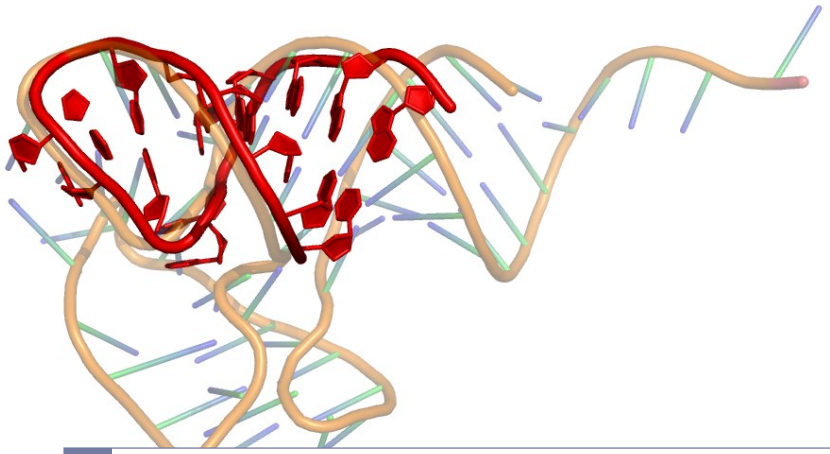
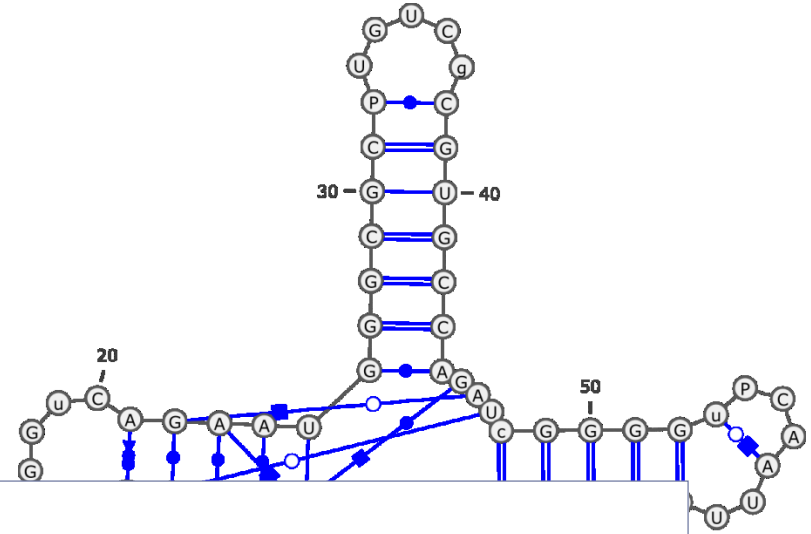




RNA Bioinformatics beyond energy minimization

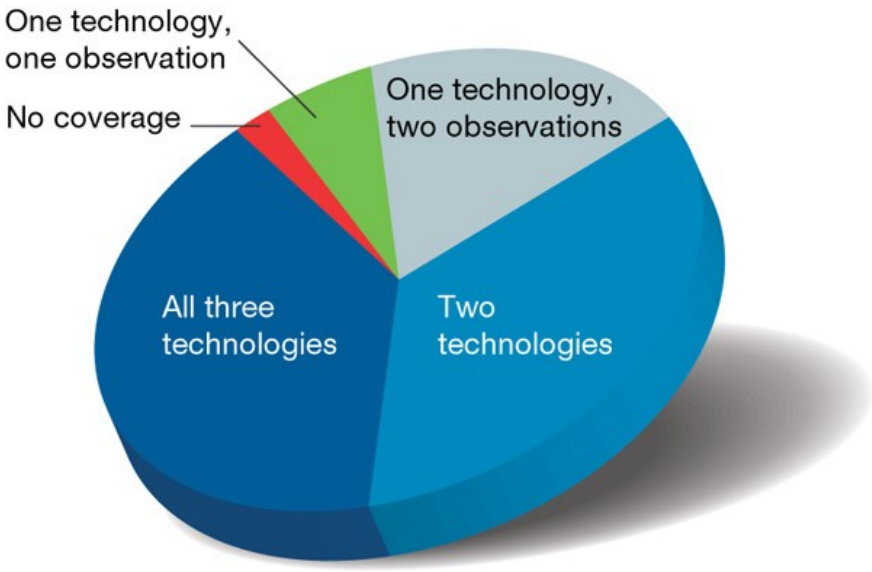
Yann PONTY
CNRS/Ecole Polytechnique/AMIB Inria Saclay

Slides @ http://goo.gl/1sc5MT

Why (non-coding) RNAs?

Why RNA is **totally awesome!**

- Ubiquitous
- Pervasively expressed

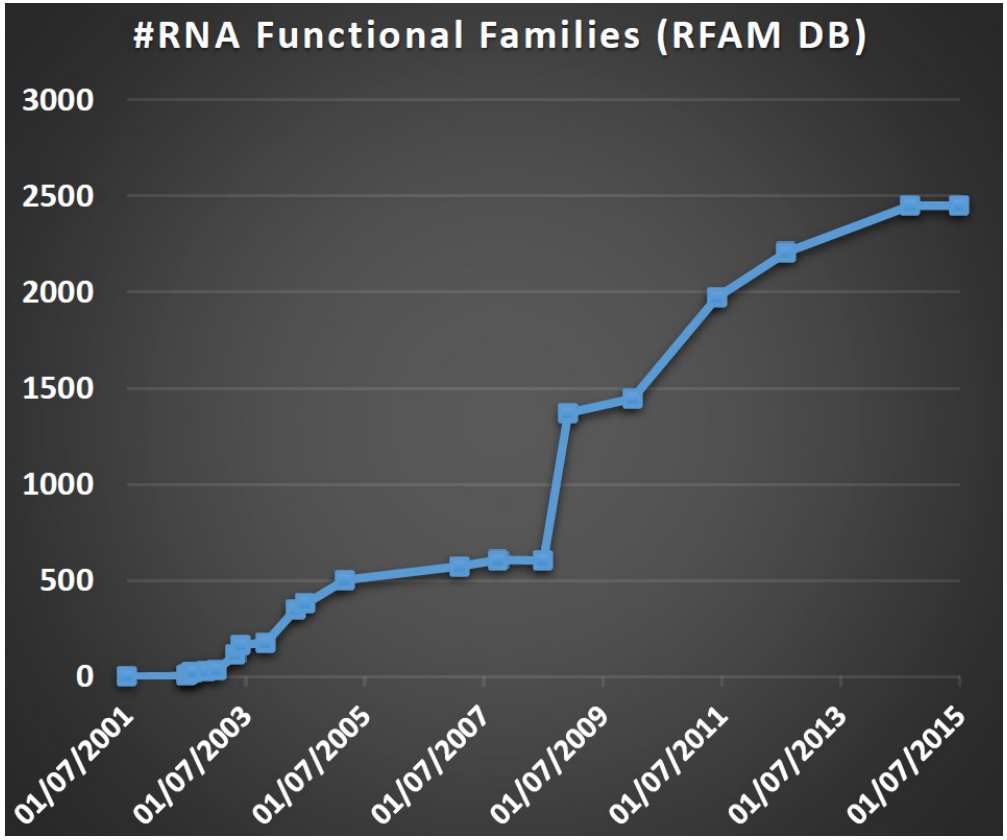


The human genome is **pervasively transcribed**, such that the majority of its bases are associated with at least one primary transcript and many transcripts link distal regions to established protein-coding loci.

ENCODE Analysis of 1% of the human genome Nature 2007

Why RNA is **totally awesome!**

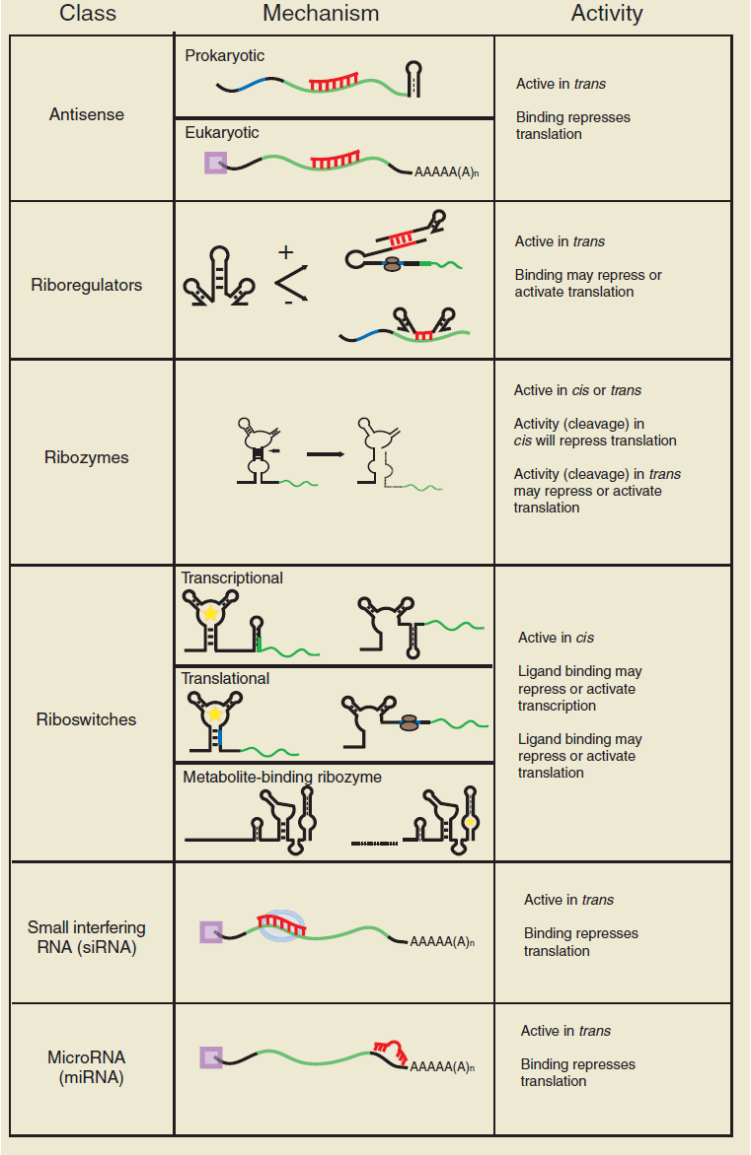
- Ubiquitous
- Pervasively expressed
- Versatile
  - Carriers
  - Transporter
  - Enzymatic
  - Processing
  - Regulatory
  - ssRNA genomes (HIV)
  - Immune system (CRISPR)
  - More soon… (lincRNAs)



Why RNA is **totally awesome!**

- Ubiquitous
- Pervasively expressed
- Versatile
- Easy to handle
  - Synthetic biology

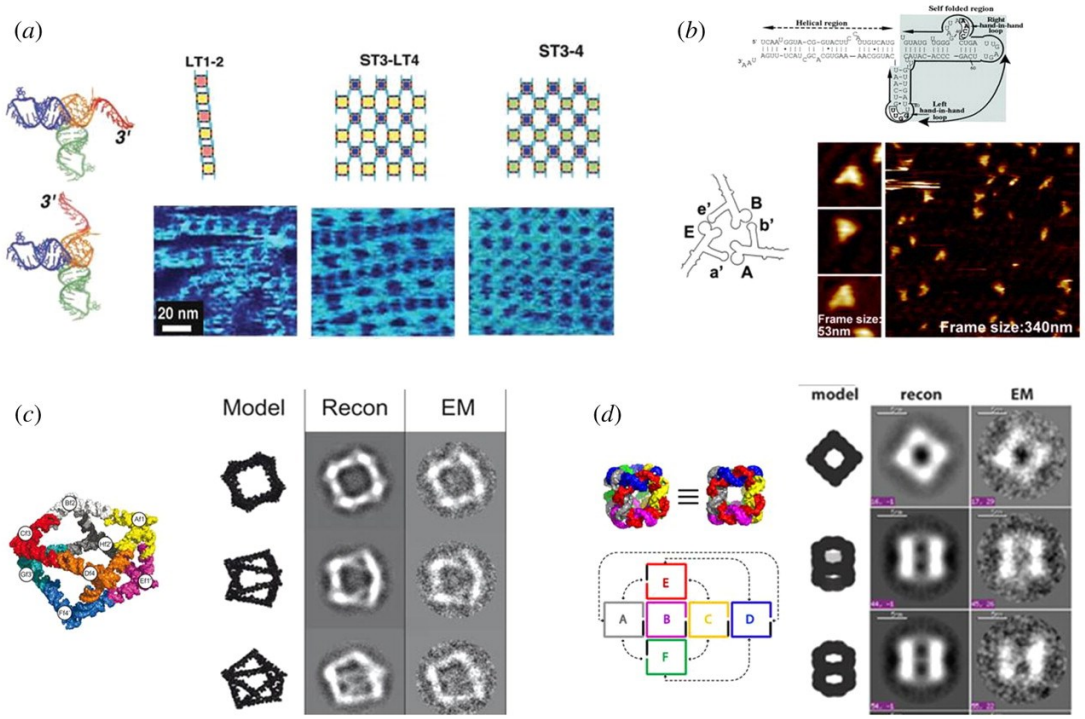
[Isaacs, F J *et al.* Nature Biotech. 2006]



Why RNA is **totally awesome!**

- Ubiquitous
- Pervasively expressed
- Versatile
- Easy to handle
  - Synthetic biology
  - Nanotechs

RNA-based Nanoarchitectures
[Li H *et al*, Interface Focus 2011]

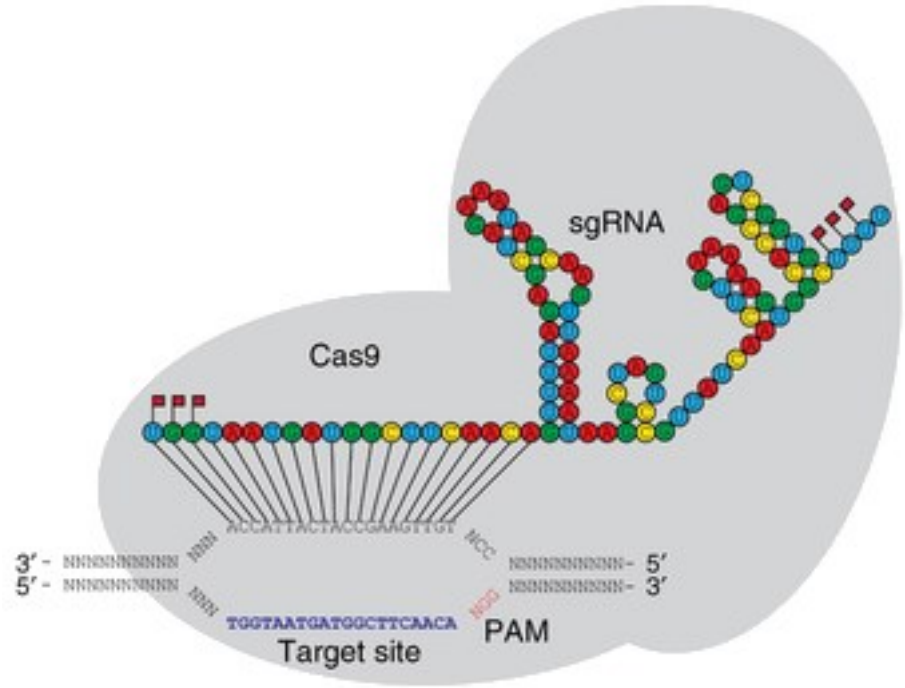


Why RNA is **totally awesome!**

- Ubiquitous
- Pervasively expressed
- Versatile
- Easy to handle
  - Synthetic biology
  - Nanotechs
  - Therapeutics and genetic engineering (CRISPR)

Blooming therapeutic RNAi…
… making way for CRISPR!

[Agrotis & Ketteler, Frontiers Genetics 2015]

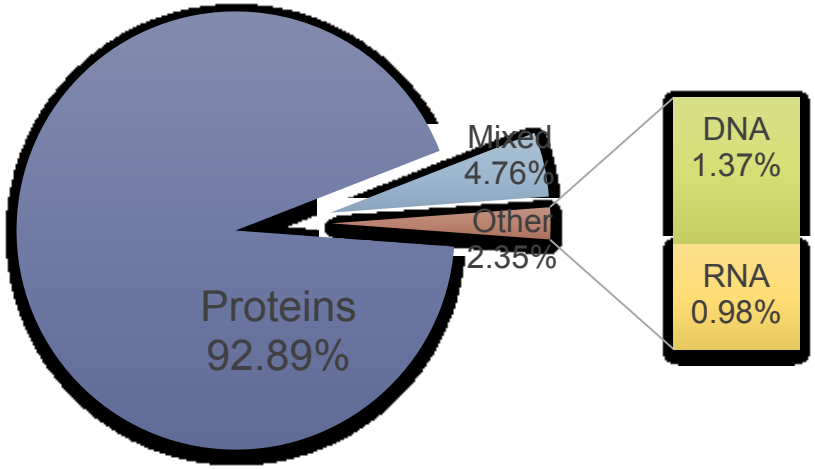


[Hendel *et al*, Nature Biotech. 2015]

Why RNA is **totally awesome!**

- Ubiquitous
- Pervasively expressed
- Versatile
- Easy to handle
  - Synthetic biology
  - Nanotechs
  - Therapeutics and genetic engineering (CRISPR)
  - Computationally fun (but still challenging)

PDB: 117,022 entries (March 2016)



(Initial) lack of structural data

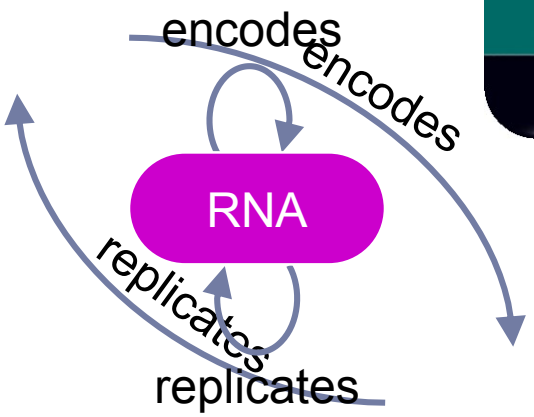
Experiment-based energy models
+ Secondary structure
+ Efficient combinatorial algorithms

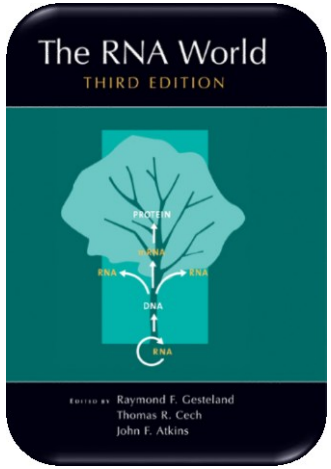
⇒ Mature *ab initio* prediction tools (Mfold, RNAfold…)

Why RNA is **totally awesome!**

- Ubiquitous
- Pervasively expressed
- Versatile
- Easy to handle
  - Synthetic biology
  - Nanotechs
  - Therapeutics and genetic engineering (CRISPR)
  - Computationally fun (but still challenging)
- RNA at the origin of life!?

