



Macroevolution Simulation of RNA

Sheng Guo, Li-San Wang, Junhyong Kim
University of Pennsylvania



Macroevolution of bio-molecules

Simplifications

- Evolve independently
- Determine the fitness of a species
- No genetic polymorphism



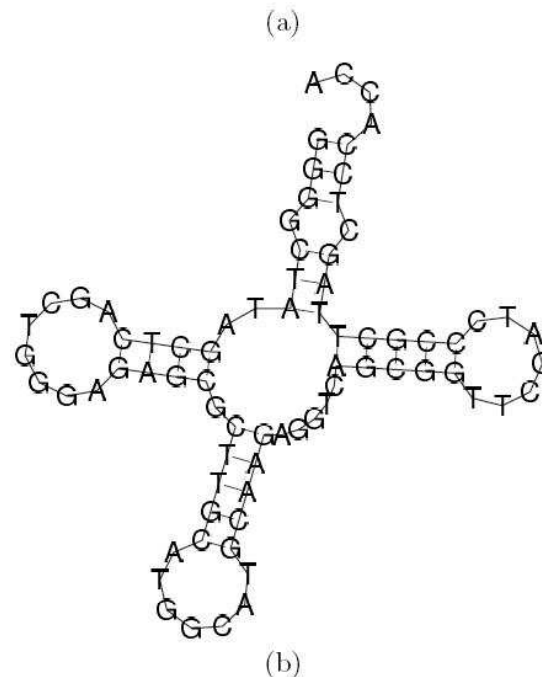
Simulation of complex evolutionary processes

Reflect more complex dynamics

- Heterogeneous rates:
 - lineage and site specific mutation rates
 - genomic context dependent rates
- Phenotypic effects
 - Selection
 - Population interaction

RNA and its secondary structure as a model system for genotype-phenotype evolution

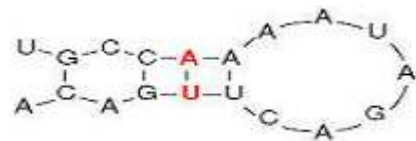
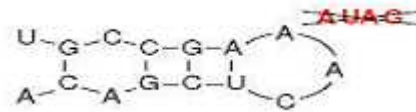
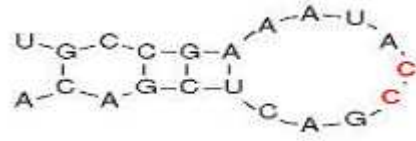
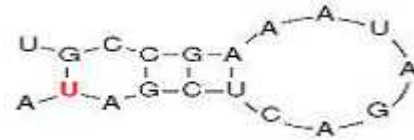
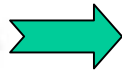
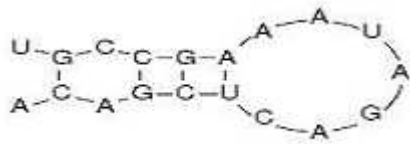
```
>Ecoli-0157-Wisc.trna75 (3314857-3314782)
GGGGCTATAGCTCAGCTGGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTTCGATCCCGCTTAGCTCCACCA
(((.((((.((((.....))))).((((.....))))). .... (((((.....)))))))).))....
```



(a). tRNA sequence and secondary structure in parenthesis format.

(b). Secondary structure drawn by *RNAPlot* from the *Vienna* package

Mutations in RNA



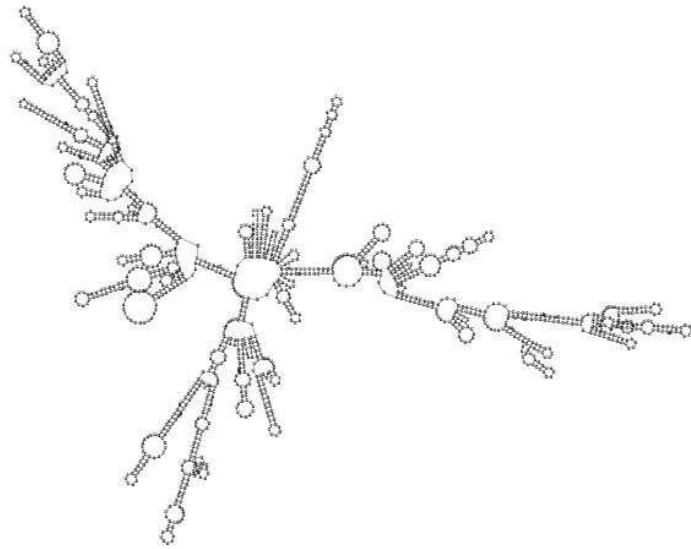
?

