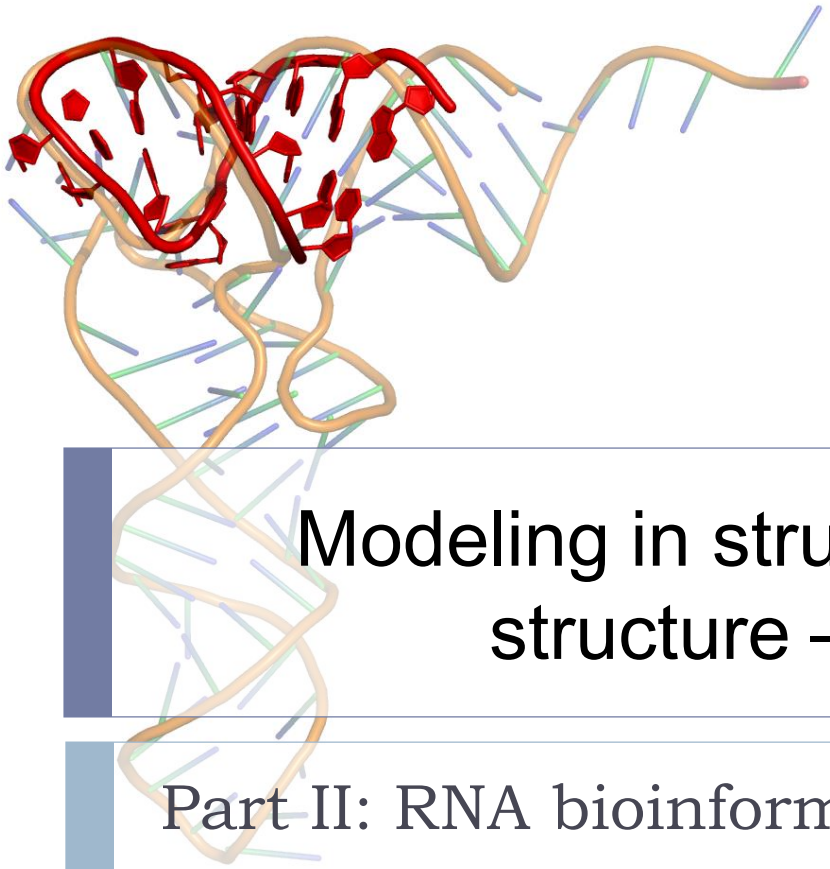
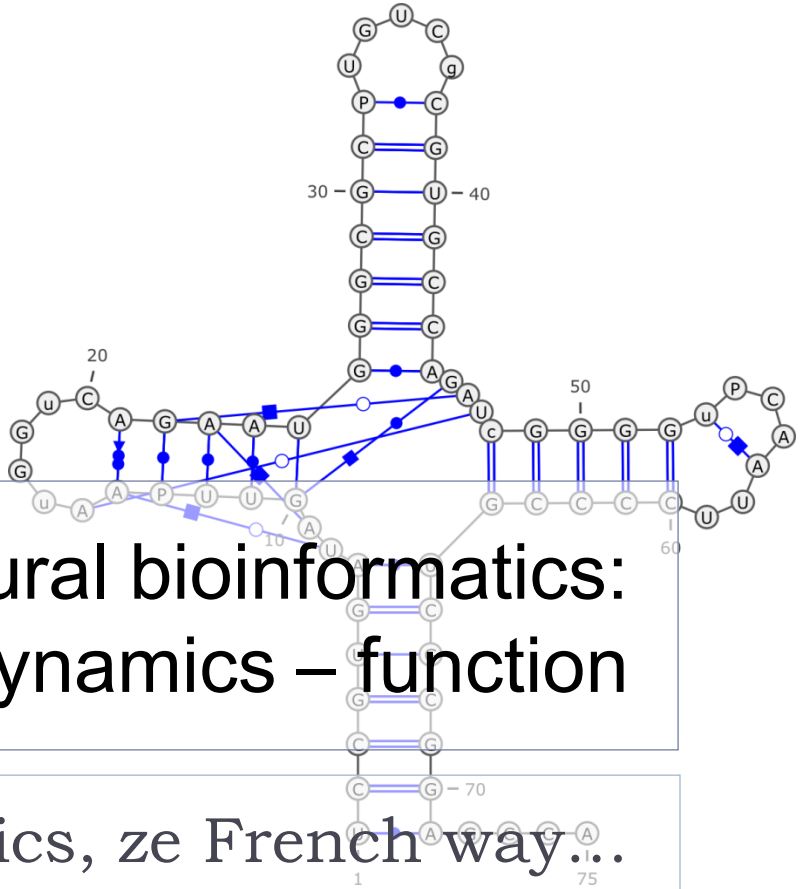
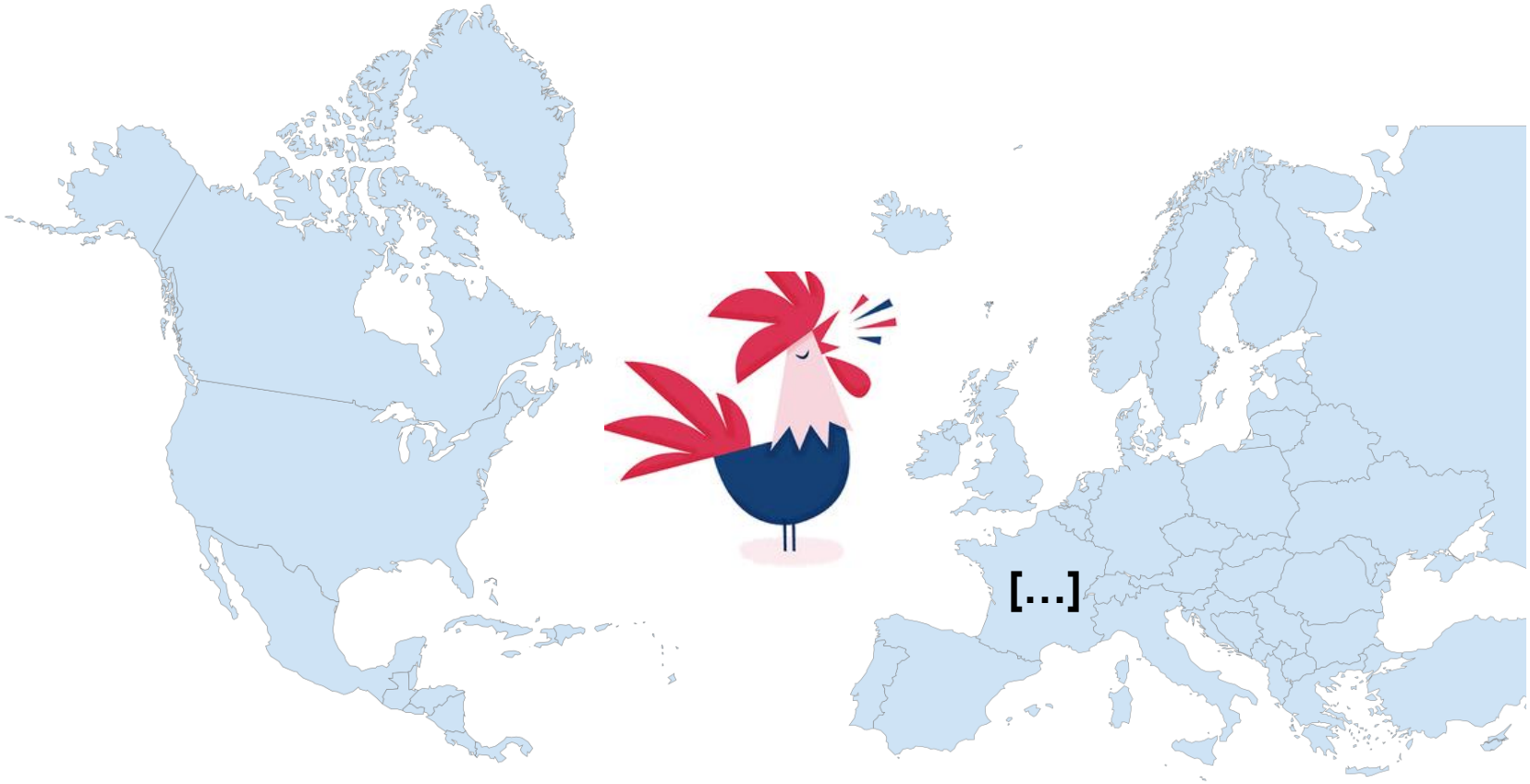




Modeling in structural bioinformatics: structure – dynamics – function

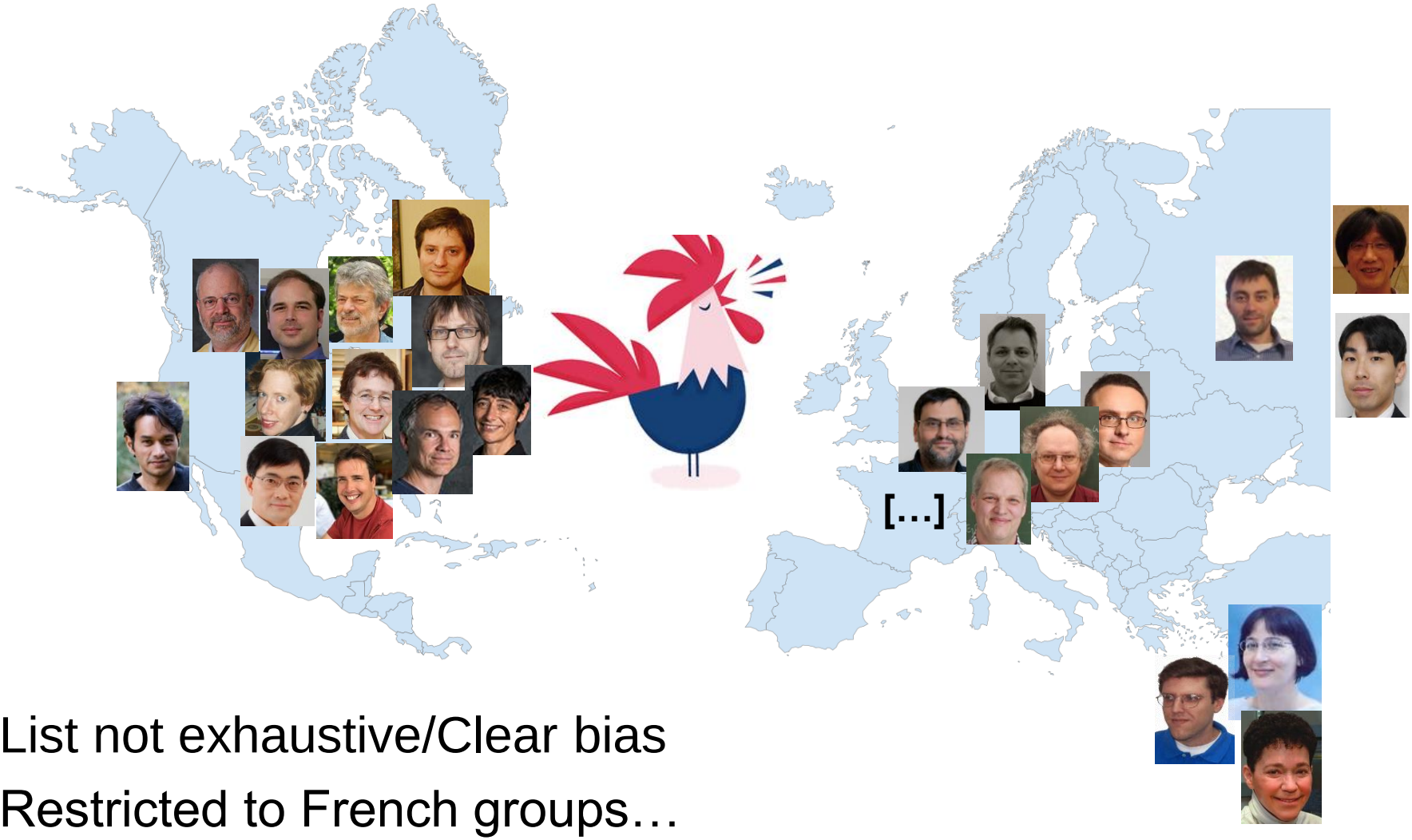
Part II: RNA bioinformatics, ze French way…

The RNA (Bioinfo) world



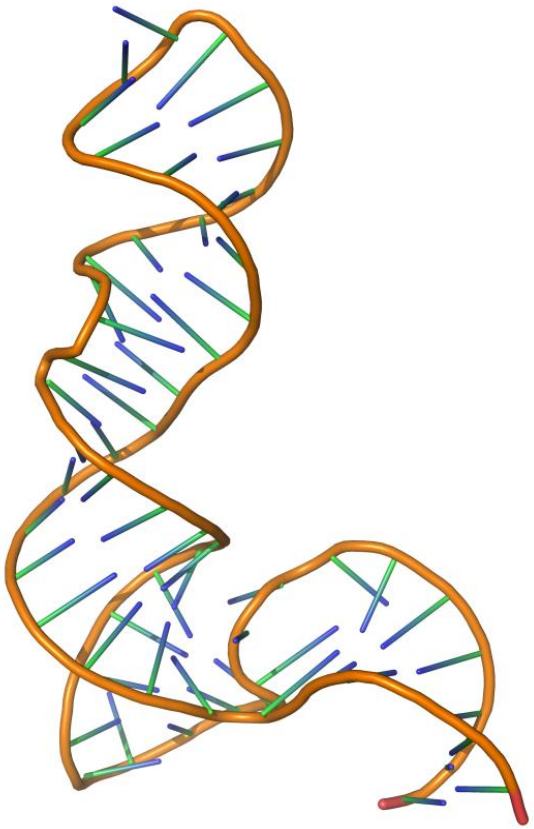
- List not exhaustive/Clear bias
- Restricted to French groups…
- … focusing on speakers at the MASIM kickoff meeting

The RNA (Bioinfo) world

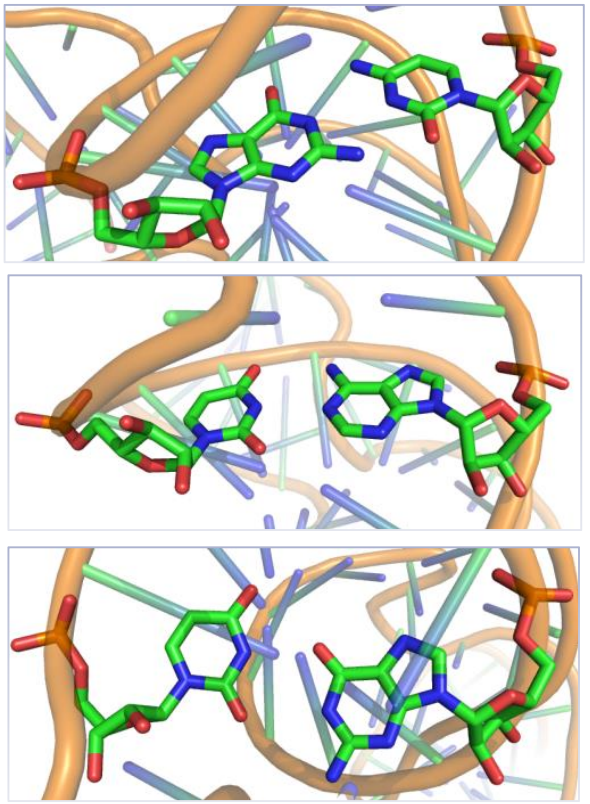


- List not exhaustive/Clear bias
- Restricted to French groups…
- … focusing on speakers at the MASIM kickoff meeting

Why RNA folds



5s rRNA (PDB ID: 1UN6)



G/C

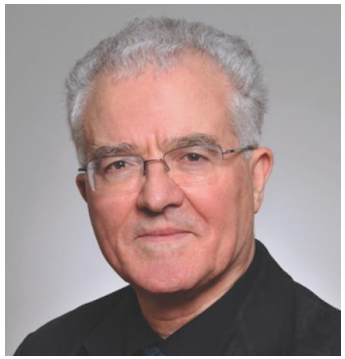
U/A

U/G

Canonical base-pairs

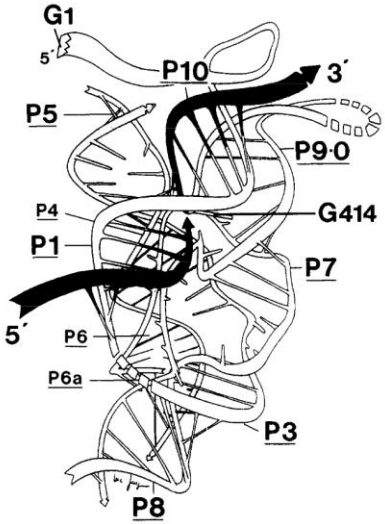
RNA folding = Hierarchical stochastic process driven by/resulting in the pairing (hydrogen bonds) of a subset of its bases.

A French pioneer in RNA Structural Modeling



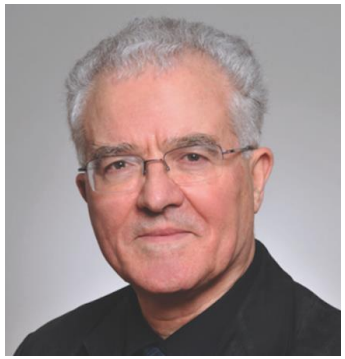
E Westhof
IBMC, Strasbourg

Tetrahymena group I intron



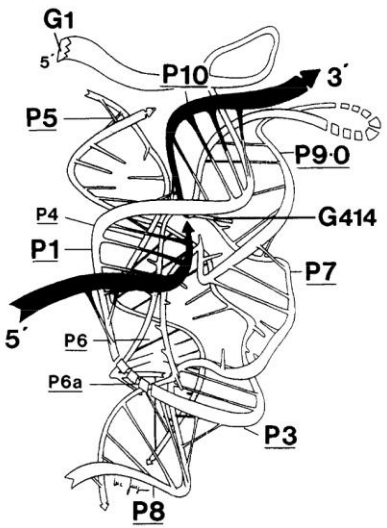
Predicted (manually)
[Westhof Michel, 1990]

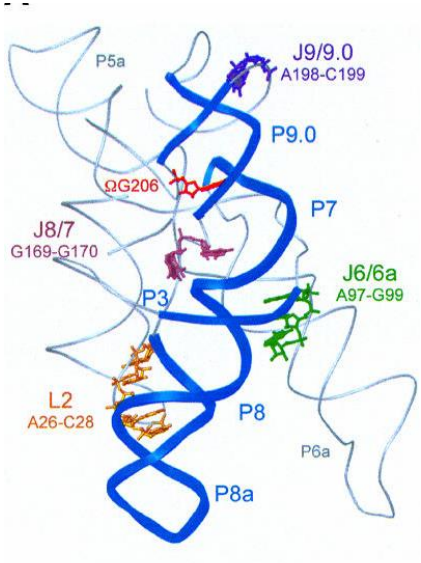
A French pioneer in RNA Structural Modeling



E Westhof
IBMC, Strasbourg

Tetrahymena group I intron

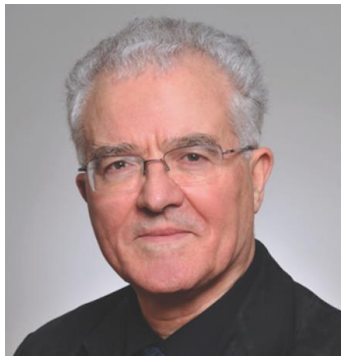




Predicted (manually)
[Westhof Michel, 1990]

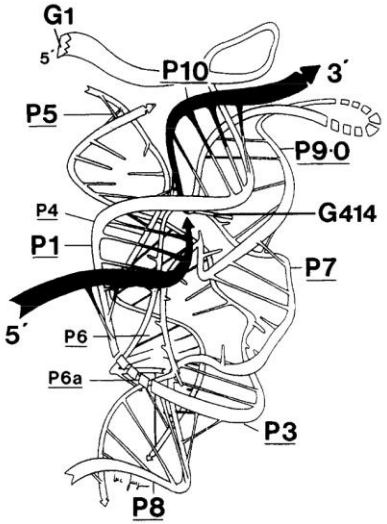
Crystal
[Adams *et al*, 2004]

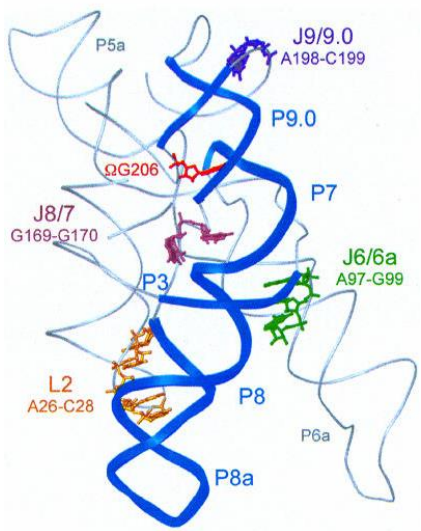
A French pioneer in RNA Structural Modeling