The *chicken vs egg* paradox at the origin of life





This is the RNA World.

*[…] **Proteins** are good at being enzymes but bad at being replicators; […] **DNA** is good at replicating but bad at being an enzyme; […] **RNA** might just be good enough at both roles to break out of the Catch-22.*

R. Dawkins. The Ancestor's tale

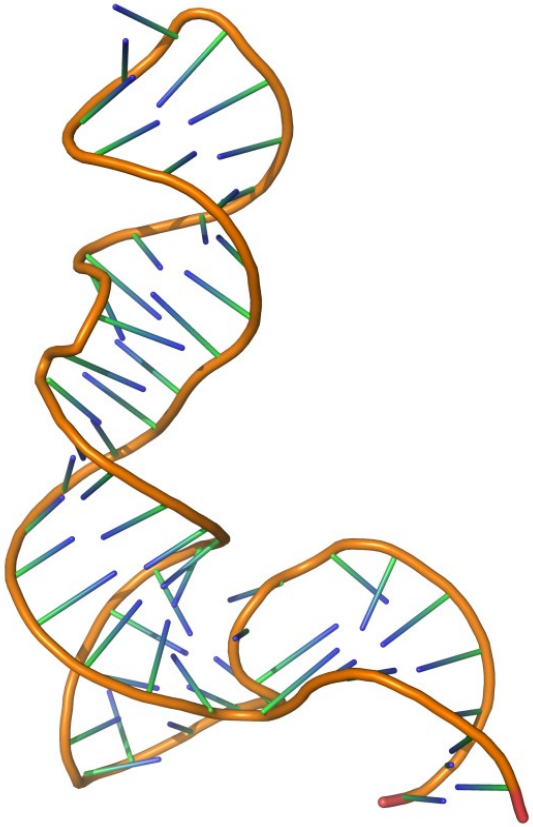
RNA Structure

Why structure matters

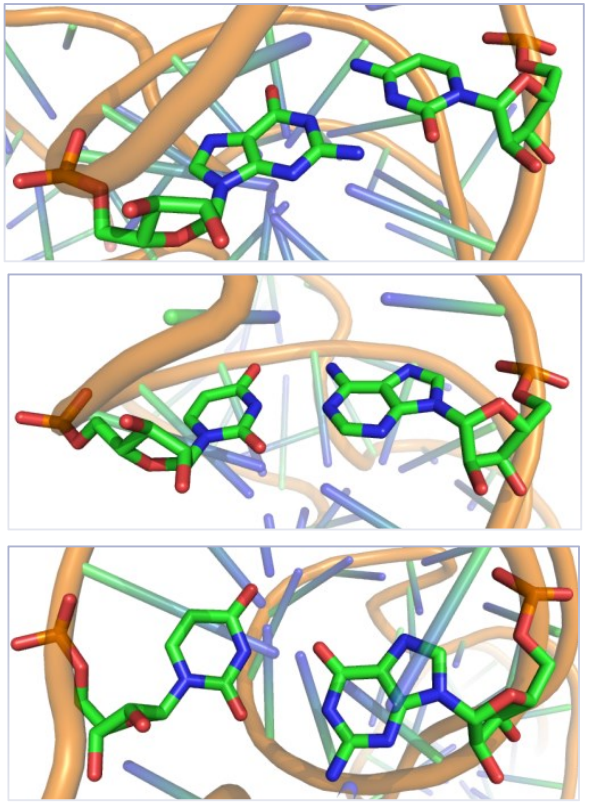
- Transcription: RNA is (mostly) single stranded
- Structurally diverse
- ncRNAs → Structure(s) typically more conserved than sequence
- Functionally versatile

Use structure as a proxy for function, to explain functional behaviors

Why RNA folds



5s rRNA (PDB ID: 1UN6)



G/C

U/A

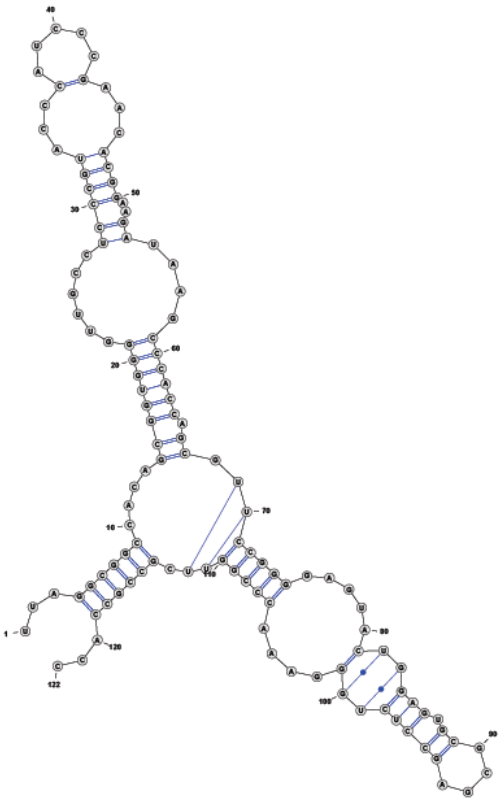
U/G

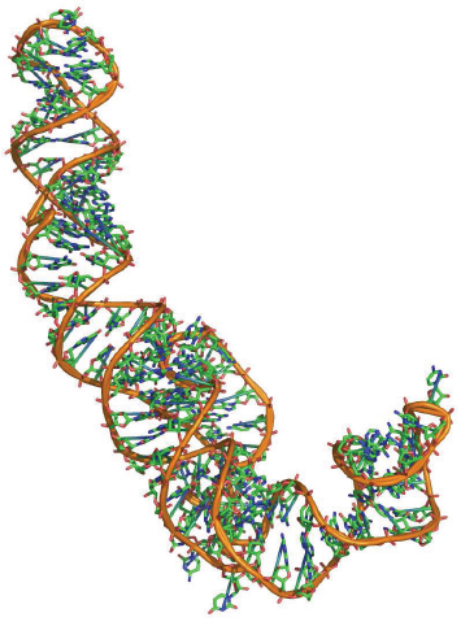
Canonical base-pairs

RNA folding = Hierarchical stochastic process driven by/resulting in the pairing (hydrogen bonds) of a subset of its bases.

Three levels of RNA structure

```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```





Primary structure

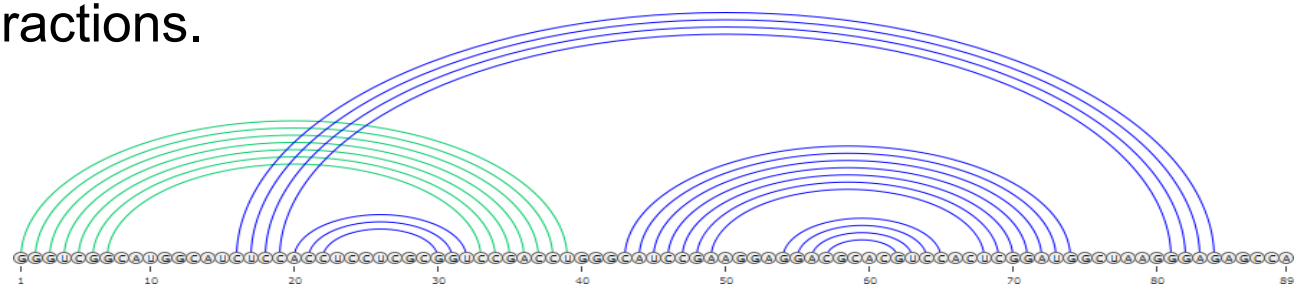
Secondary structure

Tertiary structure

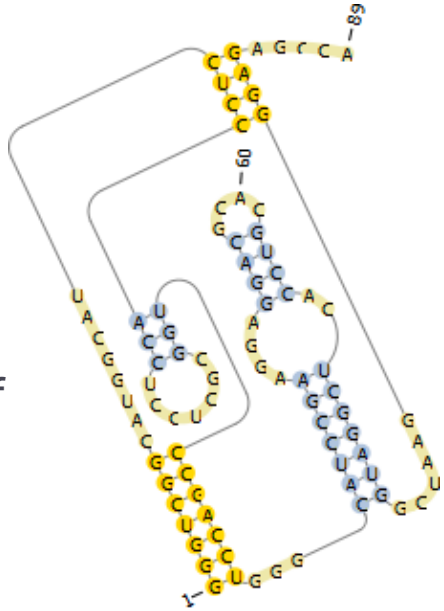
Source: 5s rRNA (PDBID: 1K73:B)

Pseudoknots

- Pseudoknots are complex topological models indicated by crossing interactions.

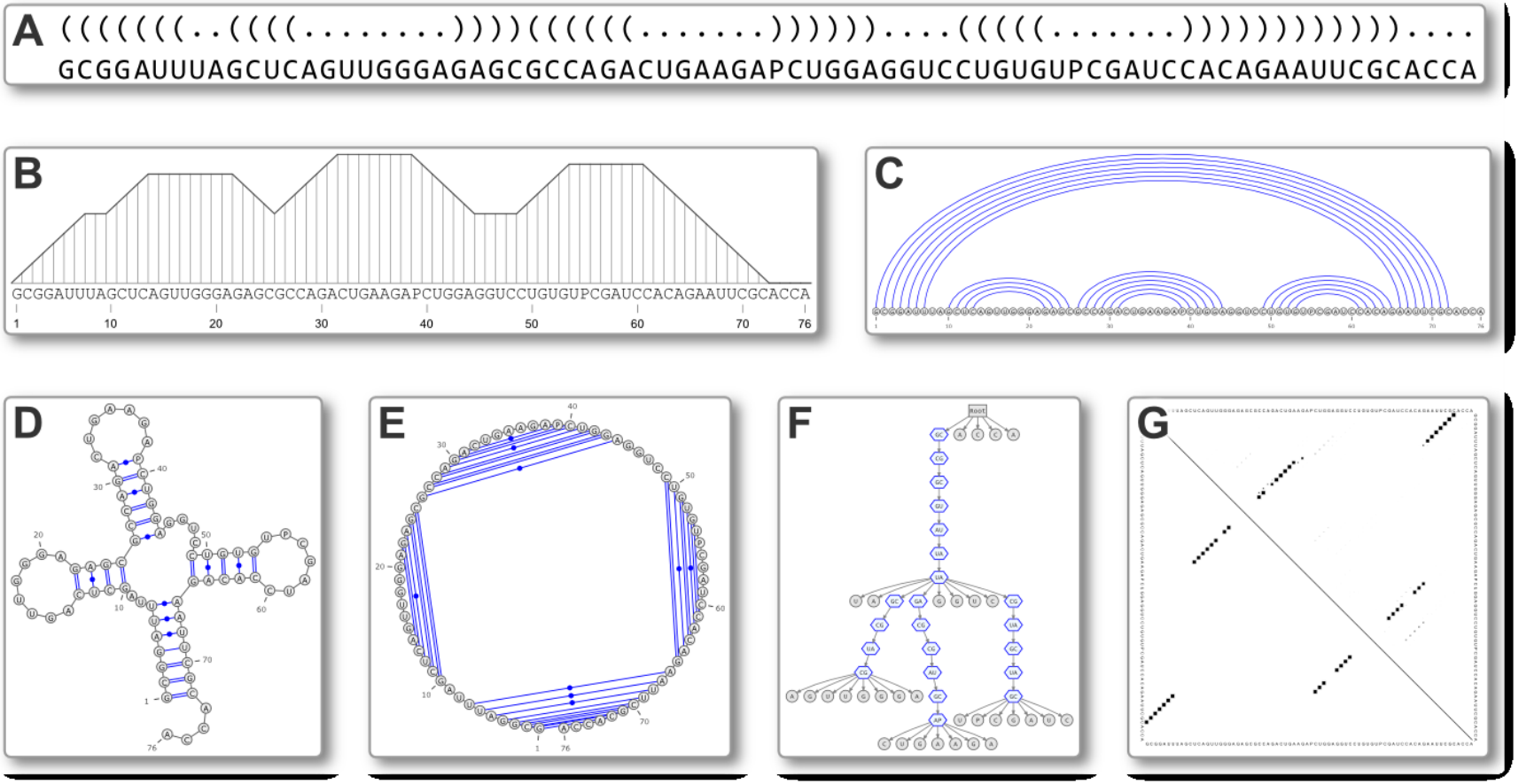


- Pseudoknots are largely ignored by computational prediction tools:
  - Lack of accepted energy model
  - Algorithmically challenging



- Yet heuristics can be sometimes efficient
  - Pknots-RG offers a reasonable time/sensitivity tradeoff

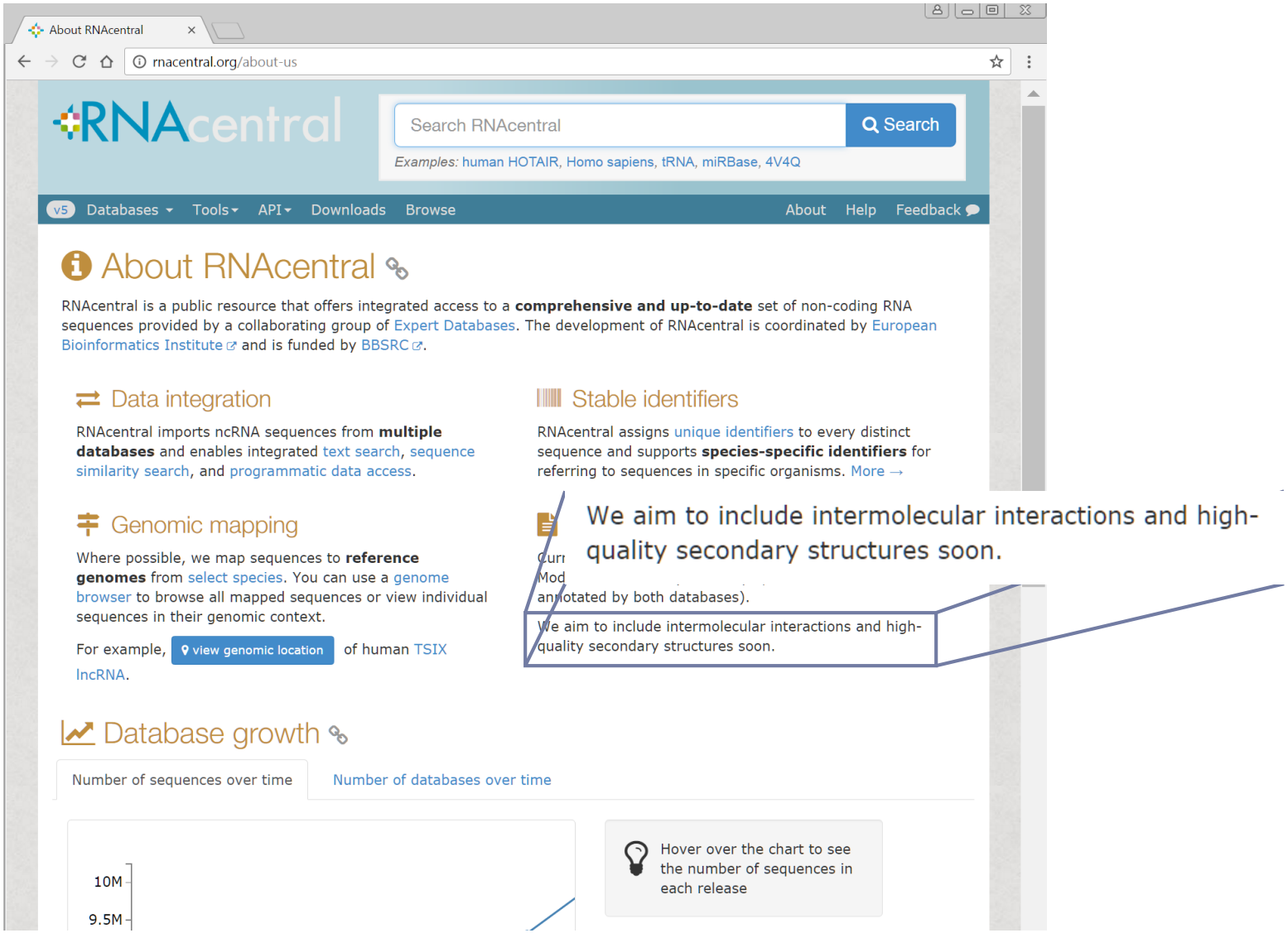
Secondary Structure representations



http://varna.lri.fr

ncRNA Data

RNACentral.org: One ID to rule them all



Sources of RNA structural data

Name	Data type	Scope	Description	File formats	#Entries	URL
PDB	All-atoms	General	RCSB Protein Data Bank – Global repository for 3D molecular models	PDB	~1,900 models	http://www.pdb.org
NDB	All-atoms, Secondary structures	General	Nucleic Acids Database – Nucleic acids models and structural annotations.	PDB, RNAML	~2,000 models	http://bit.ly/rna-ndb
RFAM	Alignments, Secondary structures[3]	General	RNA FAMilies – Multiple alignments of RNA as functional families. Features consensus secondary structures, either predicted and/or manually curated.	STOCKHOLM, FASTA	~1,973 Alignments/ structures, 2,756,313 sequences	http://bit.ly/rfam-db
STRAND	Secondary structures	General	The RNA secondary STRucture and statistical ANalysis Database – Curated aggregation of several databases	CT, BPSEQ, RNAML, FASTA, Vienna	4,666 structures	http://bit.ly/sstrand
PseudoBase	Secondary structures	Pseudoknotted RNAs	PseudoBase – Secondary structure of known pseudonotted RNAs.	Extended Vienna RNA	359 structures	http://bit.ly/pkbase
CRW	Sequence alignments, Secondary structures	Ribosomal RNAs, Introns	Comparative RNA Web Site – Manually curated alignments and statistics of ribosomal RNAs.	FASTA, ALN, BPSEQ	1,109 structures, 91,877 sequences	http://bit.ly/crw-rna
…			…			

[2012 Snapshot]

RNA file formats: Sequences (alignments)