Advantageous

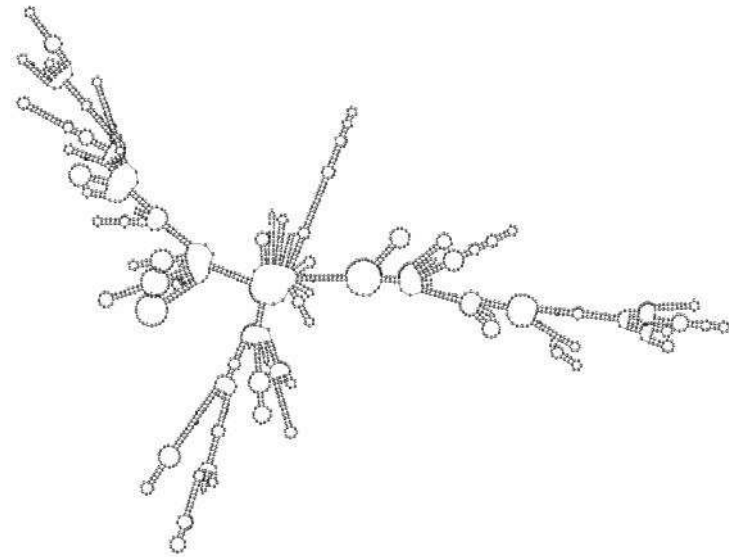
Neutral

Deleterious

Folding energy based fitness model



-491.07J/mol



-636.71J/mol



Assumption: Thermodynamically more stable structure is more fit.



Fixation probability as a function of fitness

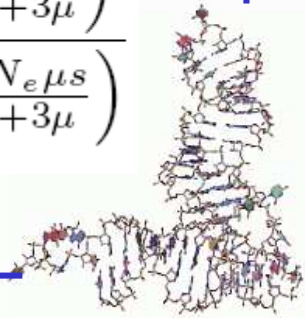
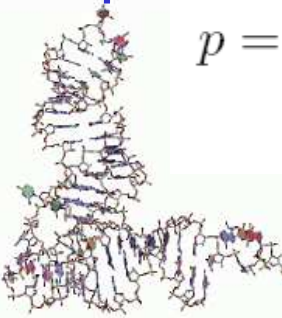
Neutral

$$p = 1/N_e$$

Advantageous ($s > 0$)/Deleterious ($s < 0$)

$$p = \frac{1 - \exp(-2s)}{1 - \exp(-2N_e s)}$$

Compensatory Mutation

$$p = N_e \frac{8N_e\mu^2 - s - 3\mu}{s + 3\mu} e^{\frac{8\mu s(N_e - 1)}{s + 3\mu}} \frac{\text{hypergeom}\left(1, -\frac{2(4N_e\mu^2 - s - 3\mu)}{s + 3\mu}, \frac{8\mu s}{s + 3\mu}\right)}{\text{hypergeom}\left(1, -\frac{2(4N_e\mu^2 - s - 3\mu)}{s + 3\mu}, \frac{8N_e\mu s}{s + 3\mu}\right)}$$


Parameters:

N_e : effective population size

μ : neutral mutation rate

s : fitness change



Macro-evolutionary change acceptance rates

Neutral

$$r = \mu.$$

Advantageous ($s > 0$) / Deleterious ($s < 0$)

$$r = N_e \mu p$$

Compensatory Mutation

$$r = N_e \mu^2 p$$

Parameters:

