott D. Kennedy[2], Marta Szabat[1], Ryszard Kierzek[1] & David H. Mathews[2✉]

Transcriptome-wide predictions with m^{6}A. To further test our m^{6}A nearest neighbor parameters and software, we predicted structures for 18,026 mRNAs that were identified as having N^{6}A methylation by whole transcriptome sequencing[61] and for which PARS structure mapping data are available[62]. We used the nearest neighbor parameters and RNAstructure package to estimate the

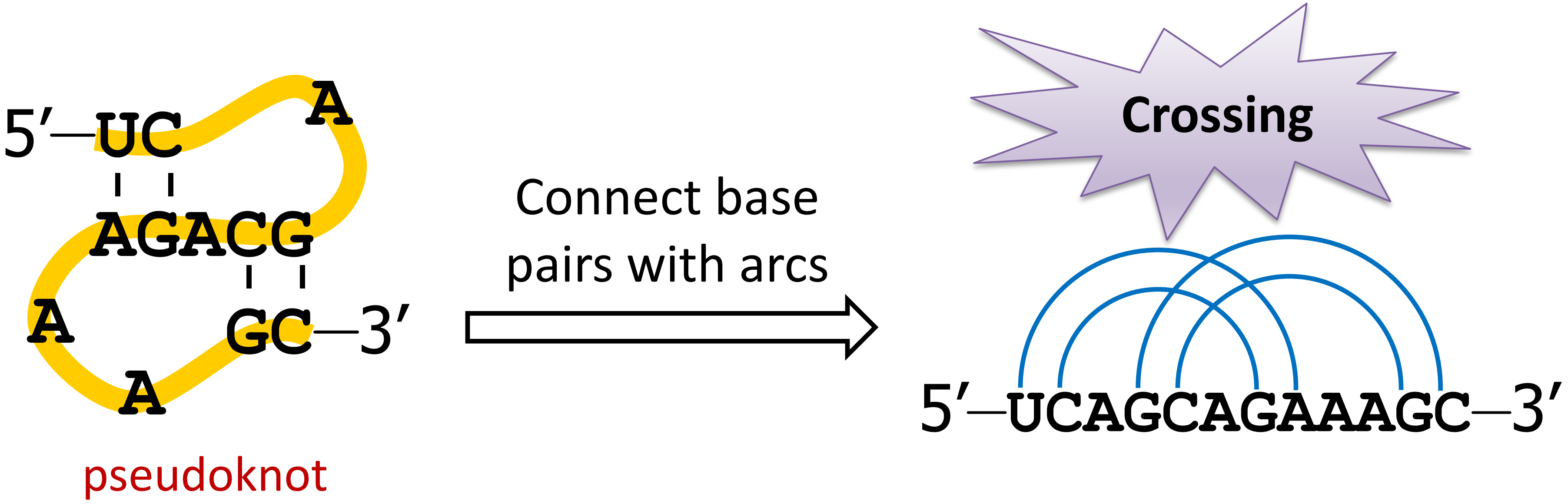
Human embryonic kidney 293T cells

Human lymphoblastoid cell lines

Table of Contents

- Overview of RNA secondary structure prediction
  - Architecture
    - Nussinov algorithm, Nearest neighbor model
  - Inference
    - MFE, MEA
  - Parametrization
    - Machine learning, Deep learning
- Future direction
  - Chemical probing
  - RNA modification
  - Pseudoknots