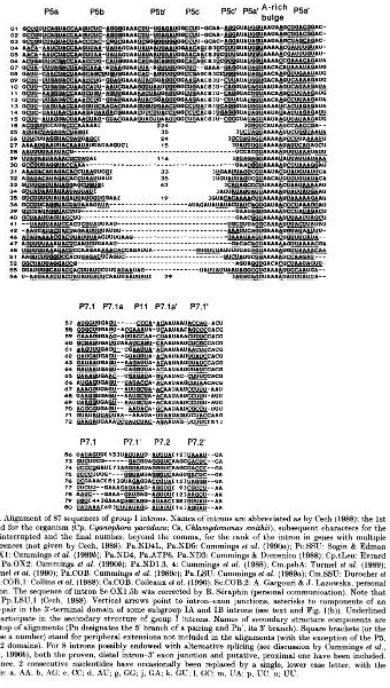


E Westhof
IBMC, Strasbourg

Tetrahymena group I intron



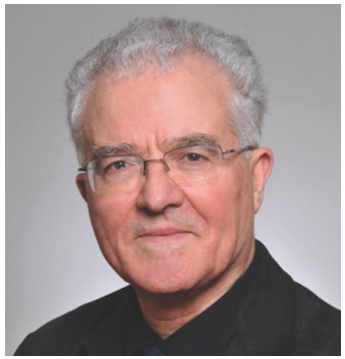




Predicted (manually) [Westhof Michel, 1990]

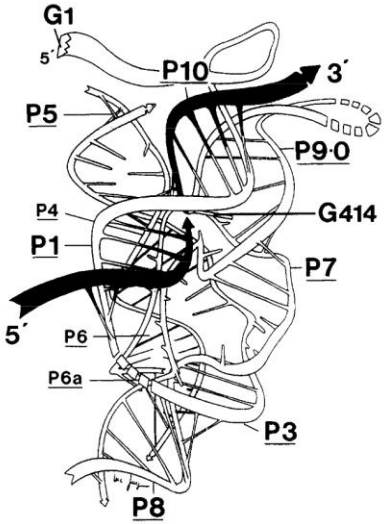
Crystal [Adams *et al*, 2004]

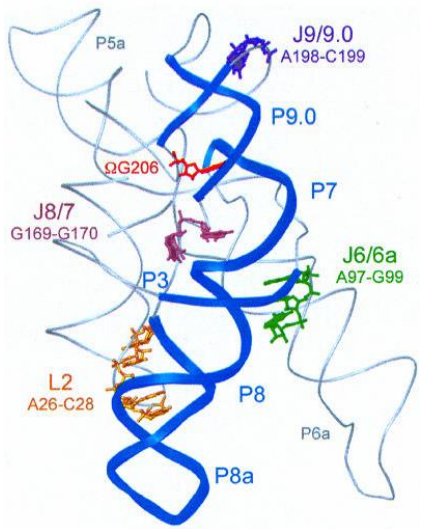
A French pioneer in RNA Structural Modeling

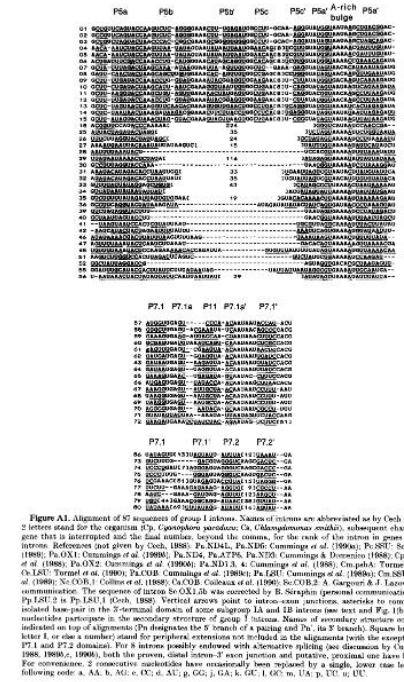


E Westhof
IBMC, Strasbourg

Tetrahymena group I intron



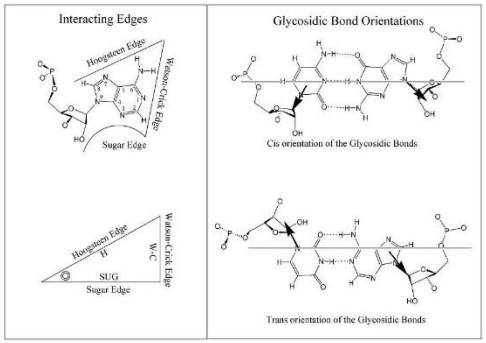


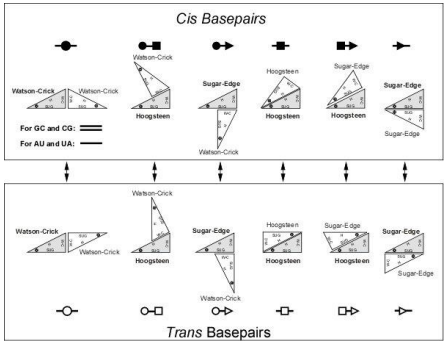


Predicted (manually) [Westhof Michel, 1990]

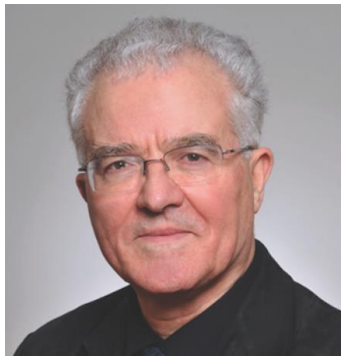
Crystal [Adams *et al*, 2004]

Non-canonical base pairs and modules Nomenclature [Leontis Westhof 2001]



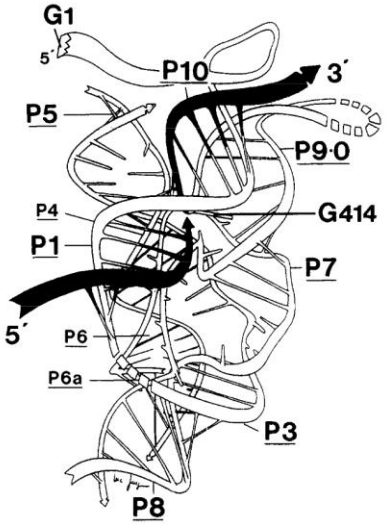


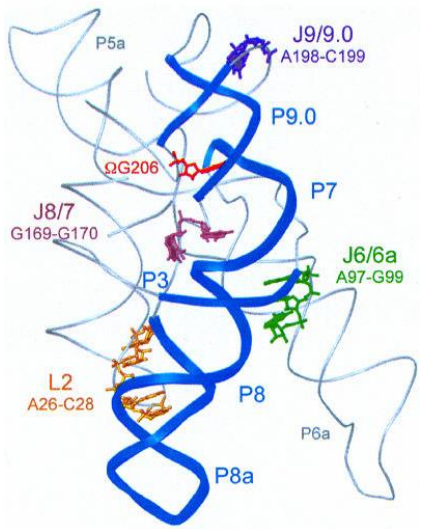
A French pioneer in RNA Structural Modeling

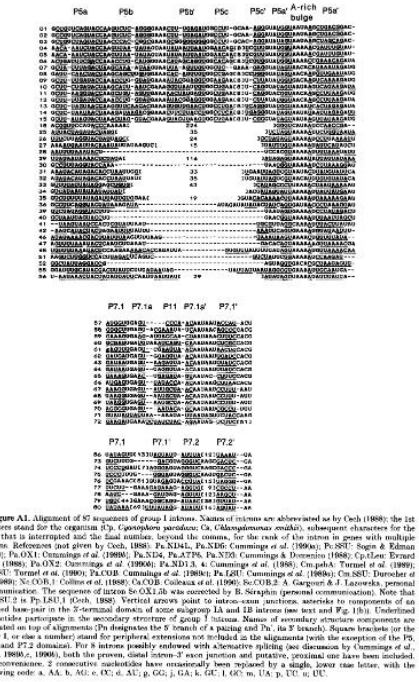


E Westhof
IBMC, Strasbourg

Tetrahymena group I intron





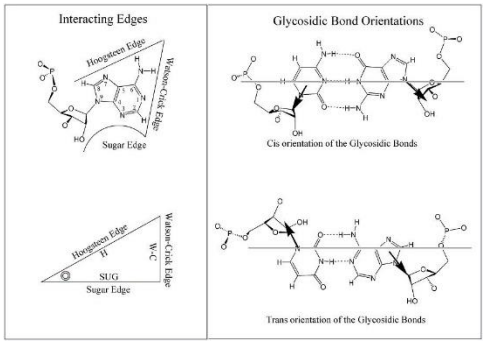


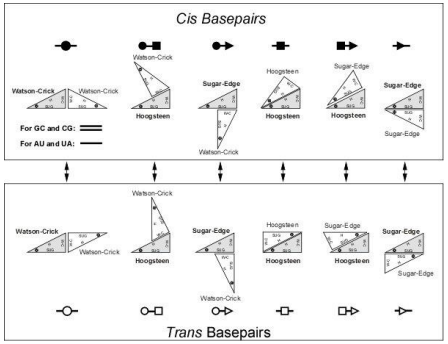
Predicted (manually) [Westhof Michel, 1990]

Crystal [Adams *et al*, 2004]

Non-canonical base pairs and modules

Nomenclature
[Leontis Westhof 2001]





A Denise
LRI, Paris-Sud

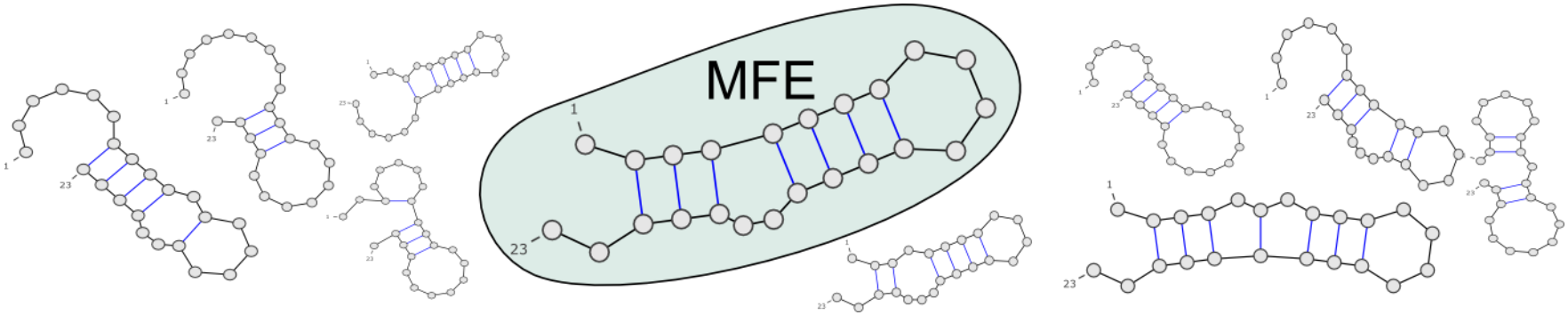
Computational paradigms and methods

Minimal Free-Energy (MFE) Folding

Despite an exponential #conformations, the Minimal Free Energy secondary structure can be predicted exactly in polynomial time in the nearest neighbor model

...**CAGUAGCCGAUCGCAGCUAGCGUA**...

RNAFold, MFold…



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

Poster-child of dynamic programming in bioinformatics

Energy-based *Ab initio* folding methods: Do they *really* work?

- Generally yes, but some variations across studies

Benchmark: 1919 non-multimer/non-pseudoknotted structures from RNAstrand
[Andronescu *et al*, BMC Bioinformatics 2008]

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

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

[Lorenz *et al*, Algo Mol Biol 2011]

Optimization methods can be sensitive to fluctuations of the energy model

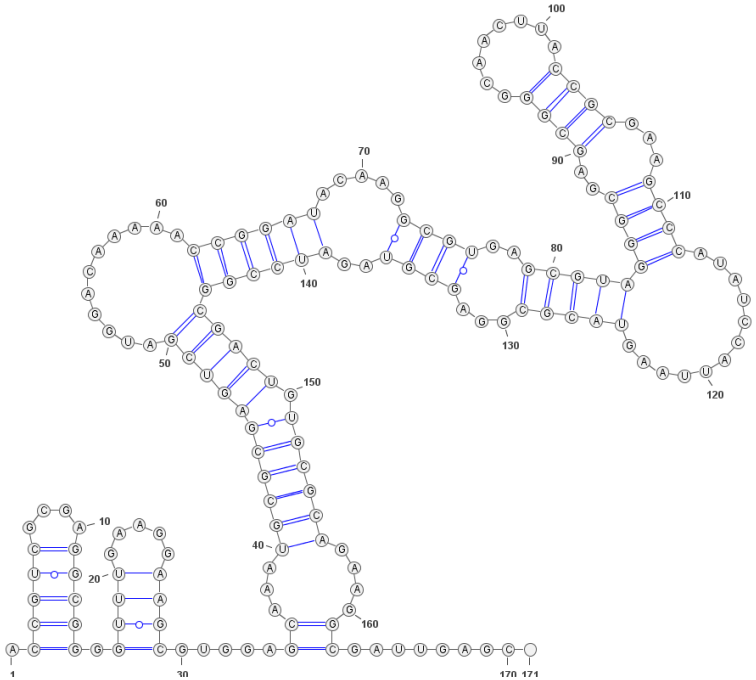
Example:

- RFAM *A. capsulatum* D1-D4 domain of the Group II intron
- Run RNAFold using:
  - Default energy model (Turner 2004)

Optimization methods can be sensitive to fluctuations of the energy model

Example:

- RFAM *A. capsulatum* D1-D4 domain of the Group II intron
- Run RNAFold using:
  - Default energy model (Turner 2004)

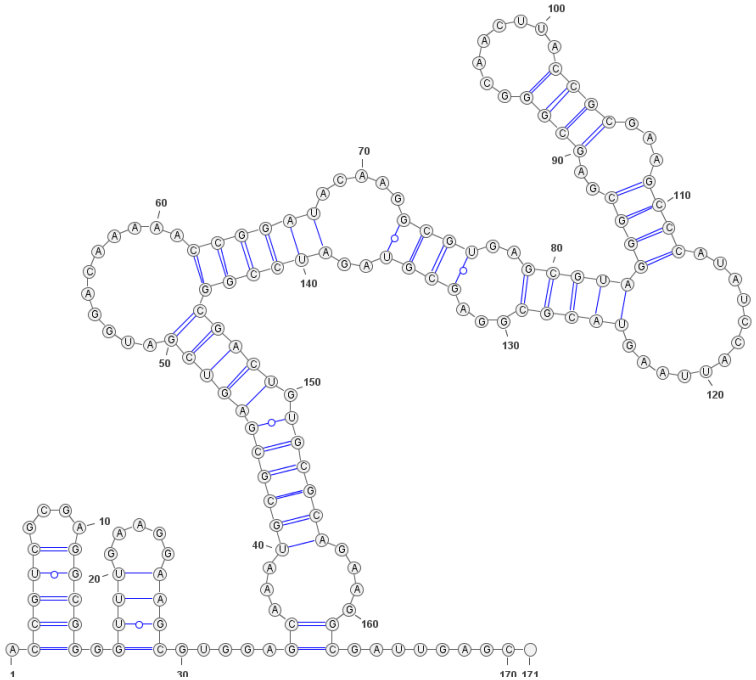


Turner 2004

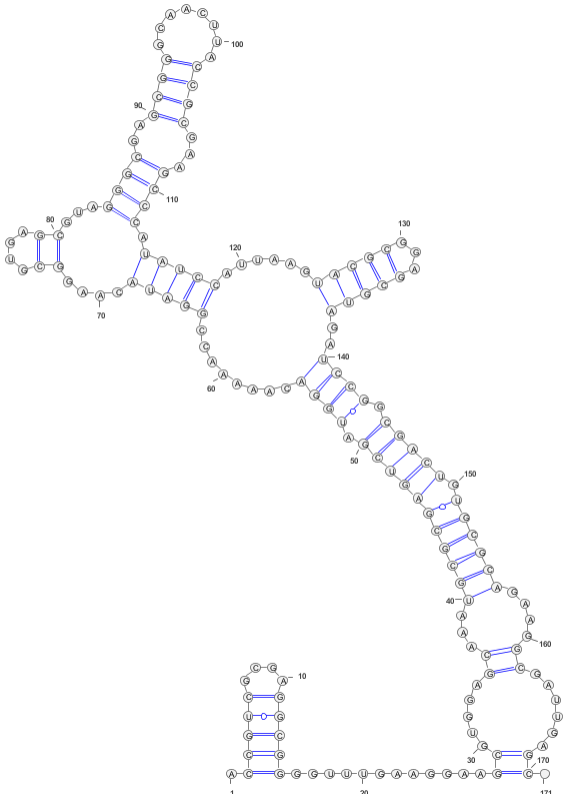
Optimization methods can be sensitive to fluctuations of the energy model

Example:

- RFAM *A. capsulatum* D1-D4 domain of the Group II intron
- Run RNAFold using:
  - Default energy model (Turner 2004)
  - Data-driven model (Andronescu 2007)



Turner 2004

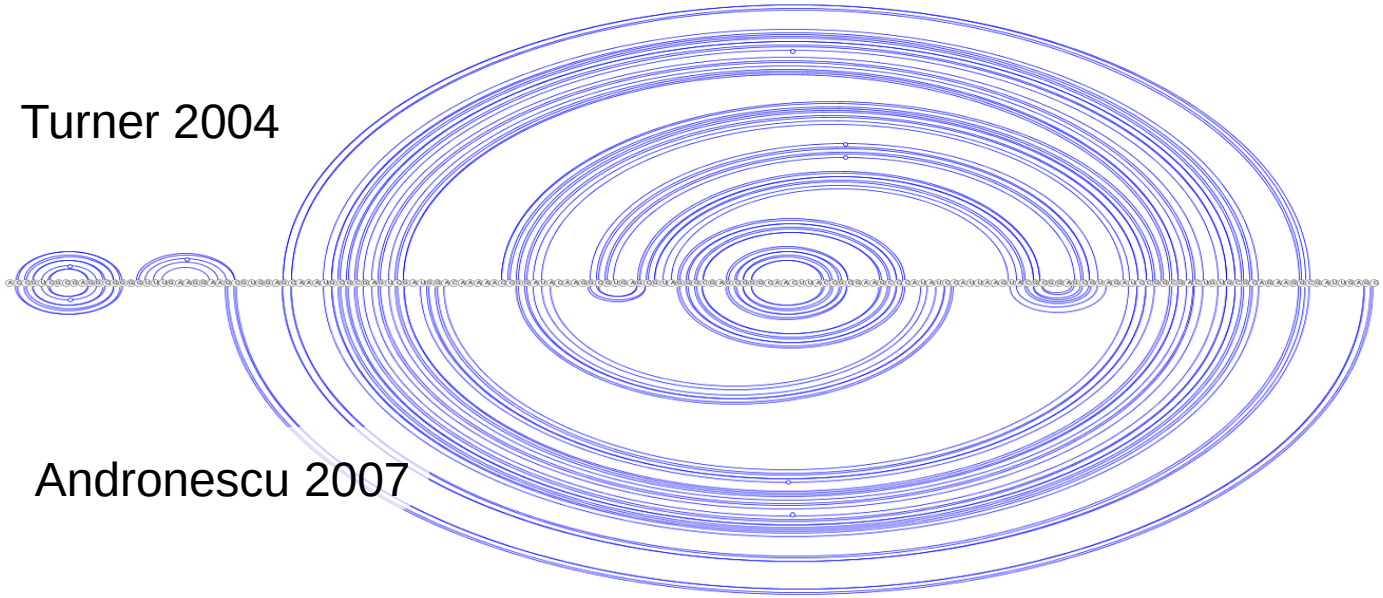


Andronescu 2007

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

Example:

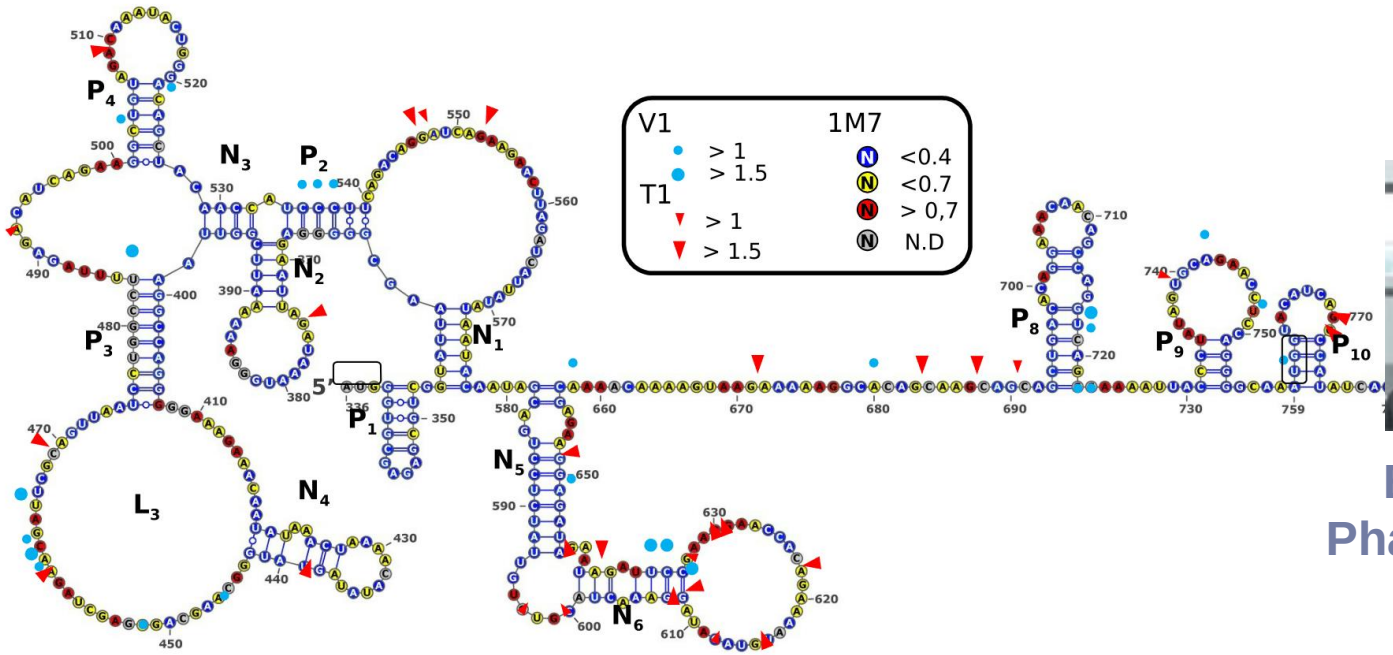
- RFAM *A. capsulatum* D1-D4 domain of the Group II intron
- Run RNAFold using:
  - Default energy model (Turner 2004)
  - Data-driven model (Andronescu 2007)



Discrepancy not as embarrassing as it first looks…
… still substantial!