```
>O.sativa.1 AJ489954.1/1-104
......UGGCUGUGACGACUAGGUGAAAUU.CAAGCUCAACAGACCAAAUCACAGGUCUC
..UCUCCAAGGCCUU.UGGAGAUGGGAUCUGUAUGCCGA.....GU..UUCCGCUC....
.AGCCG......
>O.sativa.2 AY013245.2/61987-62105
....GAUGGCAGUGACGACUUGGUAAUAUU.CAAGCUCAACAGACCAAAUCACAGGUCUU
CCUCUCUGGAUCCAC..UCCUCUGGGAUUGAUUUG..UAUGCCGAUUUUCCCGCUGAACC
GAGCCAUC....
>O.sativa.3 AJ307928.1/3-121
....GAUGGCAGUGACGACCUGGUAAUAUU.CAAGCUCAACAGACCAAAUCACAGGUCUU
..UCUCUCUGGAUCUACUCCUCAGGGAUUGAUUUG..UAUGCCGAUUUUUCCGCUGAACC
GAGCCAUC....
```

```
CLUSTAL 2.1 multiple sequence alignment

M.musculus.1      UGGCCUCGUUCAAGUAAUCCAGGAUAGG--CU--GUG-CAGGUCCCAAGGGGCCUAUUCU 55
H.sapiens.2       UGGCCUCGUUCAAGUAAUCCAGGAUAGG--CU--GUG-CAGGUCCCAAU-GGCCUAU-CU 53
H.sapiens.3       GGACCCAGUUCAAGUAAUUCAGGAUAGGUUGU--GUG-CUGU--CCAG----CCUGUUCU 51
T.rubripes.1      CAACCGGGUUCAAGUAAUCCAGGAUAGGCUCU--GUAUCUGU--CUUGG---CCUAUGCU 53
H.sapiens.1       UGGCUGGAUUCAAGUAAUCCAGGAUAGGCUGUUUCCAUCUGU--G-AGG---CCUAUUCU 54
                   ..*   .********** *********   *     . * *      .    ***.* **

M.musculus.1      UGGUUACU---UGCACGGGGAC 74
H.sapiens.2       UGGUUACU---UGCACGGGGAC 72
H.sapiens.3       CCAUUACU--UGGCUCGGGGAC 71
```

RNA file formats: Sequences (alignments)

```
# STOCKHOLM 1.0
#=GF ID    mir-22
#=GF AC    RF00653
...
O.latipes.1             CGUUG.CCUCACAGUCGUUCUUCA.CUGGCU.AGCUUUAUGUCCCACG..
Gasterosteus_aculeat.1  GGCUG.ACCUACAGCAGUUCUUCA.CUGGCA.AGCUUUAUGUCCUCAUCU
R.esox.1                AGCUGAGCACA...CAGUUCUUCA.CUGGCA.GCCUUAAGGUUUCUGUAG
...
#=GC SS_cons            .<<<<.<<..<<<<<<<<<<<<<<..<<<<..<<<<<<<.<<........
#=GC RF                 gGccg.acucaCagcaGuuCuuCa.cuGGCA.aGCuuuAuguccuuauaa

O.latipes.1             CCCCACCGUAAAGCU.GC.CAGUUGAAGAGCUGUUGUG..UGUAACC
Gasterosteus_aculeat.1  ACCAGC..UAAAGCU.GC.CAGCUGAAGAACUGUUGUG..GUCGGCA
R.esox.1                ACAGGC..UAAACCU.GC.CAGCUGAAGAACUGCUCUG..GCCAGCU
...
#=GC SS_cons            ....>>..>>>>>>>.>>.>>..>>>>>>>>>>>>>>...>>>>>>.
#=GC RF                 acaaac..UaaaGCu.GC.CaGuuGaaGaaCugcuGug..gucggCu
//
```

RNA file formats: Secondary Structures

```
> Rat Alanine tRNA
GAGGAUUUAGCUUAAUUAAAGCAGUUGAUUUGCAUUUAACAGAUGUAAGAUAUAGUCUUACAGUCCUUA
((((((...((((.....)))).(((((.......)))))...(((((((...)))))))))))))).
```

```
        1590      1600      1610      1620      1630
#       |123456789|123456789|123456789|123456789|123456
$ 1590  AAAAAACUAAUAGAGGGGGGACUUAGCGCCCCCCAAACCGUAACCCC=1636
% 1590  ::::::::::::::[[[[[[:::::(((]]]]]]::::)))::::::
```

RNA file formats: Secondary Structures

```
Filename: AM286415_b.bpseq
Organism: Yersinia enterocolitica subsp. enterocolitica 8081
Accession Numbers: AM286415
Citation and related information available at http://www.rna.ccbb.utexas.edu
1 U 0
...
117 U 0
118 U 236
119 G 235
120 C 234
121 C 233
122 U 232
123 G 231
124 G 230
...
230 C 124
231 C 123
232 A 122
233 G 121
234 G 120
235 C 119
236 A 118
...
```

```
80      dG = -33.48 [Initially -35.60]
1       U       0       2       80      1
2       G       1       3       79      2
3       G       2       4       78      3
4       G       3       5       77      4
5       A       4       6       76      5
6       U       5       7       0       6
7       G       6       8       75      7
...
75      U       74      76      7       75
76      U       75      77      5       76
77      C       76      78      4       77
78      C       77      79      3       78
79      U       78      80      2       79
80      A       79      0       1       80
```

RNA file formats: Secondary Structures

```
HEADER    RNA                                     27-JUL-09   3IGI
TITLE     TERTIARY ARCHITECTURE OF THE OCEANOBACILLUS IHEYENSIS GROUP
TITLE    2 II INTRON
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: GROUP IIC INTRON;
COMPND   3 CHAIN: A;

...

ATOM   8009  P     U A 375      19.076  79.179 370.688  1.00 66.25           P
ATOM   8010  OP1   U A 375      18.815  77.862 371.313  1.00 83.22           O
ATOM   8011  OP2   U A 375      19.869  80.203 371.409  1.00 56.32           O
...
CONECT 8654 8520
CONECT 8655 8521
CONECT 8658 8531
MASTER      717    0   66    0    0    0   69    6 8656    2  123   33
END
```

RNA file formats: Secondary Structures

```
<?xml version="1.0"?>
<!DOCTYPE rnaml SYSTEM "rnaml.dtd">
<rnaml version="1.0">
   <molecule id="xxx">
      <sequence> ... </sequence>
      <structure> ... </structure>
    </molecule>
  <interactions> ... </interactions>
</rnaml>
```

RNA file formats: Secondary Structures

```
<?xml version="1.0"?>
<!DOCTYPE rnaml SYSTEM "rnaml.dtd">
<rnaml version="1.0">
   <molecule id="xxx">
      <sequence>
         <numbering-system id="1" used-in-file="false">
            <numbering-range>
               <start>1</start><end>387</end>
            </numbering-range>
         </numbering-system>
         <numbering-table length="387">
            2    3    4    5    6    7    8...
         </numbering-table>
         <seq-data>
            UGUGCCCGGC AUGGGUGCAG UCUAUAGGGU...
         </seq-data>
         ...
      </sequence>
      <structure> ... </structure>
    </molecule>
  <interactions> ... </interactions>
</rnaml>
```

RNA file formats: Secondary Structures