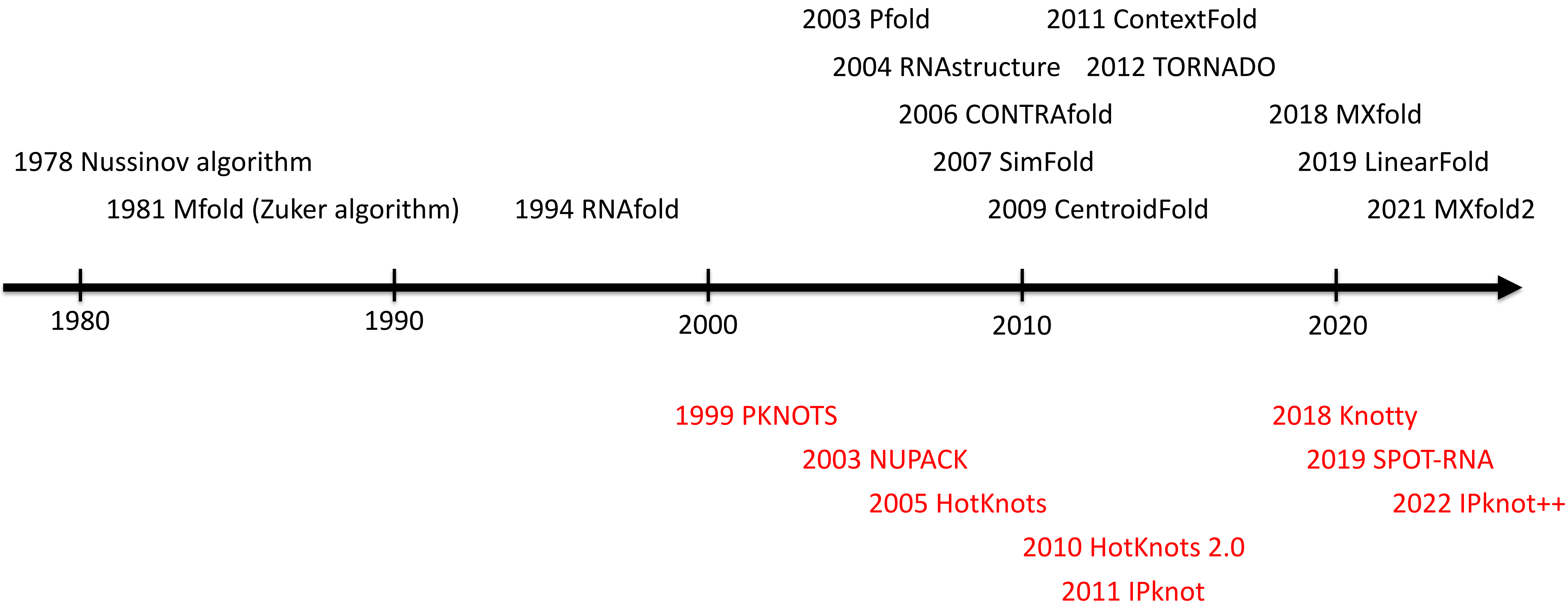
RNA pseudoknotted secondary structure



- Pseudoknots play several roles in RNA functions
  - Regulation of translation & splicing, etc.
- Pseudoknots assist the overall 3D folding