Chemical/enzymatic probing to model 2D

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

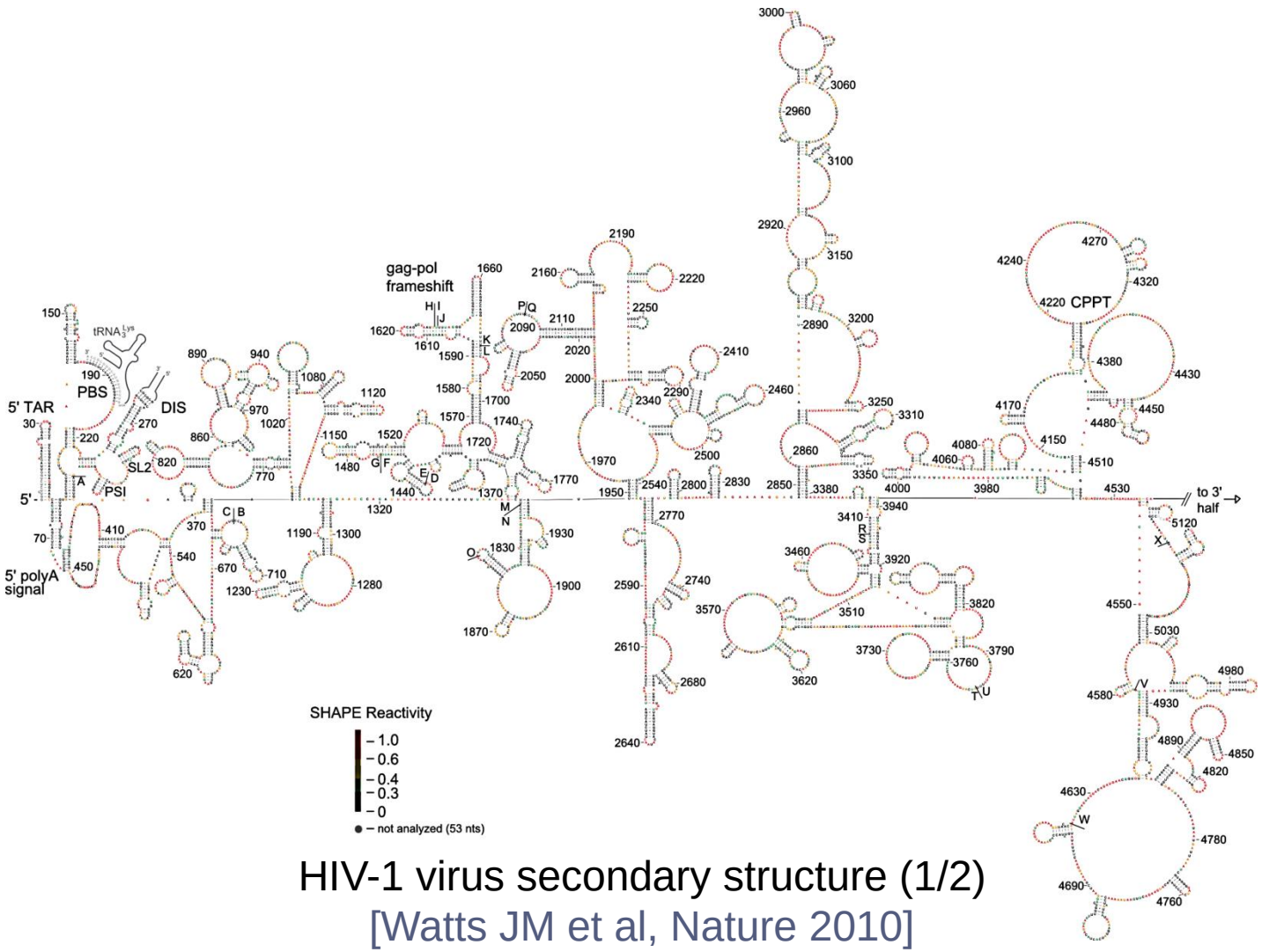


B Sargueil
Pharma,Paris V

- Inclusion as pseudo potentials within energy-models

[Lorenz *et al*, Bioinformatics 2015]

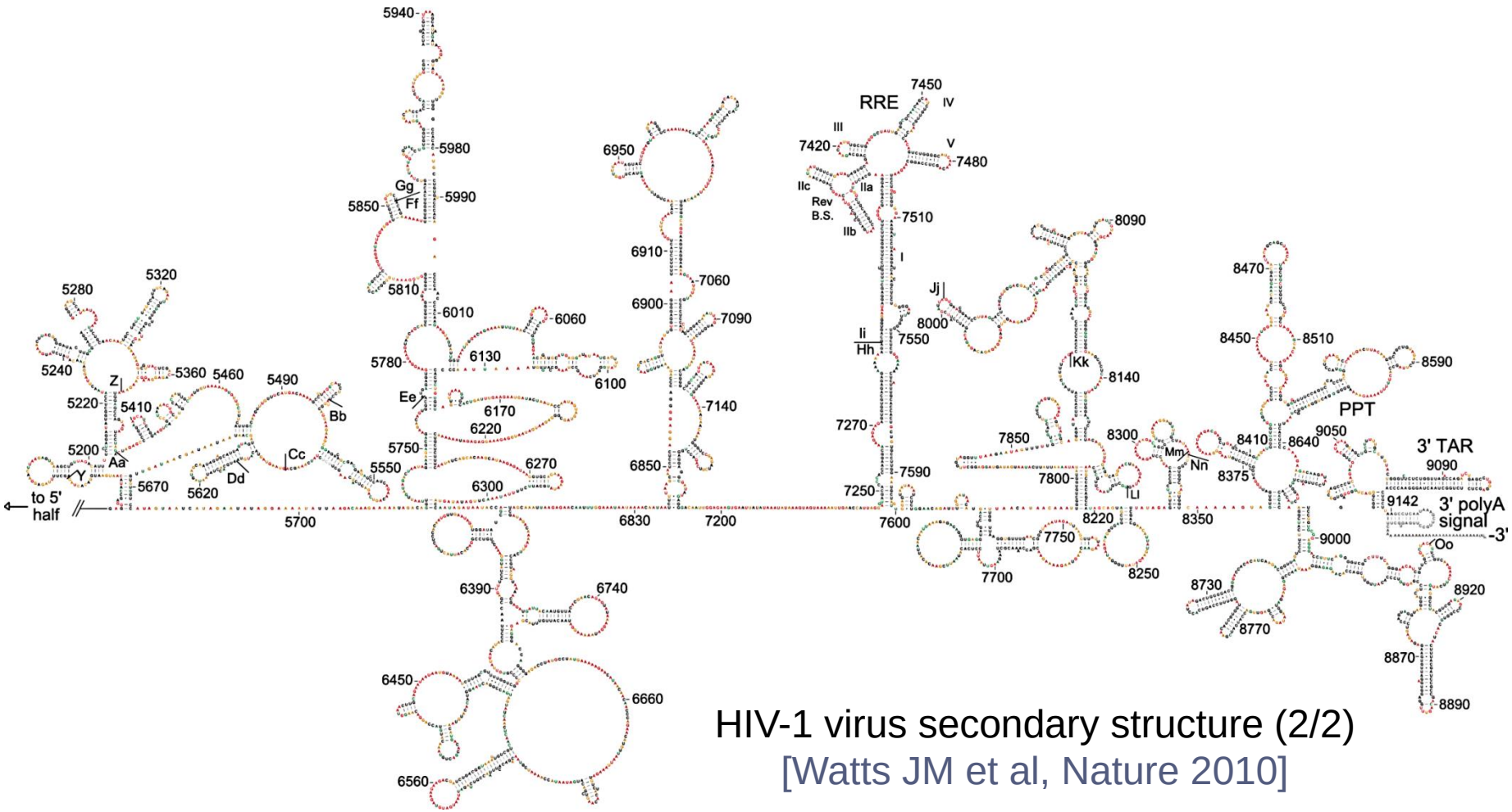
SHAPE probing to model 2D



HIV-1 virus secondary structure (1/2)

[Watts JM et al, Nature 2010]

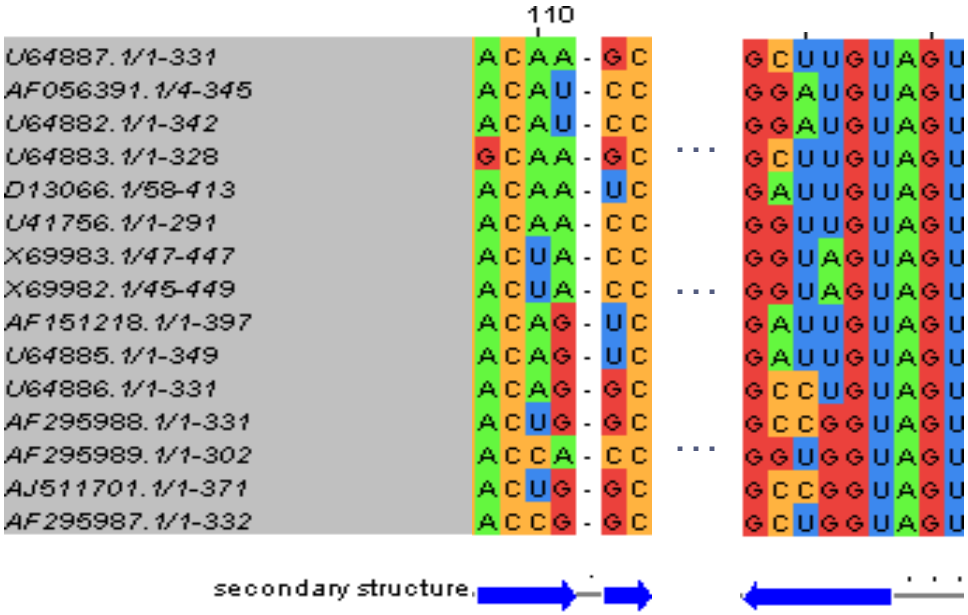
SHAPE probing to model 2D



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

Evolution to the rescue: Comparative approaches for structured RNAs

F Tahi
IBISC Evry



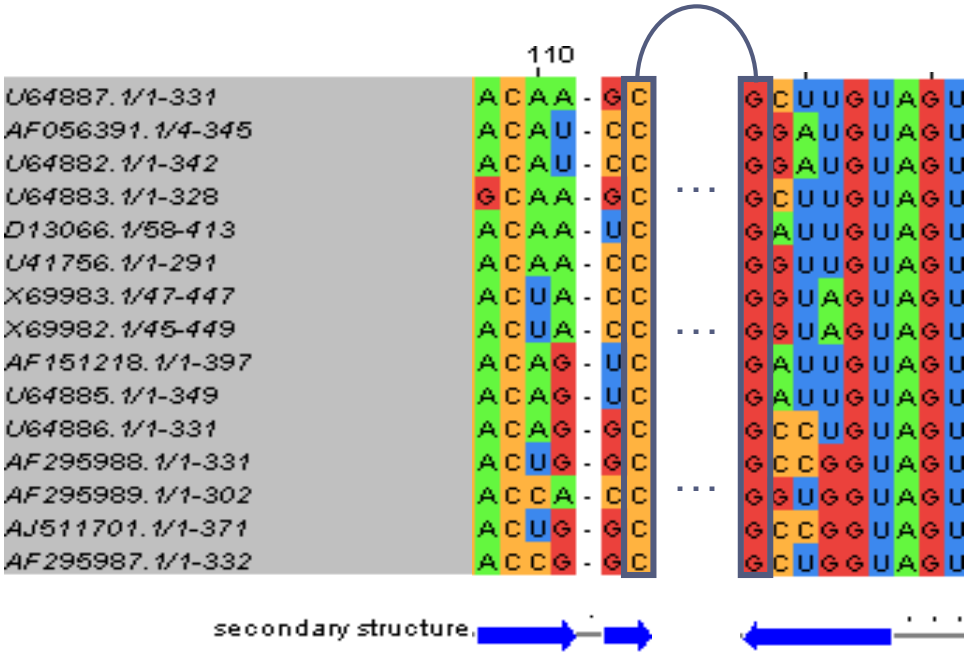
RFAM Bacterial RNase P class B Alignment
RF00011, rendered using JalView

H Touzet
Cristal, Lille I

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

Evolution to the rescue: Comparative approaches for structured RNAs

F Tahi
IBISC Evry



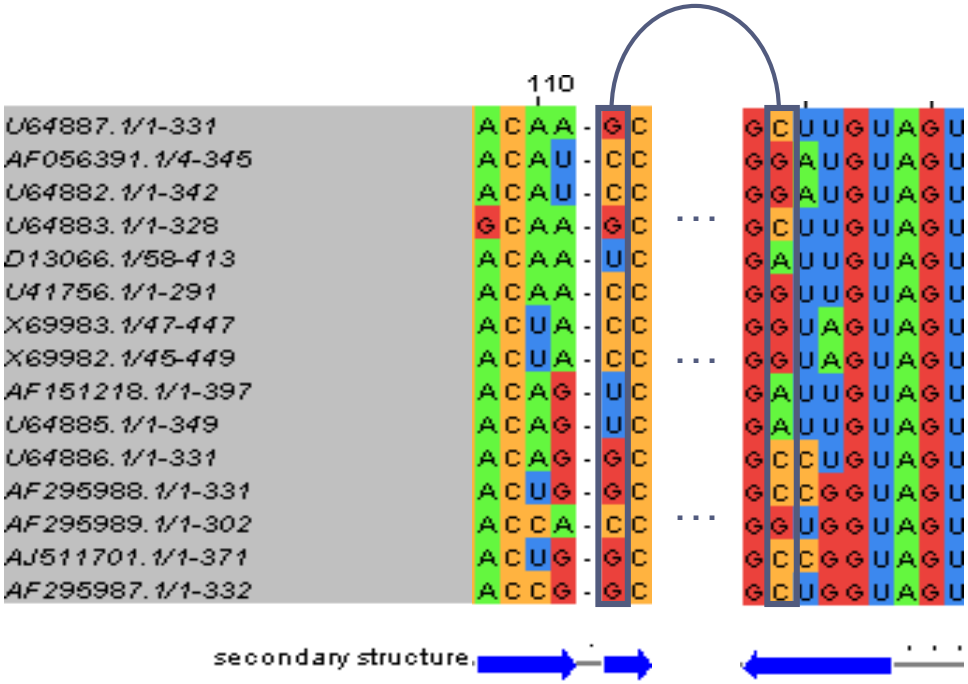
RFAM Bacterial RNase P class B Alignment
RF00011, rendered using JalView

H Touzet
Cristal, Lille I

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

Evolution to the rescue: Comparative approaches for structured RNAs

F Tahi
IBISC Evry



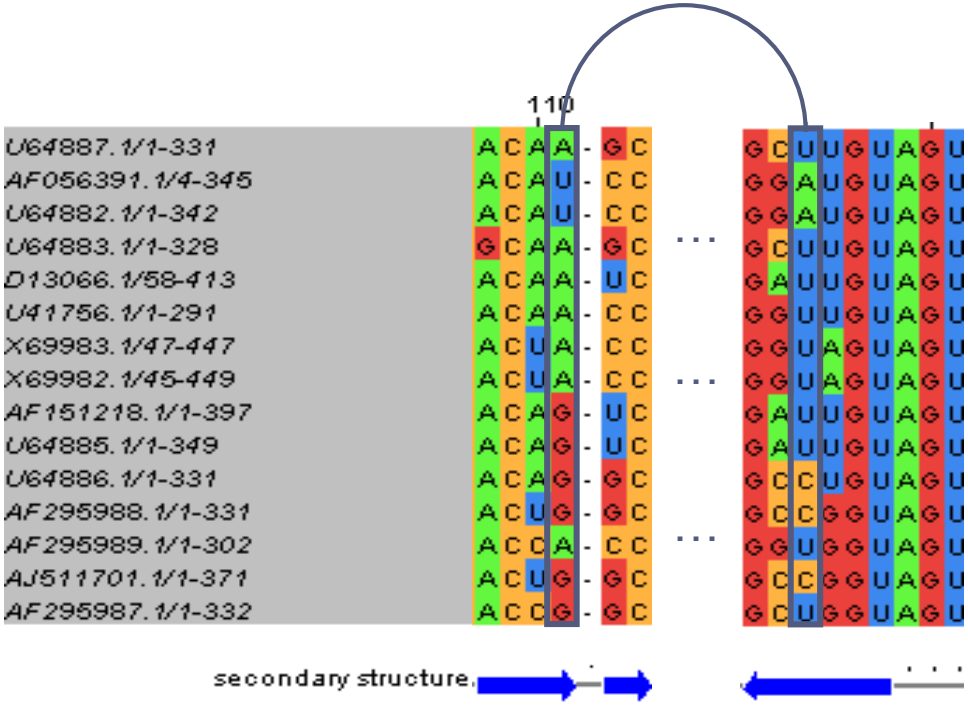
RFAM Bacterial RNase P class B Alignment
RF00011, rendered using JalView

H Touzet
Cristal, Lille I

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

Evolution to the rescue: Comparative approaches for structured RNAs

F Tahi
IBISC Evry



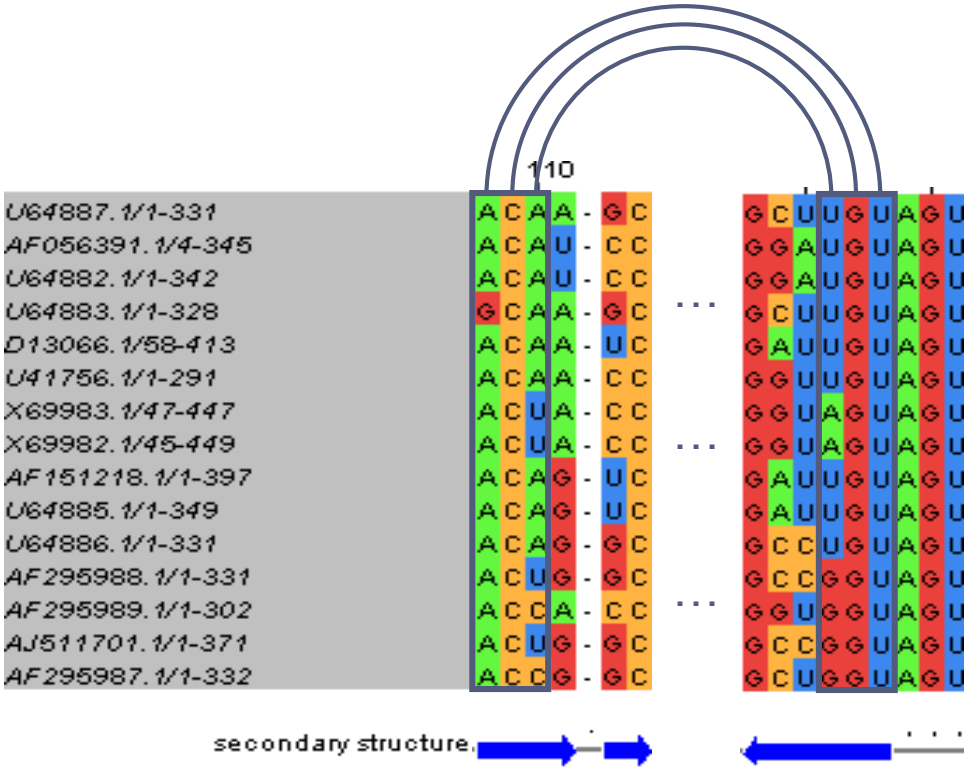
RFAM Bacterial RNase P class B Alignment
RF00011, rendered using JalView

H Touzet
Cristal, Lille I

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

Evolution to the rescue: Comparative approaches for structured RNAs

F Tahi
IBISC Evry



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

H Touzet
Cristal, Lille I

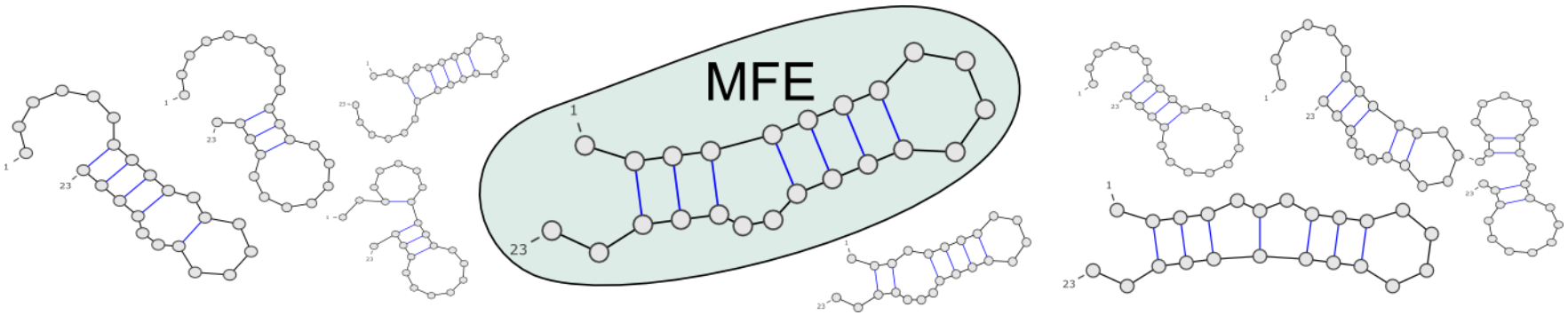
- Structure (≈ phenotype) typically more conserved than sequence
- Covariations/compensatory mutations hint towards shared structure

Ensemble approaches in RNA folding

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

...CAGUAGCCGAUCGCAGCUAGCGUA...