```
<?xml version="1.0"?>
<!DOCTYPE rnaml SYSTEM "rnaml.dtd">
<rnaml version="1.0">
   <molecule id="xxx">
      <sequence> ... </sequence>
      <structure>
         <model id="yyy">
            <base> ... </base> ...
            <str-annotation>
               ...
               <base-pair>
                  <base-id-5p><base-id><position>2</position></base-id></base-id-5p>
                  <base-id-3p><base-id><position>260</position></base-id></base-id-3p>
                  <edge-5p>+</edge-5p>
                  <edge-3p>+</edge-3p>
                  <bond-orientation>c</bond-orientation>
               </base-pair>
               <base-pair comment="?">
                  <base-id-5p><base-id><position>4</position></base-id></base-id-5p>
                  <base-id-3p><base-id><position>259</position></base-id></base-id-3p>
                  <edge-5p>S</edge-5p>
                  <edge-3p>W</edge-3p>
                  <bond-orientation>c</bond-orientation>
               </base-pair>
               ...
            </str-annotation>
         </model>
      </structure>
    </molecule>
  <interactions> ... </interactions>
</rnaml>
```

RNA Structure Prediction

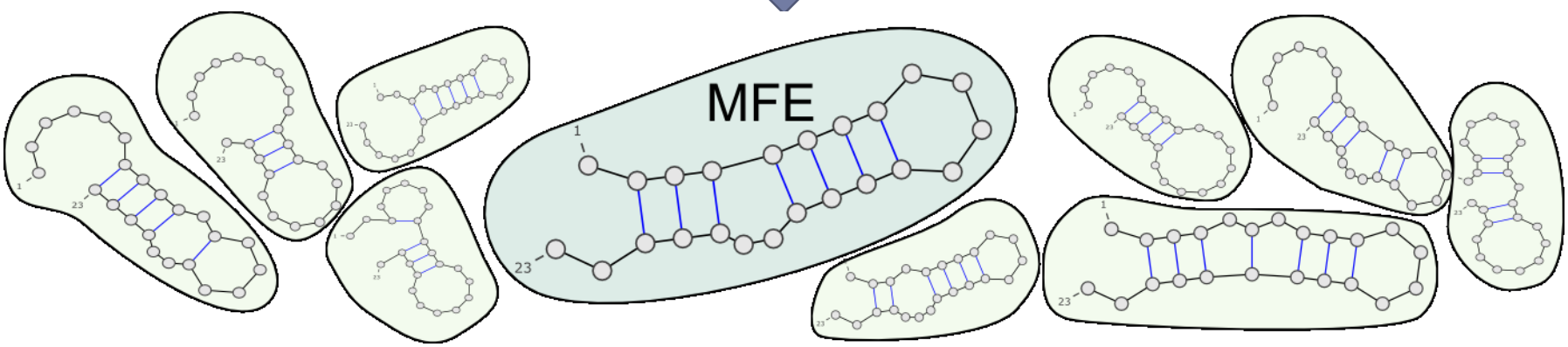
RNA structure prediction: The big picture

Biophysics → Shifting paradigms in RNA structure prediction

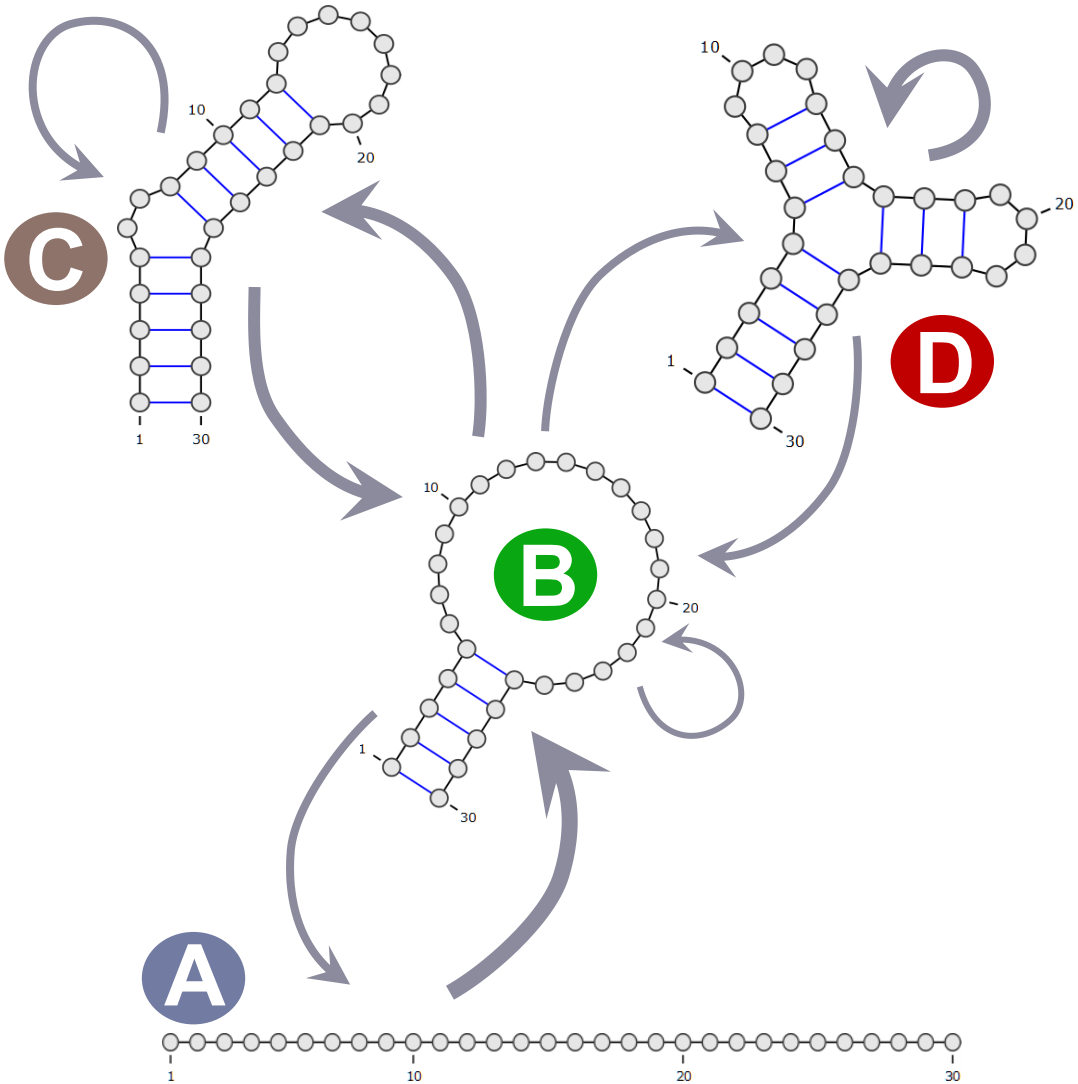
- **1970s-1990s:** Free-Energy Minimization → Maximizing **stability**
- **1990s-2010s:** Thermodynamic equilibrium → **Average** picture

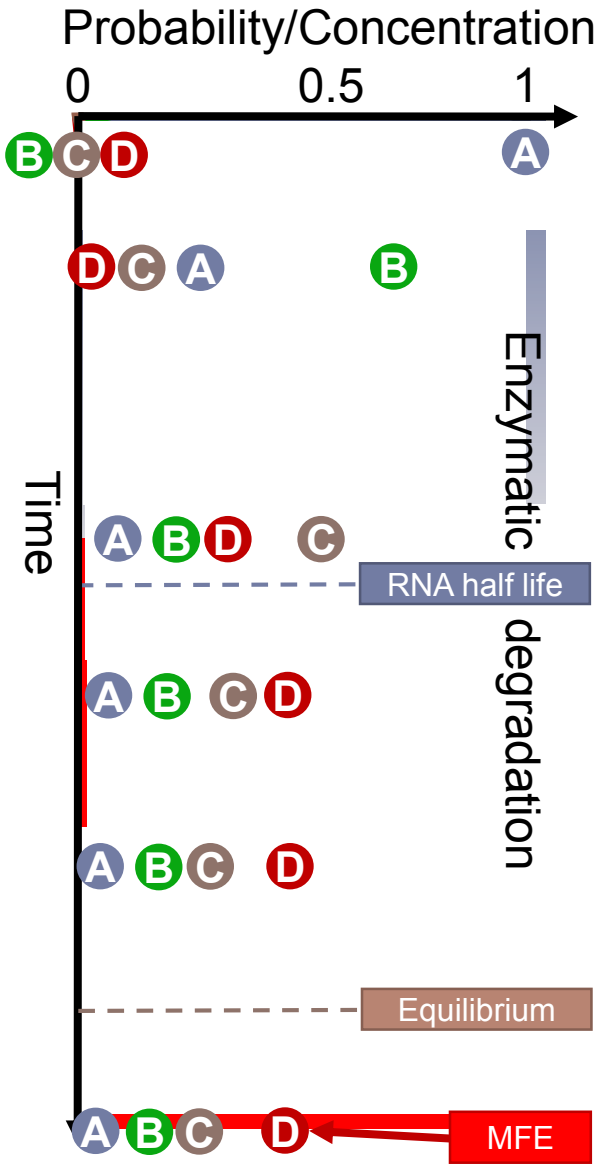
...**CAGUAGCCGAUCGCAGCUAGCGUA**...

RNAFold, MFold…



RNA kinetics: Why go through all the trouble?



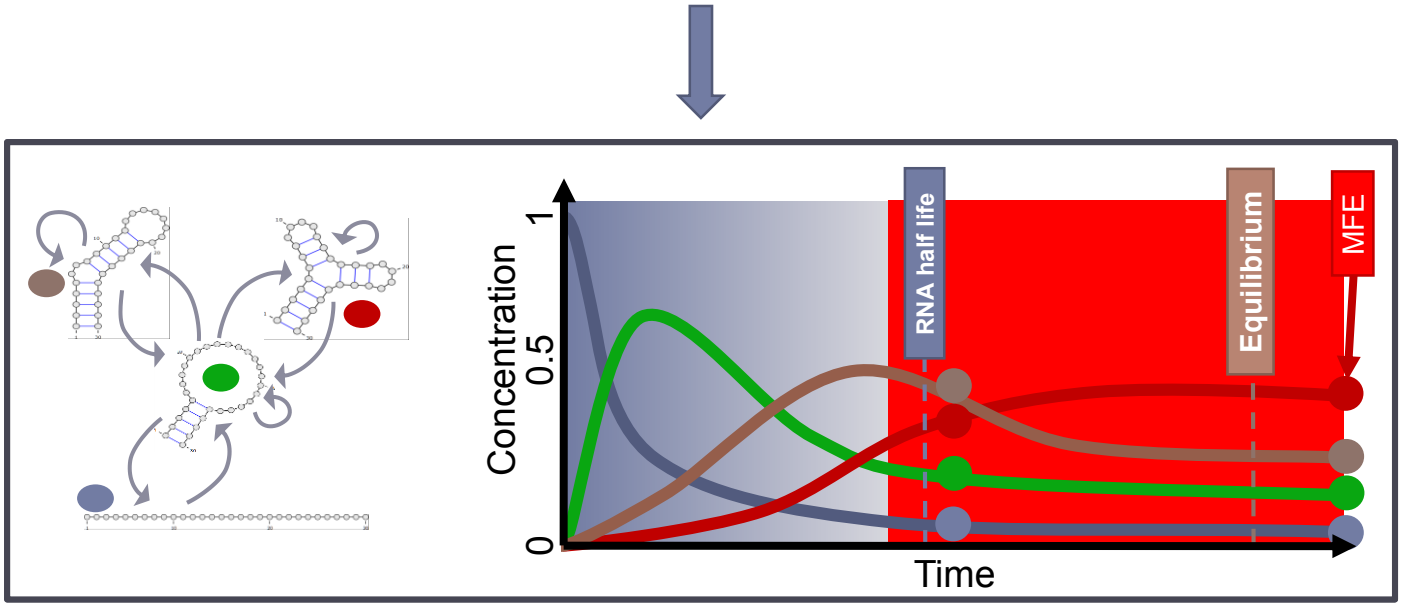


RNA structure prediction: The big picture

Biophysics → Shifting paradigms in RNA structure prediction

- **1970s-1990s:** Free-Energy Minimization → Maximizing **stability**
- **1990s-2010s:** Thermodynamic equilibrium → **Average** picture
- **2010s-???:** Kinetics → RNA folding at **finite time**

...**CAGUAGCCGAUCGCAGCUAGCGUA**...



Kinetics remains challenging **physically** and **computationally**

RNA Structure Prediction

Free-Energy Minimization (MFE)

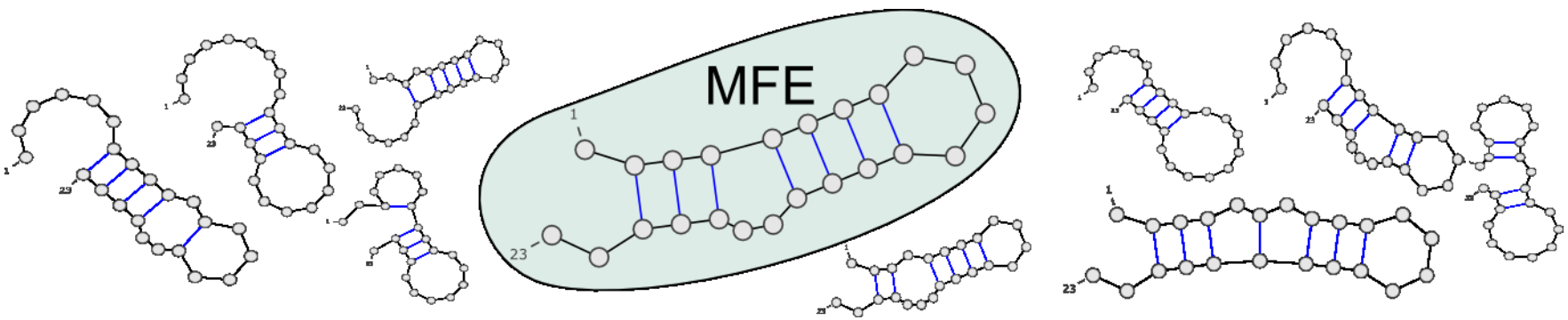
Minimal Free-Energy (MFE) Folding

Goal: Predict the functional (aka native) conformation of an RNA

- Absence of homologs/experimental evidences → Consider energy
- Turner model associates free-energies to secondary structures
- Vienna RNA package implements a $O(n^3)$ optimization algorithm for computing most stable (= min. free-energy) folding

...**CAGUAGCCGAUCGCAGCUAGCGUA**...

RNAFold, MFold…



[Nussinov & Jacobson, PNAS 1980; Zuker & Stiegler, NAR 1981]

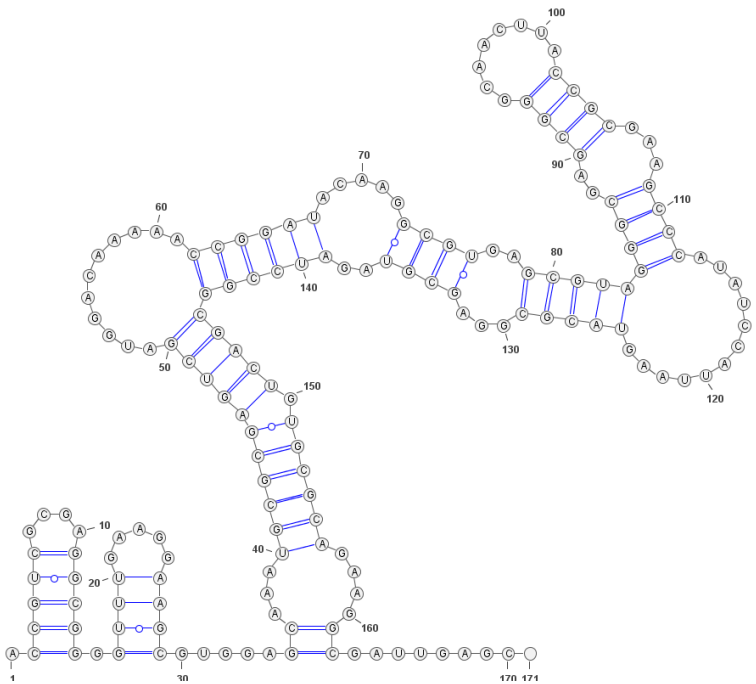
Energetic and algorithmic considerations

http://goo.gl/TSu679

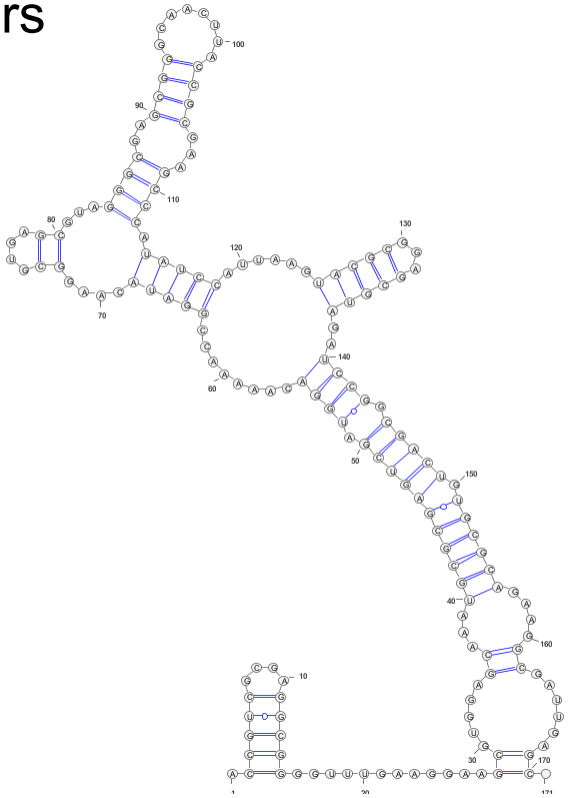
Optimization methods can be overly sensitive to fluctuations of the energy model