N_e : effective population size

μ : neutral mutation rate

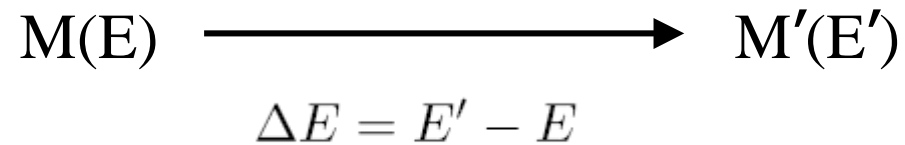
s : fitness change

p is the corresponding fixation probability







Compute s : a free-energy based schema

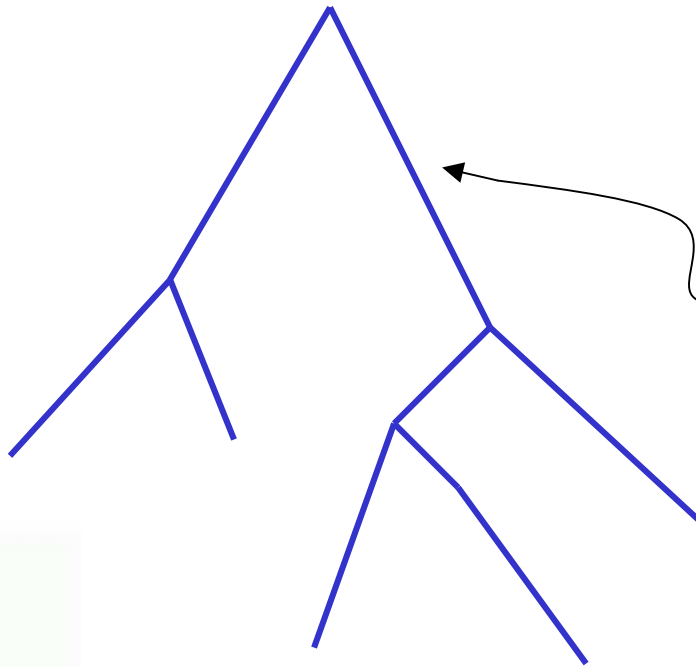
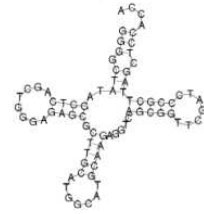


$$\Delta E \begin{cases} < 0 & \text{advantageous mutation} \\ = 0 & \text{neutral mutation} \\ > 0 & \text{detrimental mutation} \end{cases}$$

