COMP598: Advanced Computational Biology Methods and Research

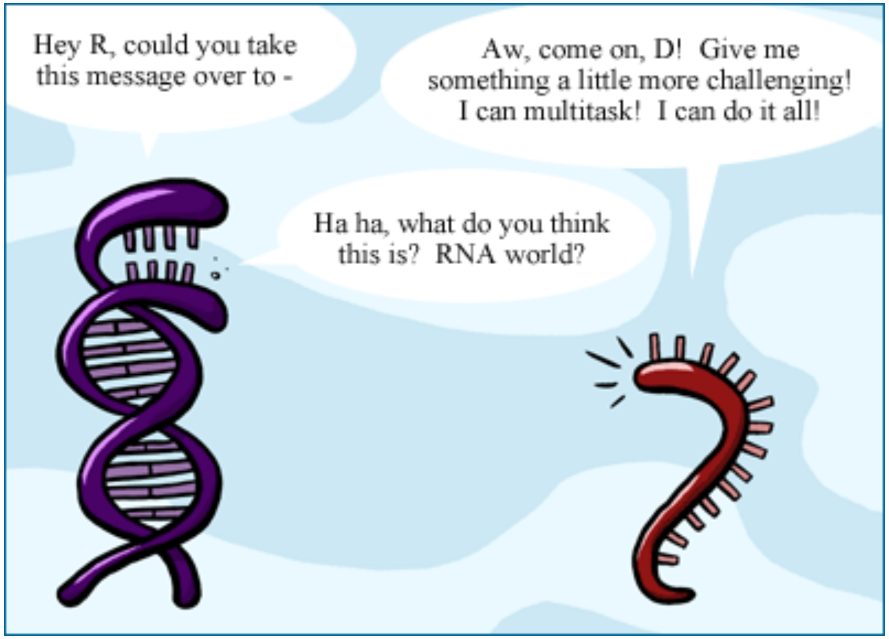
RNA in the sequence/structure network

Jerome Waldispuhl
School of Computer Science, McGill

RNA world

In prebiotic world, RNA thought to have filled two distinct roles:

1. an information carrying role because of RNA's ability (in principle) to self-replicate,
2. a catalytic role, because of RNA's ability to form complicated 3D shapes.


Hey R, could you take this message over to -
Aw, come on, D! Give me something a little more challenging! I can multitask! I can do it all!
Ha ha, what do you think this is? RNA world?

Over time, DNA replaced RNA in Its first role, while proteins replaced RNA in its second role.

Principles

Central assumptions:

- The structure of a sequence can be determined using thermodynamics principles.
- The structure determines the function.
- Evolution tends to preserve and optimize the function.

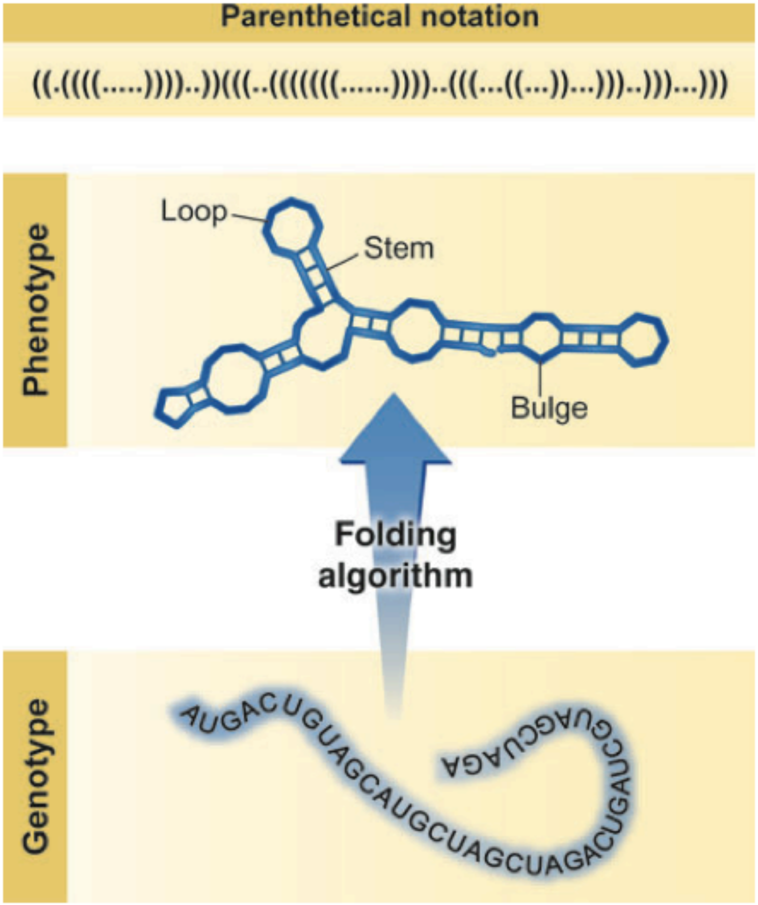


Figure from (Cowperthwaite&Meyers,2007)

Outline

- Mathematical modelling
- Characterizing the evolutionary landscape
- Evolutionary dynamics

Sequence evolution

For short sequences, the set of evolutionary operations can be restricted to:

- Insertion
- Insertion/Deletion
- Mutation

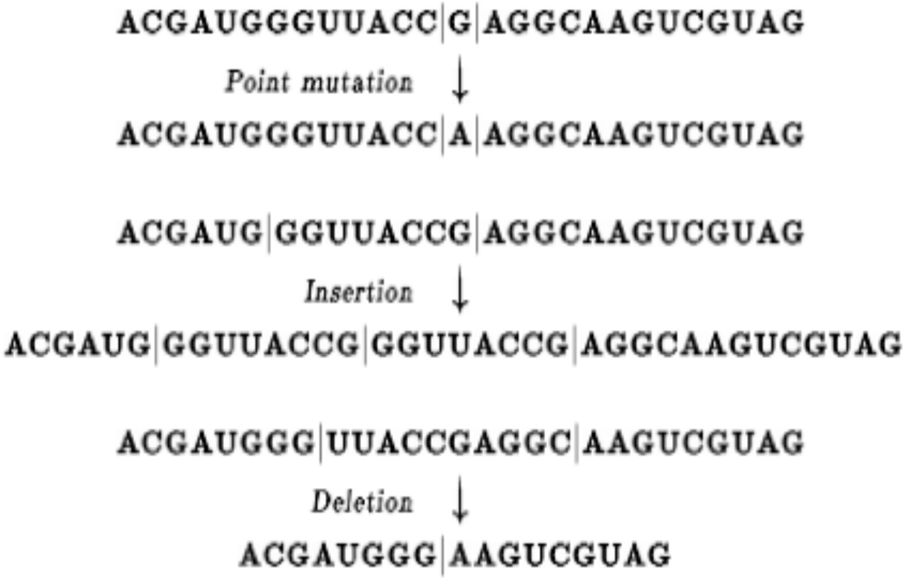

ACGAUGGGUUACC|G|AGGCAAGUCGUAG
Point mutation
ACGAUGGGUUACC|A|AGGCAAGUCGUAG
ACGAUG|GGUUACCG|AGGCAAGUCGUAG
Insertion
ACGAUG|GGUUACCG|GGUUACCG|AGGCAAGUCGUAG
ACGAUGGG|UUACCGAGGC|AAGUCGUAG
Deletion
ACGAUGGG|AAGUCGUAG

Figure from (Gobel,2000)

Mutational landscape

When the length of the sequence is fixed, the set of operations can be restricted to mutations.

The mutation landscape is represented with Hamming graphs, where nodes are the sequences and edges connect sequences differing from one single nucleotide (i.e. 1 mutation).

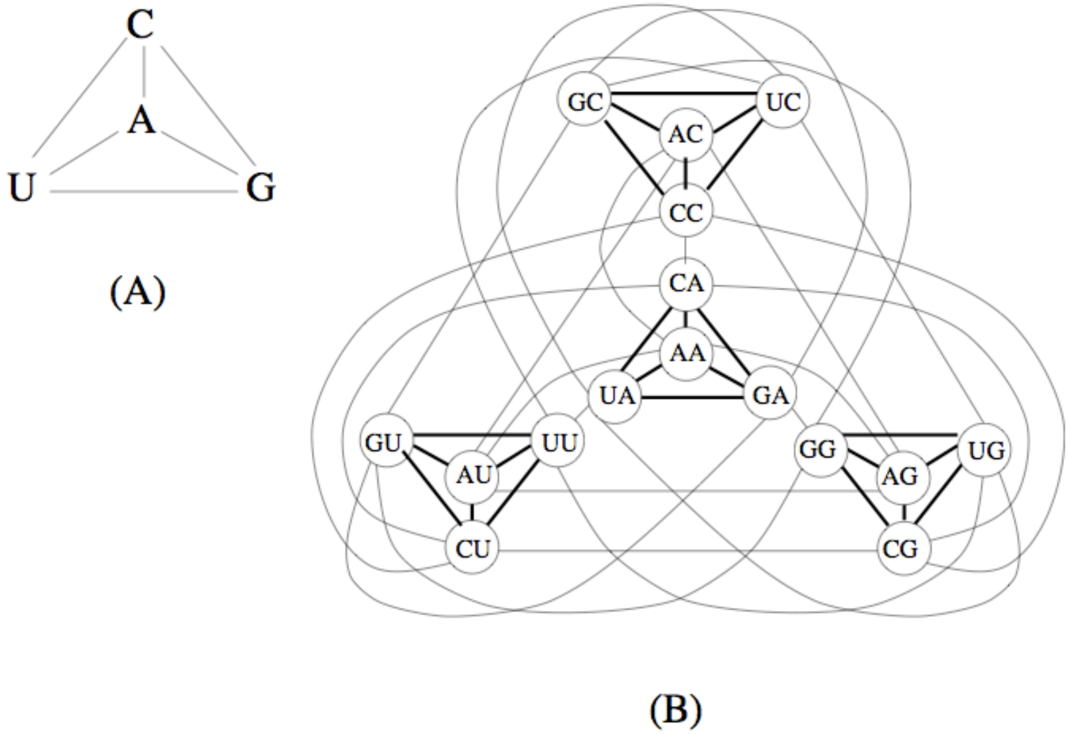


Figure from (Gobel,2000)

Assigning a Phenotype

Use folding programs (E.g. RNAfold, RNAstructure) to calculate the Phenotype.