Example:

- Get RFAM *A. capsulatum* D1-D4 domain of the Group II intron
- Run RNAFold using default parameters (Turner 2004)
- Rerun RNAFold using latest energy parameters



Turner 2004

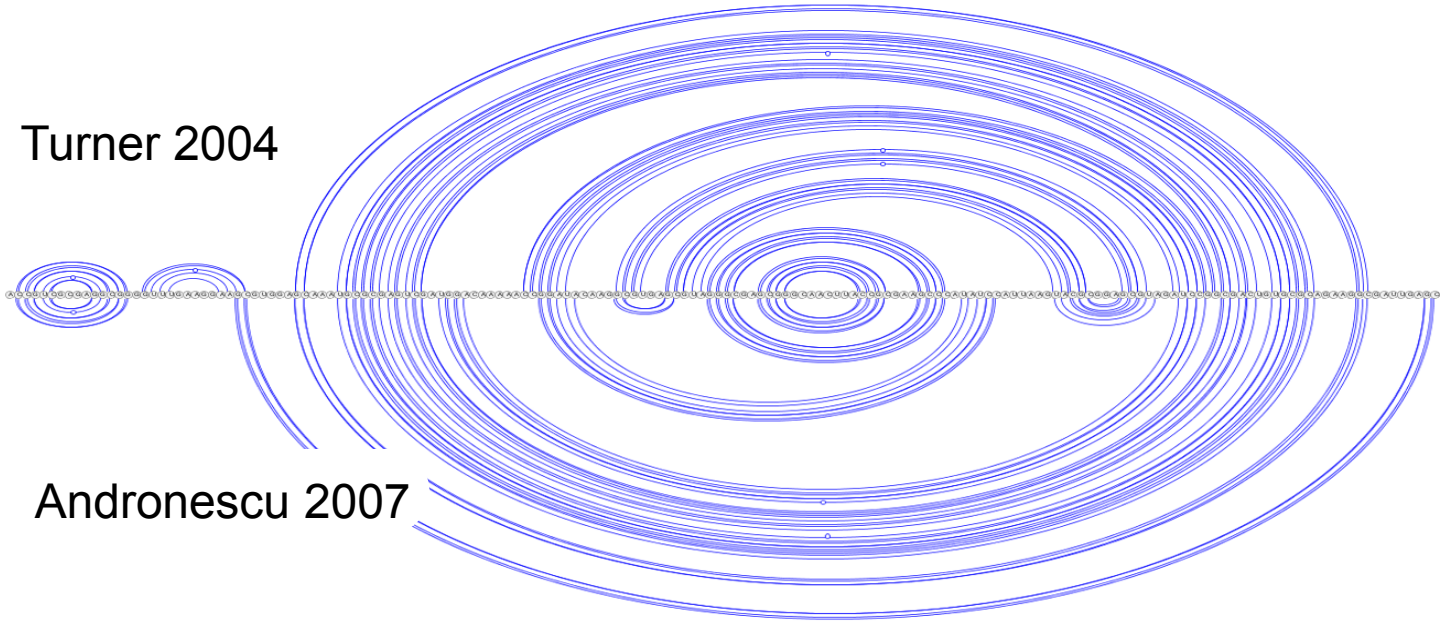


Andronescu 2007

Optimization methods can be overly sensitive to fluctuations of the energy model

Example:

- Get RFAM *A. capsulatum* D1-D4 domain of the Group II intron
- Run RNAFold using default parameters (Turner 2004)
- Rerun RNAFold using latest energy parameters

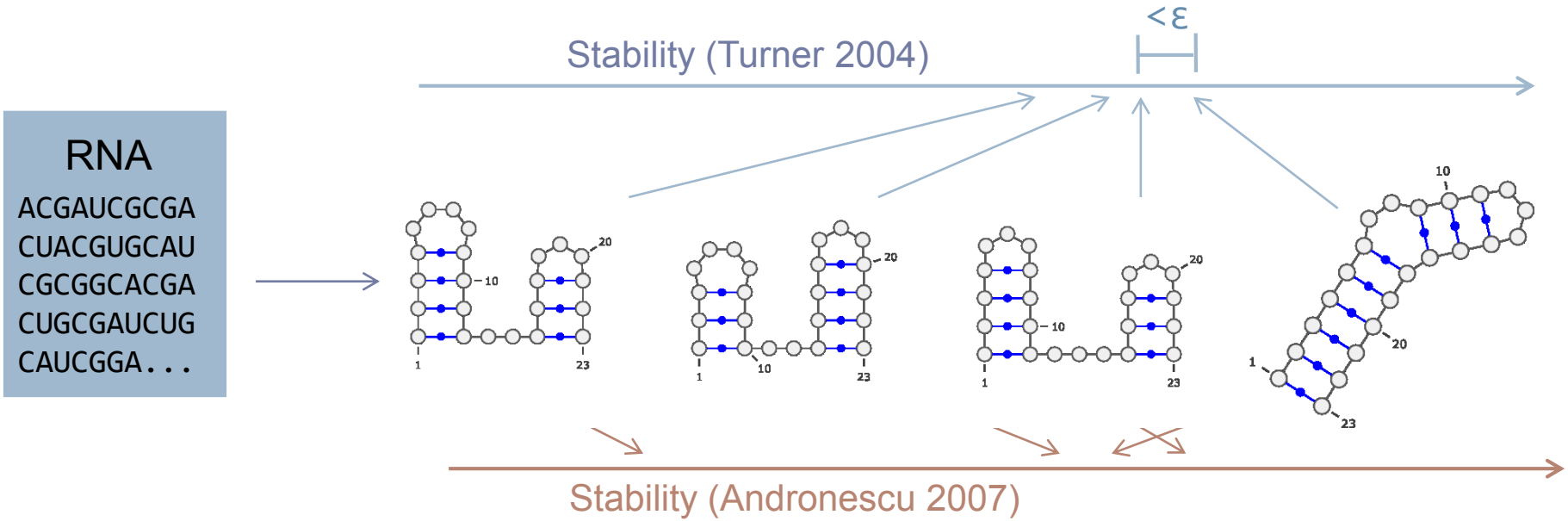


Discrepancy not as embarrassing as it first seemed…
… but still substantial!

Optimization methods can be overly sensitive to fluctuations of the energy model

Example:

- Get RFAM *A. capsulatum* D1-D4 domain of the Group II intron
- Run RNAFold using default parameters (Turner 2004)
- Rerun RNAFold using latest energy parameters



- Suboptimal structures (homogeneity, exponential growth)
- Guiding predictions with low-res/high-throughput experimental evidences

Energy-based *Ab initio* folding: Does it *really* work?

- Generally yes, but variable results for different studies

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

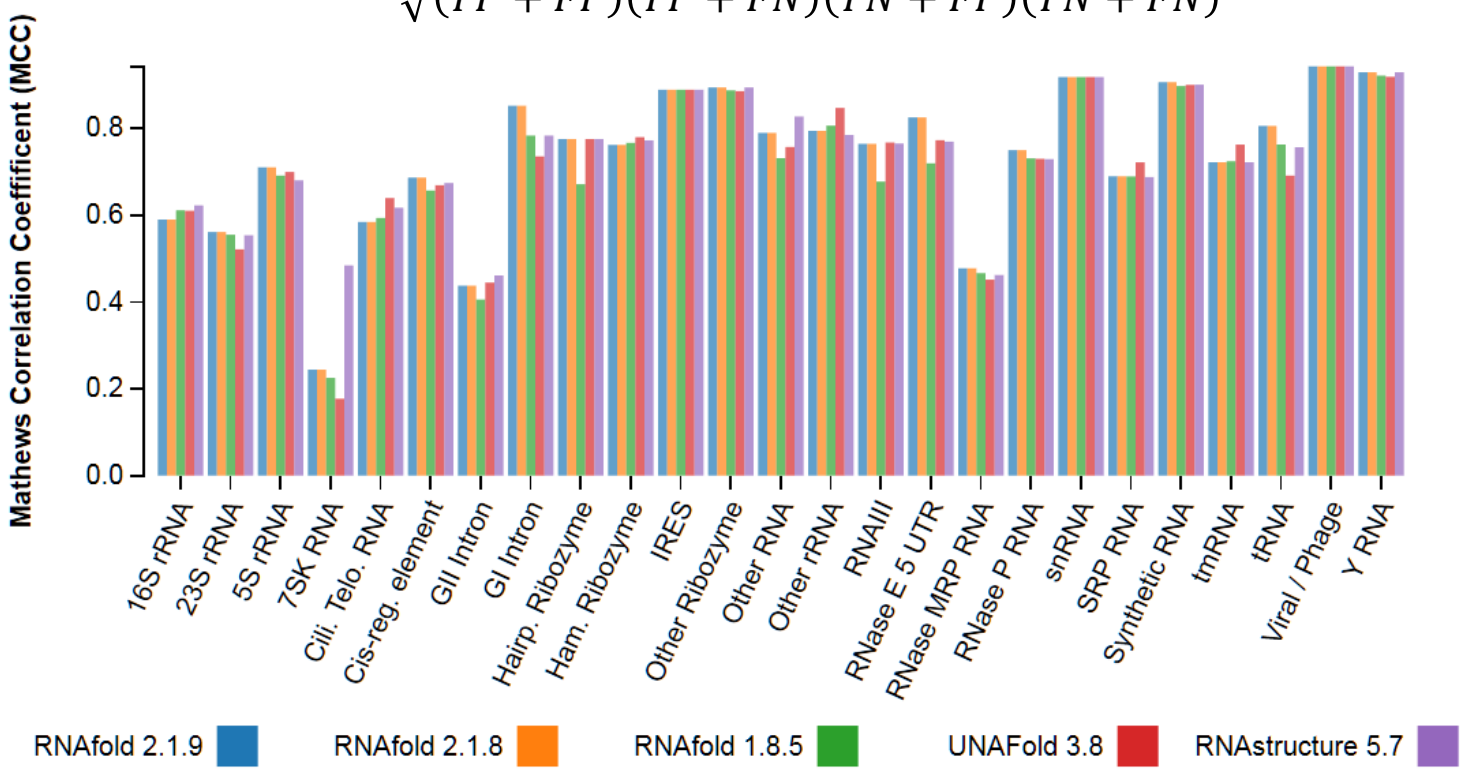
Program	Sensitivity	PPV	MCC	F-measure
RNAfold 2.1.9	0.742	0.795	0.767	0.765
RNAfold 2.1.8	0.740	0.792	0.764	0.762
RNAfold 1.8.5	0.711	0.773	0.740	0.737
UNAfold 3.8	0.693	0.767	0.727	0.725
RNAstructure 5.7	0.716	0.781	0.746	0.744

Benchmark: 1919 non-multimer/non-pseudoknotted sequence/structure pairs from the RNAstrand database (source Vienna Package web site)

Energy-based *Ab initio* folding: Does it *really* work?

- Generally yes, but variable results for different RNAs

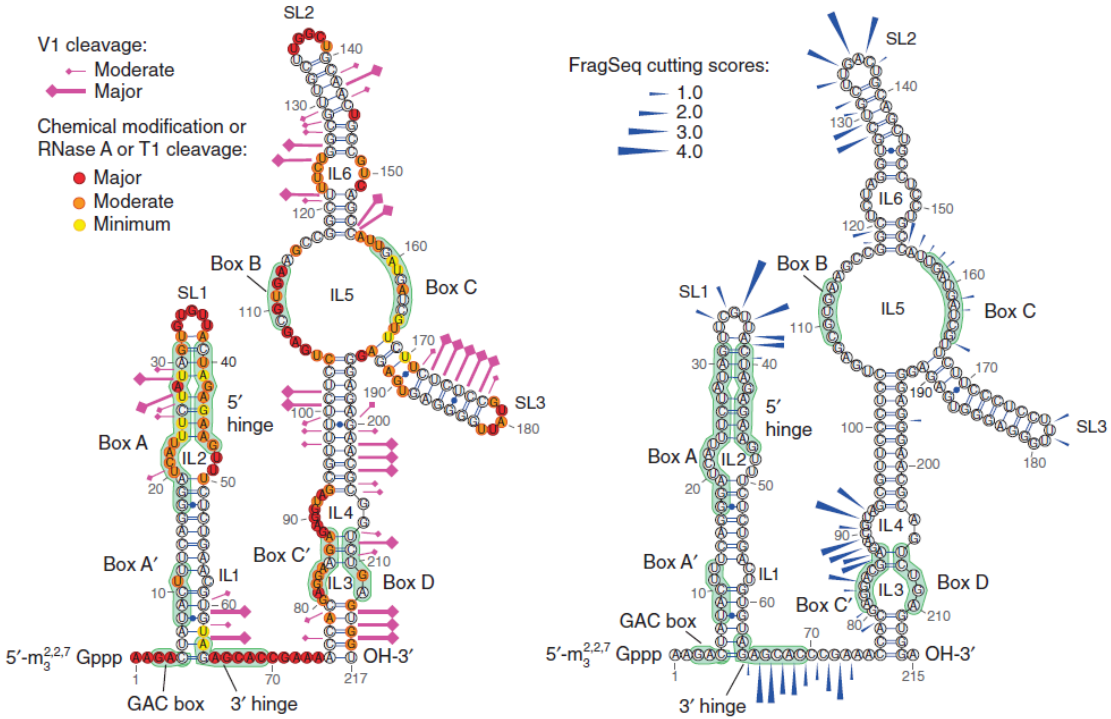
$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$



Benchmark: 1919 non-multimer/non-pseudoknotted sequence/structure pairs from the RNAstrand database (source Vienna Package web site)

Chemical/enzymatic probing to model 2D

- High-throughput secondary structure determination
- Reactivity/accessibility guide manual modeling choices

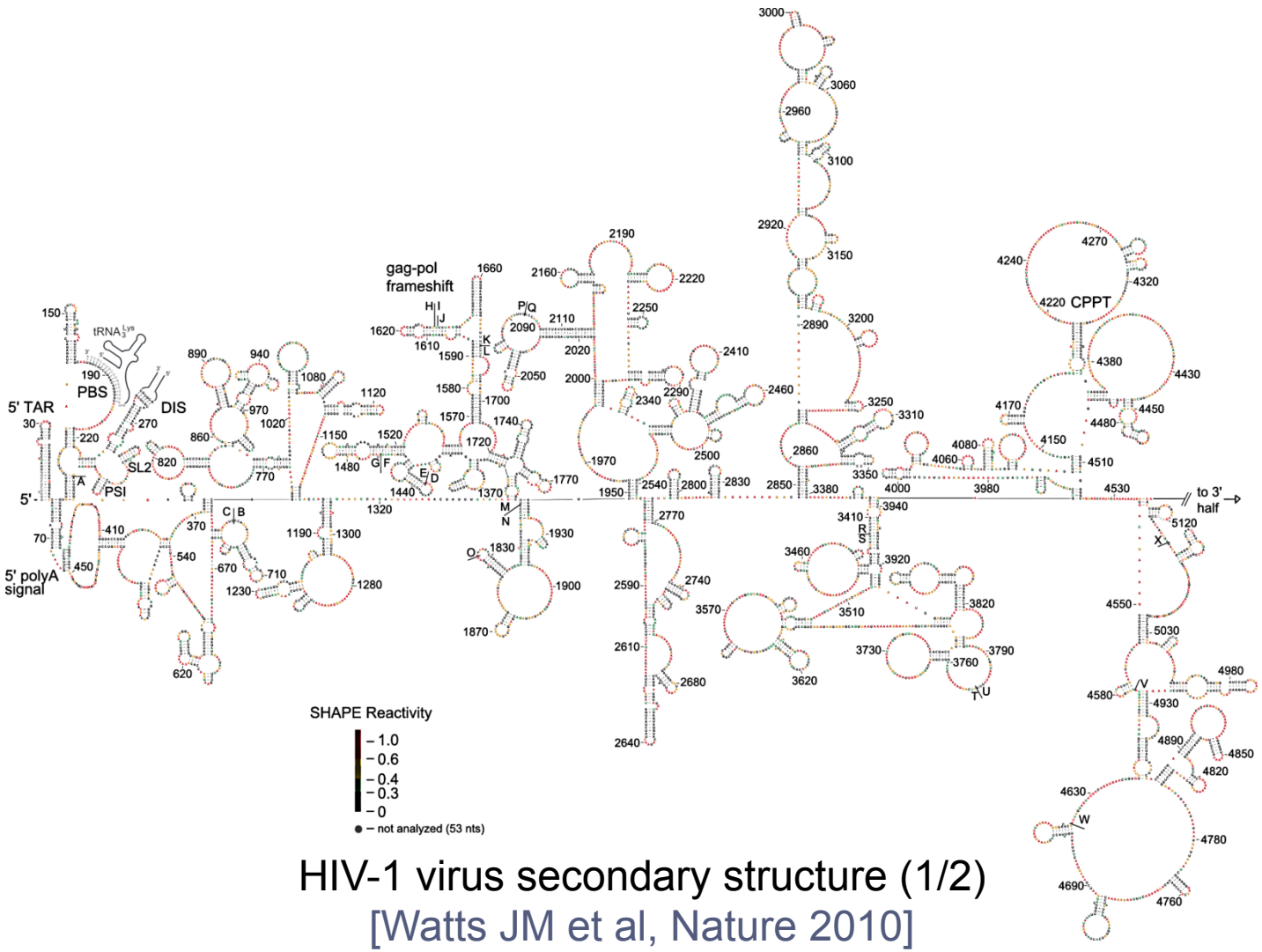


FragSeq method [Underwood *et al*, Nature Methods 2010]
(Images: VARNA)

- Inclusion as pseudo potentials within energy-models

[Lorenz *et al*, Bioinformatics 2015]

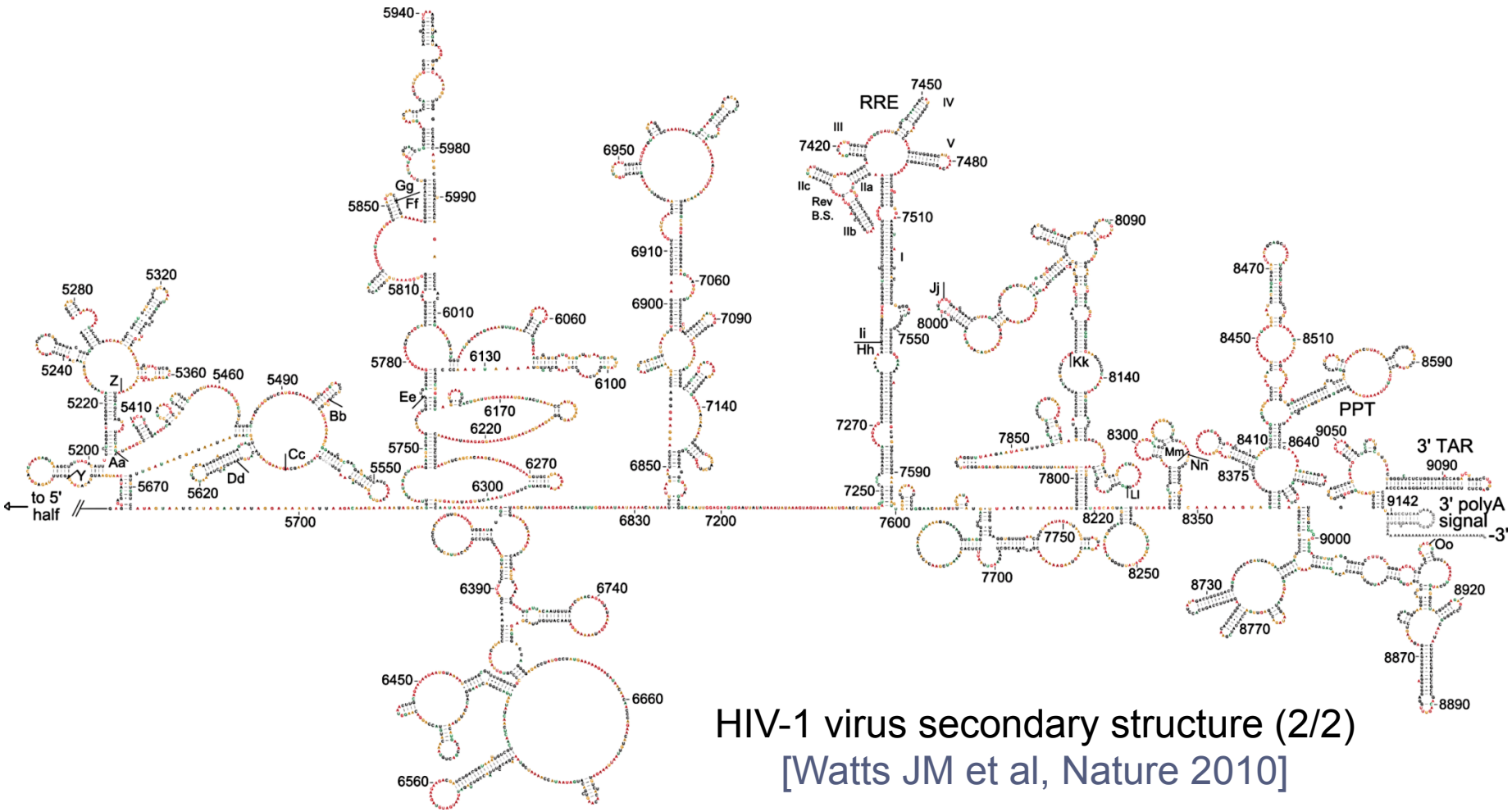
SHAPE probing to model 2D



HIV-1 virus secondary structure (1/2)

[Watts JM et al, Nature 2010]

SHAPE probing to model 2D



HIV-1 virus secondary structure (2/2)
[Watts JM et al, Nature 2010]

Lab: RNA folding basics

Write and test `Python` functions to:

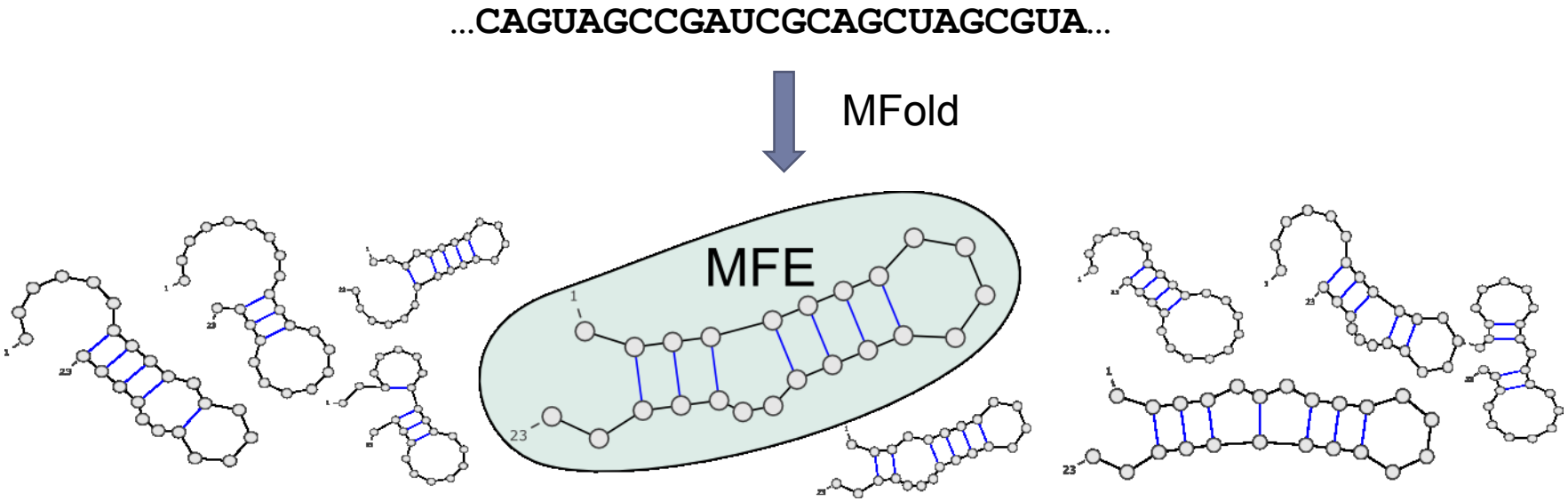
- Parse and print 2^{ary} structures
  - Dot-parenthesis notation ↔ List of base-pairs + length
  - Ex.: `“((..)(.).)”` ↔ `([(0,9),(1,4),(5,7)],10)`
- Compare alternative structures for a given RNA
  - Compute base-pair distance between two structures