Mfold, RNAfold…

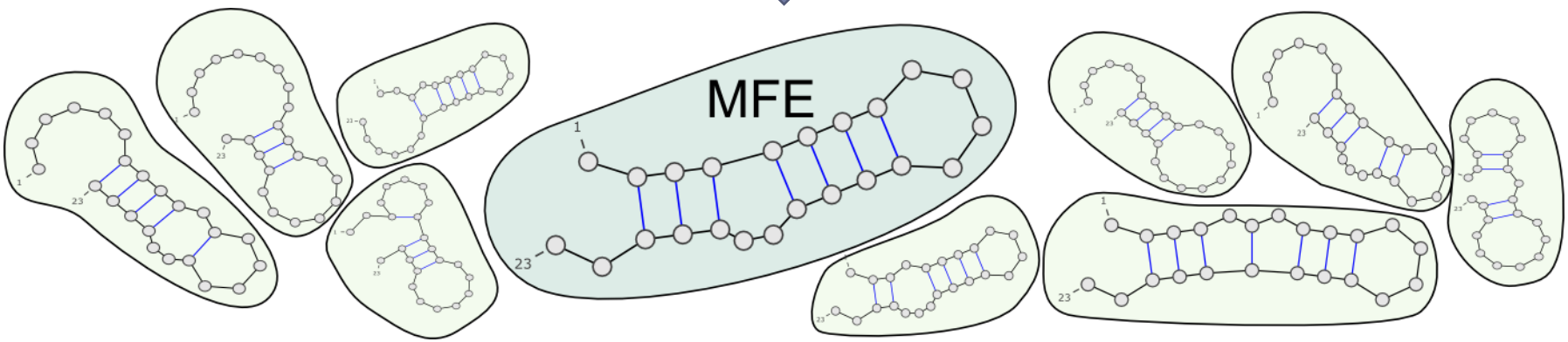


Ensemble approaches in RNA folding

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

...CAGUAGCCGAUCGCAGCUAGCGUA...

UnaFold, RNAFold, Sfold...



Thermodynamic equilibrium: Each secondary structure has probability

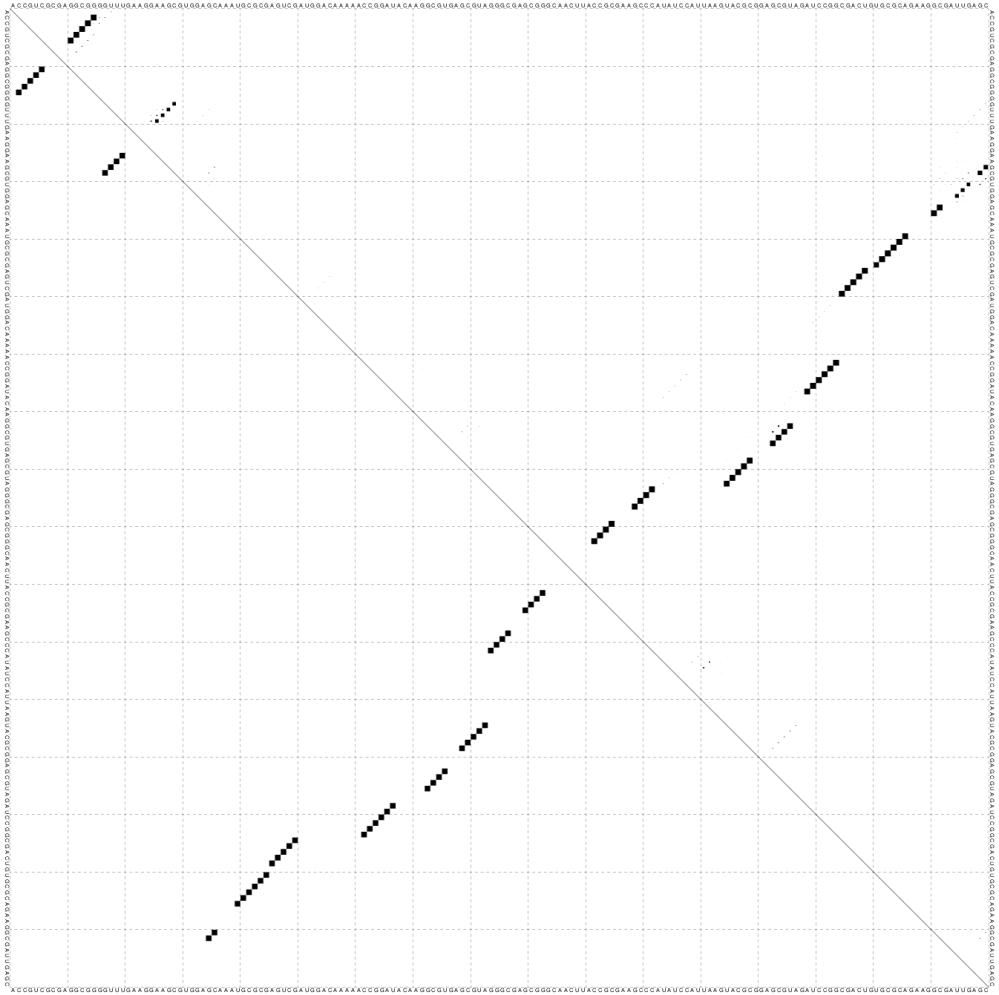
Boltzmann Probability $Prob(S) = \frac{e^{-\Delta G(S)/kT}}{Z}$

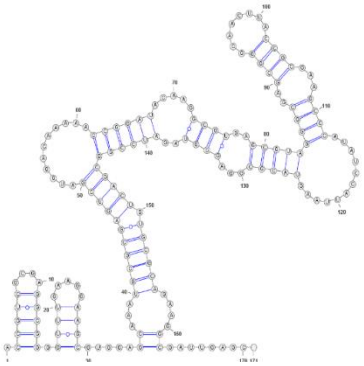
Partition Function $Z = \sum_{Struc\ S} e^{-\Delta G(S)/kT}$

[McCaskill, Biopolymers 1990]

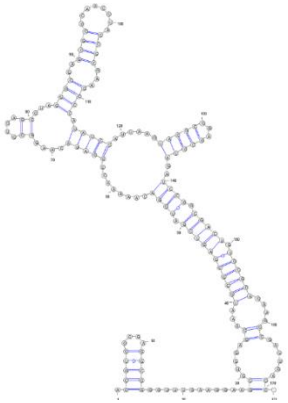
Ensemble-based methods are more robust than energy-minimizing approaches

Back to D1D4 example…





Turner 2004



Andronescu 2007

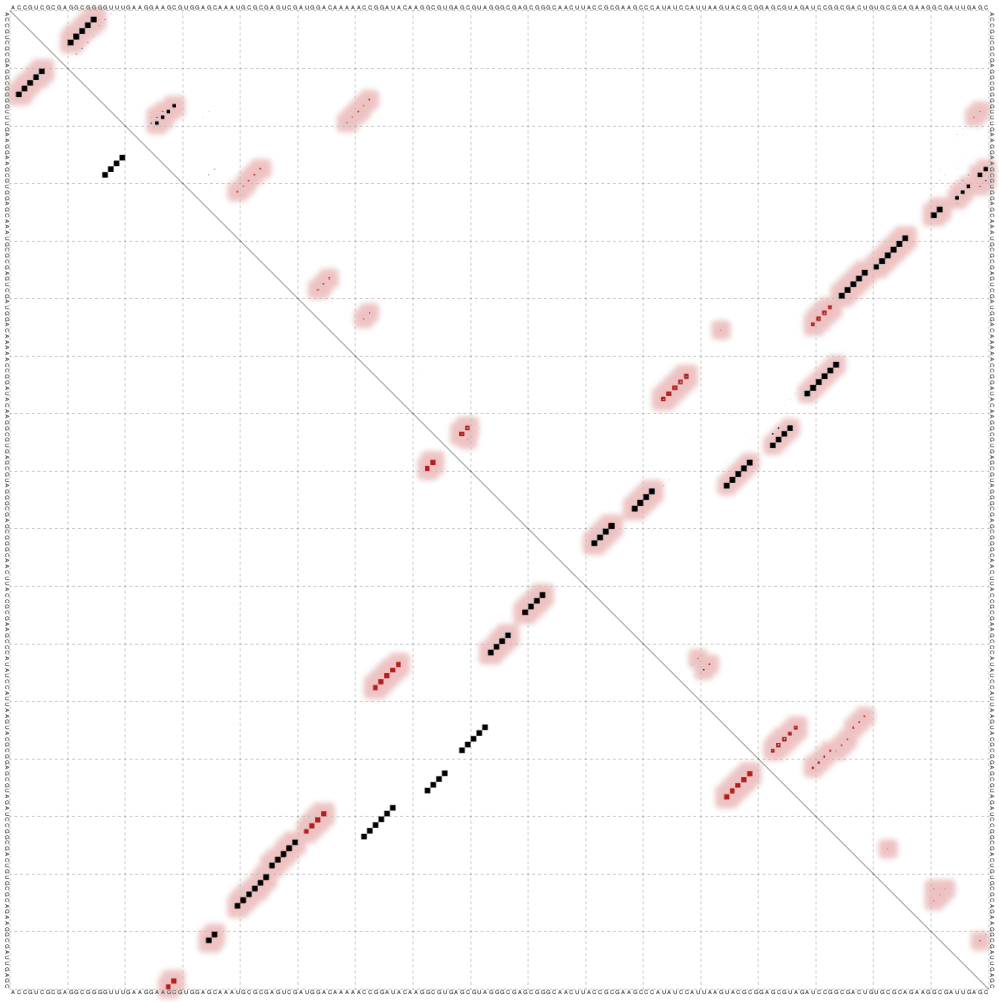
BP>99% → PPV>90%

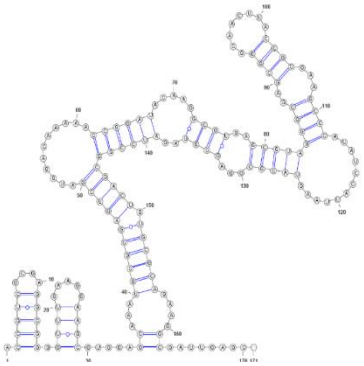
BP>90% → PPV>83%...

[Mathews, RNA 2004]

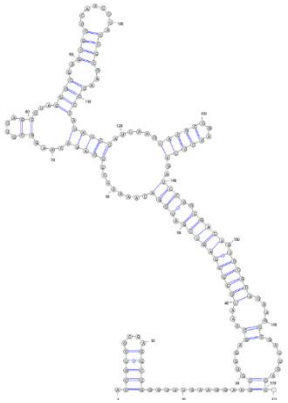
Ensemble-based methods are more robust than energy-minimizing approaches

Back to D1D4 example…





Turner 2004



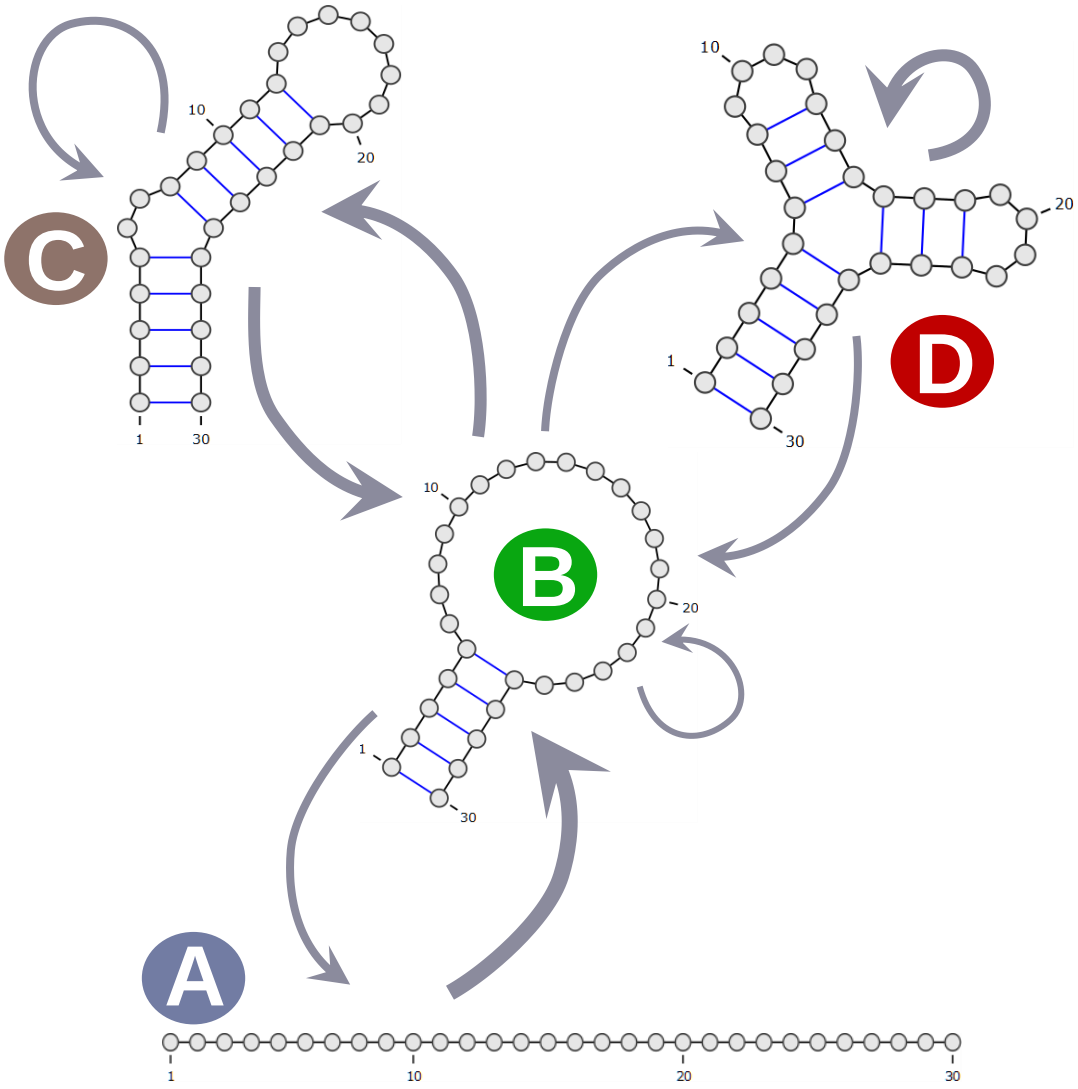
Andronescu 2007

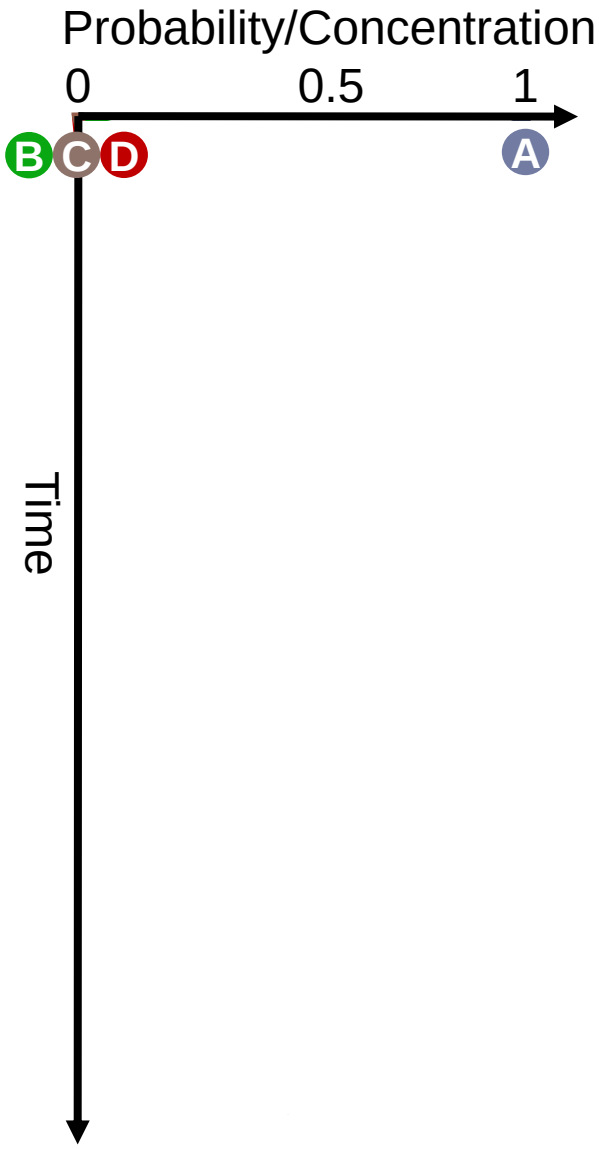
BP>99% → PPV>90%

BP>90% → PPV>83%...