Usually, we assign a single structure (the M.F.E.) to the sequence but more sophisticated model have been proposed (i.e. plastic model).

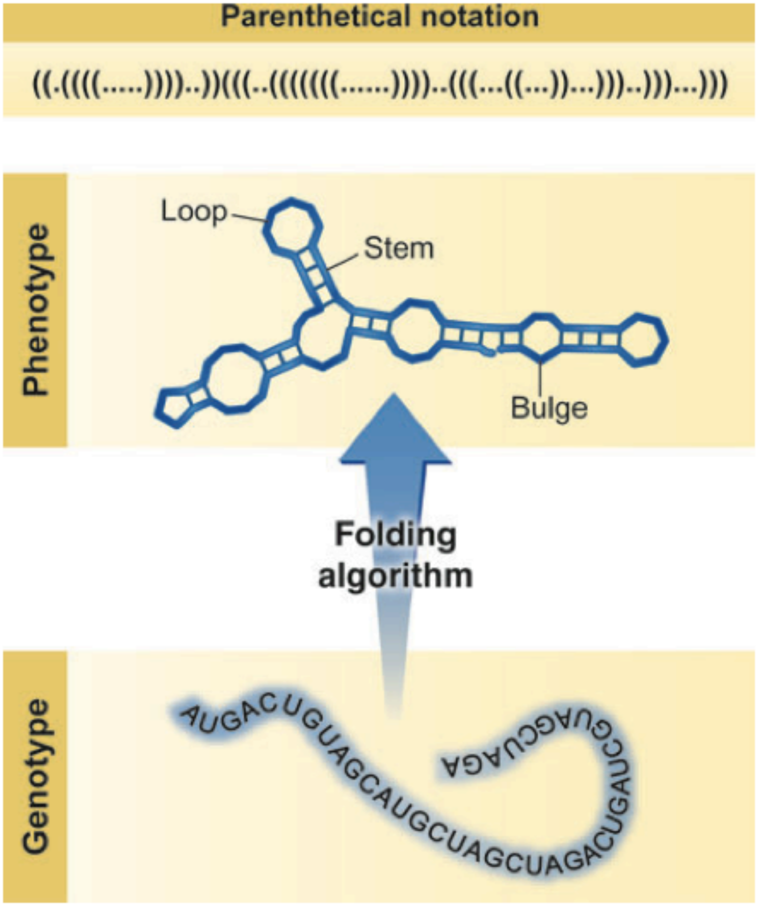
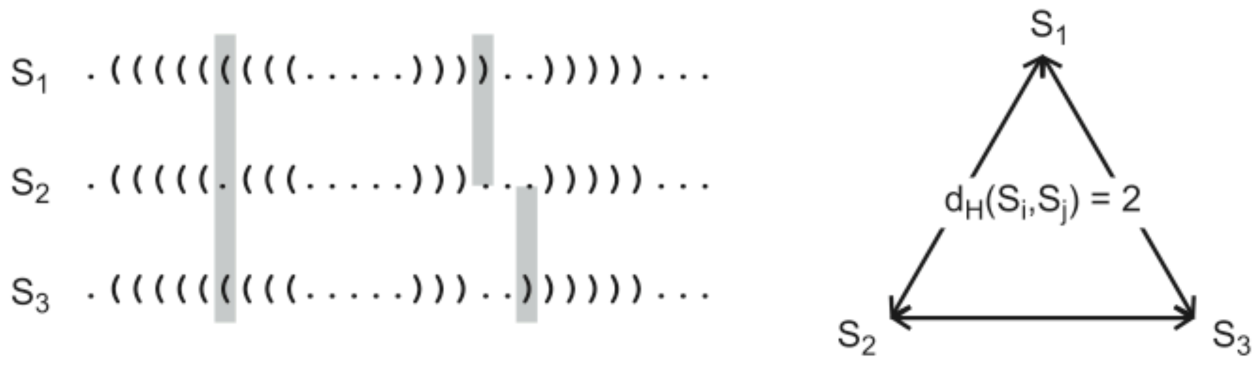


Figure from (Cowperthwaite&Meyers,2007)

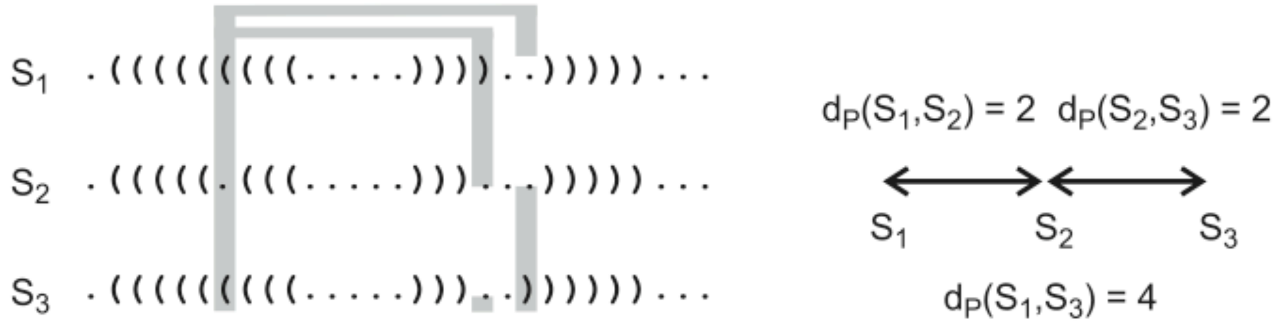
Evaluating structure similarities

Hamming distance:



S1 .(((((((((.....))))..)))))...
S2 .(((((.(((.....)))...)))))...
S3 .(((((((((.....)))..))))))...
S1
dH(Si,Sj) = 2
S2
S3

Base pair distance:

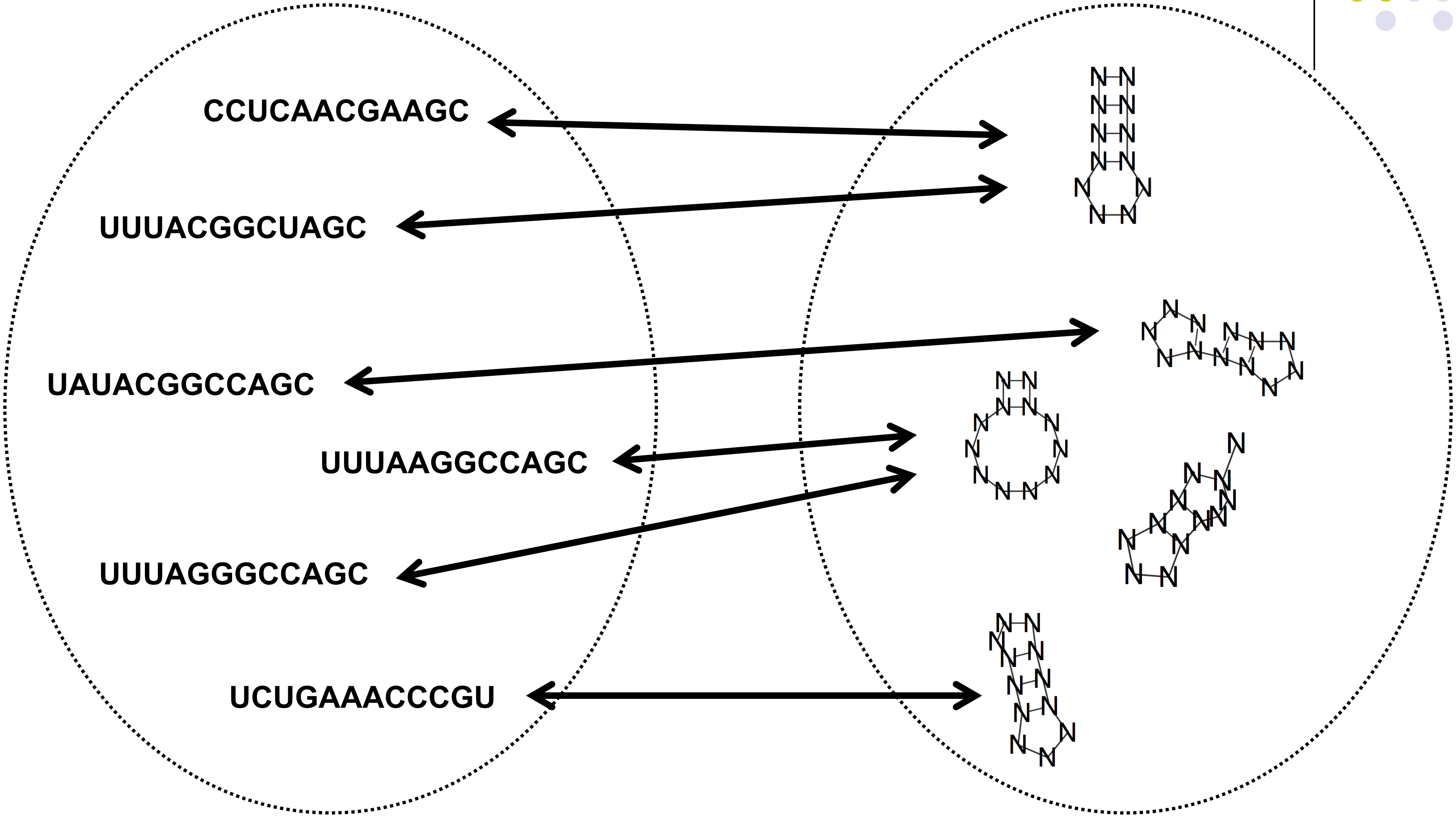


S1 .(((((((((.....))))..)))))...
S2 .(((((.(((.....)))...)))))...
S3 .(((((((((.....)))..))))))...
dP(S1,S2) = 2
dP(S2,S3) = 2
S1
S2
S3
dP(S1,S3) = 4