  - Ex.: `“(.).(.)(..)” + “((...))(..)”` → `4`
- Run `RNAfold` and retrieve its MFE structure
- Benchmark `RNAfold`
  - Download and save `http://goo.gl/l0mx9c`
  - For each sequence, predict MFE and compare to structure
  - Report average base-pair distance

RNA Structure Prediction

Boltzmann ensemble
Partition function-based methods

Ensemble approaches in RNA folding

- RNA *in silico* paradigm shift:
  - From single structure, minimal free-energy folding…

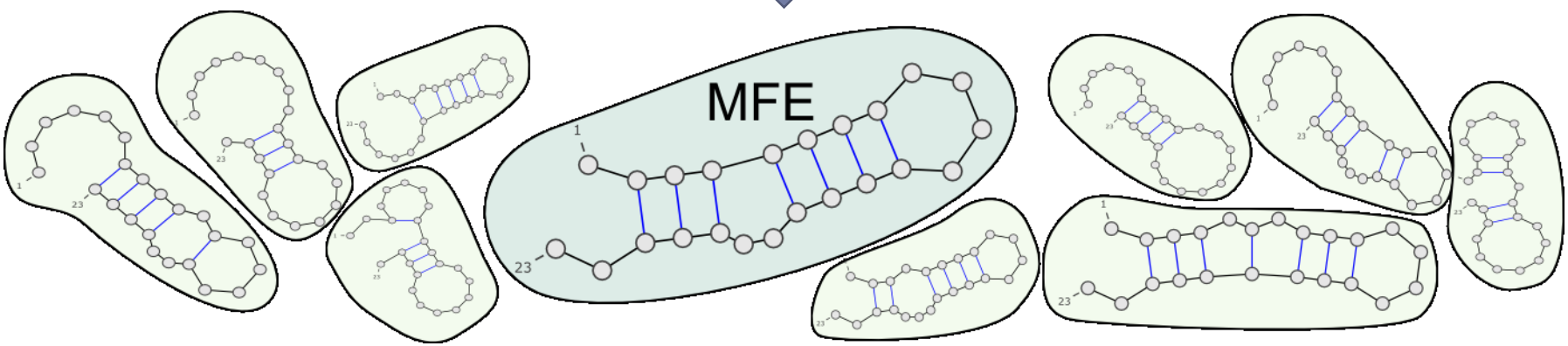


Ensemble approaches in RNA folding

- RNA *in silico* paradigm shift:
  - From single structure, minimal free-energy folding…
  - … to ensemble approaches.

...**CAGUAGCCGAUCGCAGCUAGCGUA**...

UnaFold, RNAFold, Sfold…



Thermodynamic equilibrium: Every secondary structure has probability

Boltzmann Probability $Prob(S) = \frac{e^{-\Delta G(S)/kT}}{Z}$

Partition Function $Z = \sum_{Struc\ S} e^{-\Delta G(S)/kT}$

[McCaskill, Biopolymers 1990]

→ Ensemble diversity? Structure likelihood? Evolutionary robustness?

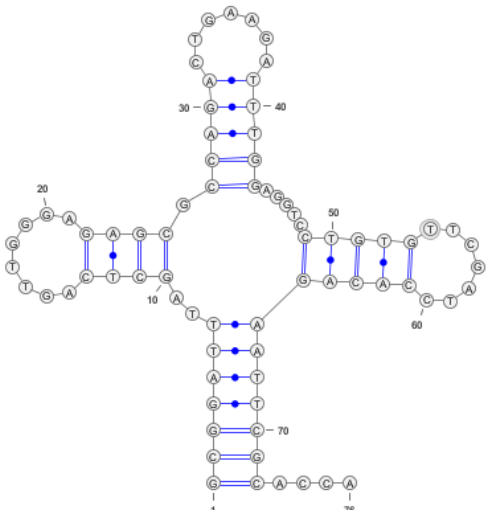
Partition function and statistical sampling

http://goo.gl/RRo6mG

Ensemble approaches indicate uncertainty and suggest alternative conformations

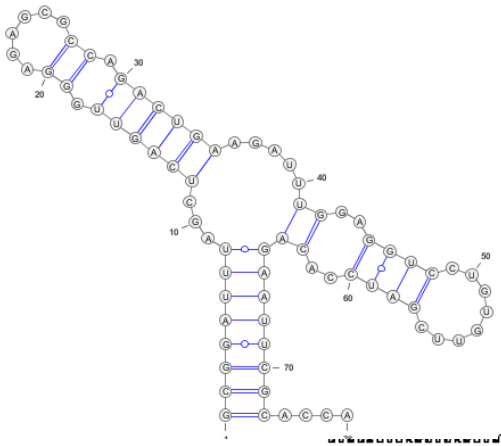
Example:

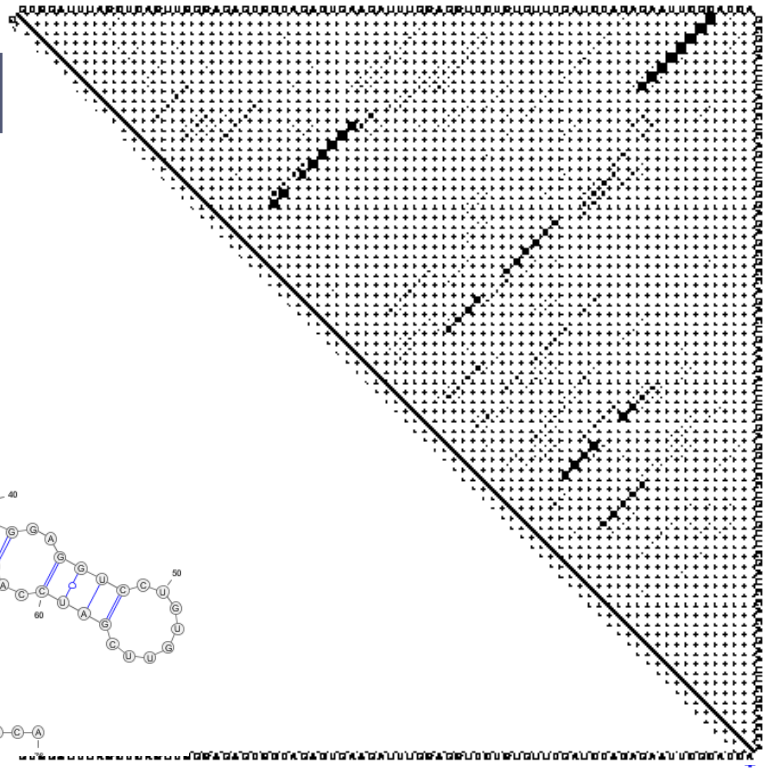
```
>ENA|M10740|M10740.1 Saccharomyces cerevisiae Phe-tRNA. : Location:1..76
GCGGATTTAGCTCAGTTGGGAGAGCGCCAGACTGAAGATTTGGAGGTCCTGTGTTCGATCCACAGAATTCGCACCA
```



Native structure

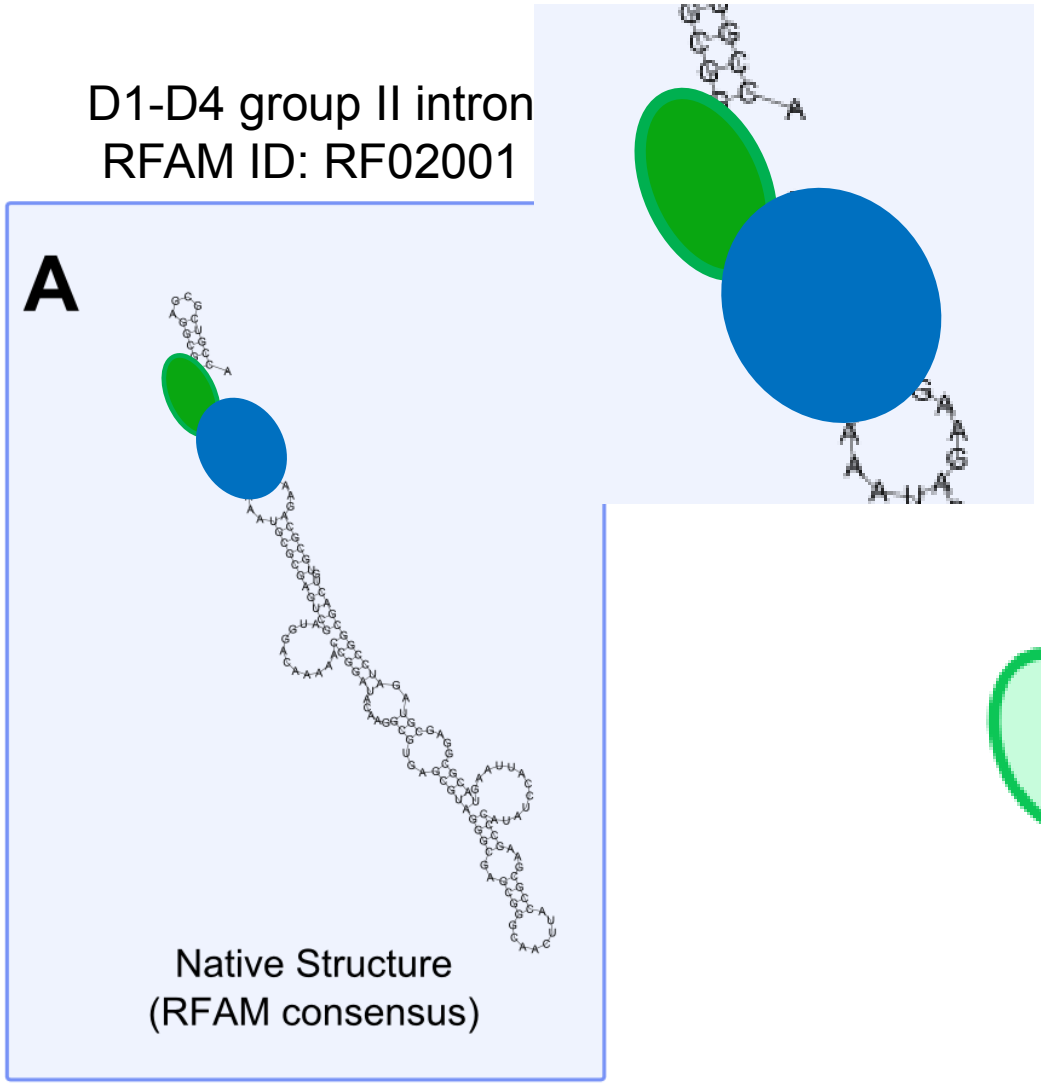
RNAFold -p

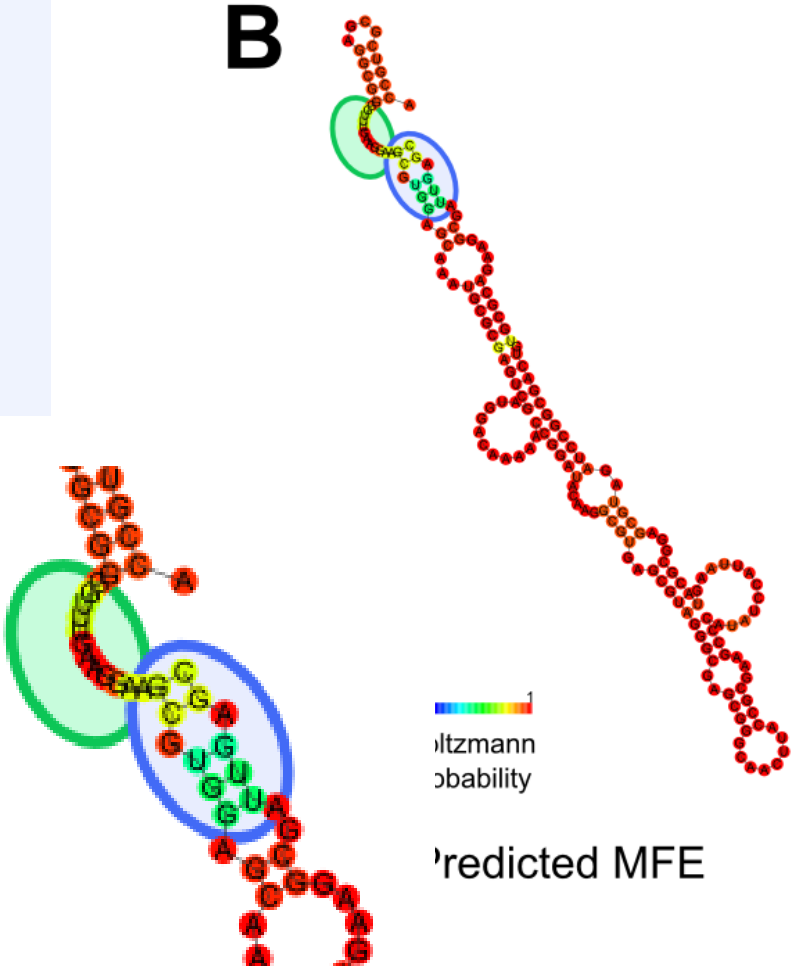




Assessing the reliability of a prediction

D1-D4 group II intron
RFAM ID: RF02001

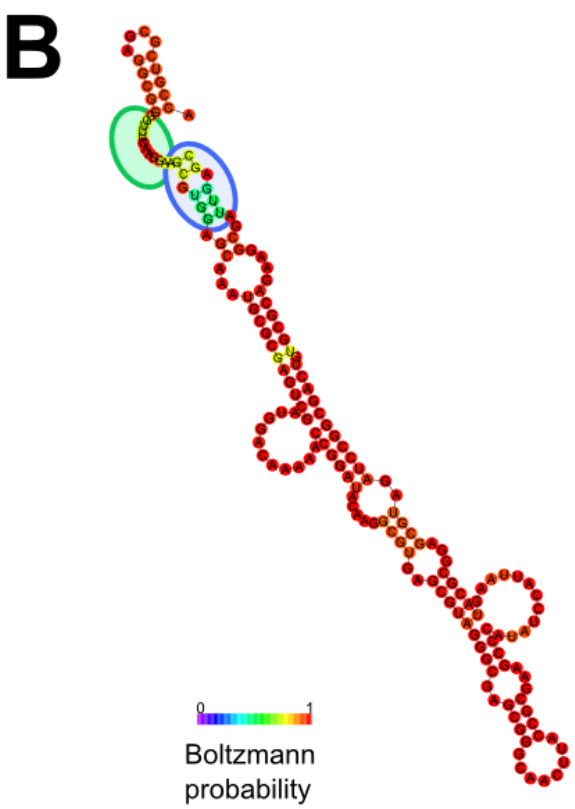




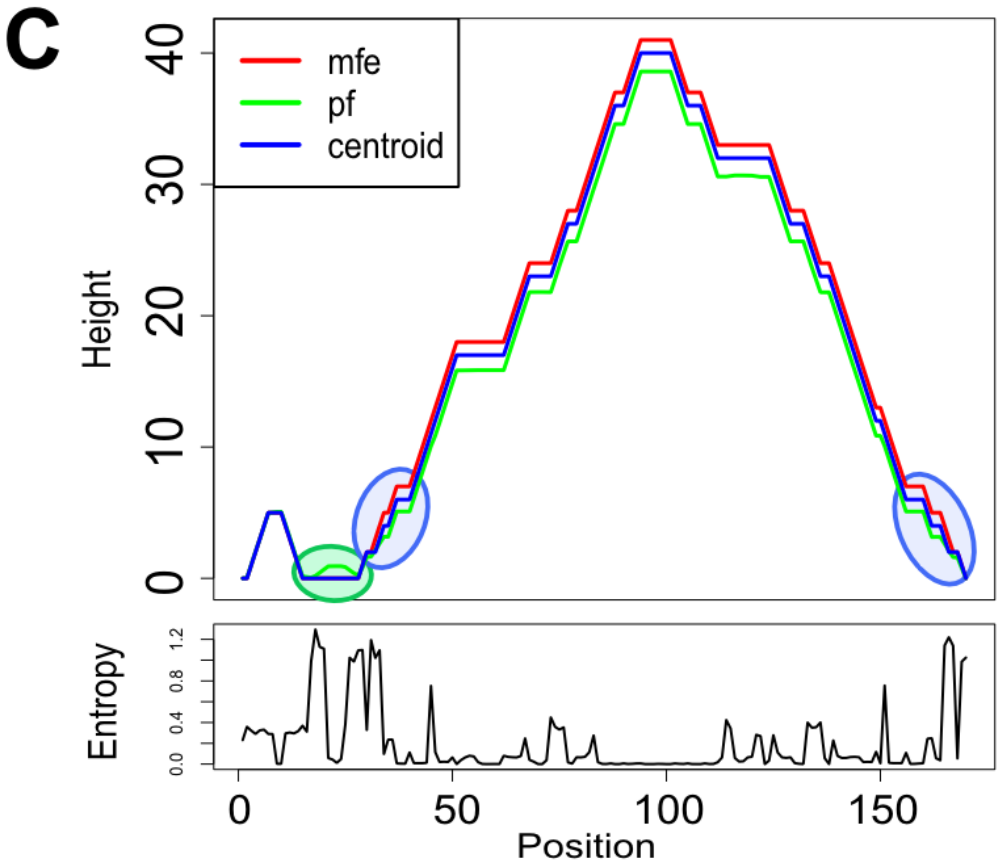
RNAFold [Gruber AR et al. NAR 2008]

Assessing the reliability of a prediction

D1-D4 group II intron
A. Capsulatum



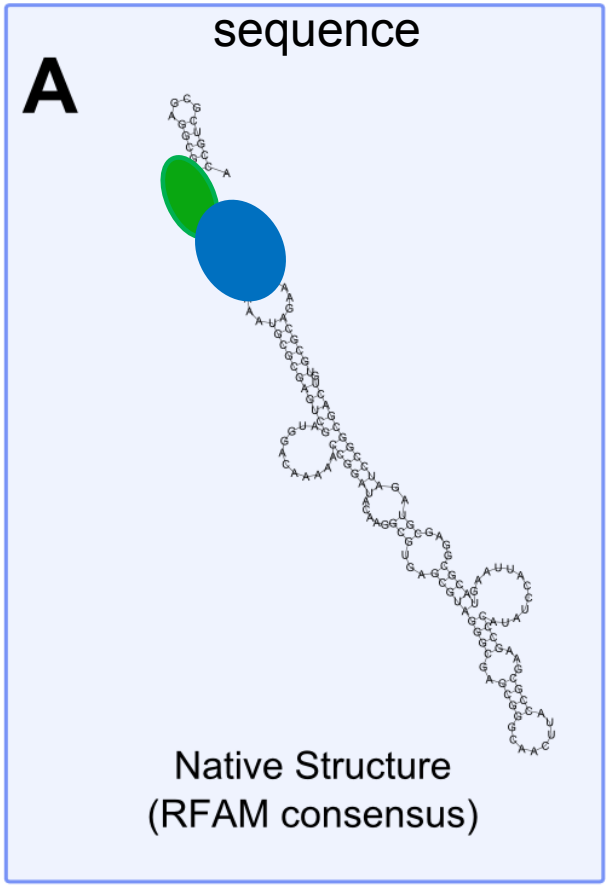
Predicted MFE

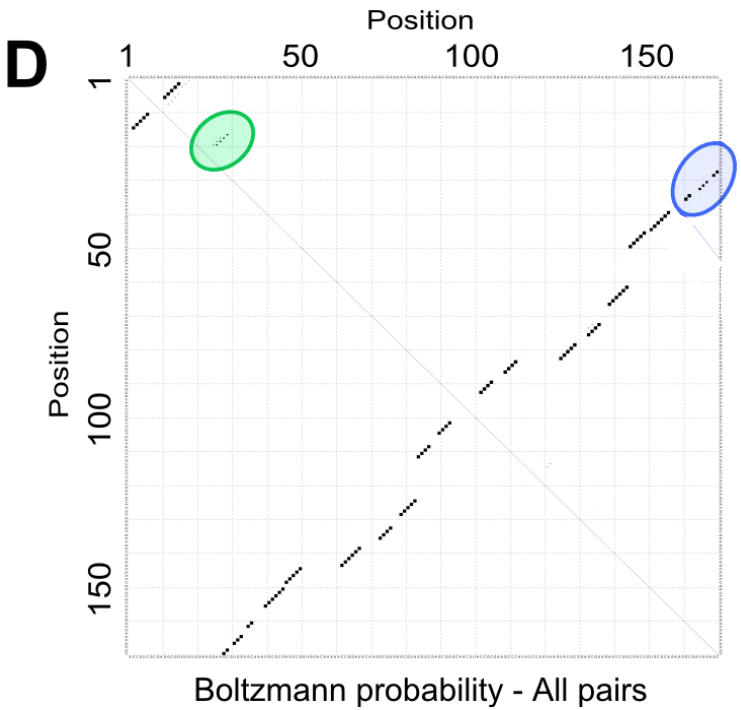


RNAFold [Gruber AR et al. NAR 2008]

Assessing the reliability of a prediction

D1-D4 group II intron
A. Capsulatum
sequence



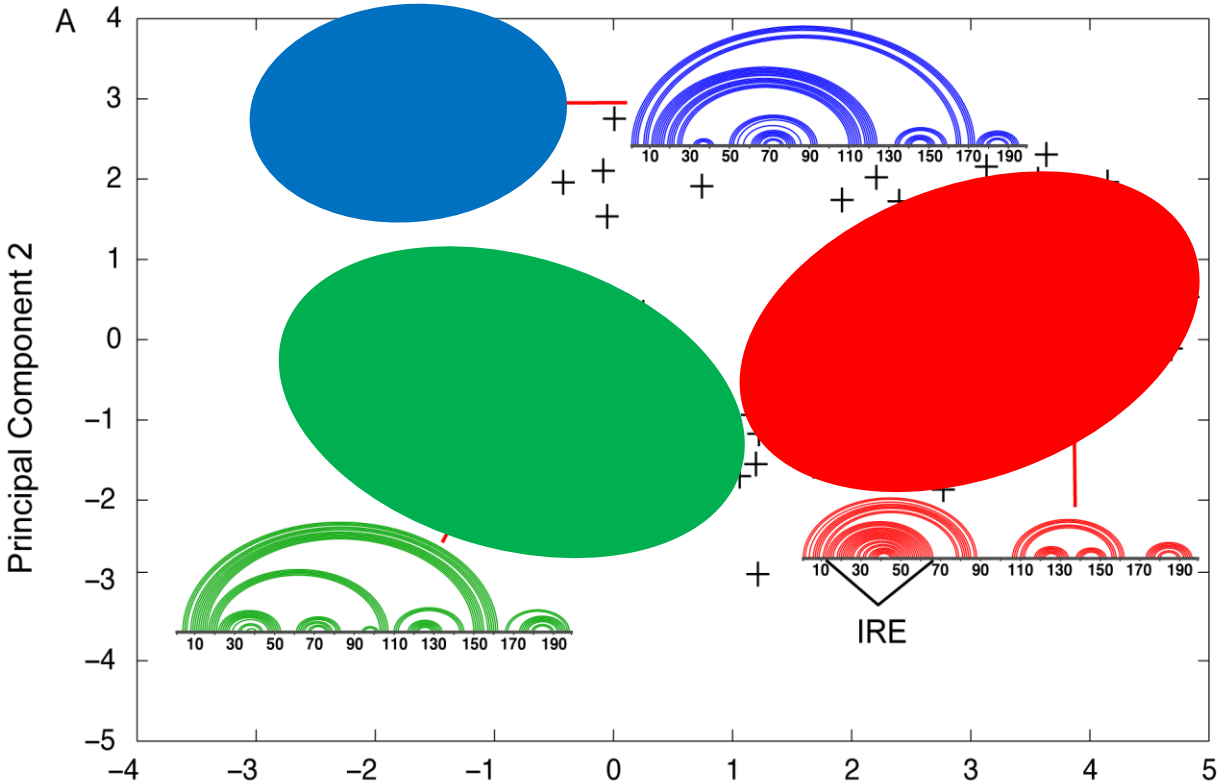


- Low BP probabilities indicate uncertain regions
- BP>99% → PPV>90% (BP>90% → PPV>83%)
  [Mathews, RNA 2004]
- Visualizing probs in the context of structure helps refining predicted structures.

RNAFold [Gruber AR et al. NAR 2008]

Sensitivity to (single-point) mutations

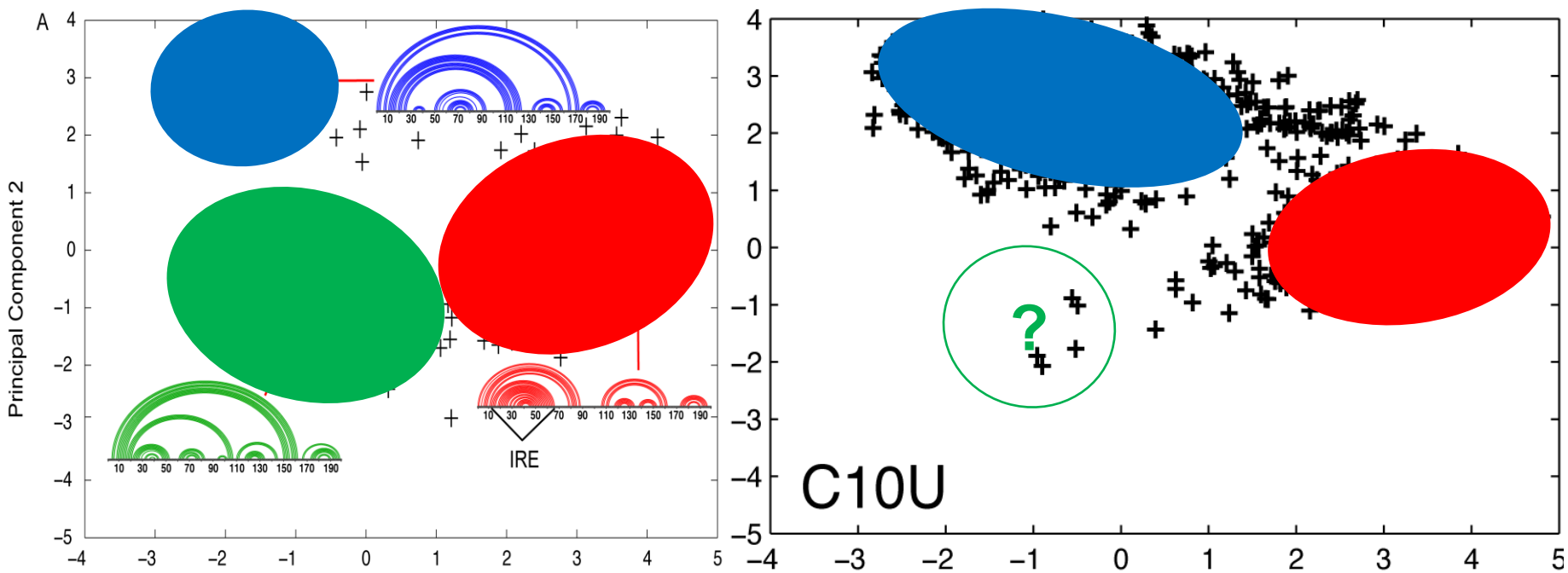
- Boltzmann Sampling → Clustering (+PCA)



[Halvorsen M *et al*, PLOS Gen 2010]

Sensitivity to (single-point) mutations

- Boltzmann Sampling → PCA → Clustering



[Halvorsen M *et al*, PLOS Gen 2010]

C10U associated with Hyperferritinemia cataract syndrome

Lab: Partition function approaches

In `Python`, implement :

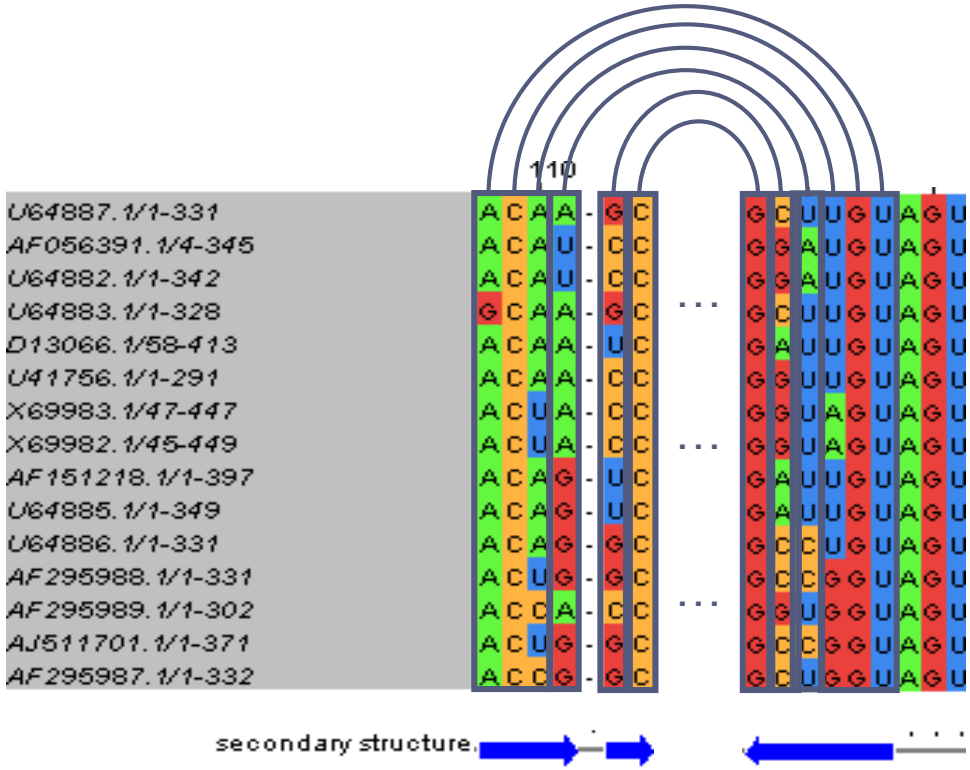
- A *Nussinov-style* DP counting algorithm