→ Pseudoknots should be considered for structural analysis

Prediction of RNA secondary structures



2003 Pfold

2011 ContextFold

2004 RNAstructure

2012 TORNADO

2006 CONTRAfold

2018 MXfold

1978 Nussinov algorithm

2007 SimFold

2019 LinearFold

1981 Mfold (Zuker algorithm)

1994 RNAfold

2009 CentroidFold

2021 MXfold2

1999 PKNOTS

2018 Knotty

2003 NUPACK

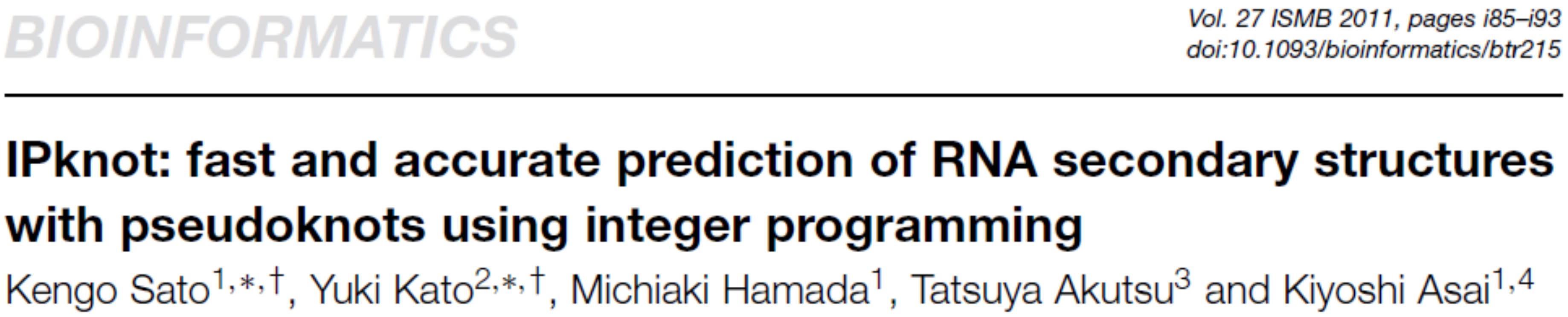
2019 SPOT-RNA

2005 HotKnots

2022 IPknot++

2010 HotKnots 2.0

2011 IPknot

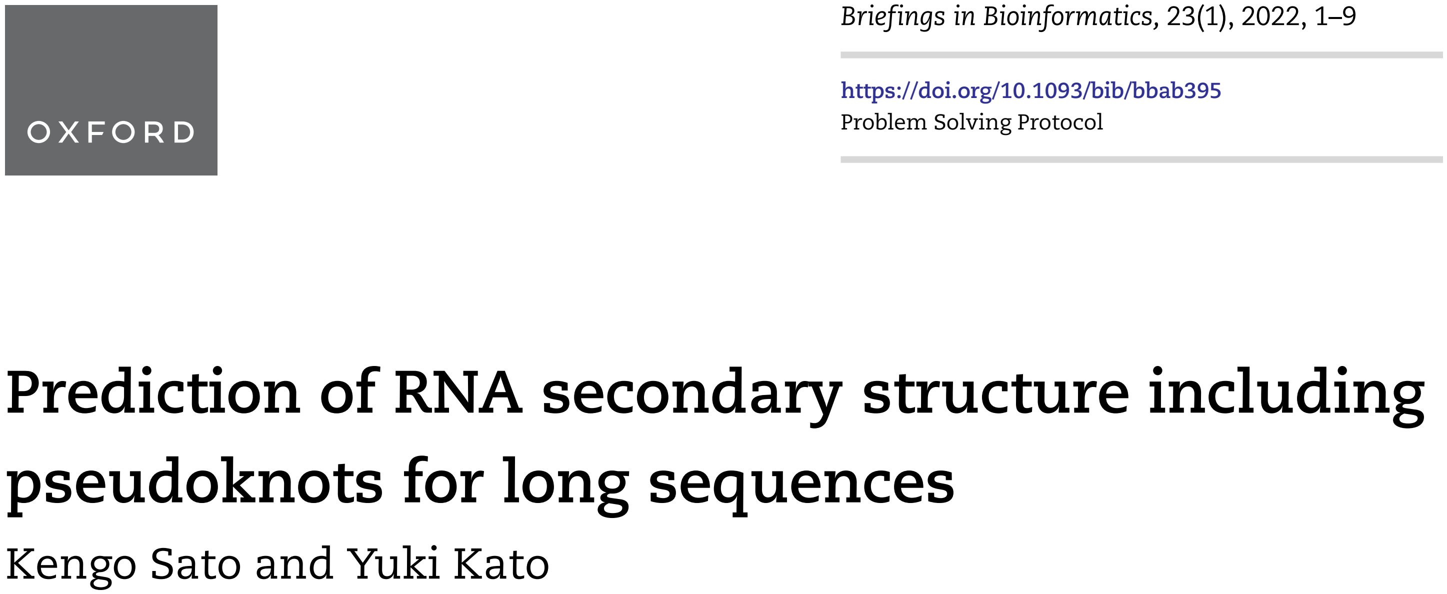


BIOINFORMATICS

Vol. 27 ISMB 2011, pages i85–i93
doi:10.1093/bioinformatics/btr215

IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming

Kengo Sato[1,*,†], Yuki Kato[2,*,†], Michiaki Hamada[1], Tatsuya Akutsu[3] and Kiyoshi Asai[1,4]