Base pair distance is the standard. It corresponds to the number of base pairs we have to remove and add to obtain one structure from the other. Both metrics have to be applied on structures of equal length.

Figure from (Schuster&Stadler,2007)

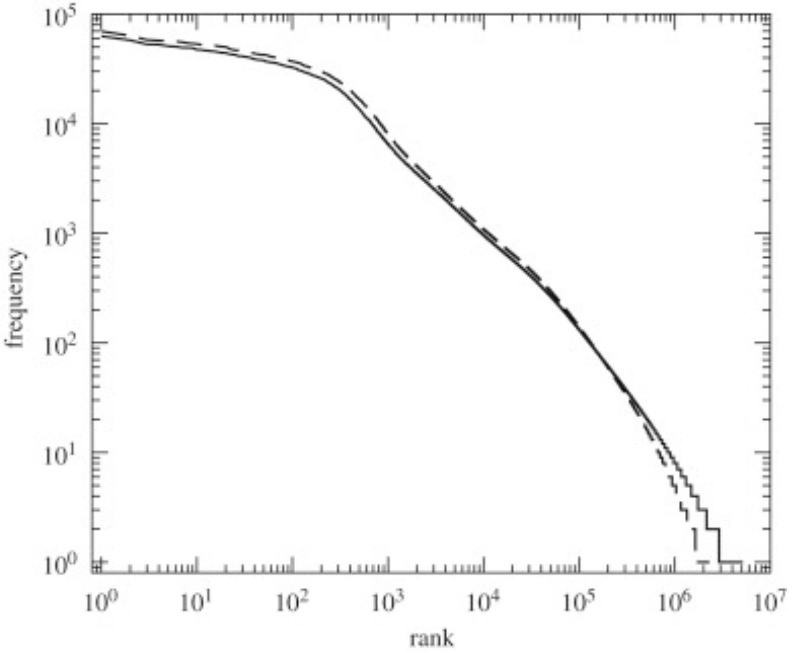
RNA sequence-structure maps

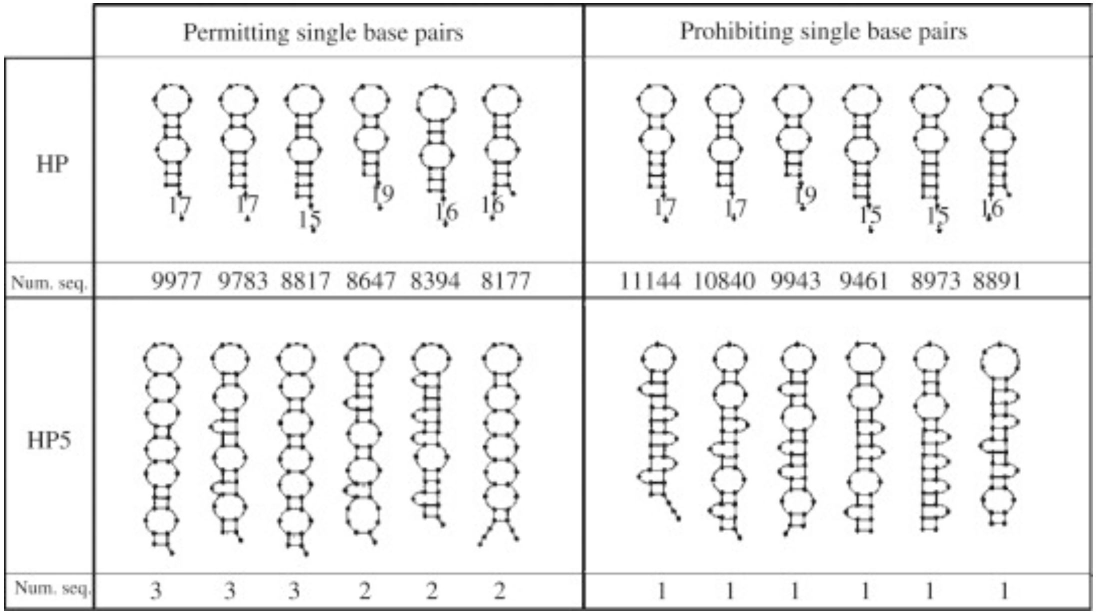


Sequence ensemble

Structure ensemble

Structural repertoire of random RNAs





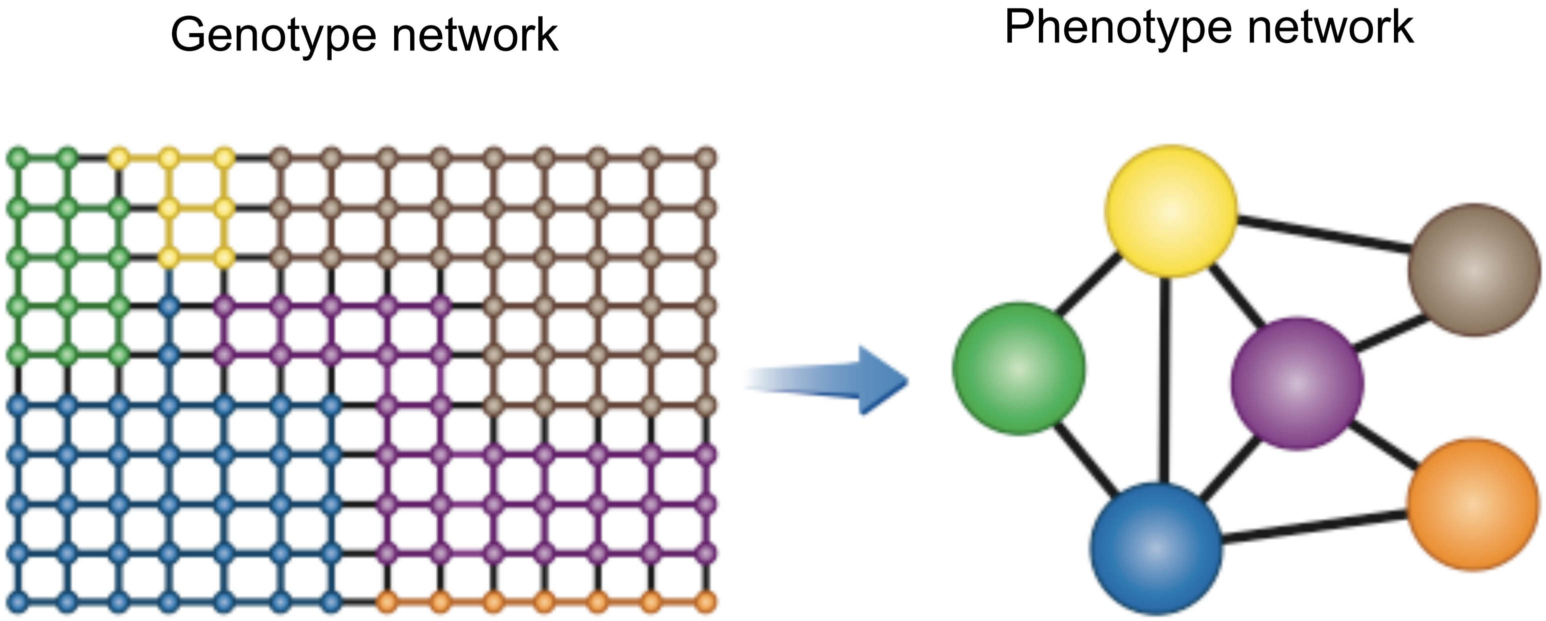
	Permitting single base pairs						Prohibiting single base pairs					
HP	17	17	15	19	16	16	17	17	19	15	15	16
Num. seq.	9977	9783	8817	8647	8394	8177	11144	10840	9943	9461	8973	8891
HP5												
Num. seq.	3	3	3	2	2	2	1	1	1	1	1	1

Abundance of structures

Most abundant structures

(Stich et al., 2008)