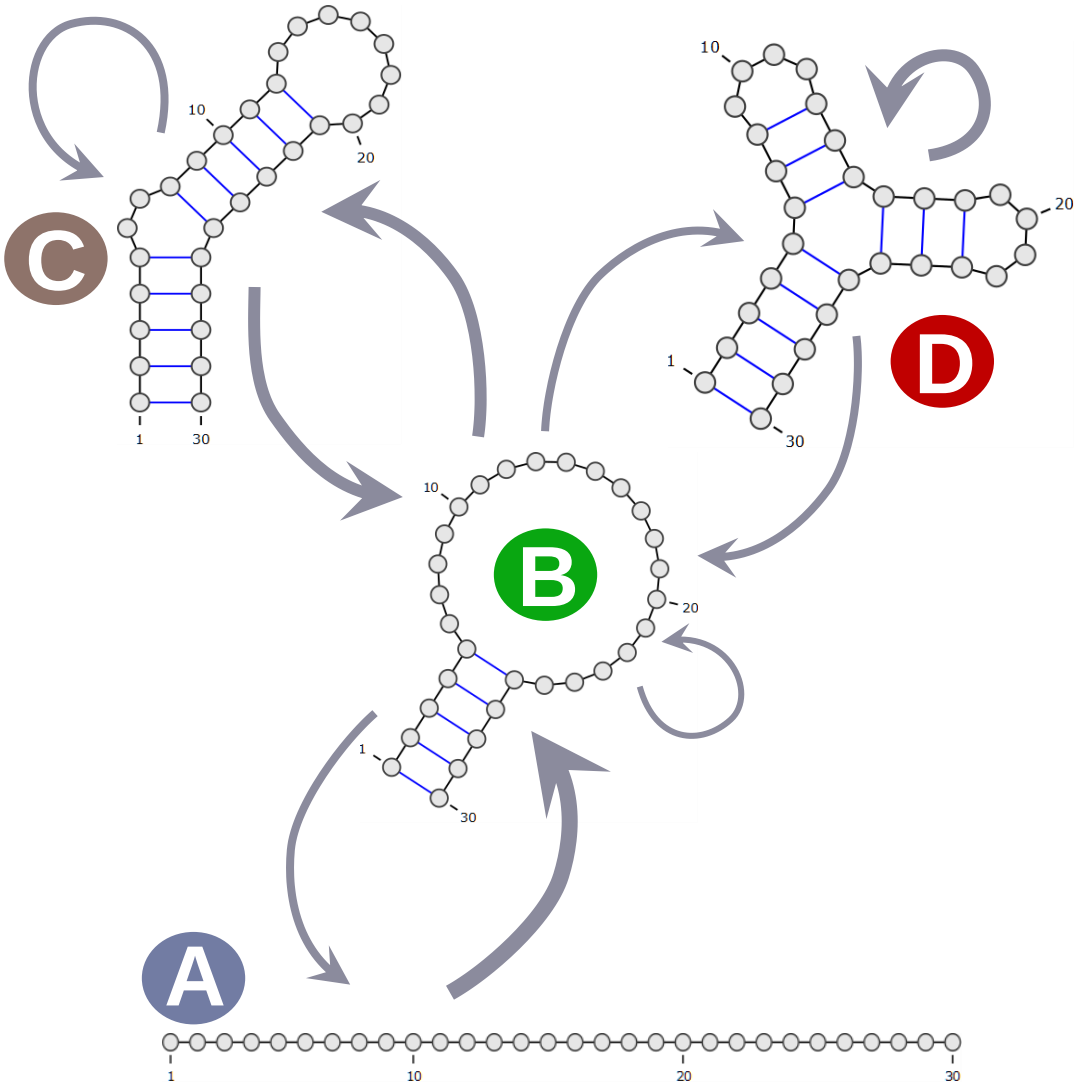
[Mathews, RNA 2004]