OXFORD

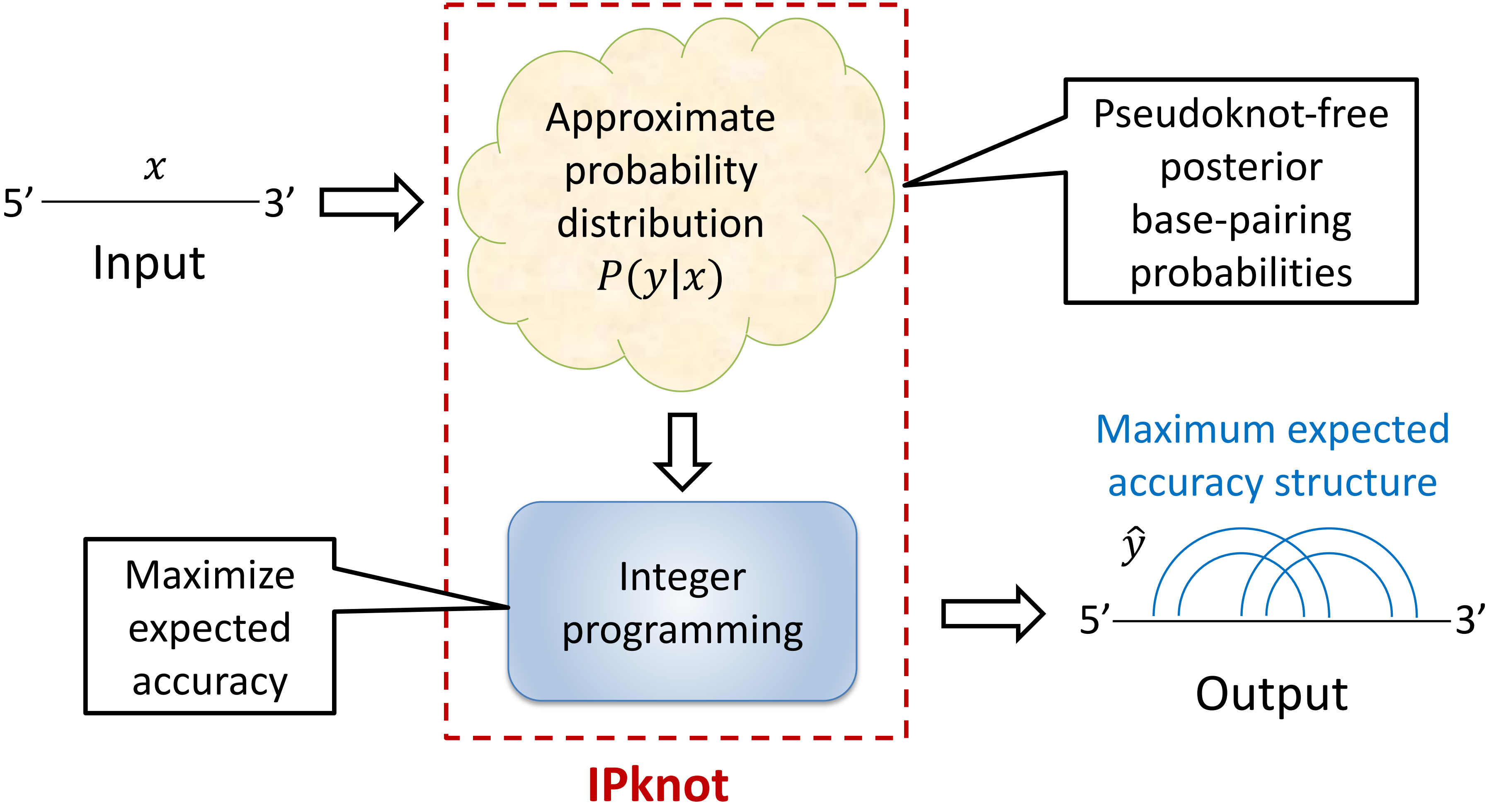
Briefings in Bioinformatics, 23(1), 2022, 1–9

https://doi.org/10.1093/bib/bbab395
Problem Solving Protocol

Prediction of RNA secondary structure including pseudoknots for long sequences

Kengo Sato and Yuki Kato

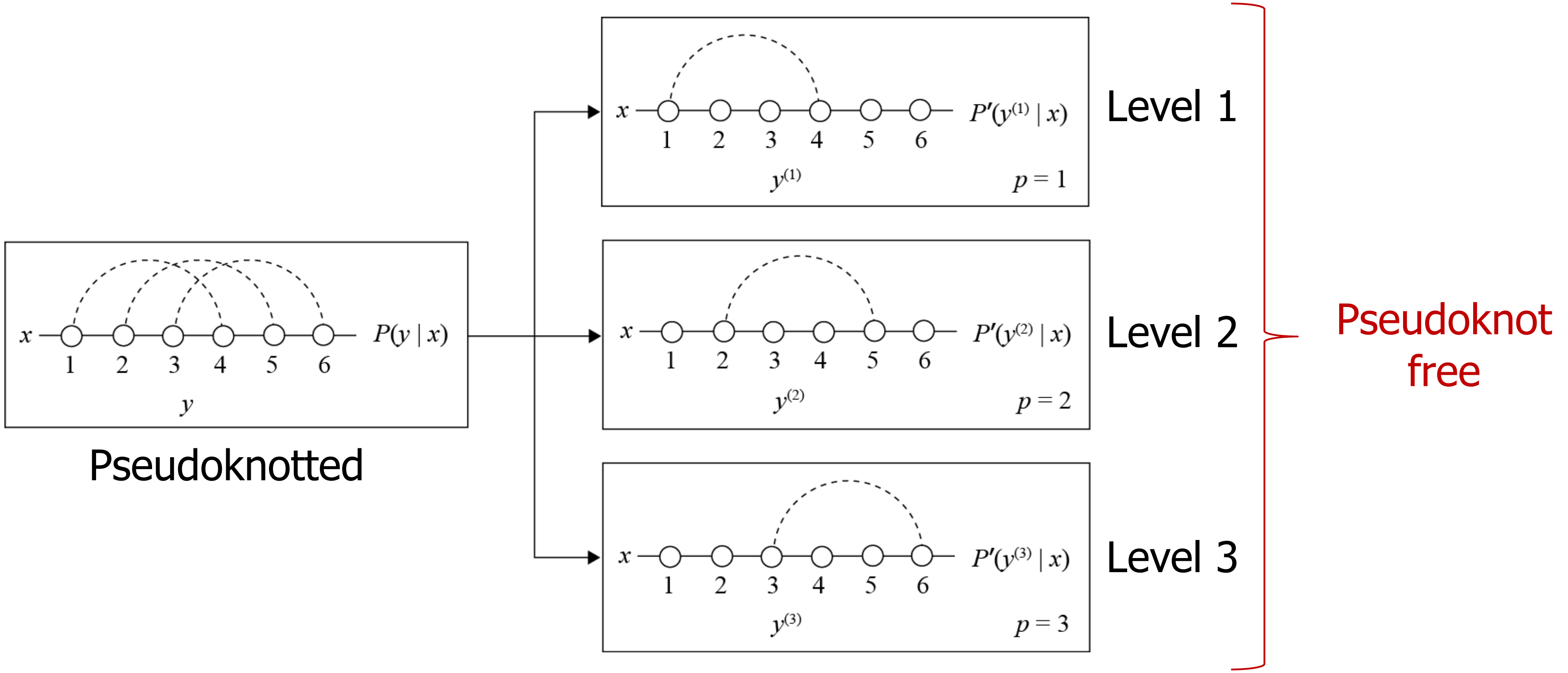
IPknot: Integer Programming-based prediction of RNA pseudoKNOTs