Neutral network



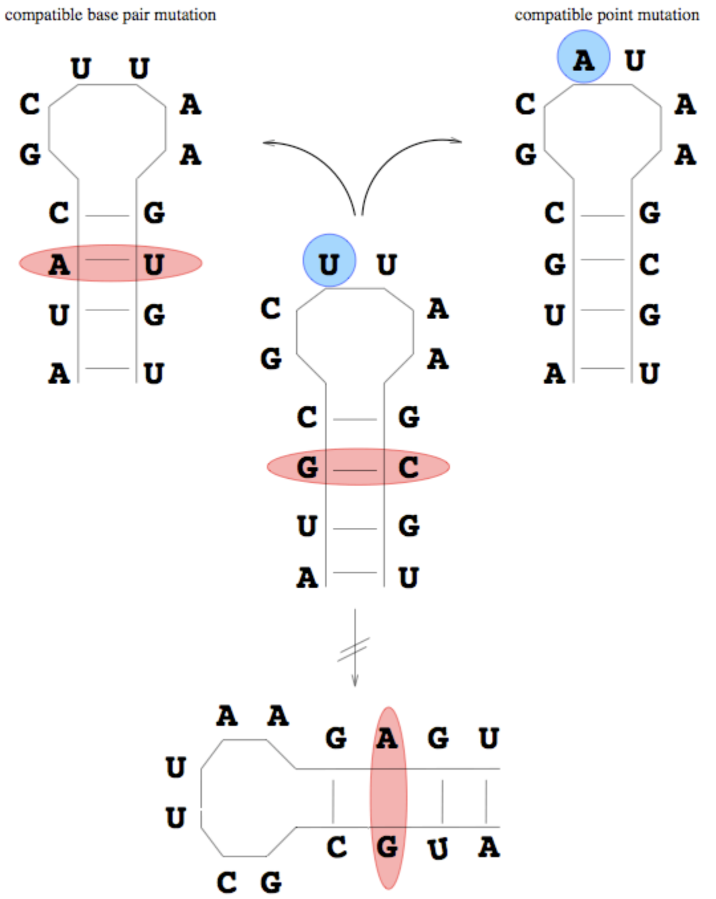
Genotype network

Phenotype network

- A structure is associated to each node (sequence) of the Hamming graph.
- Networks with the same phenotype are a neutral network.
- Introduced & studied by P.Schuster and Vienna group in 1992.

Figure from (Cowperthwaite&Meyers,2007)

Compatible mutations and structures



- Mutations in neutral networks must conserve the phenotype.
- But it is hard to decide if a mutation conserve the m.f.e. structure and hence the phenotype.
- The number of acceptable structures can be recursively computed:

$$s_{m+1}(p) = \Xi_{m+1}(p) + \phi_{m-1}(p)\,,$$

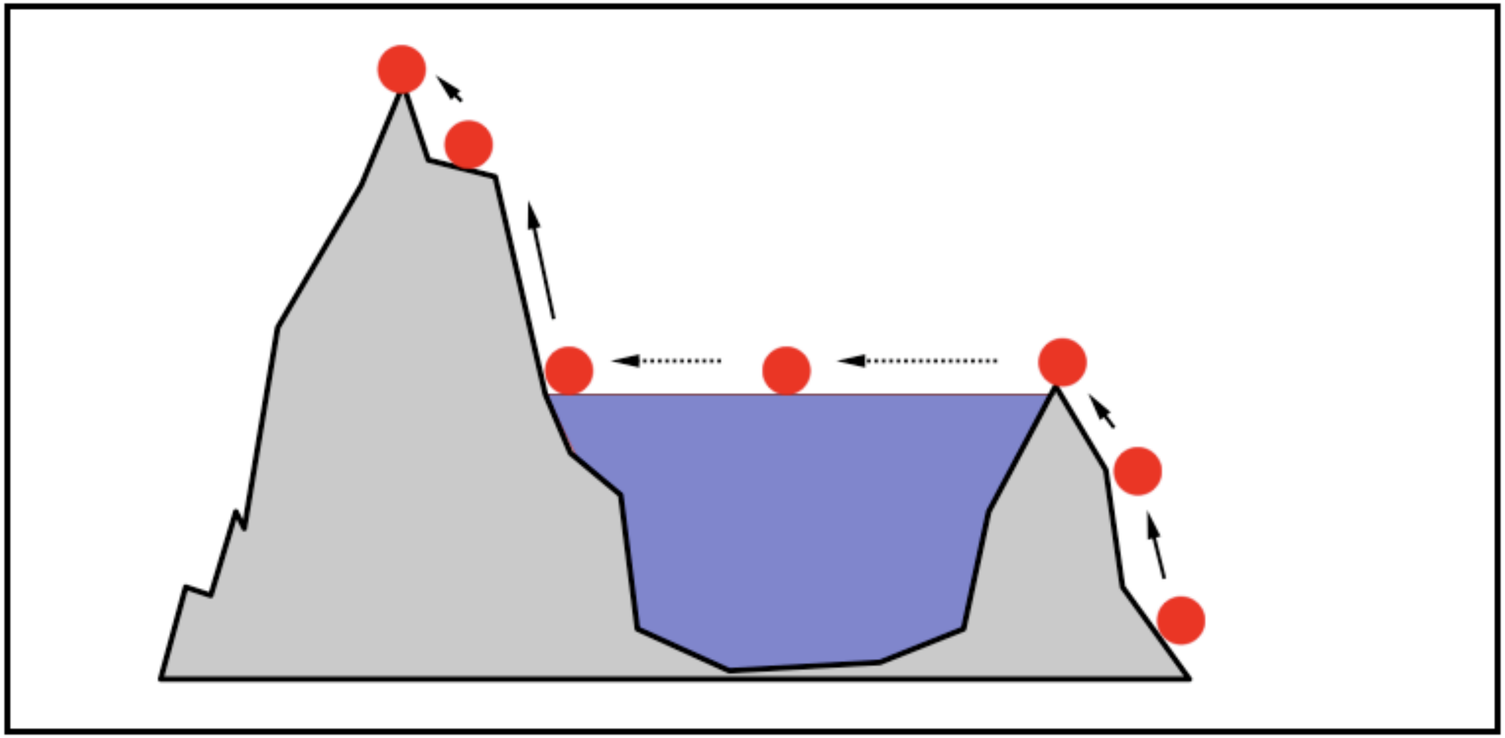
$$\Xi_{m+1}(p) = s_m(p) + \sum_{k=\lambda+2\sigma-2}^{m-2} \phi_k(p) \cdot s_{m-k-1}(p)$$

$$\phi_{m+1}(p) = p \sum_{k=\sigma-1}^{\lfloor (m-\lambda+1)/2 \rfloor} \Xi_{m-2k+1}(p) \cdot p^k$$

Hairpin minimum length λ required and length of stacks bounded σ.

Figure from (Gobel,2000)

Role of neutral networks



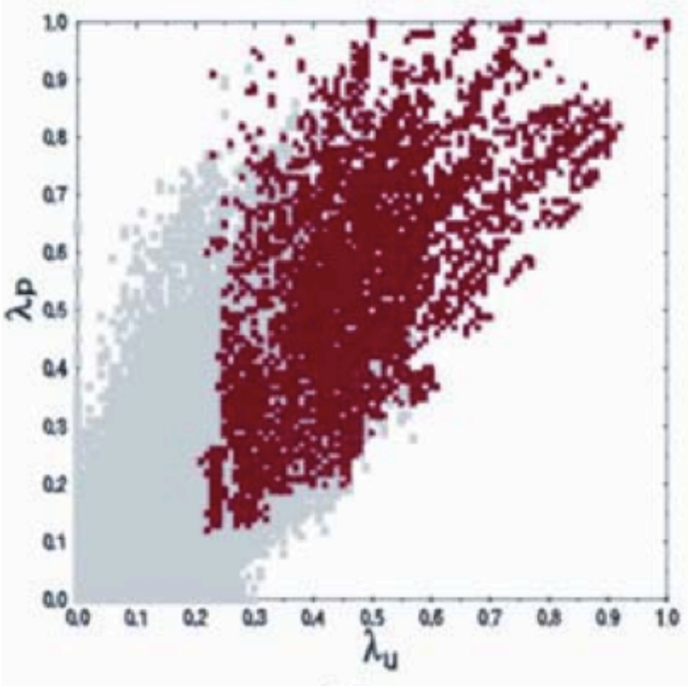
- Evolution tends to select mutations improving the structure.
- A smooth landscape (few maxima) favors the strategy.
- Facilitate evolution by allowing populations to explore genotype space while structure is preserved.