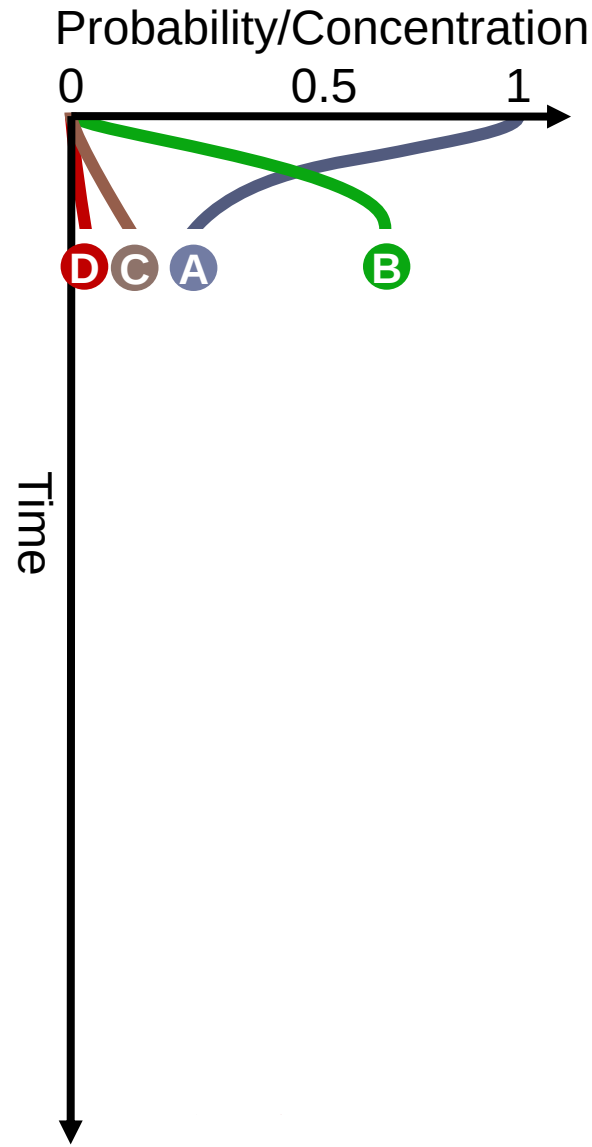
RNA kinetics



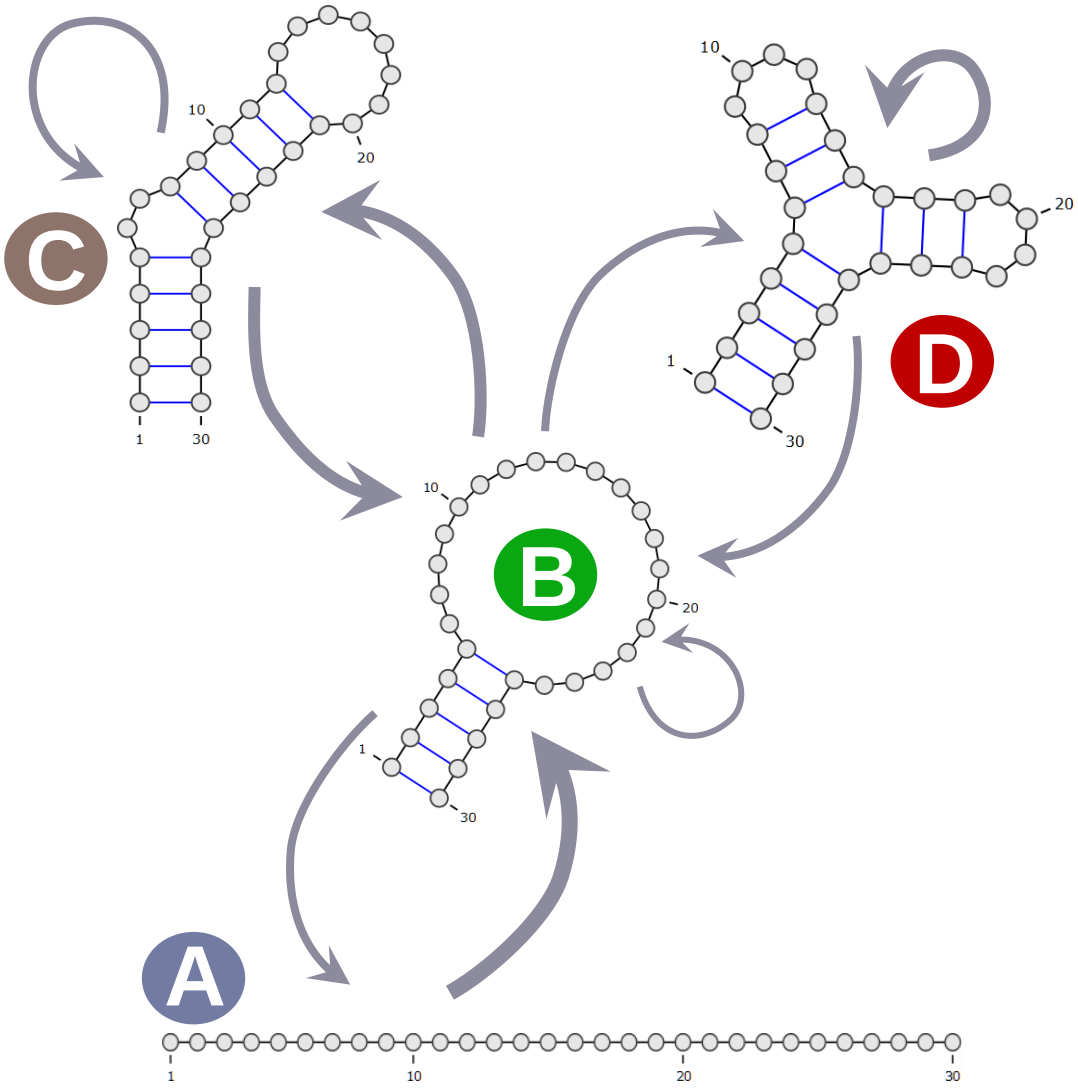


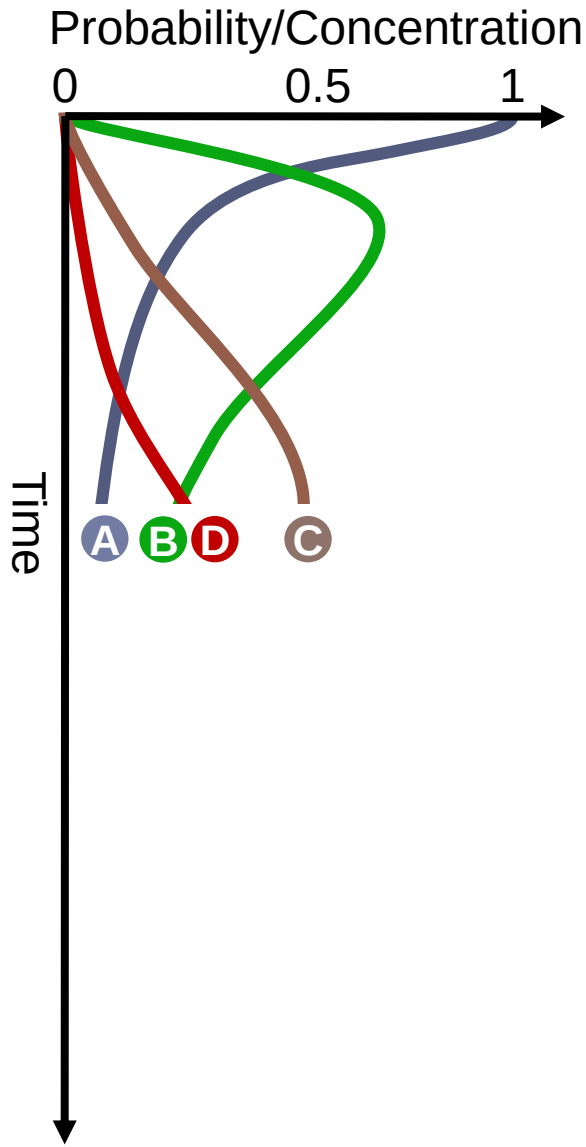
RNA kinetics



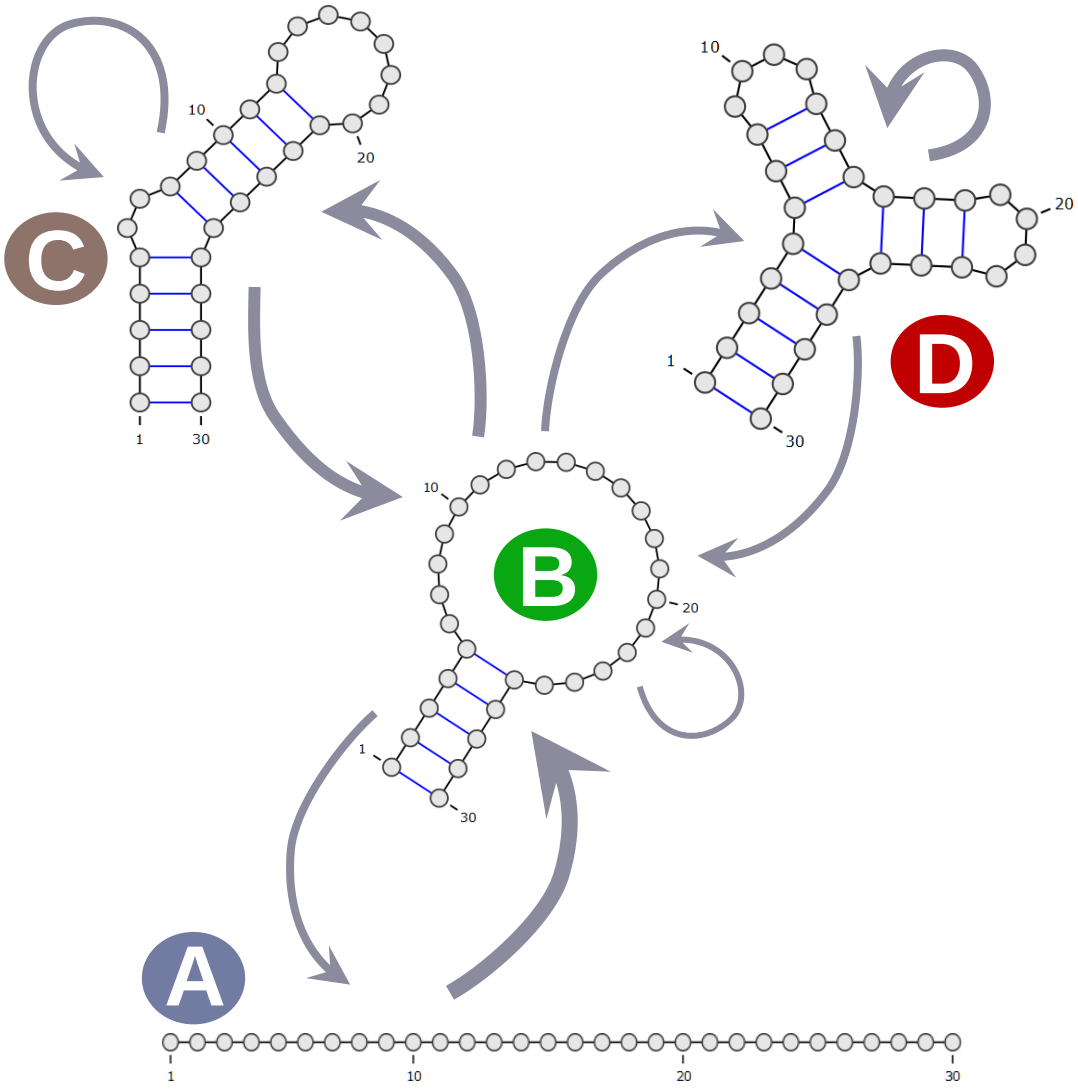


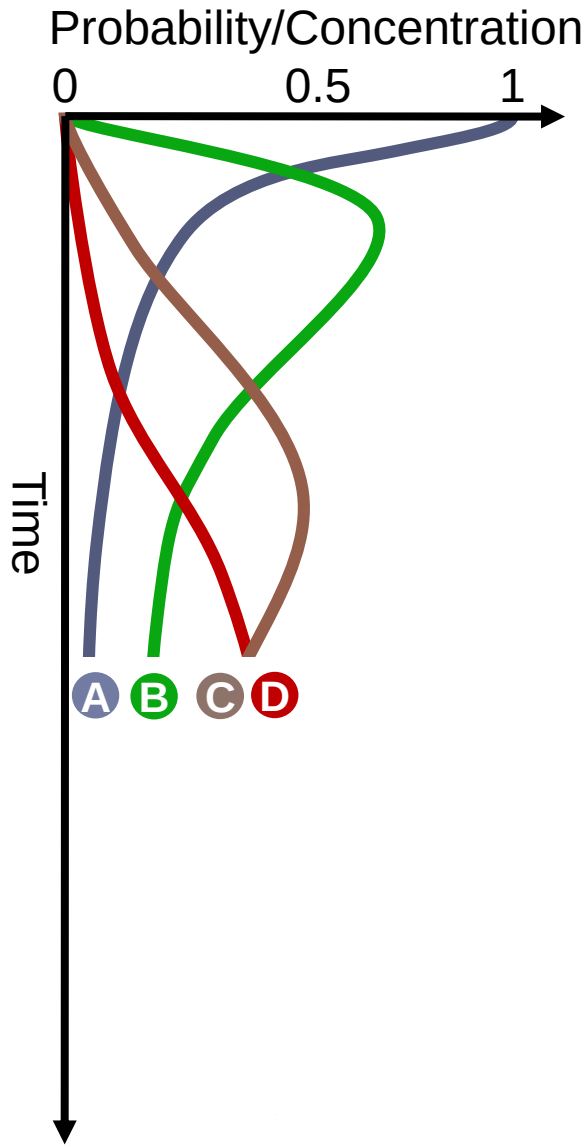
RNA kinetics



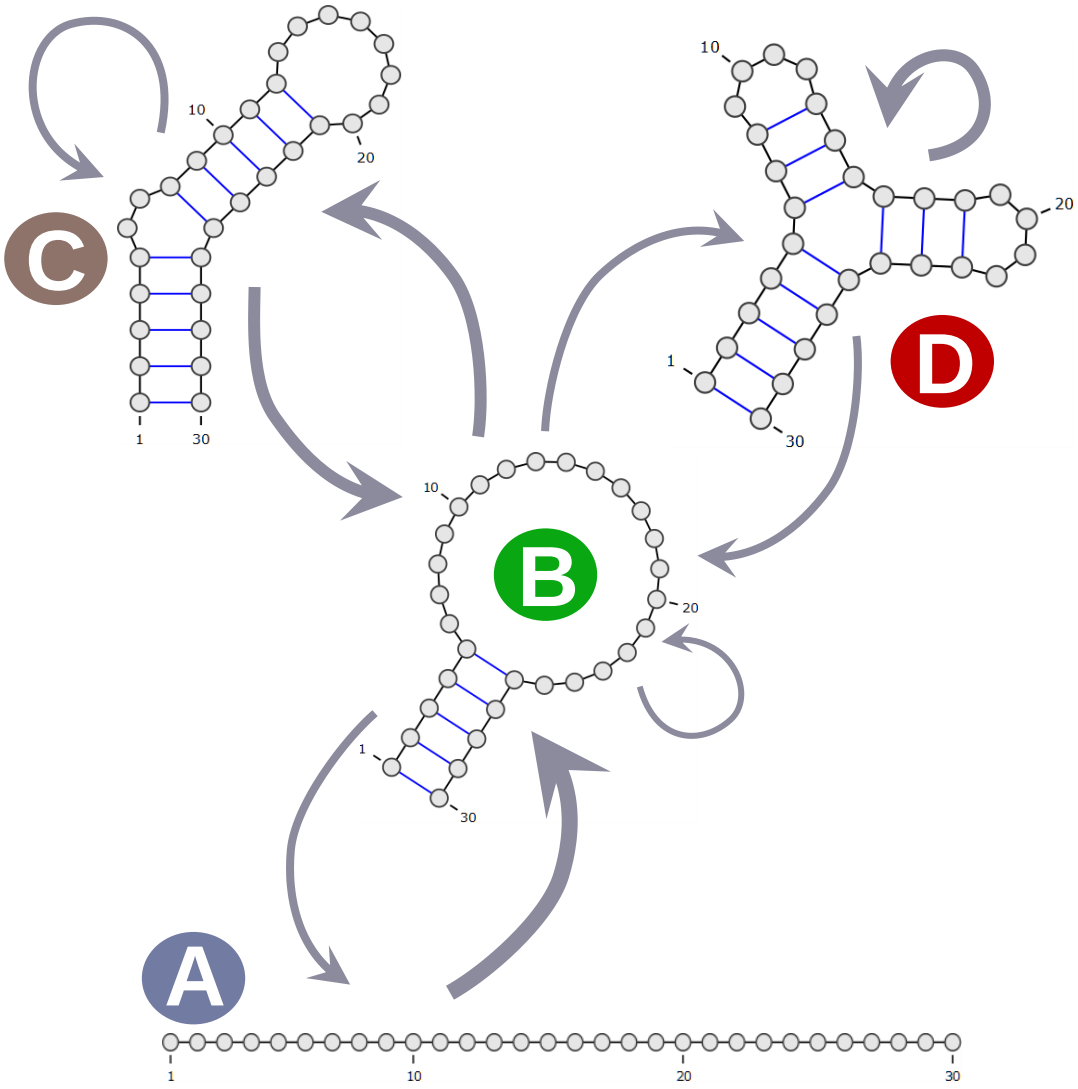


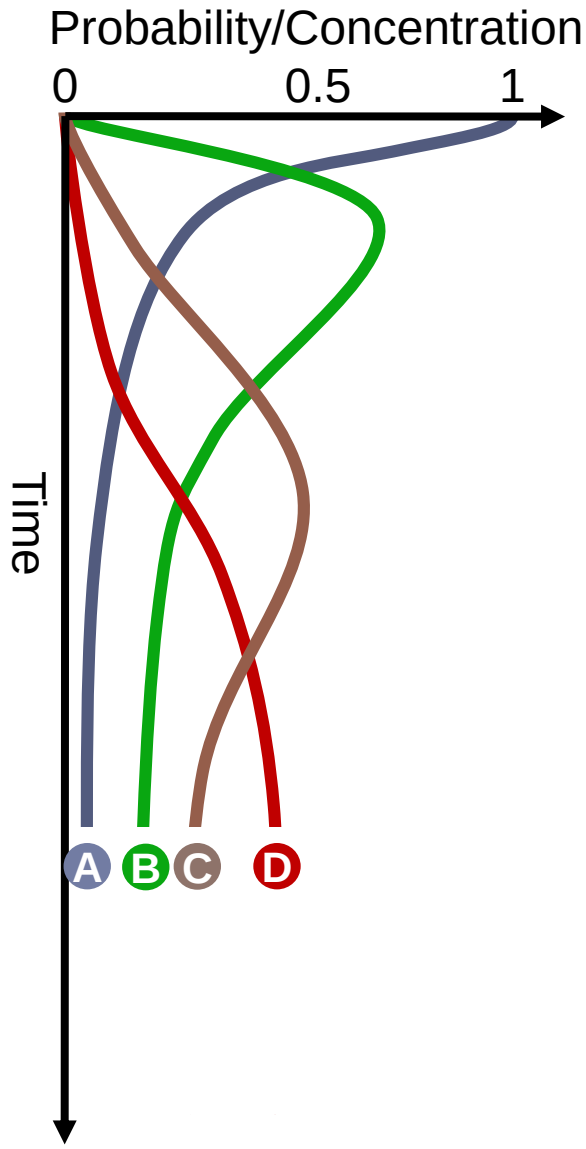
RNA kinetics



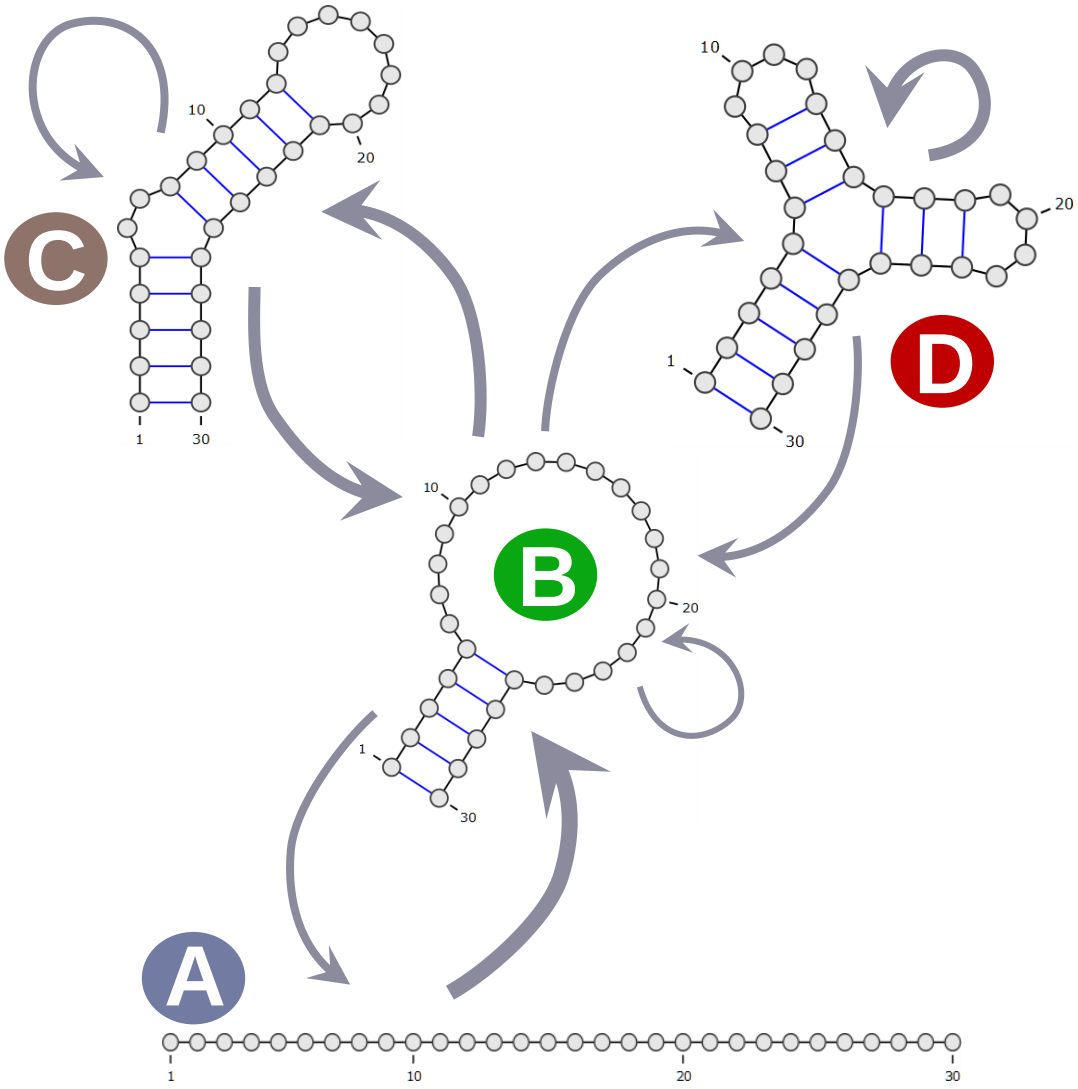


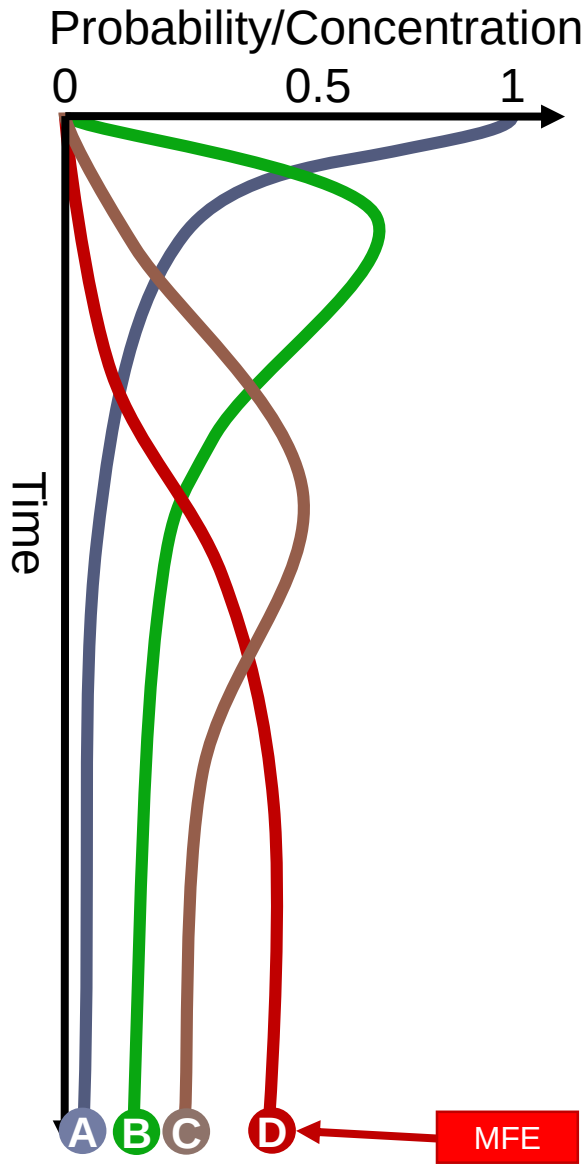
RNA kinetics



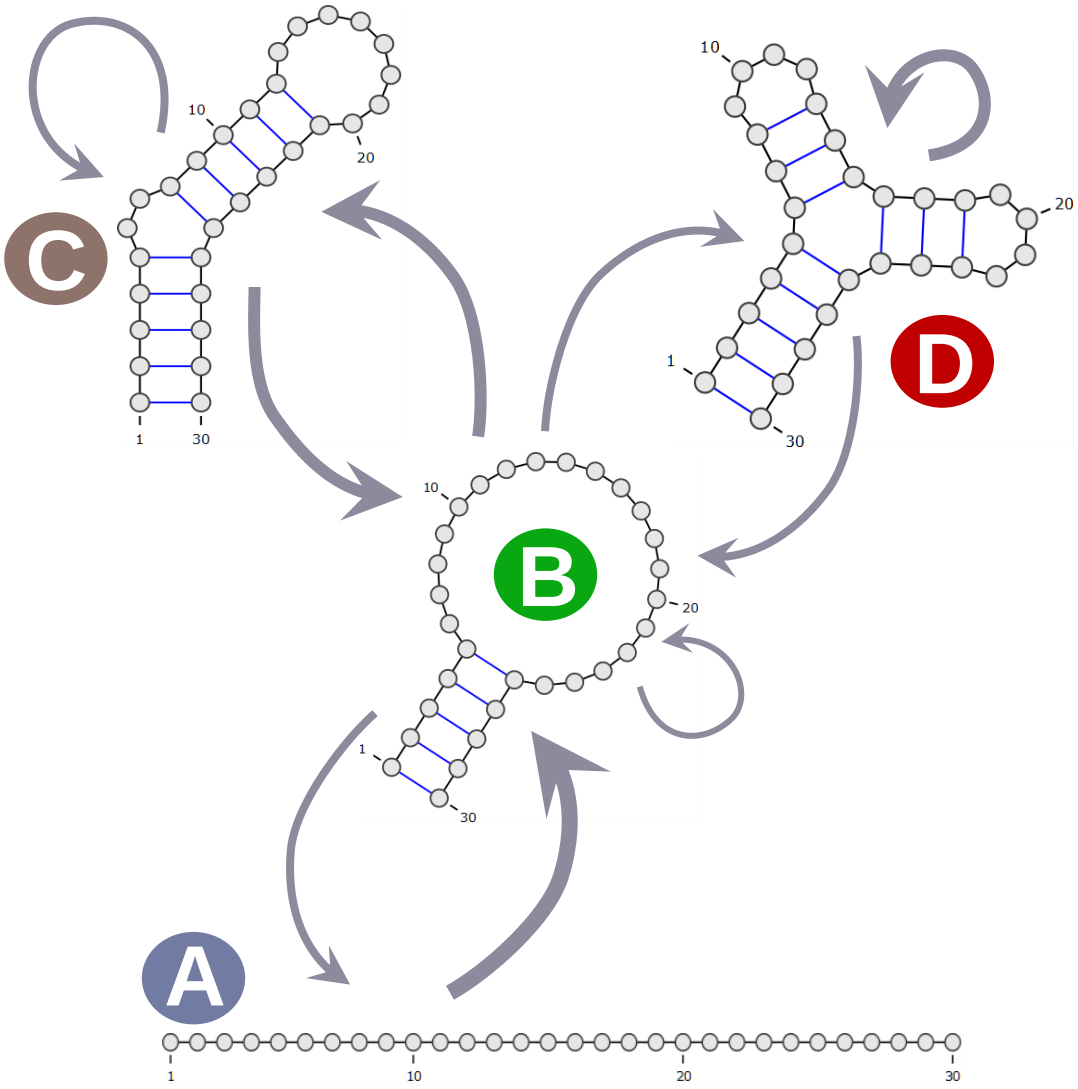


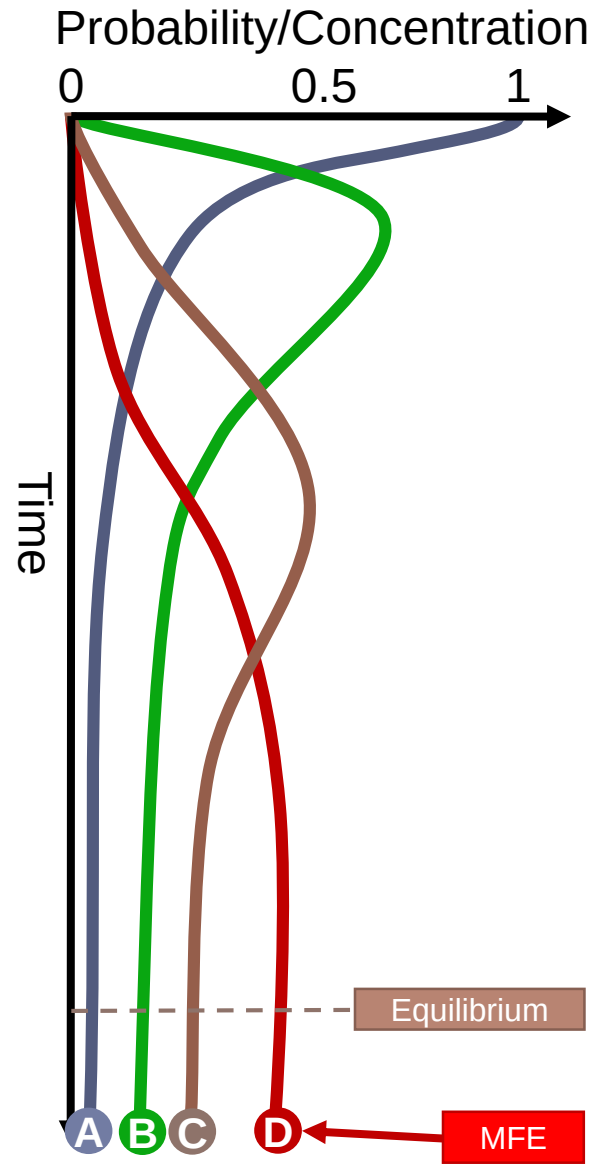
RNA kinetics



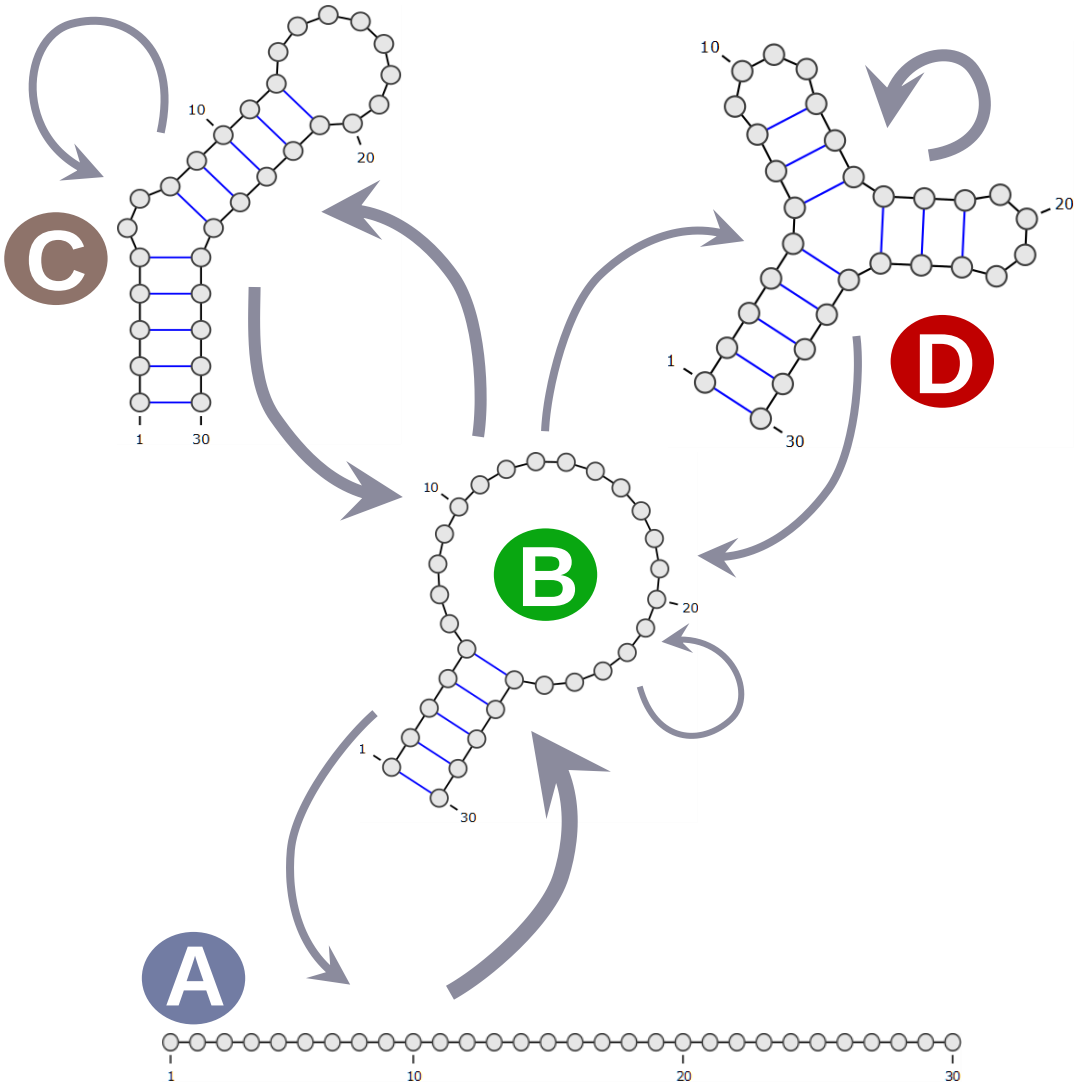


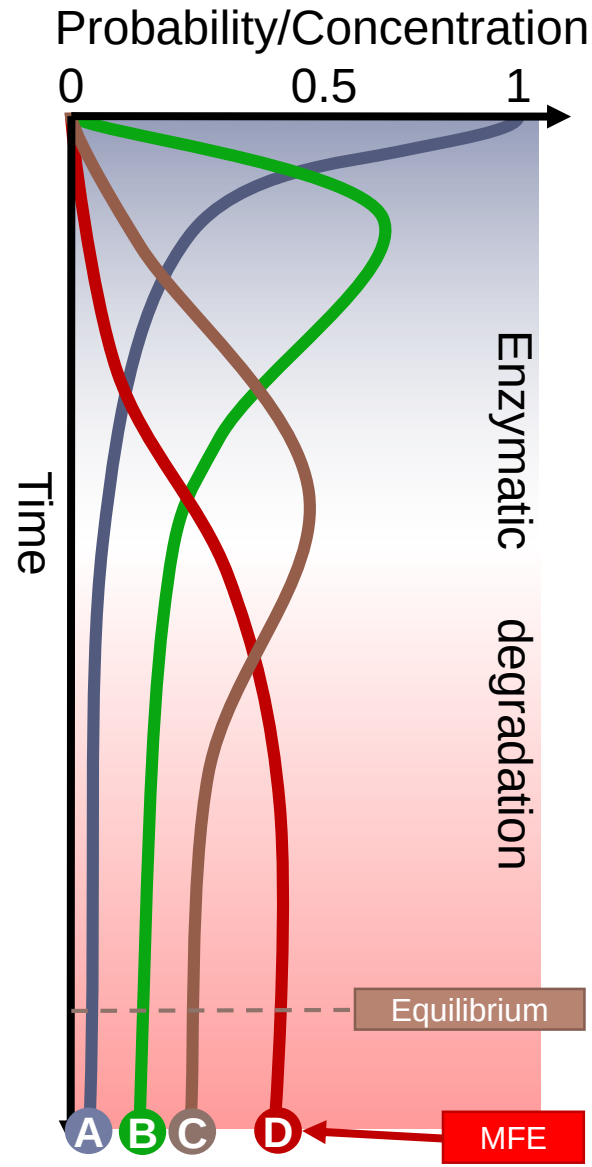
RNA kinetics



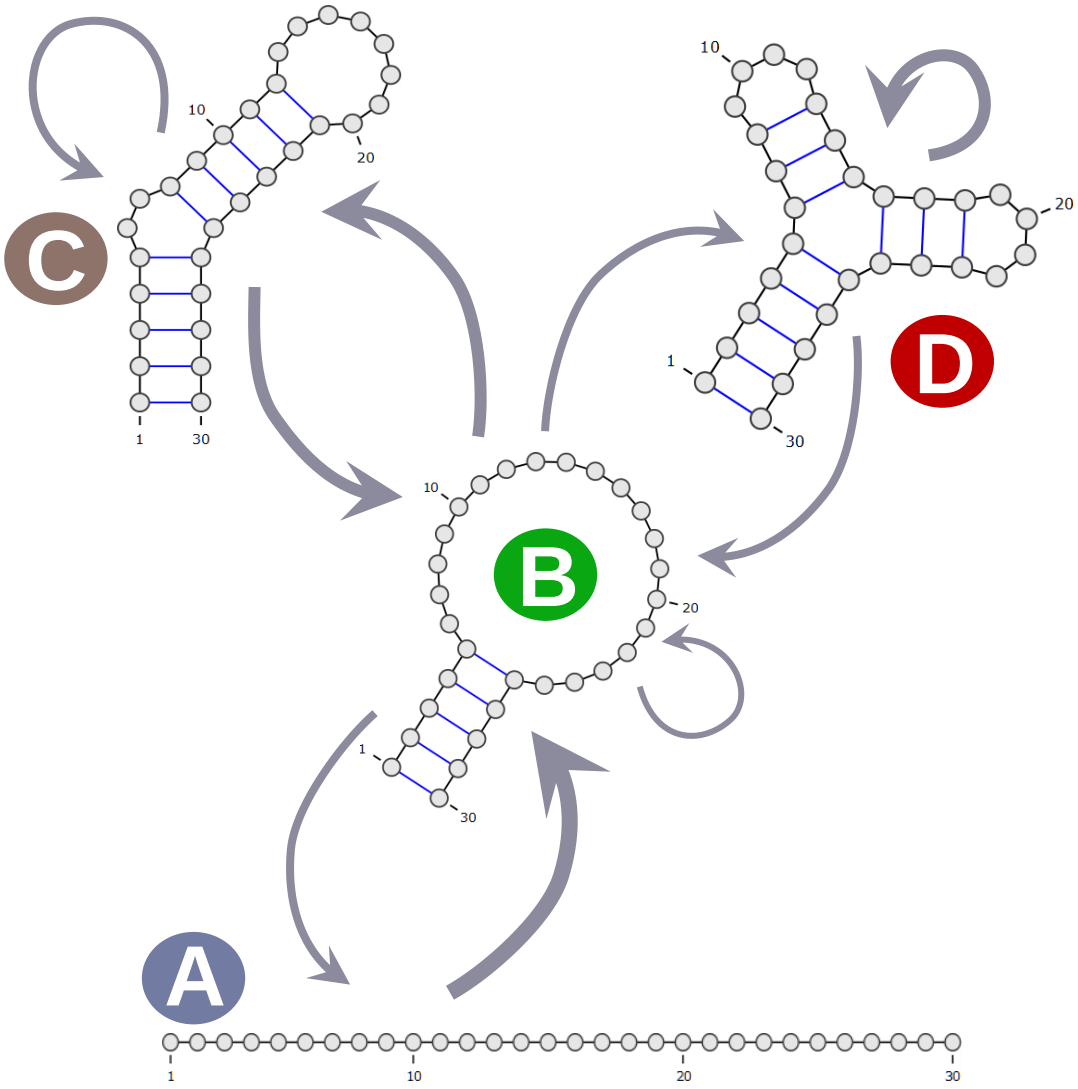


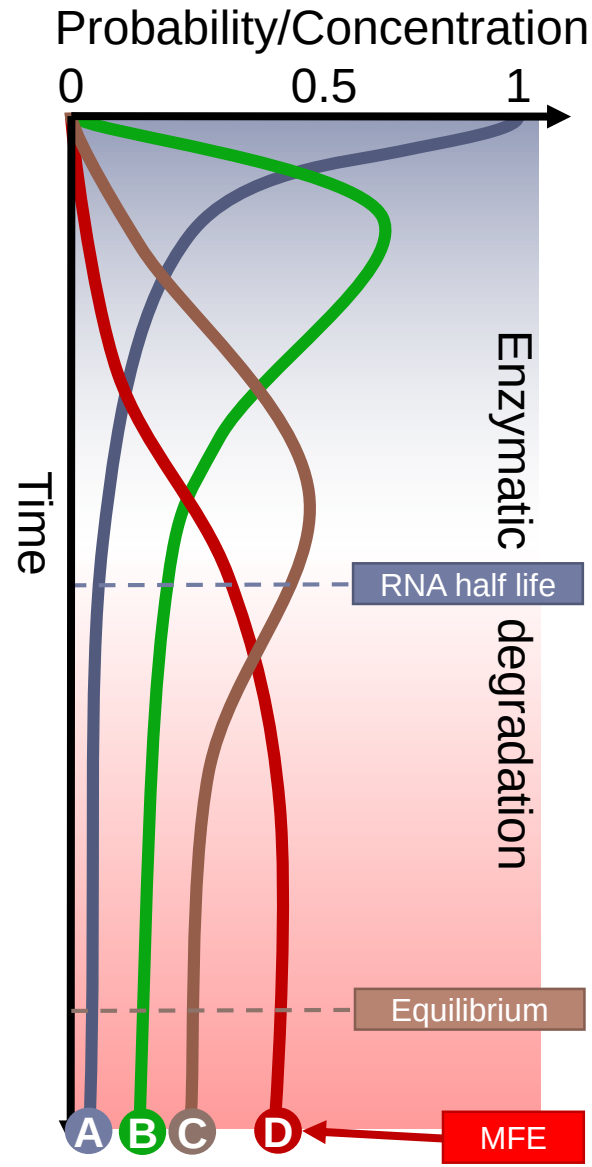
RNA kinetics





RNA kinetics





Beyond 2D

- Molecular dynamics (coarse-grained)

S Pasquali
Pharma, Paris Descartes

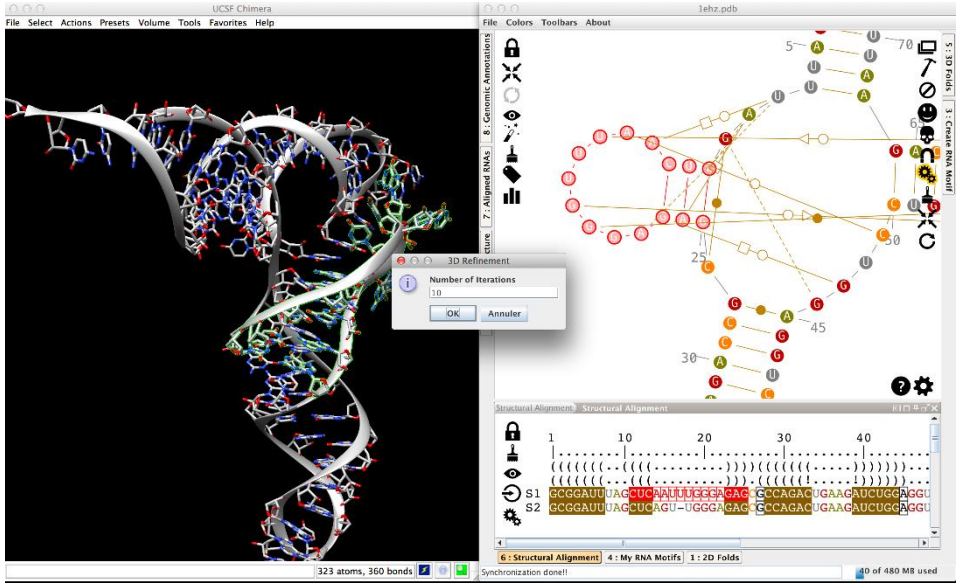
Beyond 2D

- Molecular dynamics (coarse-grained)

S Pasquali
Pharma, Paris Descartes

- Computer-assisted modeling (Assemble)

F Jossinet
Univ. Strasbourg



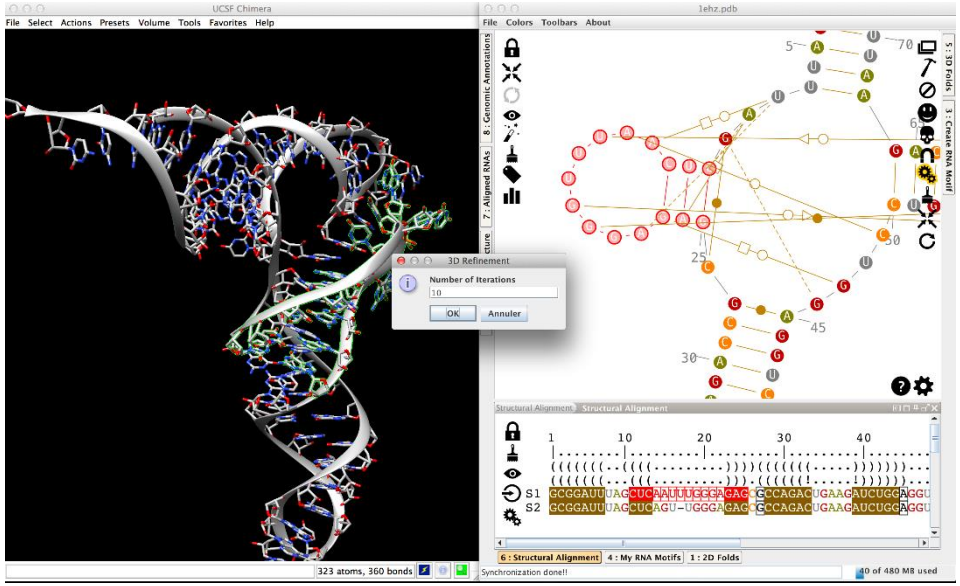
Beyond 2D

- Molecular dynamics (coarse-grained)

S Pasquali
Pharma, Paris Descartes

- Computer-assisted modeling (Assemble)