$$t = e^{-\alpha \Delta E} \quad (\text{wildtype has fitness } 1)$$

$$s = t - 1 = e^{-\alpha \Delta E} - 1$$


Macroevolution on a phylogenetic tree



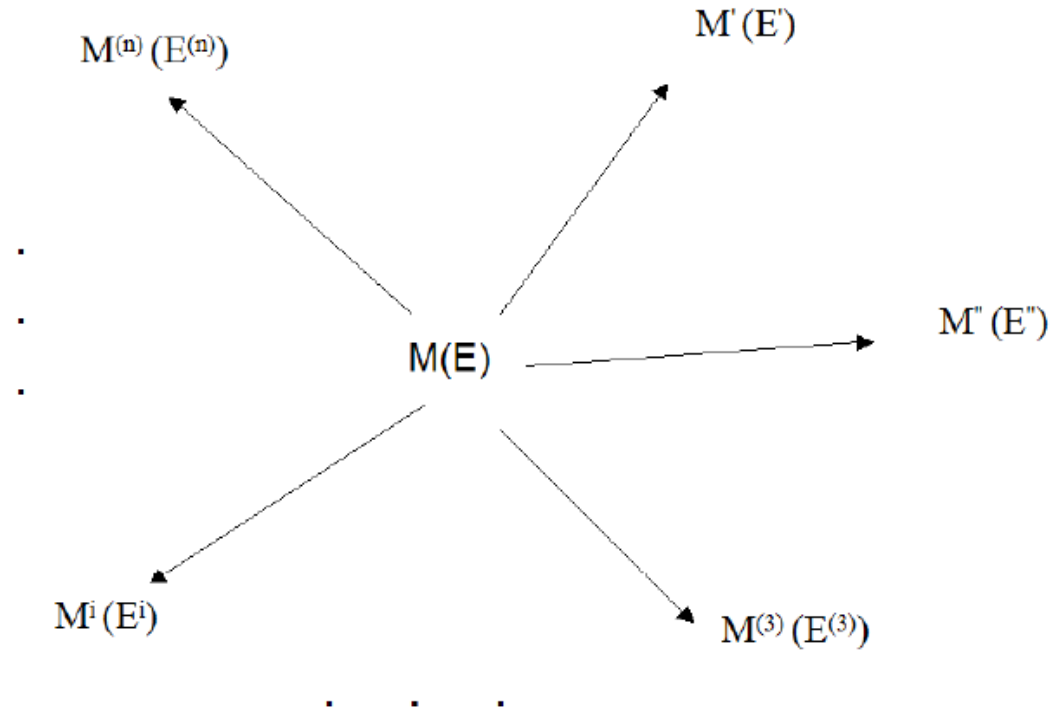
$\{M, T, Ne, \mu, \alpha\}$

M: ancestor RNA

*T: rooted phylogeny
with edge length.*

Context dependent
evolutionary rates: Event
rate depends on the state
of the molecular at any
given time

One-step mutation ensemble of a RNA



The one-step mutant of a wildtype may have different free energy than the wildtype, hence may change the fitness of RNA.



Simulate mutation of RNA

Naïve algorithm:

Give RNA molecule M , randomly draw a mutant, compute its fixation probability p . *Accept this mutant with probability p .*

problem: p is very small ($<10^{-5}$)

Improvement:

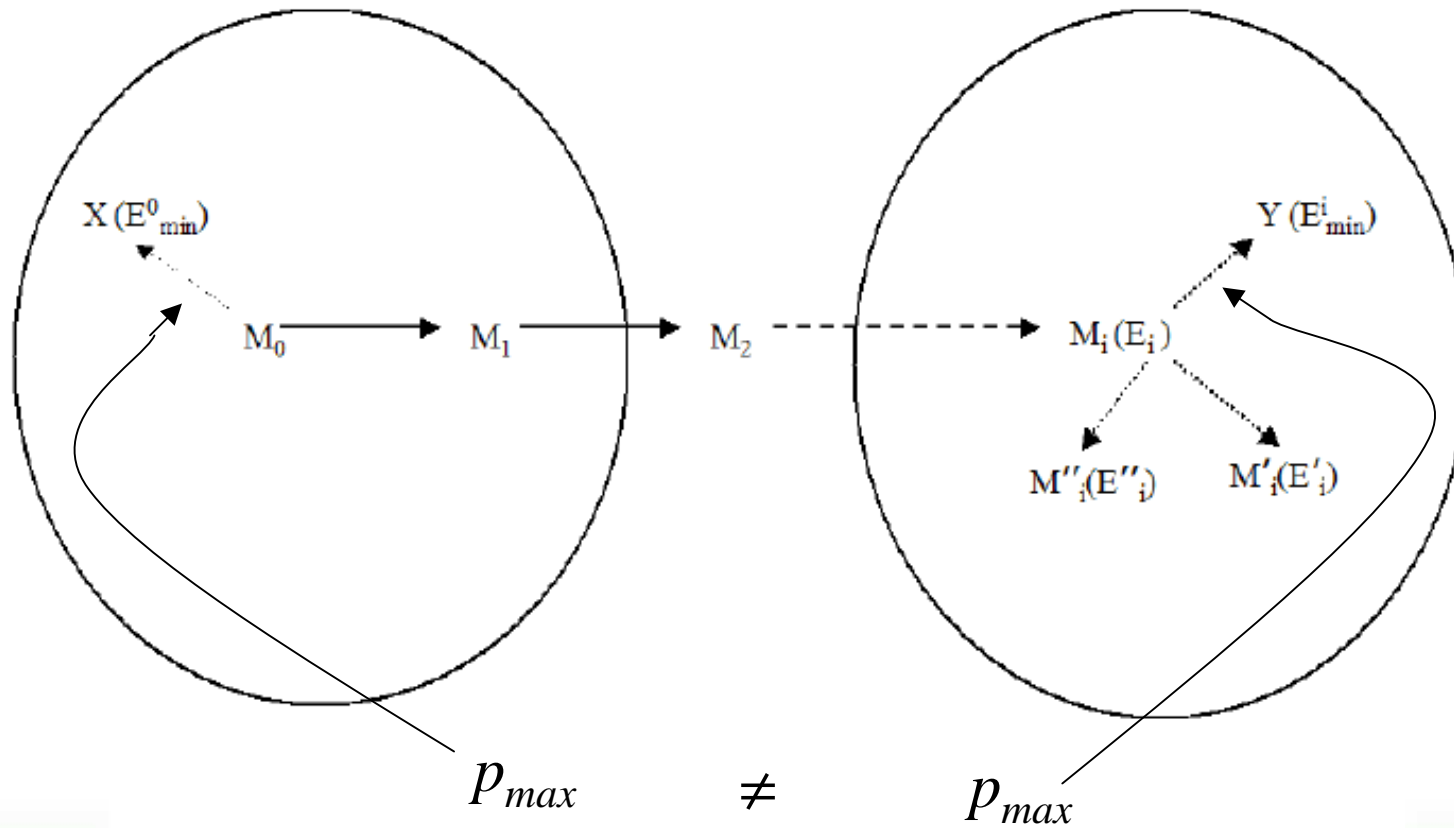
Compute the maximal fixation probability p_{max} for the whole mutant ensemble. *Accept a mutant with probability p/p_{max}*