Figure from (Gobel,2000)

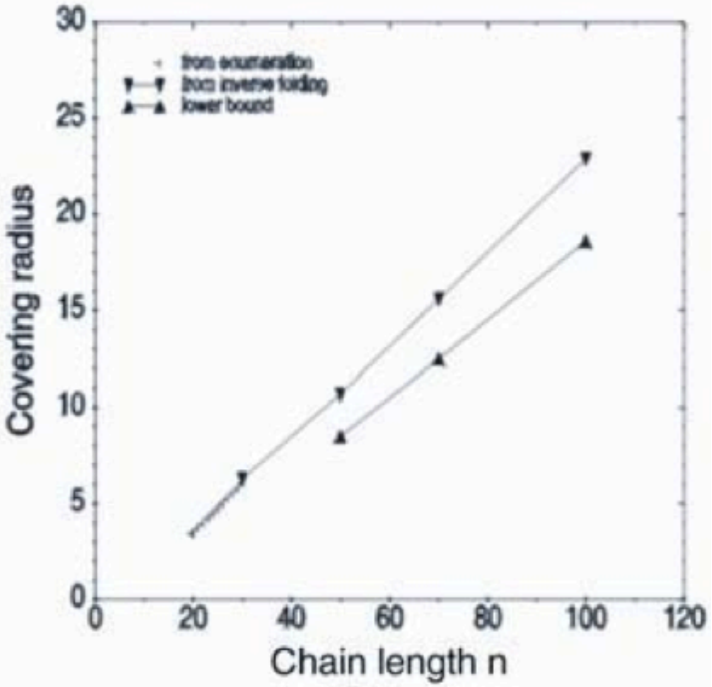
Properties of neutral networks

- More sequences than structures.
- Few common and many rare structures.
- Distribution of neutral genotype is approximately random.
- Neutral networks are connected unless specific features of RNA structure.
- The fraction of neutral neighbors $<\lambda>$ characterizes the neutral networks. Theory predicts a phase transition in their structures with $\lambda_c = 1 - k^{-1/(k-1)}$.
  - $<\lambda> < \lambda_c$: many isolated parts and one giant component.
  - $\lambda_c < < \lambda>$: generally connected.
- Few mutations almost certainly lead to a change of the structure.
- The number of disjoint components in a phenotype's neutral network does not appear to correlate with its abundance.

Neutral network and shape space covering: Examples



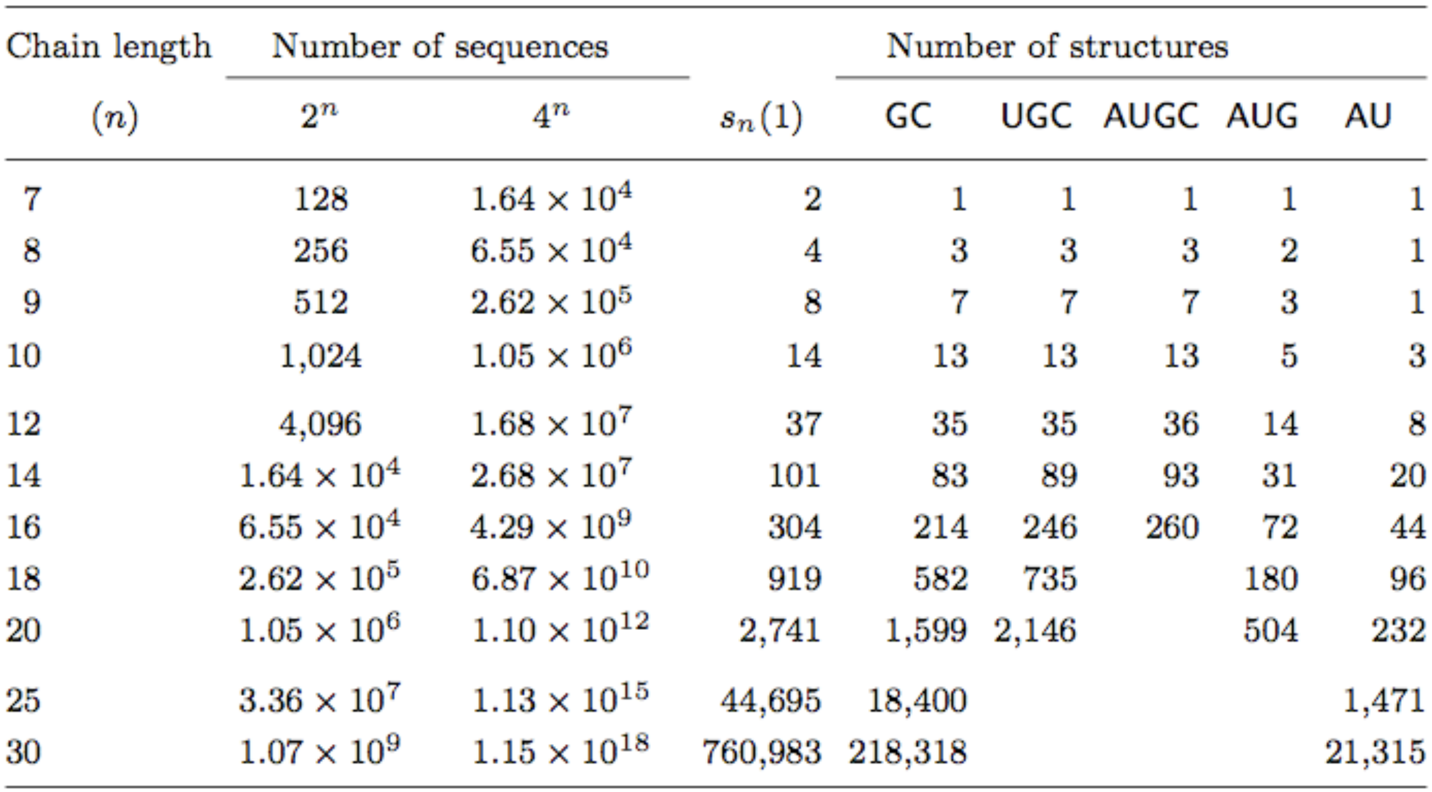
Full neutral network of GC sequence space with length=30.
λ_u: fraction of neutral mutations in unpaired regions.
λ_p: fraction of neutral mutations in paired regions.
Grey: fragmented networks (λ_x below threshold).
Red: 1-4 connected components (λ_x above threshold).



Shape space covering radius (radius of sphere containing in average at least one sequence per possible structure)

Data from (Gruner et al.,1999)
Figure from (Hofacker&Stadler,2006)

Comparison of exhaustively folded sequence spaces

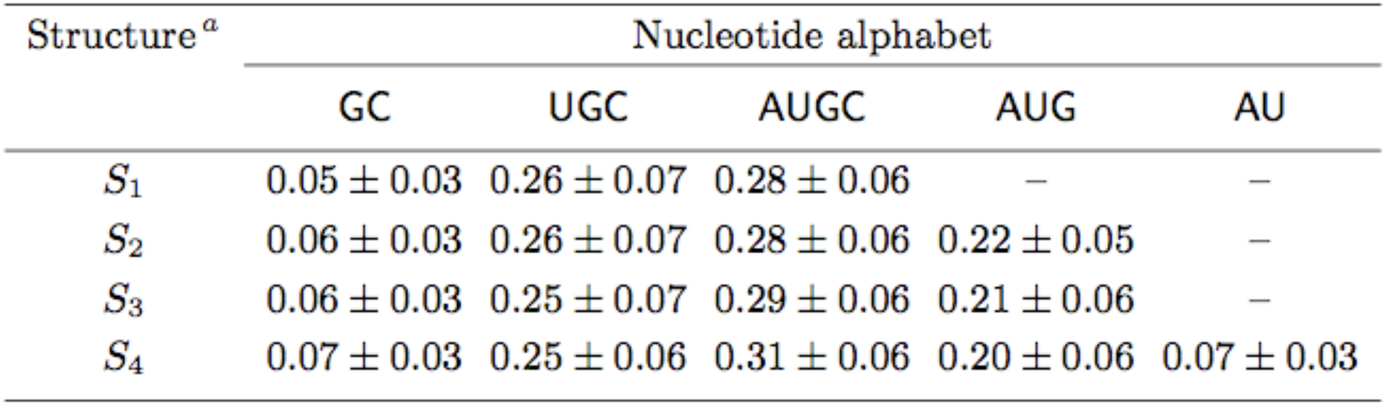


Chain length	Number of sequences			Number of structures				
(n)	2^n	4^n	$s_n(1)$	GC	UGC	AUGC	AUG	AU
7	128	1.64×10^4	2	1	1	1	1	1
8	256	6.55×10^4	4	3	3	3	2	1
9	512	2.62×10^5	8	7	7	7	3	1
10	1,024	1.05×10^6	14	13	13	13	5	3
12	4,096	1.68×10^7	37	35	35	36	14	8
14	1.64×10^4	2.68×10^7	101	83	89	93	31	20
16	6.55×10^4	4.29×10^9	304	214	246	260	72	44
18	2.62×10^5	6.87×10^{10}	919	582	735		180	96
20	1.05×10^6	1.10×10^{12}	2,741	1,599	2,146		504	232
25	3.36×10^7	1.13×10^{15}	44,695	18,400				1,471
30	1.07×10^9	1.15×10^{18}	760,983	218,318				21,315

Values computed on five different alphabets: GC, UGC, AUG, AU.
Structures with a single base pair are excluded from the enumeration.

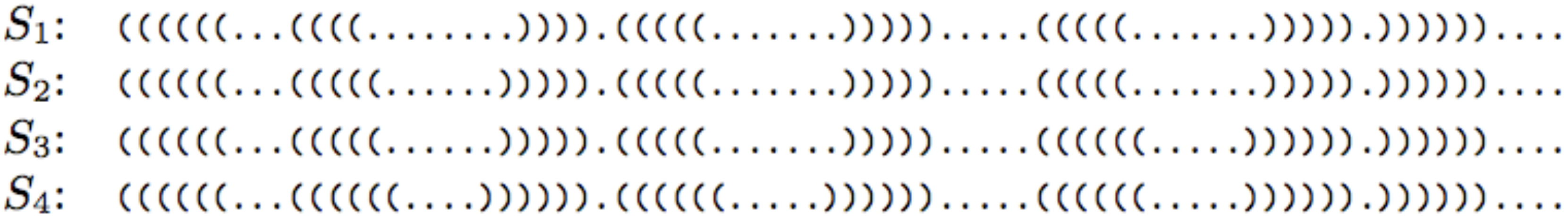
Data from (Schuster&Stadler,2007)

Degree of neutrality of tRNAs



Structure[a]	Nucleotide alphabet				
	GC	UGC	AUGC	AUG	AU
S_1	0.05 ± 0.03	0.26 ± 0.07	0.28 ± 0.06	–	–
S_2	0.06 ± 0.03	0.26 ± 0.07	0.28 ± 0.06	0.22 ± 0.05	–
S_3	0.06 ± 0.03	0.25 ± 0.07	0.29 ± 0.06	0.21 ± 0.06	–
S_4	0.07 ± 0.03	0.25 ± 0.06	0.31 ± 0.06	0.20 ± 0.06	0.07 ± 0.03

Fraction of neutral neighbors (degree of neutrality) computed from 1,000 random sequences fitting the structures using an inverse folding algorithm.