  - Input: RNA sequence *w* + Min. base pair distance *theta*
  - Output: #Secondary structures compatible with (*w*,*theta*)
  - Ex.: "AU",0 → 1 "AU",1 → 0 "ACU",1 → 1
    "GGGAAACCC",3 → 20
- (Uniform) stochastic backtrack
  - Propose a validation procedure
- A basic agglomerative clustering procedure
  - At each step pick the closest structures and merge them
  - Stop when k=10 clusters are found
- Benchmark `RNAsubopt -p` + Clustering

Comparative methods and the pitfalls of benchmarks

The BRaliBase dent—a tale of benchmark design and interpretation

[Löwes, Chauve, Ponty, Giegerich, Brief Bioinfo 2016]

Evolution to the rescue: Comparative approaches for structured RNAs



RFAM Bacterial RNase P class B Alignment
RF00011, rendered using JalView

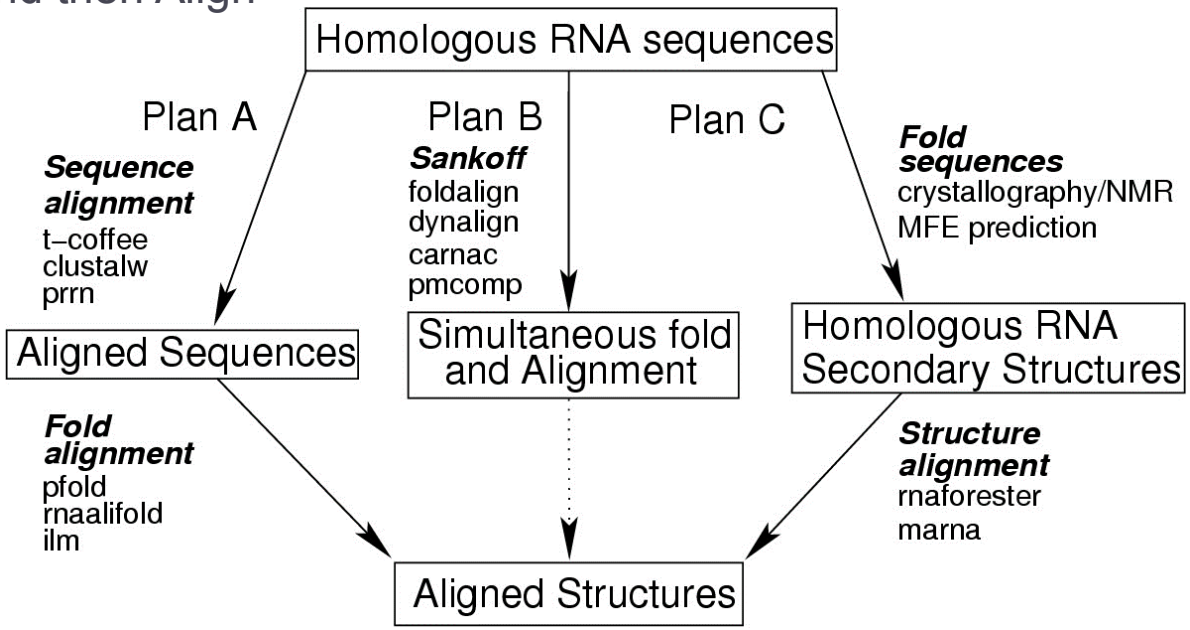
- Structure (=phenotype) more typically conserved than sequence
- Covariations/compensatory mutations hint towards shared structure

Evolution to the rescue: Comparative approaches for structured RNAs

- **Idea:** If Sequence Alignment available, then **fold** columns!

  RNAAlifold [Bernhardt et al, BMC Bioinfo 2008]

- From unaligned sequences, chicken and egg paradox (again!)
  - Align and then Fold
  - Fold and align simultaneously (Sankoff) → $\Theta(n^{3m})/\Theta(n^{2m})$ time/memory
  - Fold and then Align

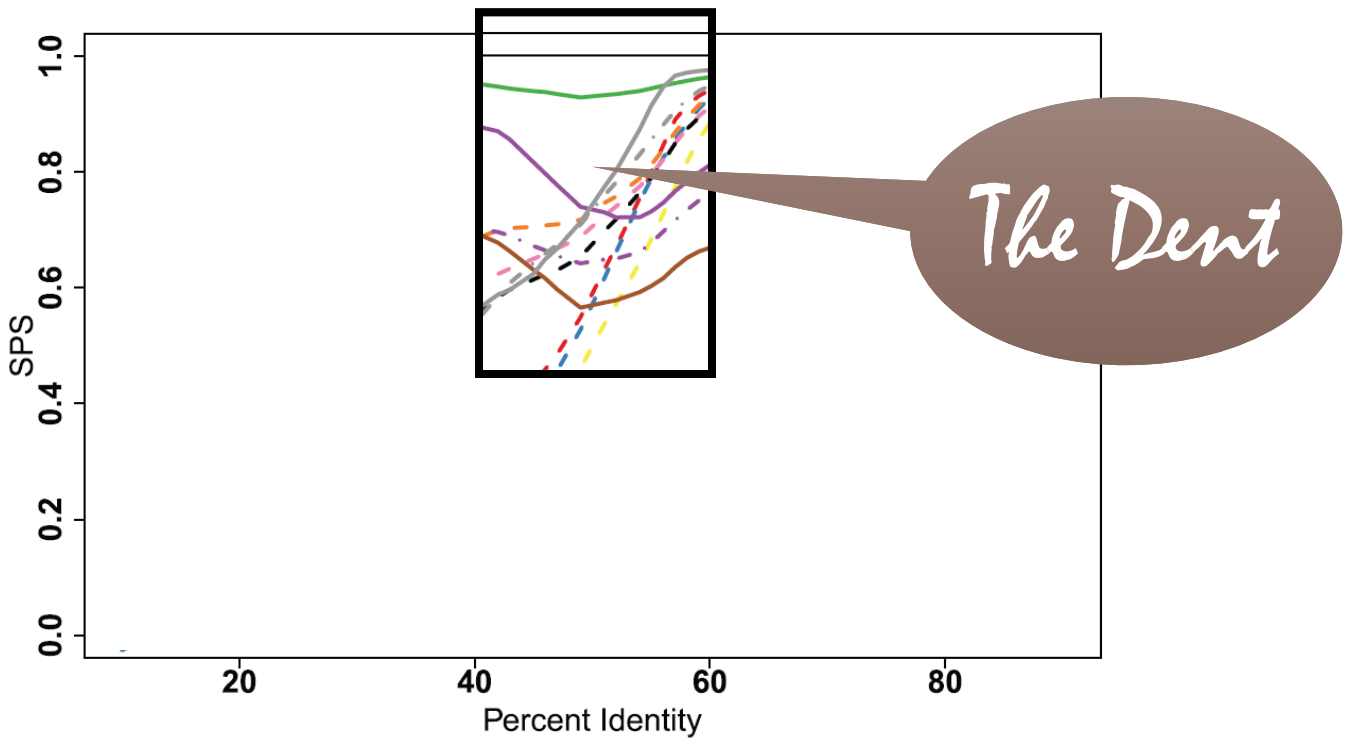


[Gardner & Giegerich, BMC Bioinfo 2004]

BRAliBase

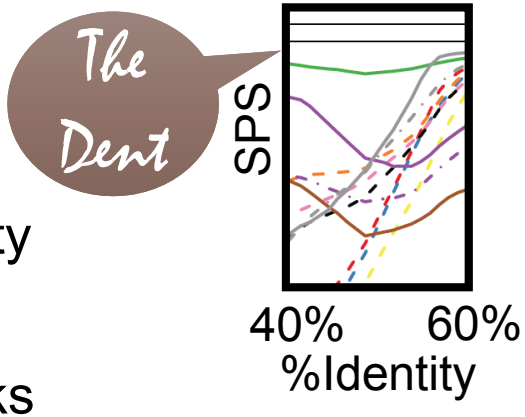
[Gardner & Giegerich, BMC Bioinfo 2004]
[Gardner,Wilm & Washietl, NAR, 2005]
[Wilm, Mainz & Steger, Alg Mol Biol, 2006]

- Benchmark of sequence/alignment since 2004-2005
- Cited ~800 times, *de facto* standard for new tools
- Based on sequence/structure alignments for several RNA families

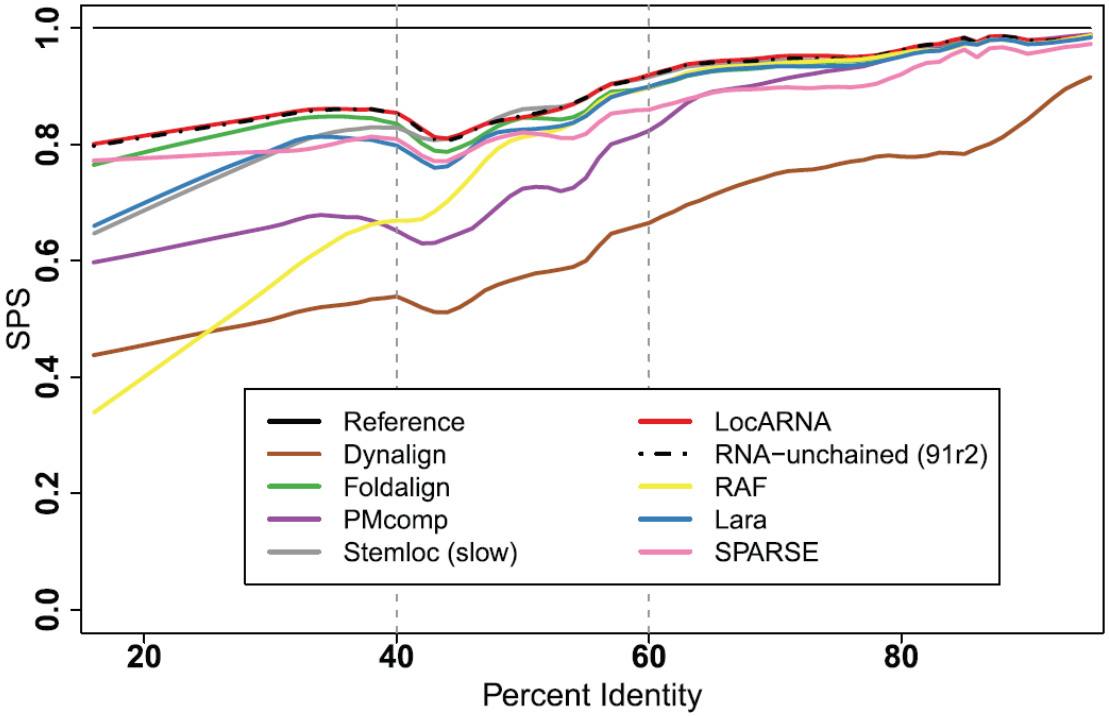


Quality Score: Sum-of-Pairs Score (SPS) = $\frac{\text{\#Correctly predicted chars pairs}}{\text{\#Chars pairs in curated alignments}}$

The BRAliBase dent

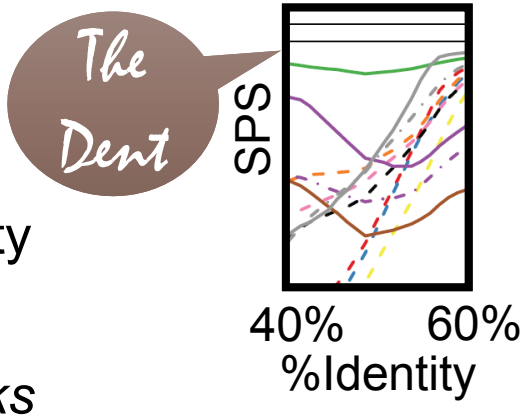


- The Dent = Quality drop in 40%-60% sequence identity
- Tool-independent phenomenon found in 2005
- Reproduced by following tools & improved benchmarks
- Inspiration for new algorithms, creative conjectures…

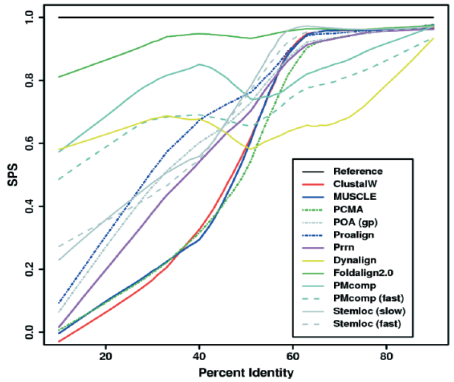


[Gardner, Wilm & Washietl, NAR, 2005]
[Lindgreen, Gardner & Krogh, Alg. Mol. Biol., 2006]

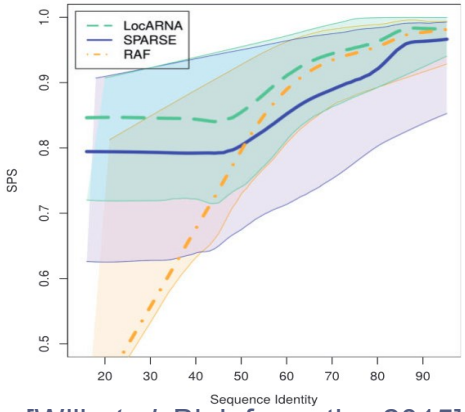
The BRAliBase dent



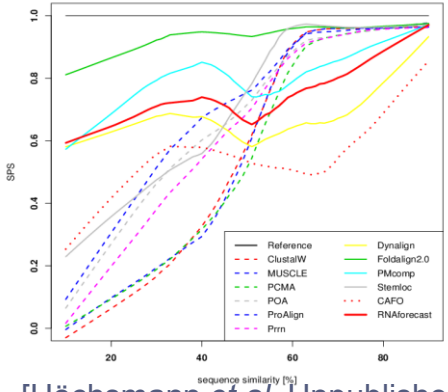
- The Dent = Quality drop in 40%-60% sequence identity
- Tool-independent phenomenon found in 2005
- Reproduced by *following tools & improved benchmarks*
- Inspiration for *new algorithms*, creative conjectures…



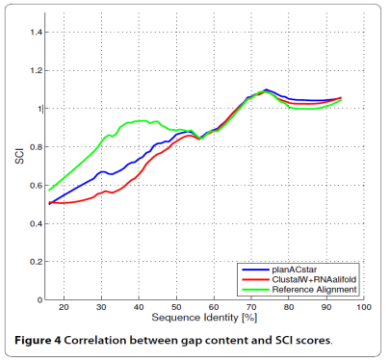
[Gardner *et al*, NAR 2005]



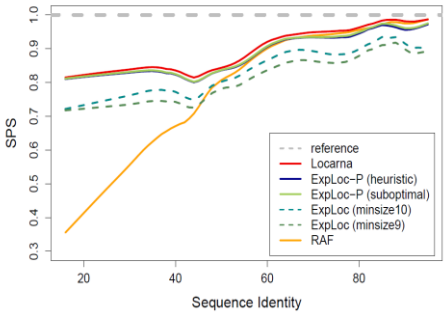
[Will *et al*, Bioinformatics 2015]



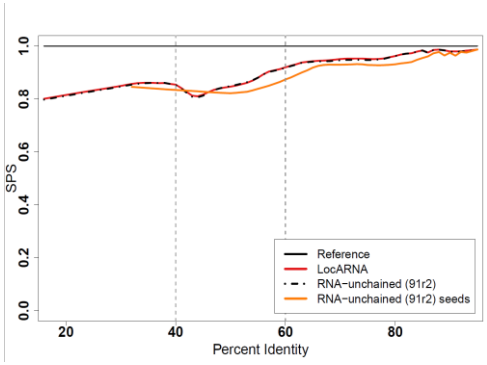
[Höchsmann *et al*, Unpublished]