x
5'
3'
Input
Approximate probability distribution $P(y|x)$
Pseudoknot-free posterior base-pairing probabilities
Integer programming
Maximize expected accuracy
Maximum expected accuracy structure
$\hat{y}$
5'
3'
Output
IPknot

Approximate probability distribution

- Approximate a probability distribution over pseudoknotted structures by its factorization

$$P(y \mid x) \simeq \prod_{1 \leq p \leq m} P'(y^{(p)} \mid x)$$



Objective function (expected accuracy)

$$\text{maximize} \quad \sum_{1\le p\le m} \alpha^{(p)} \sum_{i<j} \Big[\underbrace{(\gamma^{(p)}+1)p_{ij} - 1}_{\text{To be positive}}\Big] \underbrace{\hat{y}_{ij}^{(p)}}_{\text{Predicted base pair}} + C \qquad (*)$$

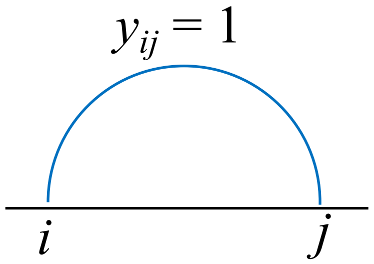
- Consider only base pairs whose pairing probabilities are larger than **thresholds**

→Threshold cut

Find $y = (y^{(1)},\ldots, y^{(m)})$ that maximizes (*)

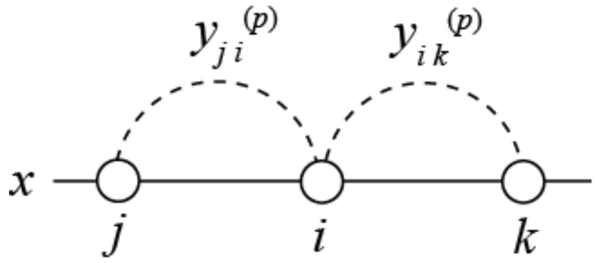
$$\left\{\begin{array}{l} \bullet\ \ y_{ij}^{(1)} \text{ such that } \ p_{ij} > \theta^{(1)} = 1/(\gamma^{(1)}+1) \\ \qquad \vdots \\ \bullet\ \ y_{ij}^{(m)} \text{ such that } \ p_{ij} > \theta^{(m)} = 1/(\gamma^{(m)}+1) \end{array}\right.$$