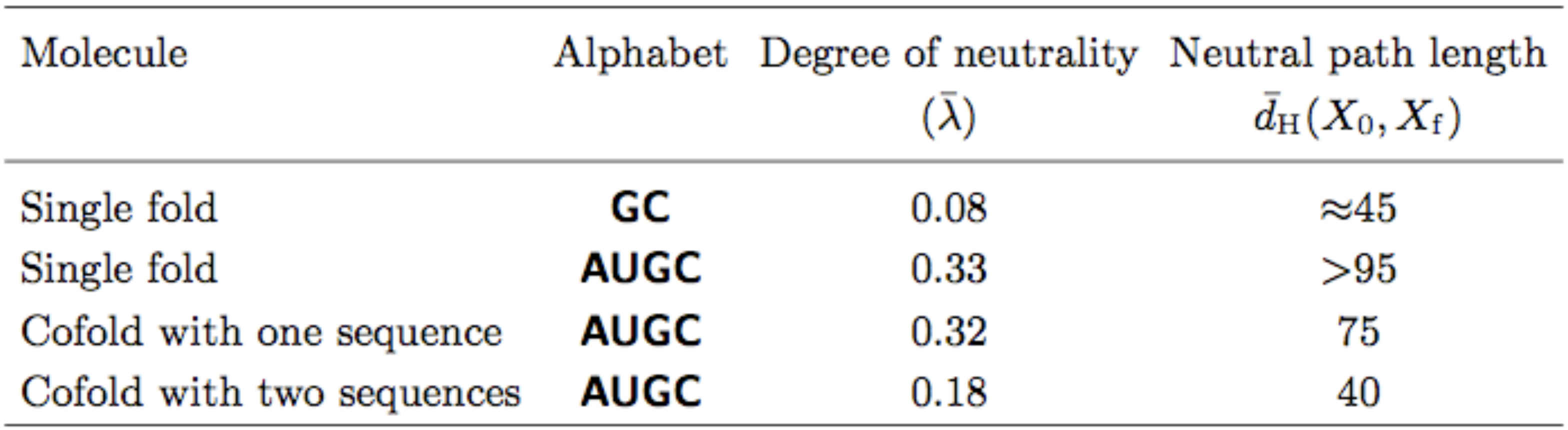
S_1: ((((((...((((........)))).(((((.......))))).....(((((.......))))).))))))....
S_2: ((((((...(((((......))))).(((((.......))))).....(((((.......))))).))))))....
S_3: ((((((...(((((......))))).(((((.......))))).....((((((.....)))))).))))))....
S_4: ((((((...((((((....)))))).((((((.....)))))).....((((((.....)))))).))))))....

- Different network structures for 2 and 4-letter alphabets.
- Weak structure depence.

Data from (Schuster&Stadler,2007)

Length of neutral paths

- Neutral paths connects neutral sequences differing with 1 mutations.
- Hamming distance from the origin strictly increase along the path.
- Path ends when all neighbors are closer to the reference sequence.

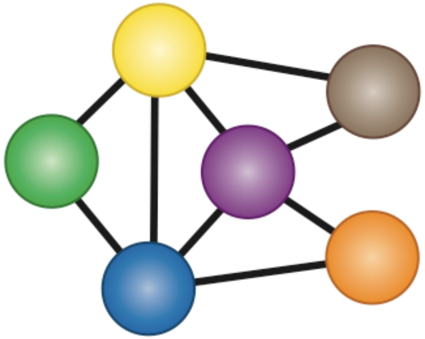


Molecule	Alphabet	Degree of neutrality $(\bar{\lambda})$	Neutral path length $\bar{d}_{\mathrm{H}}(X_0, X_{\mathrm{f}})$
Single fold	**GC**	0.08	$\approx$45
Single fold	**AUGC**	0.33	$>$95
Cofold with one sequence	**AUGC**	0.32	75
Cofold with two sequences	**AUGC**	0.18	40

Data computed from 1,200 random sequences of length 100.

Data from (Schuster&Stadler,2007)

Properties of phenotype networks



- Nodes are structures.
- Connect two nodes A,B if it exists 2 sequences a,b with phenotypes A,B that differ from 1 mutation.

- Highly irregular, with few nodes connected to many others and most nodes connected to few others.
- Abundant shapes are connected to almost every other shapes.
- The degree of mutational connectivity is not a binary properties. It exists some preferential connections. Moreover, these connections are always asymmetrical.
- Plastic model showed that neutral networks are not homogeneous. Probability of the m.f.e. structure in the low-energy ensemble varies. Most thermodynamically stable sequence lies in the center of the neutral network.

Fitness model

Objective: Evaluate the dynamic of the evolution of shapes.

Requirement: a metric to compare a predicted structure and a target shape.

Models:

- simple: The predicted structure is the m.f.e. structure.
- plastic: Suboptimal structures can be considered.

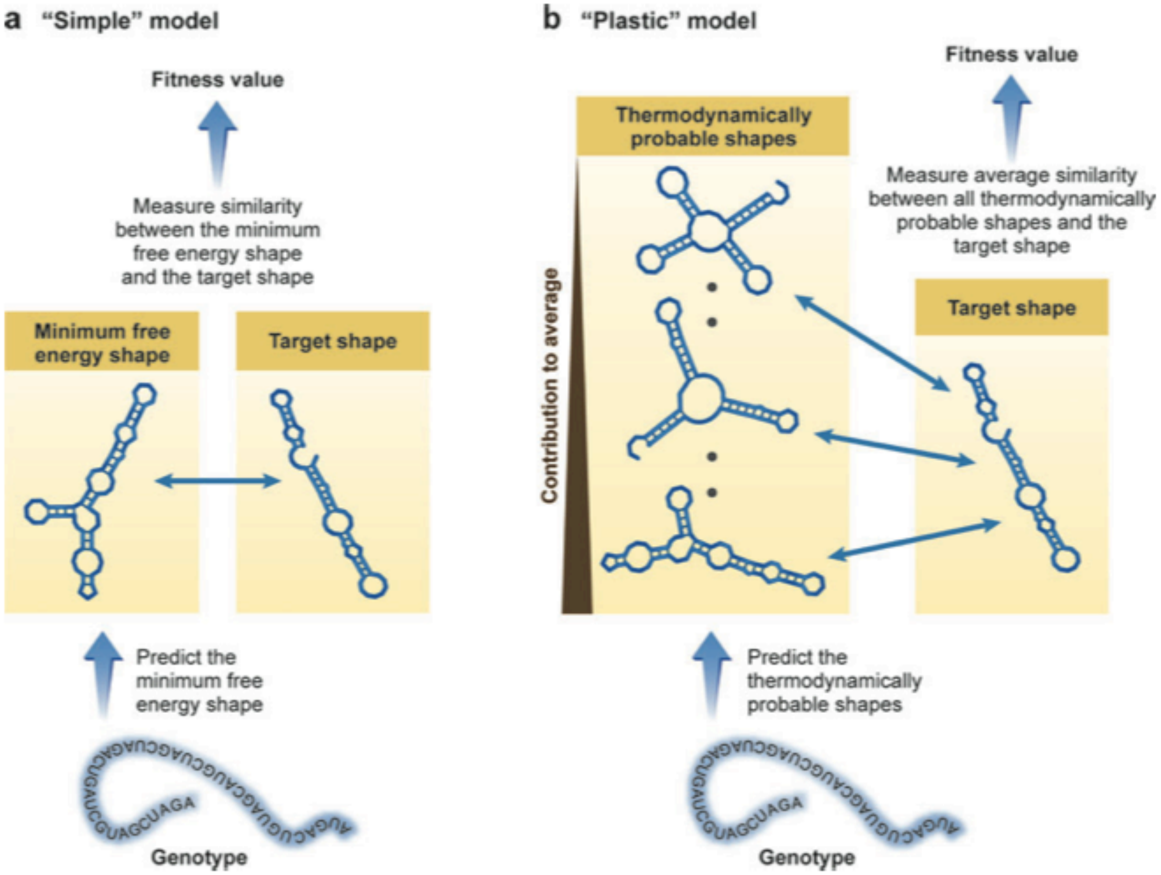

a “Simple” model
Fitness value
Measure similarity between the minimum free energy shape and the target shape
Minimum free energy shape
Target shape
Predict the minimum free energy shape
Genotype
b “Plastic” model
Fitness value
Thermodynamically probable shapes
Measure average similarity between all thermodynamically probable shapes and the target shape
Target shape
Contribution to average
Predict the thermodynamically probable shapes
Genotype