p/p_{max} is usually in the order of $10^{-1} \sim 10^{-2}$

Still problematic, why?



Simulate mutation of RNA



M_0 is the ancestral RNA, $M_1 \dots M_i$ are the intermediate RNAs leading to the descendant RNA.

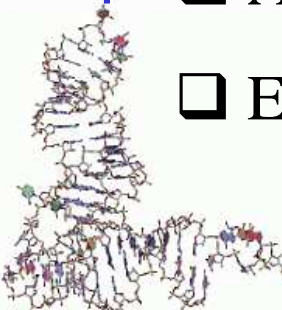


Simulate mutation of RNA--Heuristics

Use a globe p_{max} , for the whole tree.

- Equivalently, we use a global ΔE_{max} .
- To estimate ΔE_{max} we *only* enumerate all mutants of the root RNA, and set ΔE_{max} to be the free energy difference between root RNA and its optimal mutant.
- We *update* ΔE_{max} when we see a better one later on.

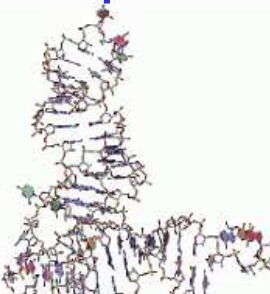

Further improvements

- Add a small slack constant, use $\Delta E_{max} + \xi$ as a safeguard.
 - Estimate ΔE_{max} from more RNA's in the tree.
- 