Thresholds

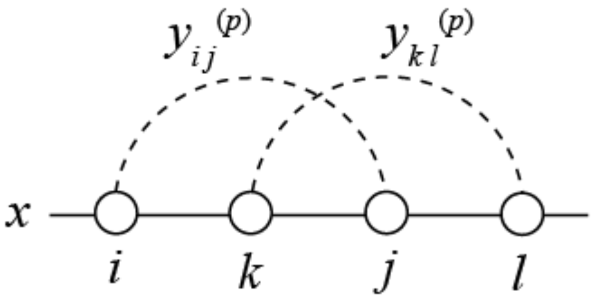


Constraints

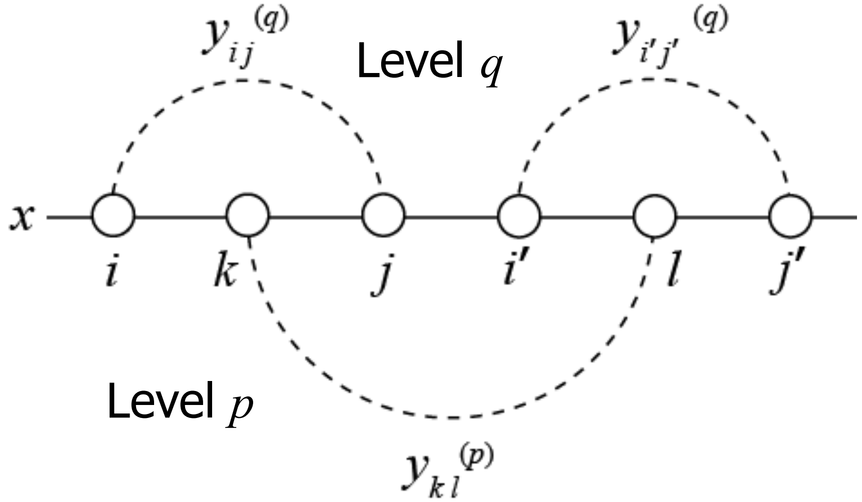
- The following hold for all levels p $(1 \leq p \leq m)$ and q $(< p)$