Figure from (Cowperthwaite&Meyers,2007)

Evolutionary Dynamics

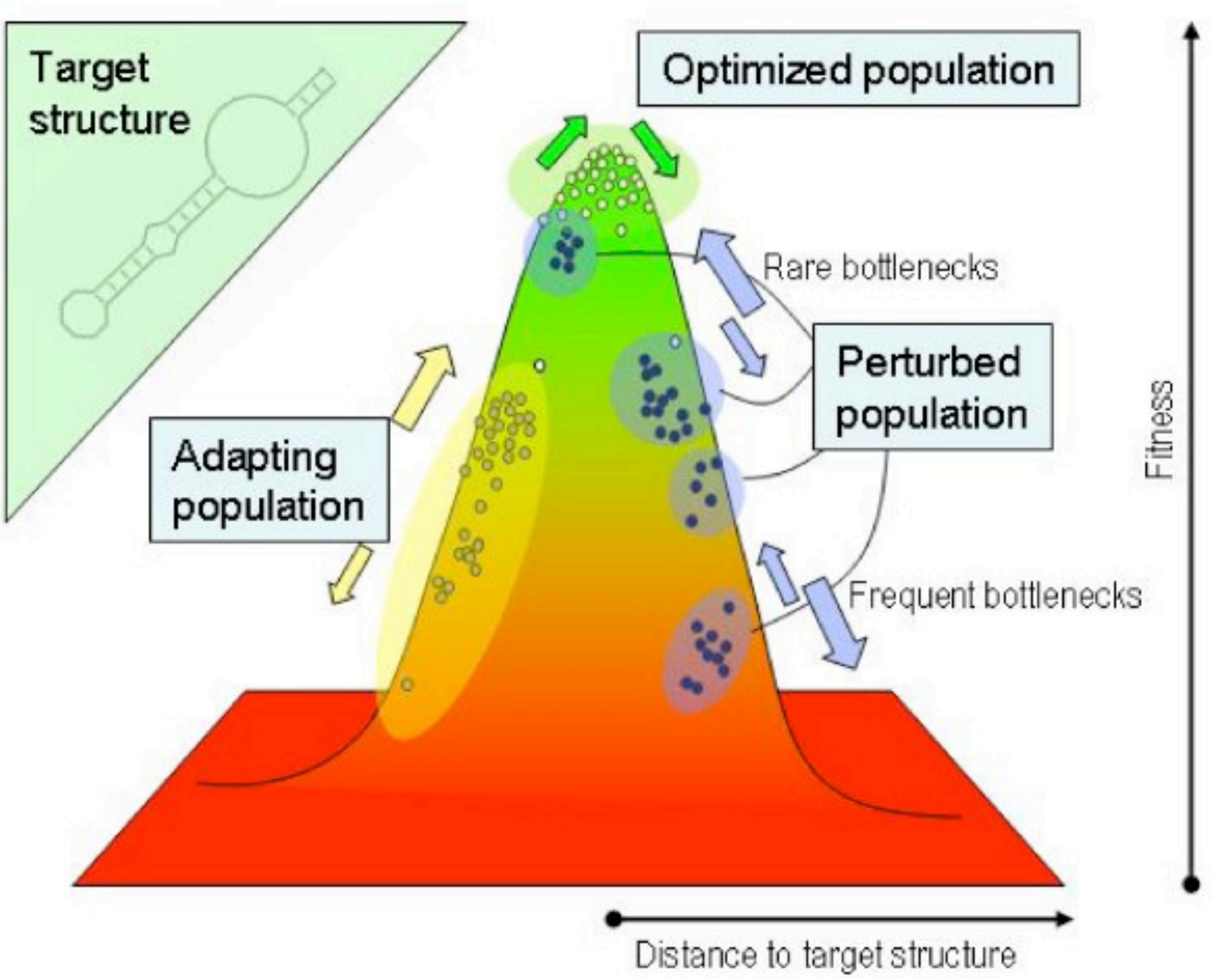
Start with a random population. Choose a target S. Each molecule *i* in the population replicate with probability:

$$P(d_i) = \frac{\mathrm{e}^{-\beta \frac{d_i}{l}}}{Z_i}$$

where d_i is the distance between the structure corresponding to sequence *i* and the target structure *S*.

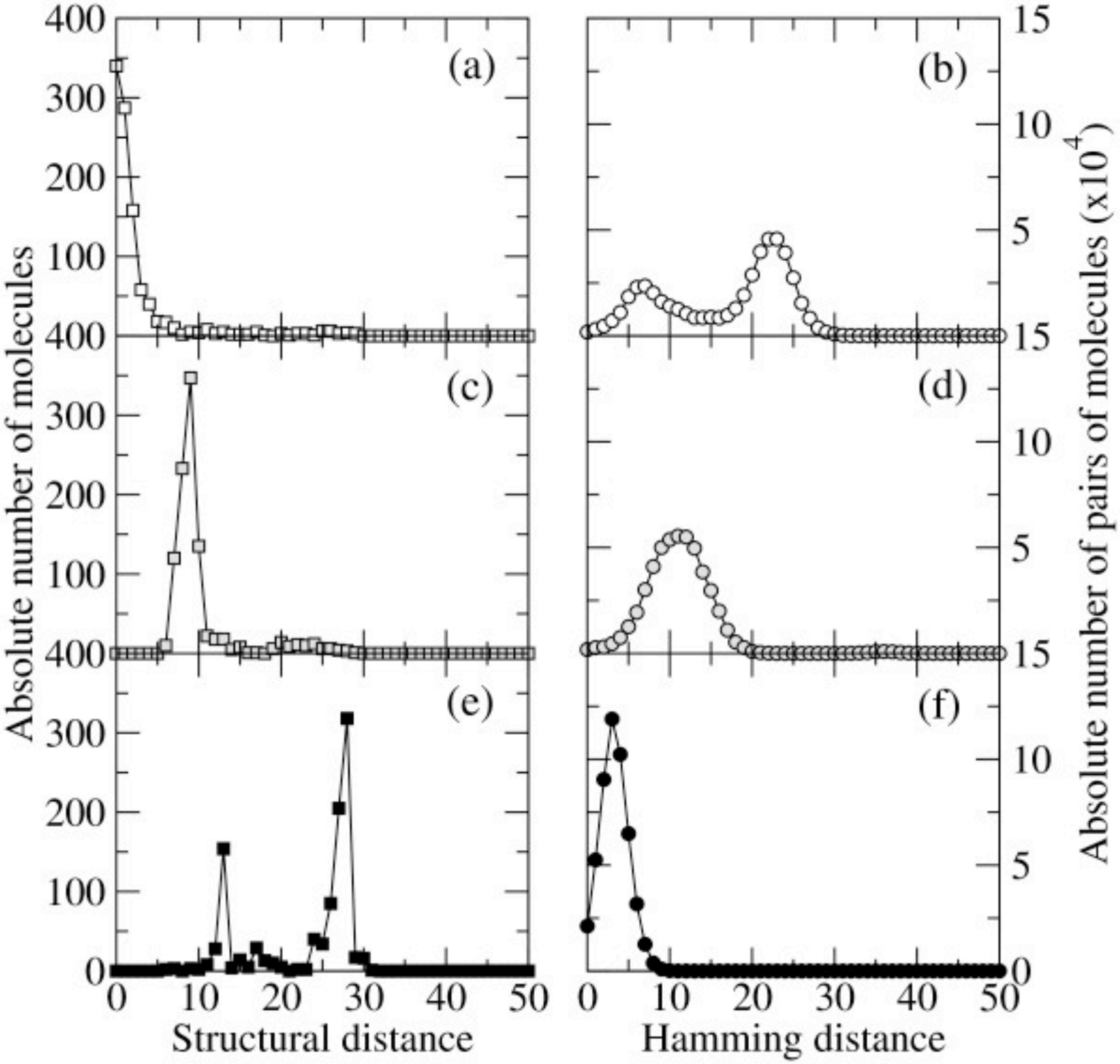
Replication happens with errors (i.e. mutations).