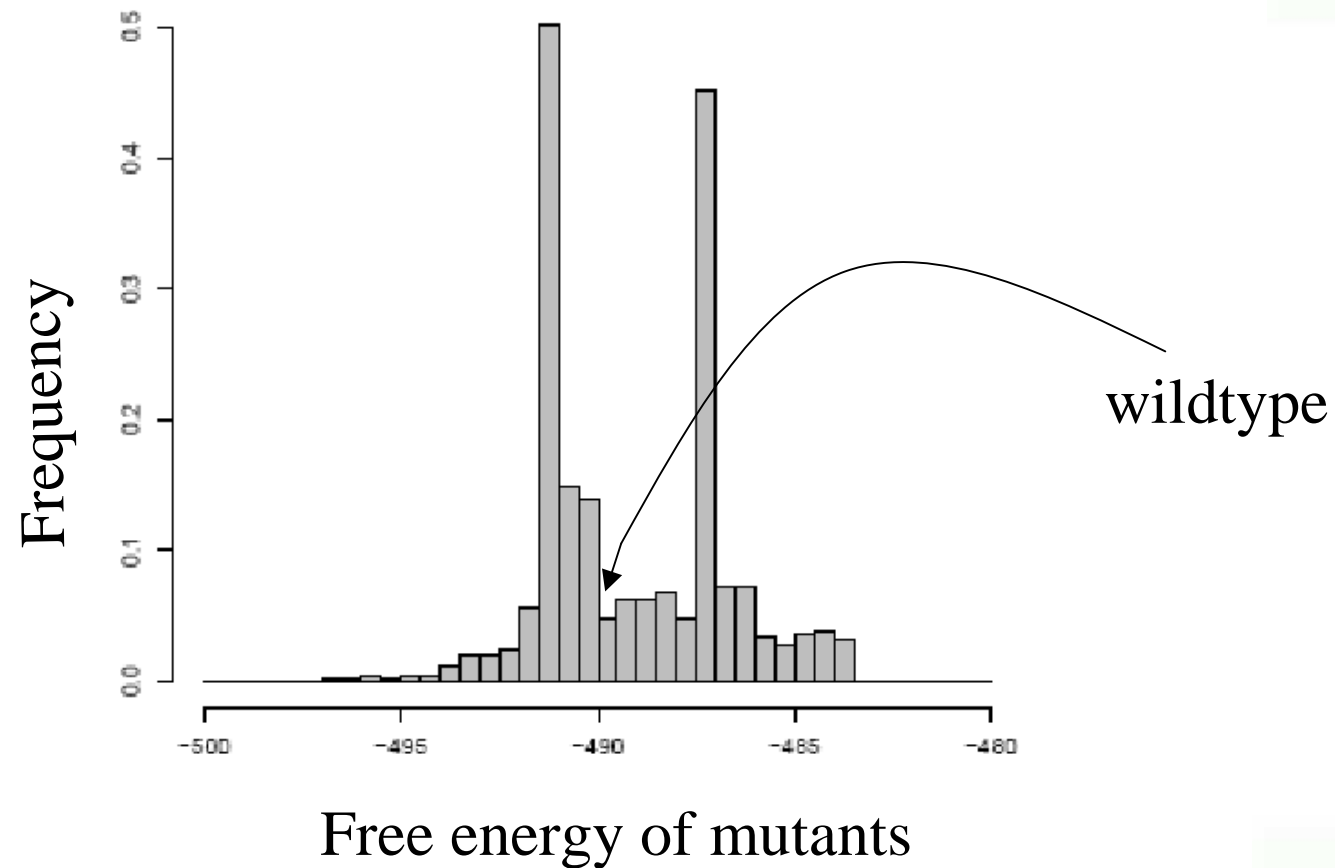


Simulate mutation of RNA--Heuristics

Pitfalls

- If ΔE_{max} is too big, then p_{max} is too big, it results too many futile samplings (acceptance probability is p/p_{max})
 - If ΔE_{max} is too small, it violates the acceptance method. ($p/p_{max} > 1$, while in principle $p/p_{max} \leq 1$)
- 
- 

Simulate mutation of RNA--Heuristics



Free energy change is small, example uses *E.coli* small subunit rRNA (1542nt), there are ~9,000 mutants



Simulate mutation of RNA—A Case Study

Settings:

1. 5,000-taxon binary phylogenetic tree
2. $N_e = 10,000$; $\mu = 1.0 \times 10^{-6}$; $\alpha = 1.0 \times 10^{-4}$
3. Root sequence : *E.coli* ssuRNA 1542nt with free energy -491.07J/mol

Platform:

1. Single Pentium IV 1.5 GHz CPU with 512M RAM.
- 
- 

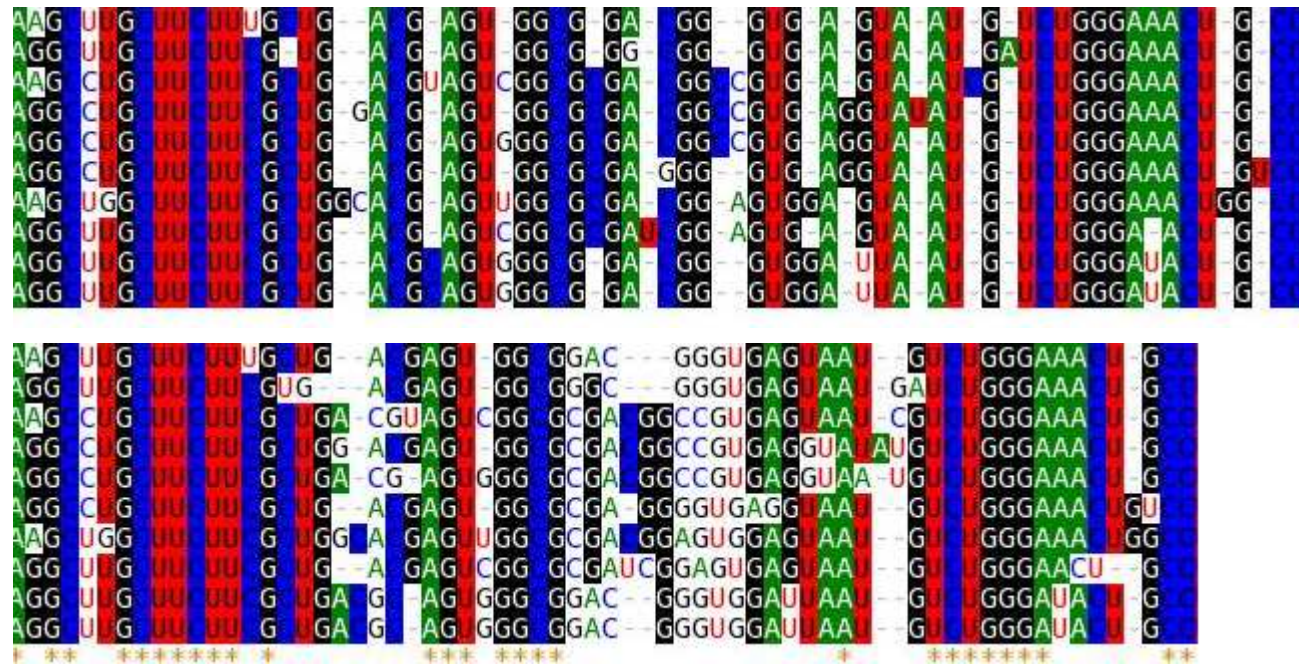
The slide features four RNA secondary structure diagrams, one in each corner. These diagrams are complex, branching structures with various colored nodes (red, green, blue) and lines representing base pairs. A blue border frames the central text area.

Simulate mutation of RNA—A Case Study

Running Time: 940 seconds (~16 minutes).

1. Total number of fixed mutations is 81,349. On average, each RNA(ancestral+extant) was hit by 8 mutations.
2. Total number of sampled mutations is 2,933,500, so acceptance ratio is 2.77% (This ratio is different for distinct branches, the variation is $< 0.3\%$).
3. Free energy difference for the root and its optimal mutant is -7.1J/mol , the global maximal difference is -8.1J/mol . {So the slack constant $\xi = 1$, indeed small }
4. All leaf RNA's are more stable than root RNA, the most stable one is -696.19J/mol (root = -491.07 ; $\sim 42.5\%$ decrease)

Simulate mutation of RNA—A Case Study



Top is homologous alignment. Bottom is Clustalw alignment.

First sequence is root RNA, others are randomly chosen leaf RNA's



Summary

- ❑ A principled simulator for the macroevolution of the RNA genotype-phenotype model
- ❑ Few parameters and flexible settings (N_e, μ, α)
- ❑ Running time is linear-scalable, suitable for large-scale (million-taxon phylogeny) simulation



Acknowledgements

Penn

- Stephen Fisher
- Yifeng Zheng
- Susan Davidson

UT Austin

- Tracy Heath
- David Hillis

NSF/CIPRES

Penn Genomics and Computational Biology Group