Each base can be paired with at most one base

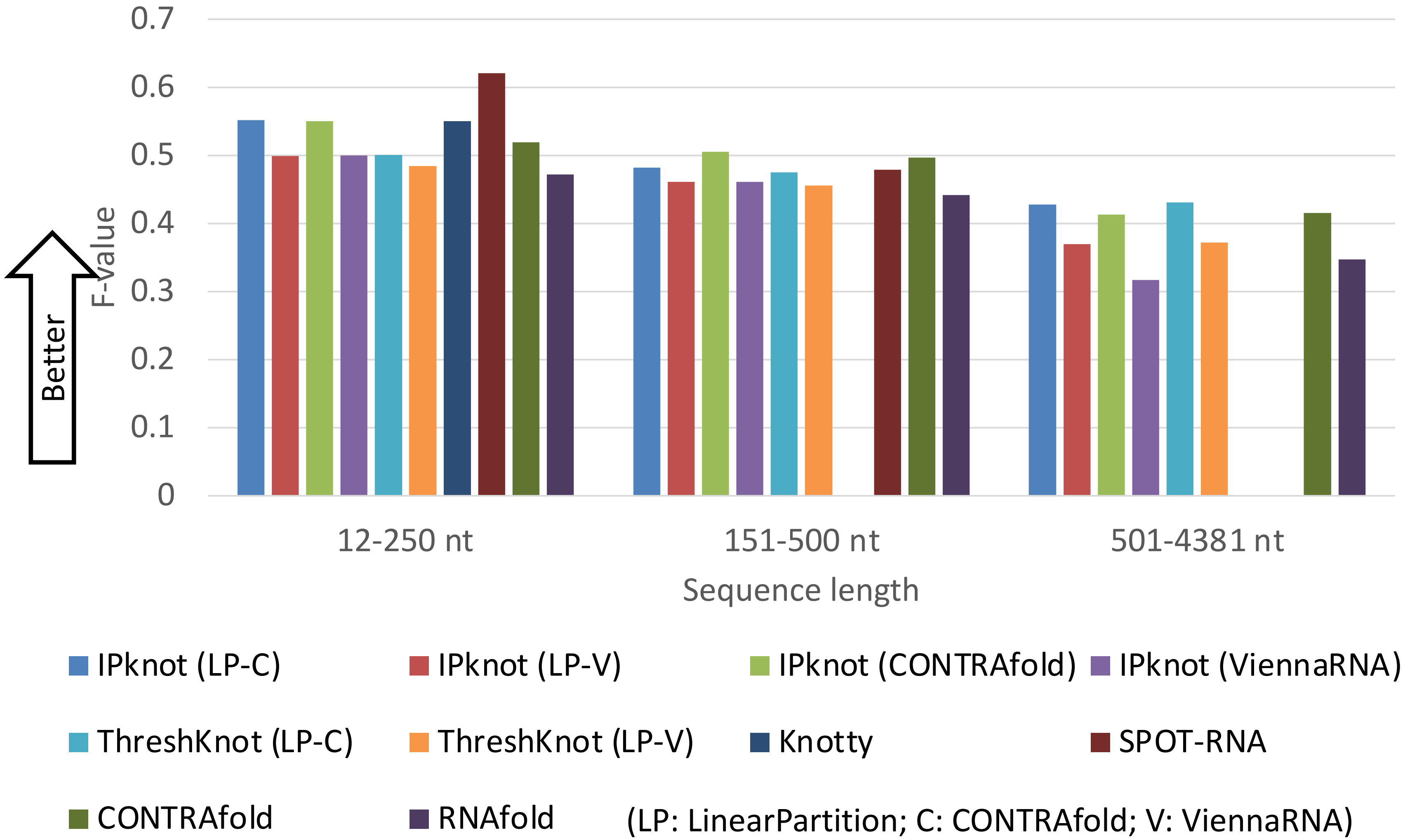


No pseudoknots



Each base pair at the level p is pseudoknotted to at least one base pair at the lower level q

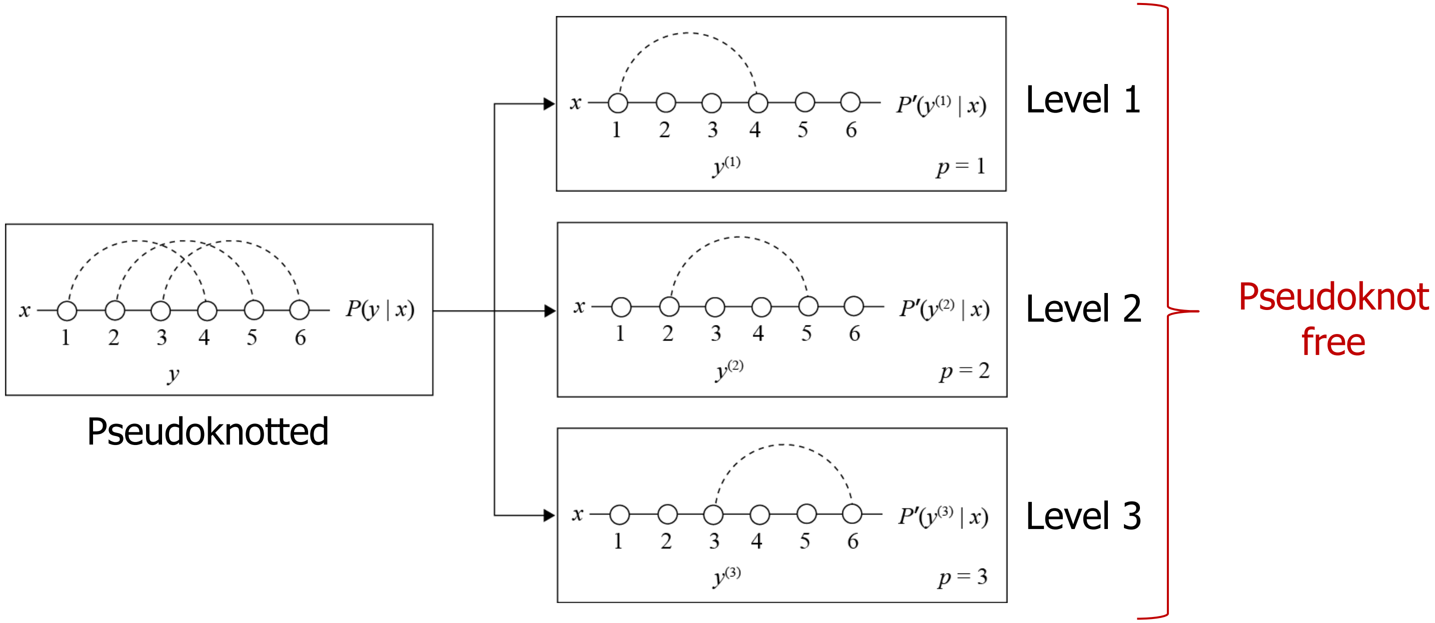
Prediction accuracy for pseudoknotted structures



Test data was compiled from bpRNA-1m and Rfam 14.5.

IPknot integrated with MXfold2

- IPknot approximates a probability distribution over pseudoknotted structures by its factorization of pseudoknot-free structures:
  - (2011 version) CONTRAfold model, ViennaRNA model, NUPACK model
  - (2022 version) LinearFold-C model, LinearFold-V model
- We implemented new IPknot that integrates MXfold2 as a probability distribution over pseudoknot-free structures.