[Bremges *et al*, BMC Bioinfo, 2010]

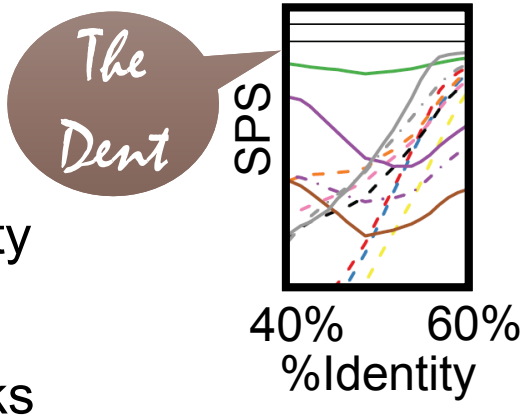


[Schmiedl *et al*, RECOMB 2012]



[Bourgeade *et al*, J Comp Biol, 2015]

The BRAliBase dent



- The Dent = Quality drop in 40%-60% sequence identity
- Tool-independent phenomenon found in 2005
- Reproduced by following tools & improved benchmarks
- Inspiration for new algorithms, *creative conjectures*…

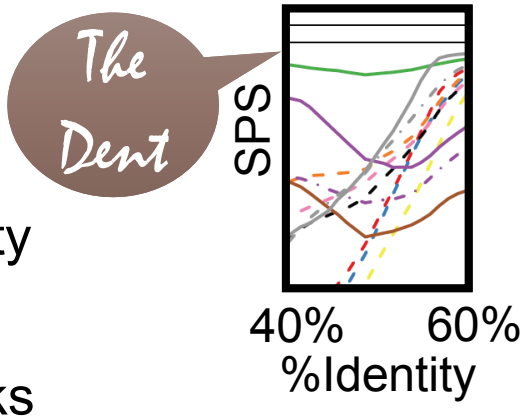
The dent marks the transition between sequence and structure-driven alignments

The dent identifies inconsistent practices by alignment curators

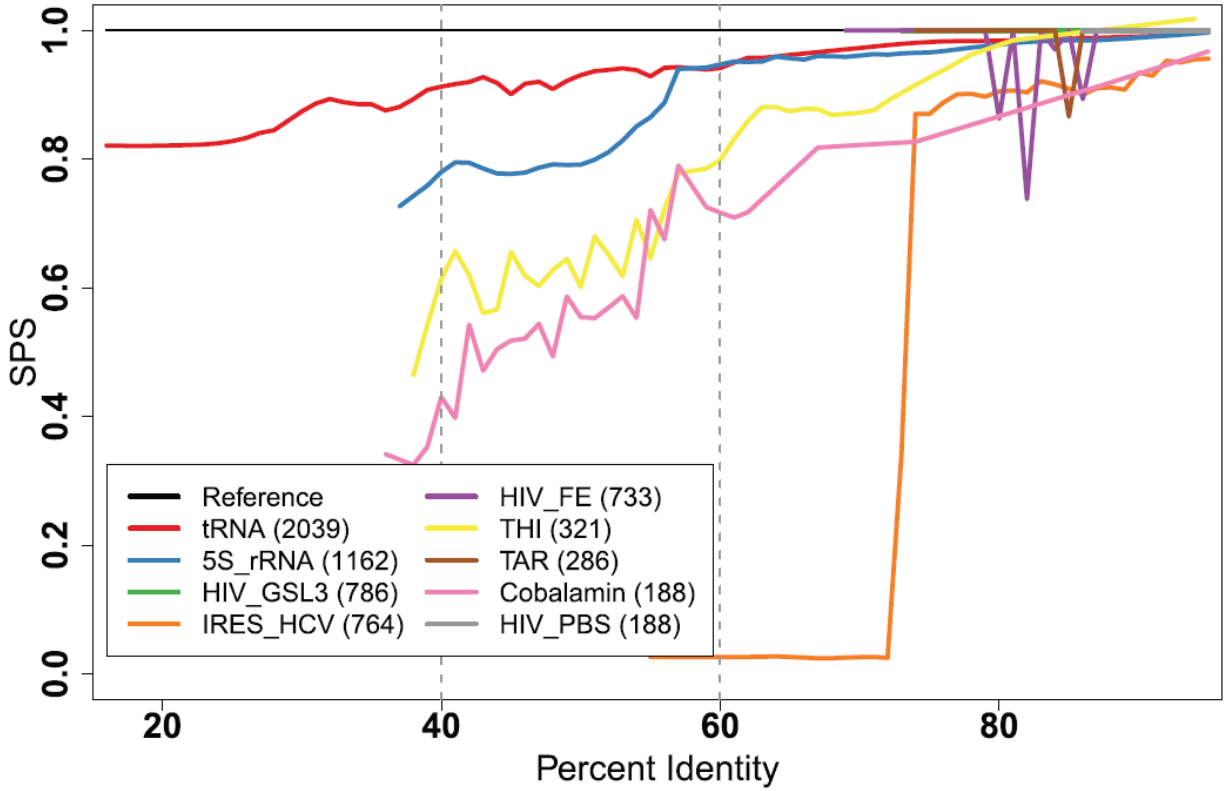
The dent undeniably proves the existence of the great spaghetti monster in the sky…

(Very) probably not…

The BRAliBase dent

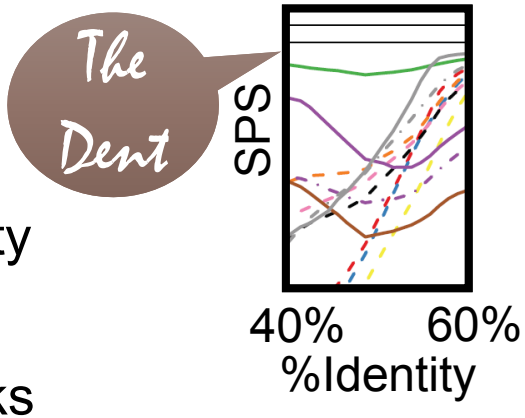


- The Dent = Quality drop in 40%-60% sequence identity
- Tool-independent phenomenon found in 2005
- Reproduced by following tools & improved benchmarks
- Inspiration for new algorithms, creative conjectures…

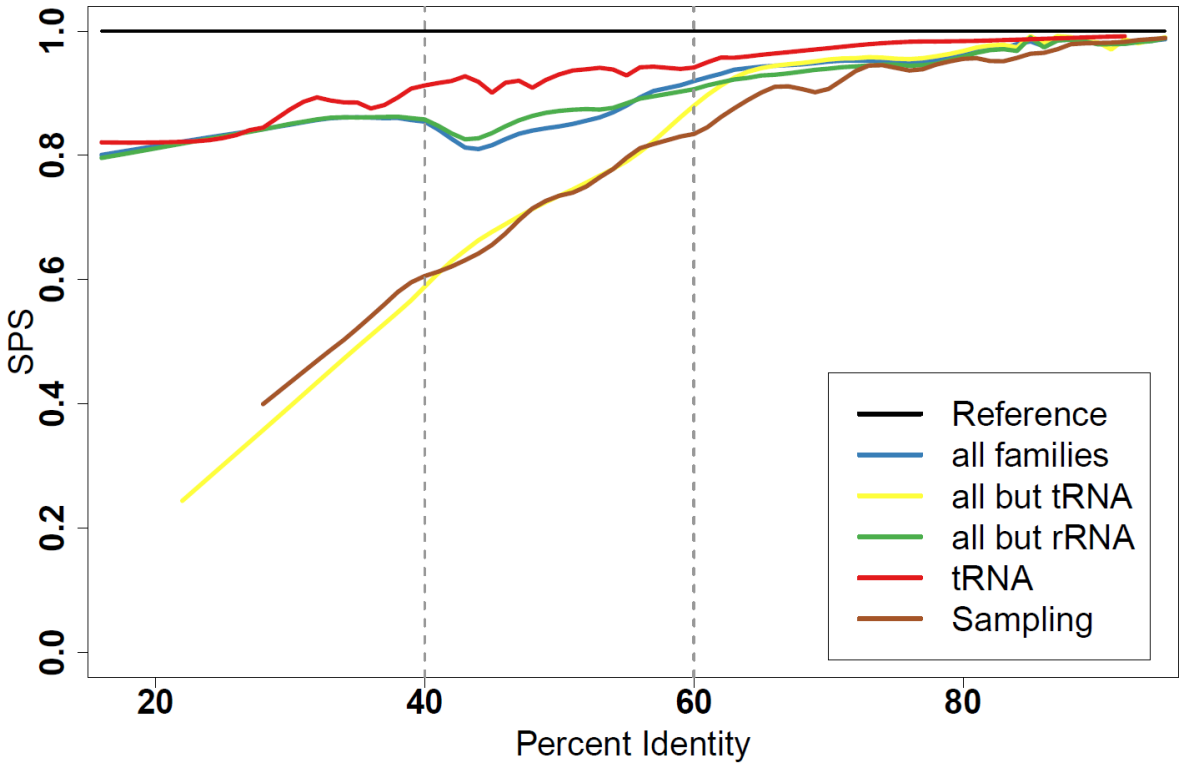


M'kay…
so what?
(still no dent)

The BRAliBase dent



- The Dent = Quality drop in 40%-60% sequence identity
- Tool-independent phenomenon found in 2005
- Reproduced by following tools & improved benchmarks
- Inspiration for new algorithms, creative conjectures…
- … purely an artifact due to heavy bias towards well-predicted tRNAs!



tRNAs are overly dominant for low identities and very well-predicted

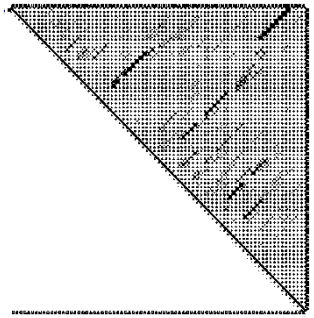
The dent simply occurs when they cease to dominate.

Conclusion

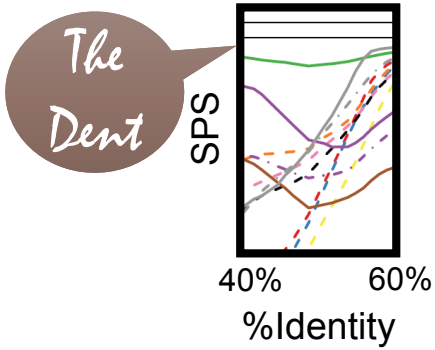
Conclusion

- More to RNA than single-structure prediction methods
- Most methods run in a few seconds, and are available online!

- Thermodynamic equilibrium: Making statements about the complete (exponential) (sub)optimal space (in polynomial time)
  - Assess reliability (Boltzmann probability)
  - Detect presence of alternative conformers (Dot-plot)
  - Identify dominant structures (Boltzmann sampling + clustering)



- Comparative approaches: Mature methods (LocARNA) significantly outperform single-sequence predictions
  - Avoid using structure-agnostic sequence MSAs
  - Benchmarks must be taken with a grain of salt…
  - … and should not be the sole driving force for methodological development!



The future

- RNA Design
  - Inverse folding = Synthesize RNA folding into a predefined structure
  - Gap between theory (almost nothing) and practice (design of regulatory networks)
  - Many software, hard to decide which one to choose for a given task

- RNA Kinetics: Boltzmann ensemble approaches postulate equilibrium … but RNAs may have short life span (+co-transcriptional folding)
  - Probably no efficient *ab initio* combinatorial approaches (NP-hard problems)
  - Tools to study of RNA >100nts will require collaborations between App. Maths, biochemistry and computer science