F Jossinet
Univ. Strasbourg



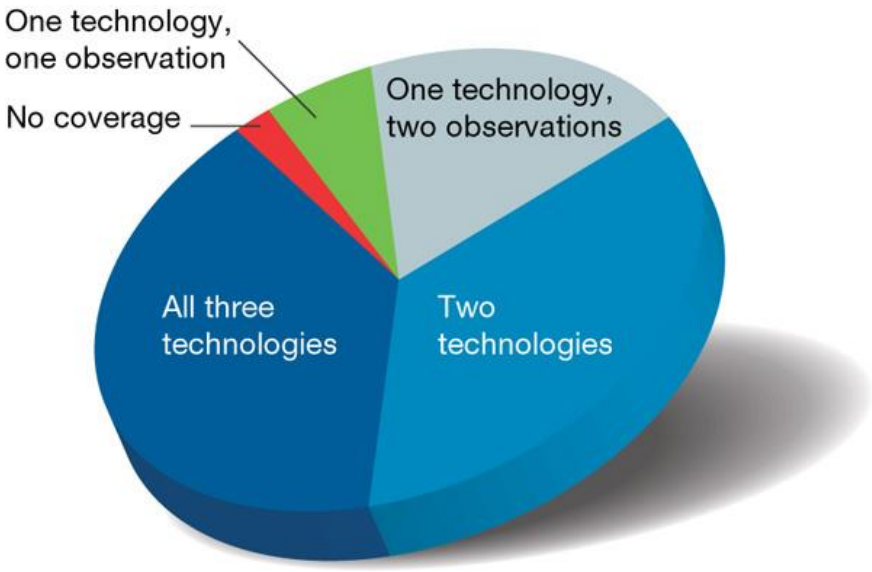
- RNA-Protein Interactions

F Leclerc
I2BC, Paris-Sud

Sounds good, but …
why should we care?

Why RNA is **totally awesome!**

- Ubiquitous
- Pervasively expressed

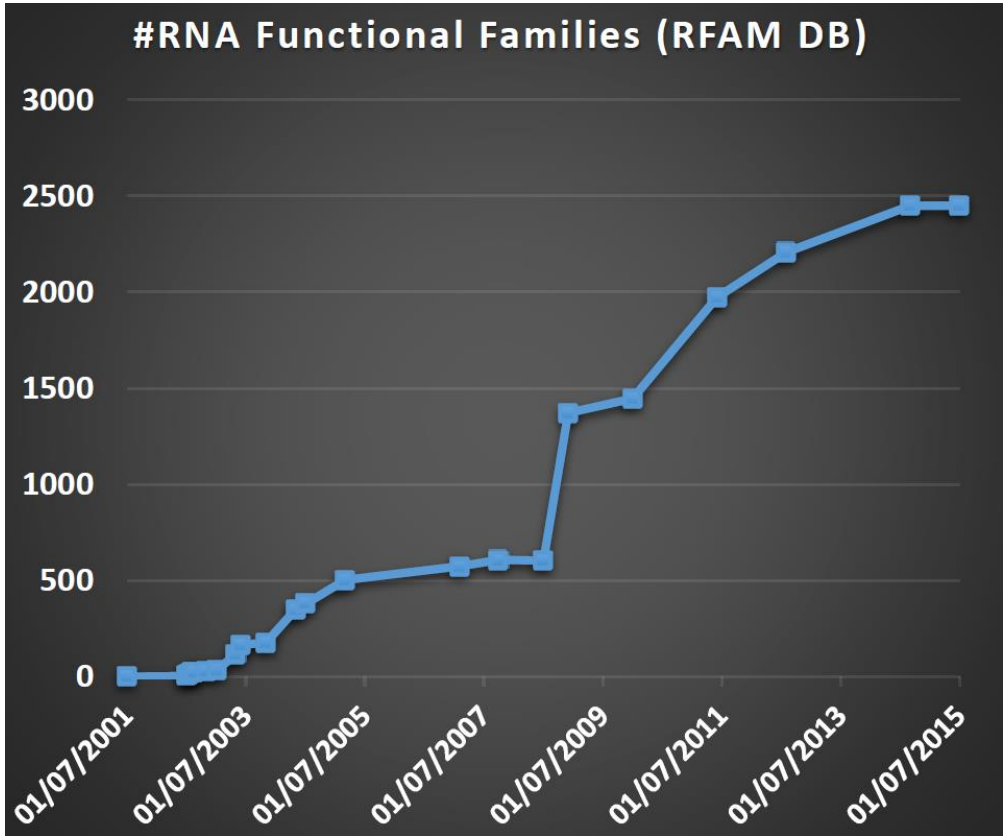


The human genome is **pervasively transcribed**, such that the majority of its bases are associated with at least one primary transcript and many transcripts link distal regions to established protein-coding loci.

ENCODE Analysis of 1% of the human genome Nature 2007

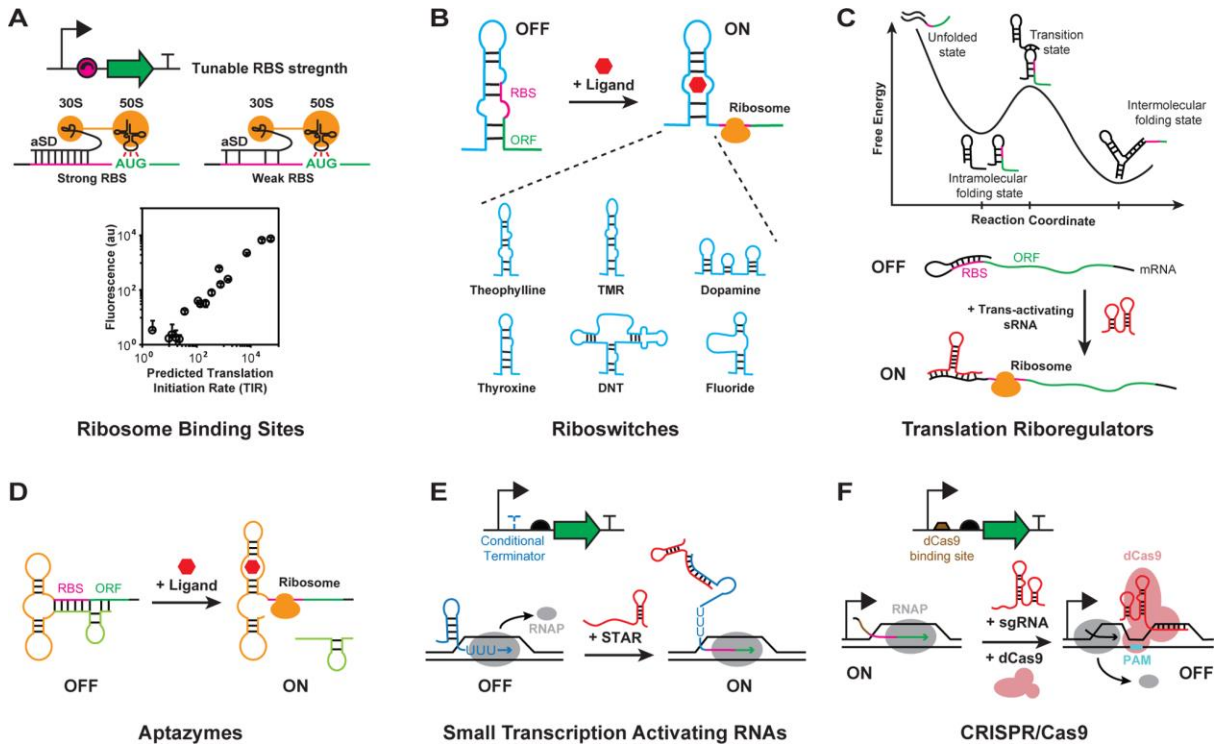
Why RNA is **totally awesome!**

- Ubiquitous
- Pervasively expressed
- Versatile
  - Carriers
  - Transporter
  - Enzymatic
  - Processing
  - Regulatory
  - ssRNA genomes (HIV)
  - Immune system (CRISPR)
  - And more



Why RNA is **totally awesome!**

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

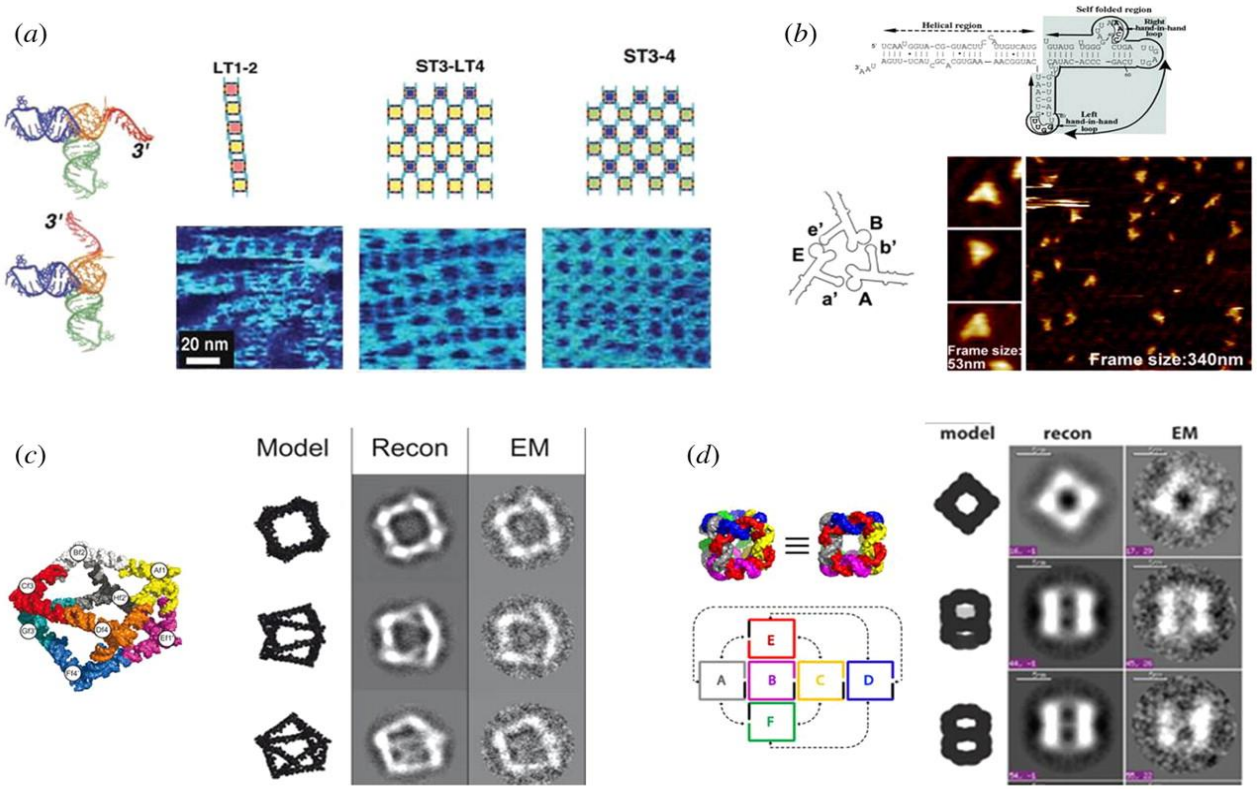


[Kushwaha *et al*, ACS Synthetic Bio. 2016]

A Jaramillo
Evry/Warwick

Why RNA is **totally awesome!**

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



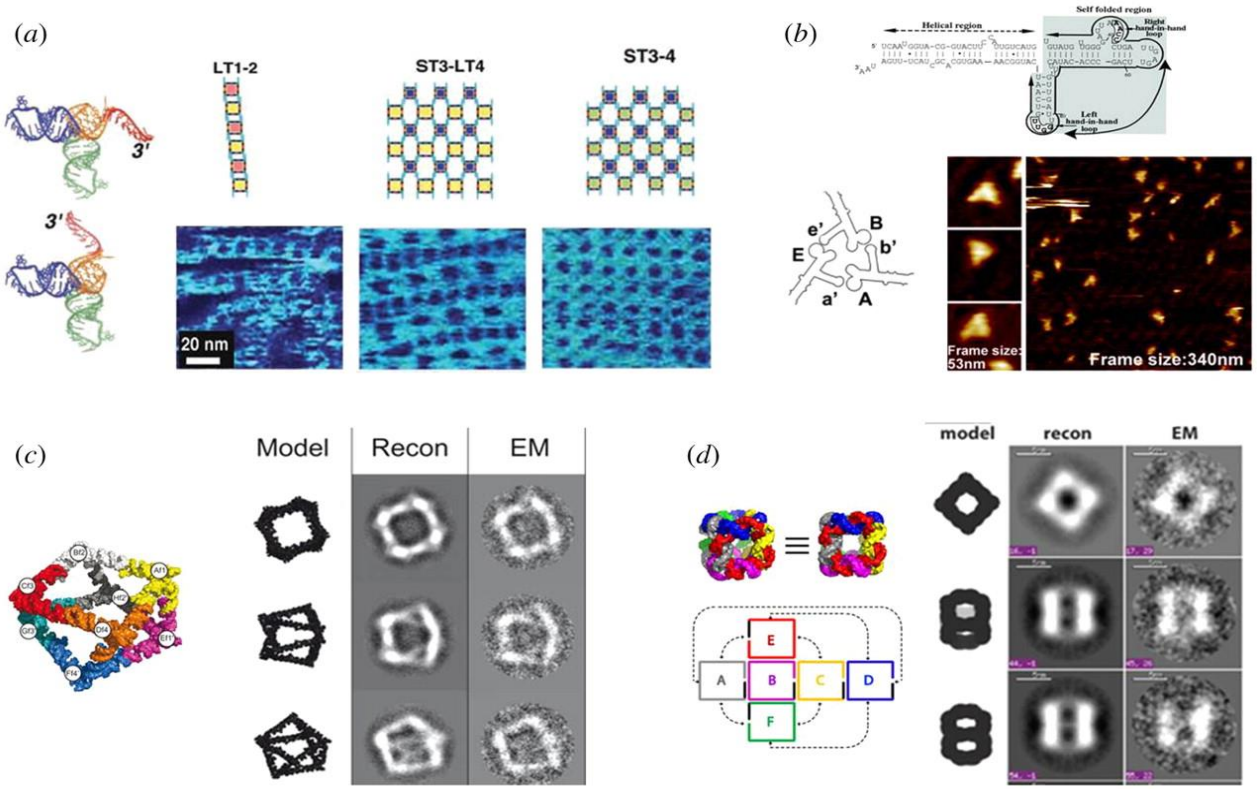
RNA-based Nanoarchitectures
[Li H *et al*, Interface Focus 2011]