Pseudoknotted

Level 1

Level 2

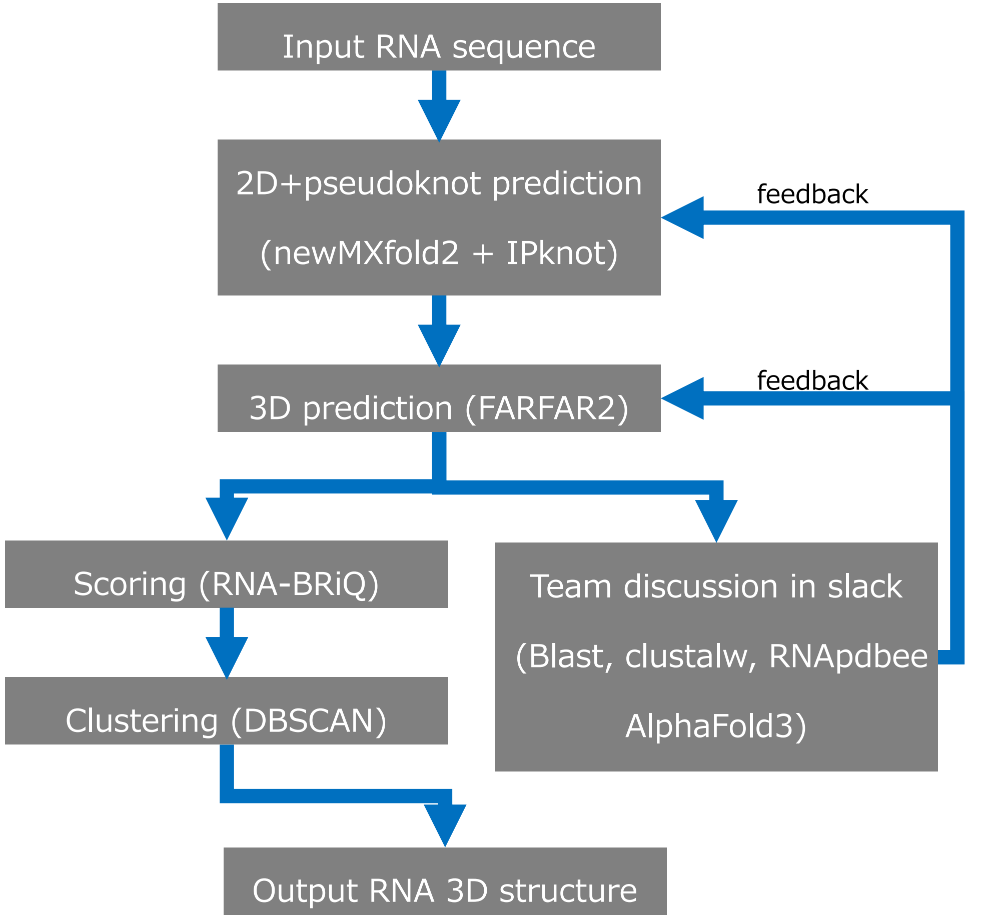
Level 3

Pseudoknot free

IPknot integrated with MXfold2

- We are participating in CASP16 as **RNA_Dojo** team with a workflow based on the new IPknot with MXfold2, FARFAR2, and RNA-BRiQ.

Prediction workflow of **RNA_Dojo** team



Input RNA sequence

2D+pseudoknot prediction

(newMXfold2 + IPknot)

3D prediction (FARFAR2)

Scoring (RNA-BRiQ)

Clustering (DBSCAN)

Team discussion in slack

(Blast, clustalw, RNApdbee

AlphaFold3)

Output RNA 3D structure

Table of Contents

- Overview of RNA secondary structure prediction
  - Architecture
    - Nussinov algorithm, Nearest neighbor model
  - Inference
    - MFE, MEA
  - Parametrization
    - Machine learning, Deep learning
- Future direction
  - Chemical probing
  - RNA modification
  - Pseudoknots

Acknowledgements

- IPknot, CentroidFold
  - Yuki Kato (Osaka U)
  - Kiyoshi Asai (U Tokyo)
  - Tatsuya Akutsu (Kyoto U)
  - Michiaki Hamada (Waseda U)
  - Hisanori Kiryu (U Tokyo)
  - Toutai Mituyama (AIST)
- MXfold2
  - Yasubumi Sakakibara (Keio U)
  - Manato Akiyama (Keio U)
- RNA Dojo in CASP16
  - Junichi Iwakiri (U Tokyo)
  - Takumi Otagaki (U Tokyo)
  - Kazuteru Yamamura (U Tokyo)
  - Shunsuke Sumi (U Tokyo)
  - Ikuo Kurisaki (Waseda U)
  - Jiro Kondo (Sophia U)

科研費 KAKENHI

CREST

GteX 革新的GX技術創出事業