Why RNA is **totally awesome!**

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

Rational Design

AMIBio
LIX Polytechnique /Inria Saclay



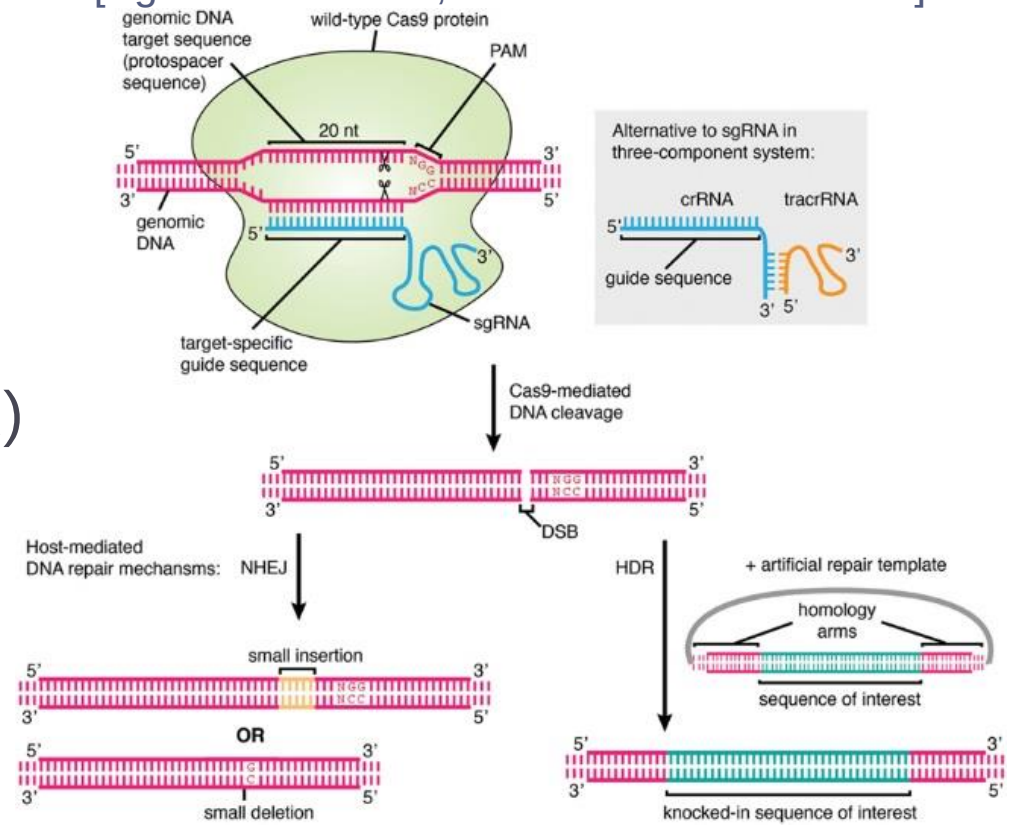
RNA-based Nanoarchitectures
[Li H *et al*, Interface Focus 2011]

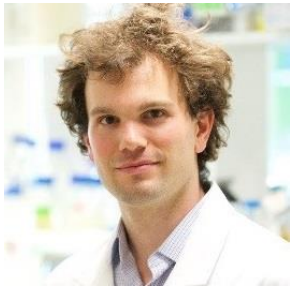
Why RNA is **totally awesome!**

- Ubiquitous
- Pervasively expressed
- Versatile
- Easy to handle
  - Synthetic biology
  - Nanotechs
  - Genome editing (CRISPR)

Blooming therapeutic RNAi…
… making way for CRISPR!

[Agrotis & Ketteler, Frontiers Genetics 2015]





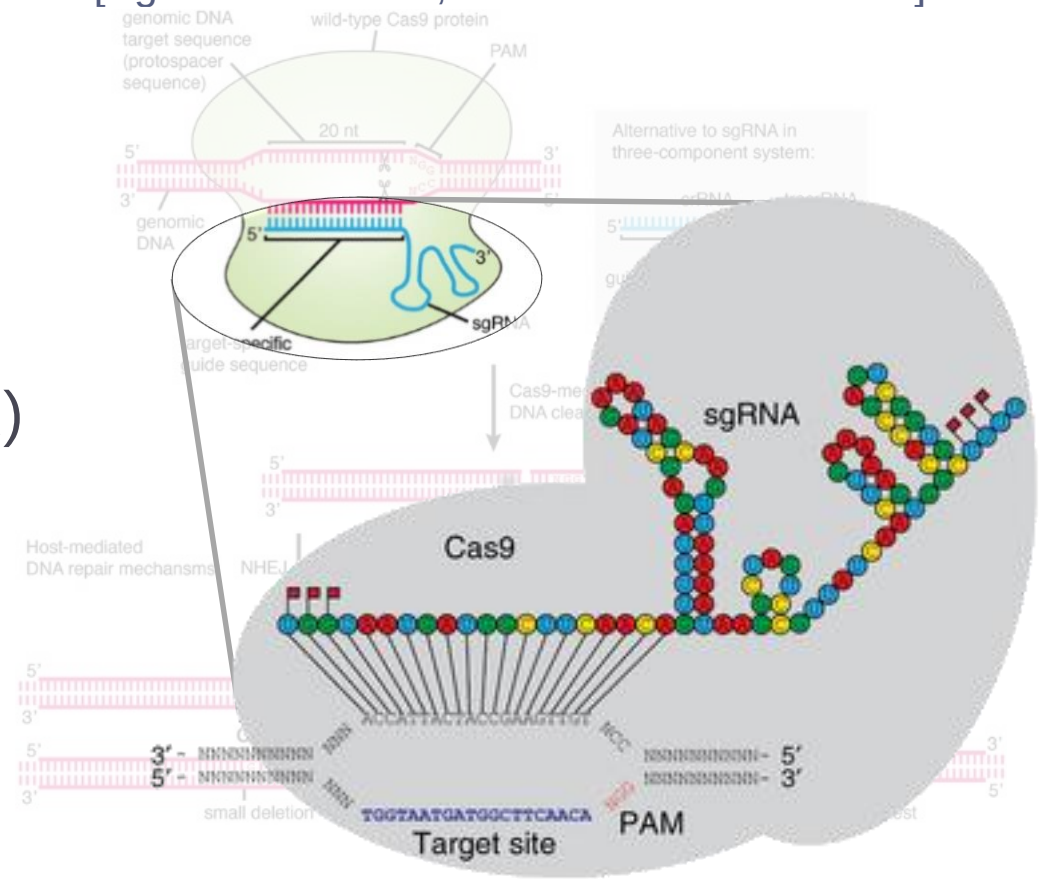
D Bikard
Pasteur

Why RNA is **totally awesome!**

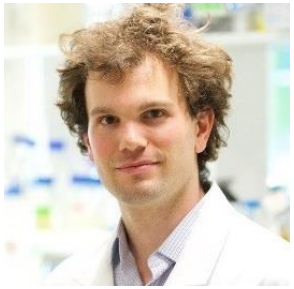
- Ubiquitous
- Pervasively expressed
- Versatile
- Easy to handle
  - Synthetic biology
  - Nanotechs
  - Genome editing (CRISPR)

Blooming therapeutic RNAi…
… making way for CRISPR!

[Agrotis & Ketteler, Frontiers Genetics 2015]



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



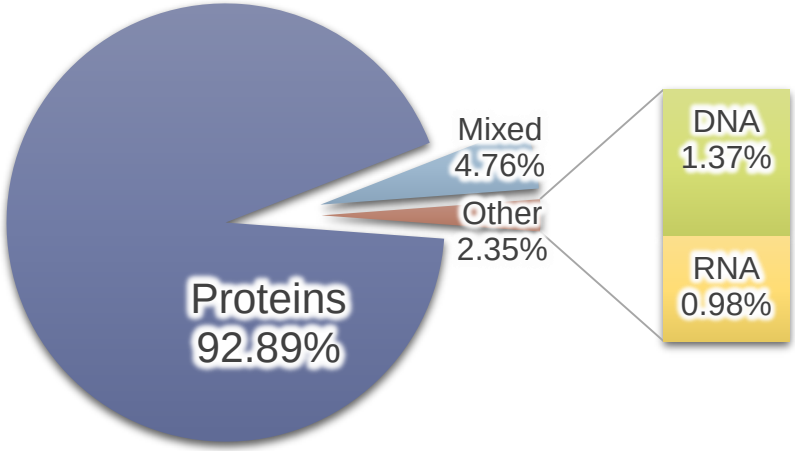
D Bikard
Pasteur

Why RNA is **totally awesome!**

- Ubiquitous
- Pervasively expressed
- Versatile
- Easy to handle
  - Synthetic biology
  - Nanotechs
  - Genome editing (CRISPR)
  - Computationally fun
    (but still challenging…)

D Barth
DAVID, UVSQ

PDB: 117,022 entries (March 2016)



(Initial) lack of structural data

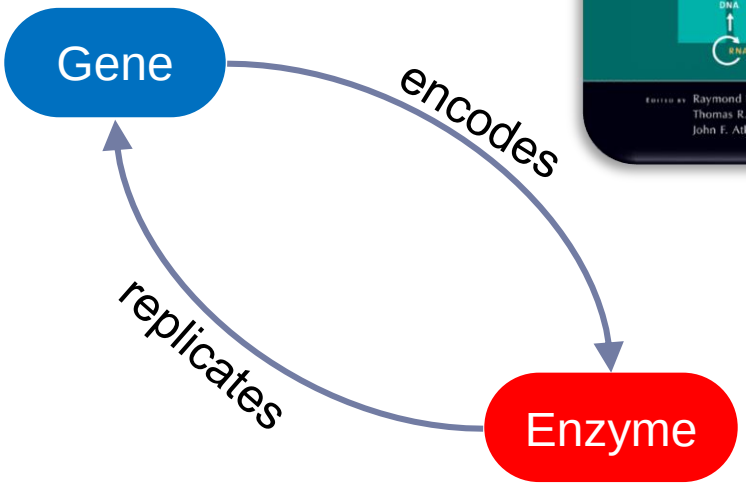
Experiment-based energy models
+ Secondary structure
+ Efficient combinatorial algorithms

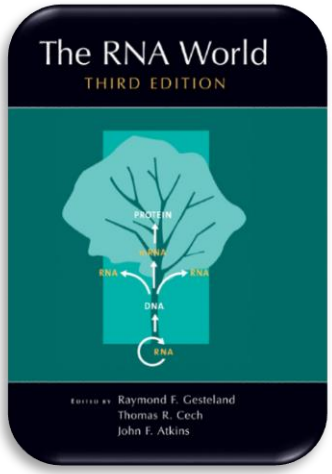
⇒ Mature *ab initio* prediction tools
(Mfold, RNAfold…)

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!?

Chicken vs egg paradox at the origin of life





This is the RNA World:

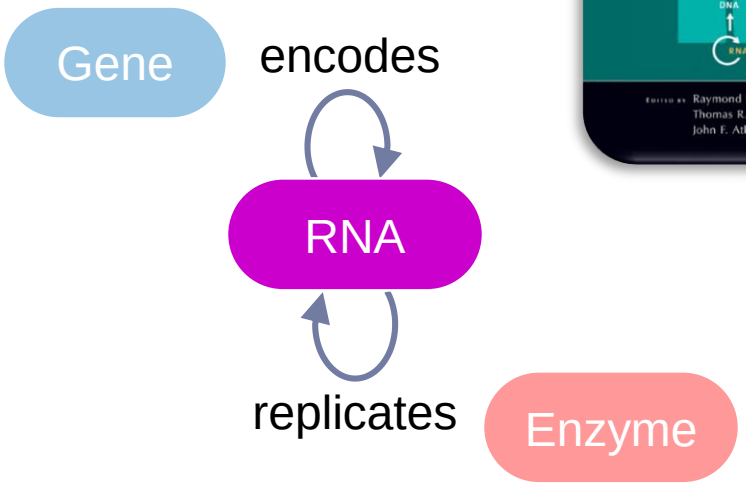
- **Proteins** are good at being enzymes but bad at being replicators;
- **DNA** is good at replicating but bad at being an enzyme;
- **RNA** might just be good enough at both roles to break out of the Catch-22.

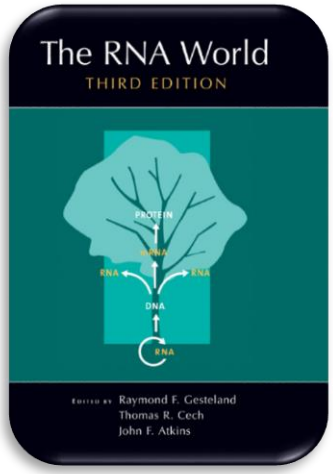
R. Dawkins. The Ancestor's tale

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!?

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

GT MASIM

Méthodes et Algorithmes pour les
Structure(s) et Interactions des Macromolécules

GT MASIM

Depuis Sept. 2017

Objectifs :

Accompagner un dialogue interdisciplinaire
entre méthodologistes et praticiens
en bioinformatique structurale

- Participer à l'animation d'une communauté nationale de (jeunes) chercheurs en bioinformatique structurale;
- Accompagner le développement et la diffusion de méthodologies pérennes, basées sur des fondements algorithmiques bien maîtrisés;
- Recenser l'offre logicielle au niveau national ainsi que les bonnes pratiques tant au niveau développement qu'utilisation;
- Catalyser les échanges avec l'industrie
- …

Pas besoin d'être affilié au GDR BIM pour participer !

GT MASIM

Quelques problèmes et leurs méthodes:

- Analyse cristallographiques : géométrie, apprentissage (régression, classification)
- Analyse RMN: géométrie des distances, modèles Bayésiens
- Analyse comparative : information mutuelle, corrélations structure - séquence
- Analyse de structures cryo-EM : analyse d'images, clustering
- Comparaison de structures : matching et graphes produits, recherche de structure
- Classification de structures tertiaires : classification, clustering
- Modélisation d'assemblages (cryo-EM, spectrométrie de masse) : optimisation multi-critères, inférence de contacts, et algo/théorie des graphes
- Recherche de structures stables - paysages énergétiques: algorithmes d'échantillonnage, optimisation, analyse de processus Markoviens
- Prédiction de structures secondaires (ARN,proteines, chromatine?) : modèles discrets (graphes, arbres), énumération, programmation dynamique
- Modèles pour la thermodynamique : algorithmes d'échantillonnage (dynamique moléculaire et variantes, méthodes de Monte Carlo et variantes), réduction de dimension, modèles harmoniques, analyse spectrale; régression pour la prédiction d'affinité
- Modèles pour la cinétique : modèles de Markov multi-échelle, path sampling
- Parallélisation des codes de calcul
- …

Contacts : Frédéric Cazals (Inria Sophia) & Yann Ponty (LIX)

GT MASIM

Actions :

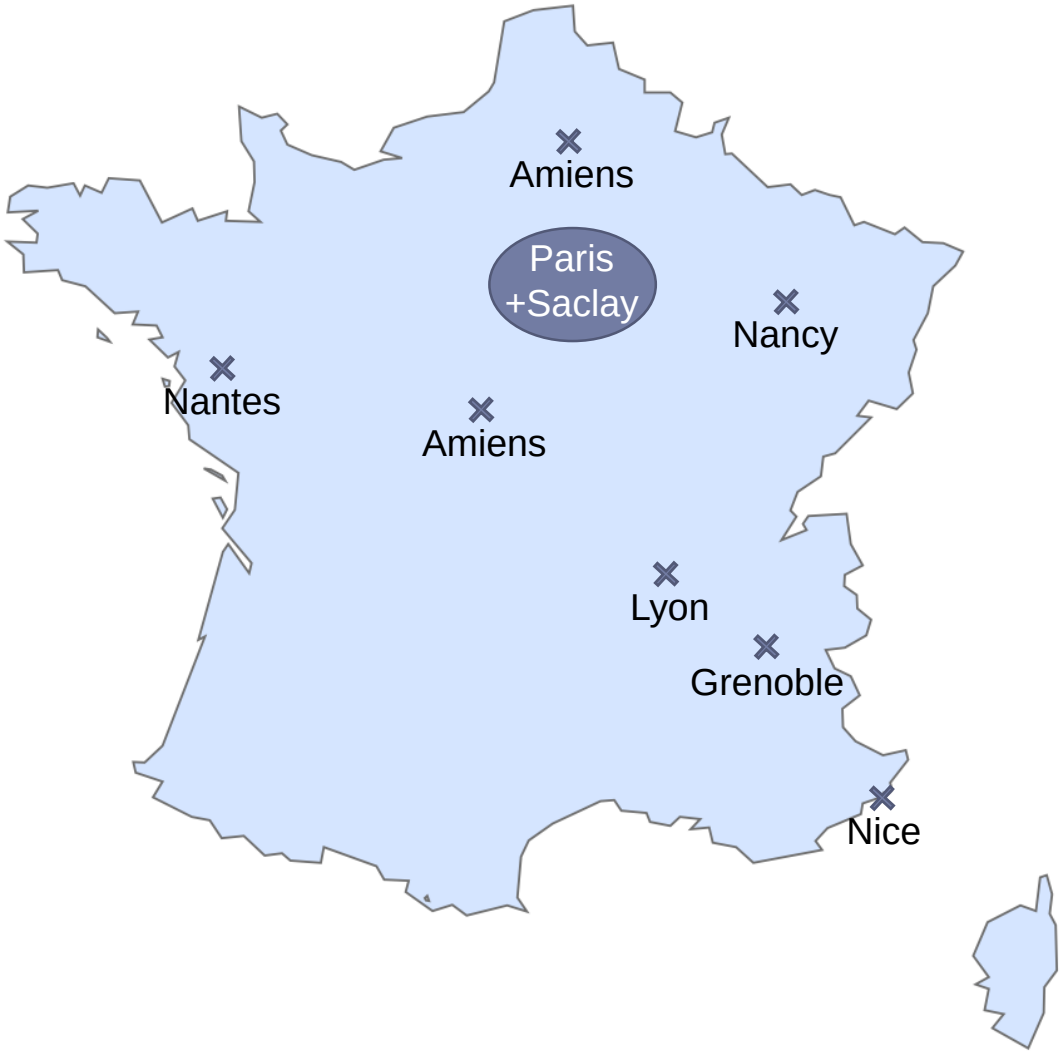
- ✓ Site Web dédié
- ✓ Journées annuelles du GT
- ✓ Mailing list communautaire
- ⏲ Accompagnement organisation école thématiques
- ⏲ Répertoire de l'offre logicielle nationale
- ✗ Newsletter (*highlights* scientifiques)
- Et ??? …

En coordination avec le GGMM !

Contacts : Frédéric Cazals (Inria Sophia) & Yann Ponty (LIX)

Journées de lancement 16-17 Nov 2017

- Inria Paris
- ~50 participants
- 27 exposés
- 7 sessions
- 18+ litres de café
- 2 organisateurs épuisés
- …



Journées de lancement 16-17 Nov 2017

- Inria Paris
- ~50 participants
- 27 exposés
- 7 sessions
- 18+ litres de café
- 2 organisateurs épuisés
- …

But : Couverture nationale de la communauté

Journées de lancement 16-17 Nov 2017

- Inria Paris
- ~50 participants
- 27 exposés
- 7 sessions
- 18+ litres de café
- 2 organisateurs épuisés
- …

But : Couverture nationale de la communauté