Fitness Landscape



(Stich et al., 2010)

Genotype distribution of adapting populations



Optimized population

Adapting population

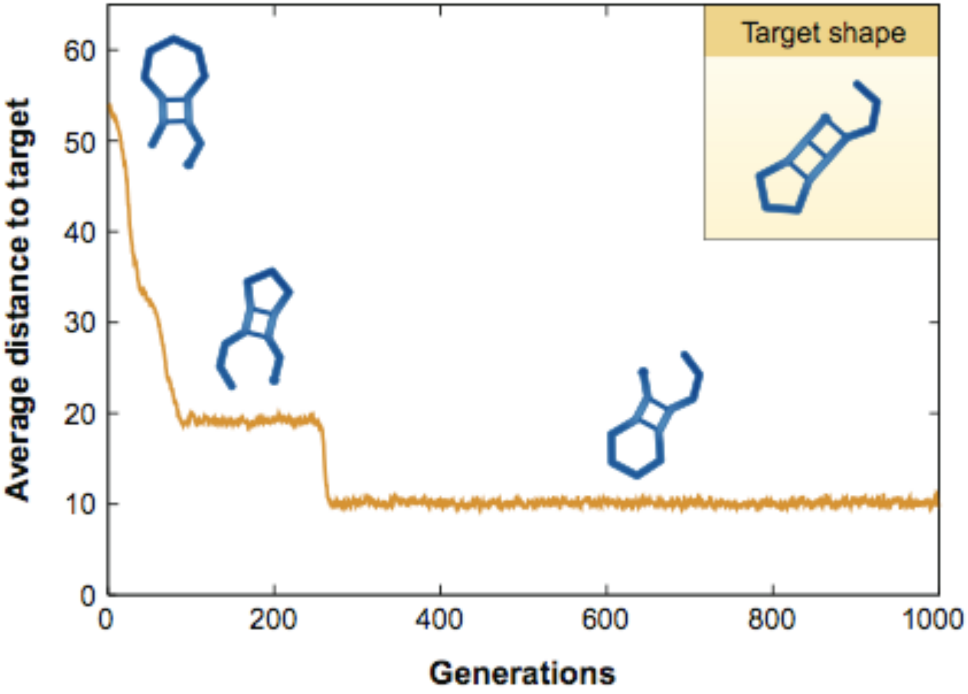
Perturbed popupation

(Stich et al., 2010)

Some Results from Computational Simulations

- Exploration of the sequence/structure network through simulations.
- Populations evolving toward a target shape experience long period of phenotypic stasis and short periods of rapid changes.
- On large neutral networks, the population subdivides in several subpopulations exploring different regions of the network.
- Size of neutral network increase the probability of evolving to this particular phenotype and/or from this phenotype to another one.
- The needle in the haystack: Population evolving on large neutral network do not adapt more quickly than those evolving on smaller networks (due to a larger search space).

Evolutionary dynamics



- Model favors mutations evolving toward the target shape.

- Short period of rapid phenotypic changes are punctuated by long period of stasis.

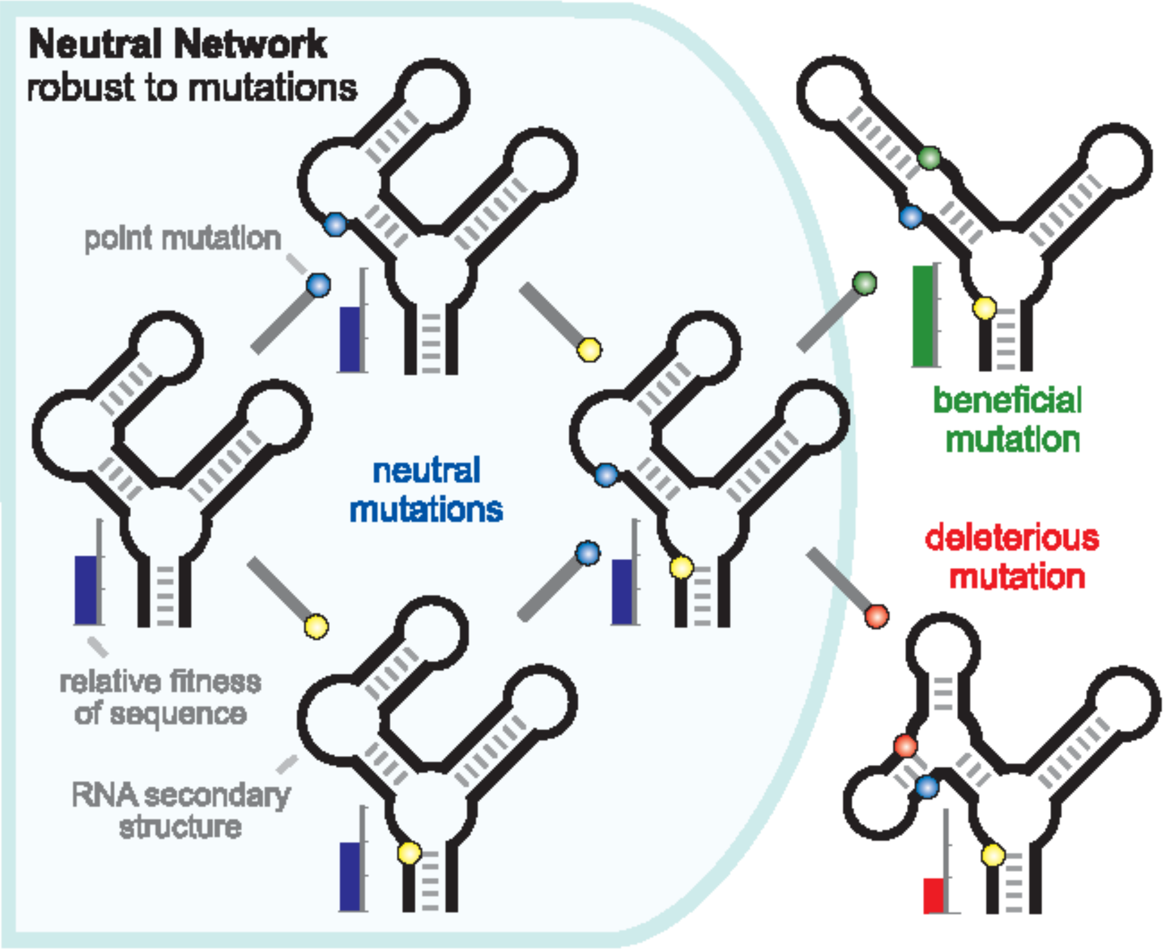
- Two types of transitions: Continuous (nearby phenotypes) and Discontinuous (radical change).

- Continuous transitions appear essentially in initial period of the simulation, while discontinuous transitions are predominant later.

- Phenomena mediated through neutral drifts (genotype that can change radically the phenotype through a single mutation). But these sequence are hard to find.

Figure from (Cowperthwaite&Meyers,2007)

Mutational Robustness



(Lenski et al., 2006)

Genetic robustness: Results

- Sequences carrying phenotypes should be robust to environmental and genetic perturbations.
- Unlike Environment robustness, genetic robustness is hard to justify. 3 potential scenario:

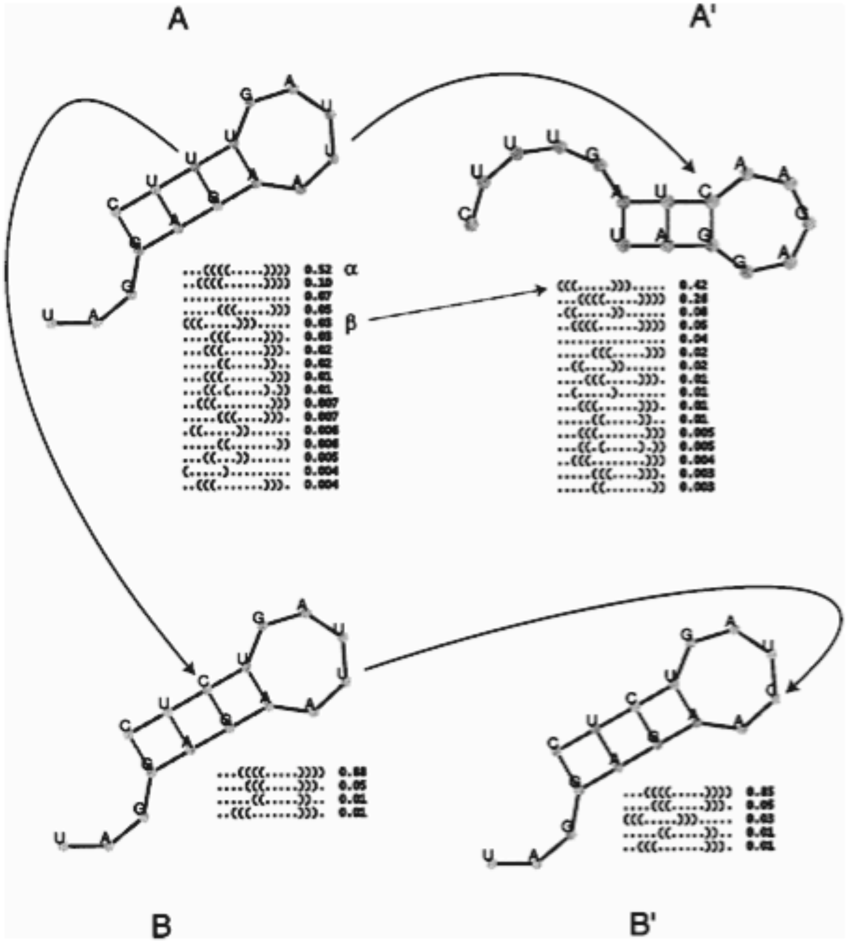
  a. Adaptive robustness: natural selection.

  b. Intrinsic robustness: correlated byproduct of character selection.

  c. Congruent robustness: correlated byproduct of selection for environmental robustness.

- Adaptive robustness (a) is possible. Trans-generational cost of deleterious mutations drives sequence in the heart of neutral network.
- Congruent robustness (c) is tested using the plastic model. Simulations showed that models targeting a shape lead to a reduction of plasticity. Also, they highlight a slow-down and possible halting of the evolutionary process.
- Reduction of plasticity leads to an extreme modularity (side-effect?).

Plastogenetic congruence



List suboptimal structures and weight them by the time spent by the molecule in that fold (energy).

(1) A→A' : makes β the m.f.e.
(2) A→B: makes α stronger, exits β.
(3) B→B' : same mutation brings back β,
but keeps α on top.

(1) correlates structures in the plastic repertoire to mutational neighbors.
(3) shows the *epistatic* control of neutrality. The more time spent in m.f.e., the higher the fraction of neutral neighbors.

- "plastogenetic congruence": the set of shapes realized by a sequence correlates to the m.f.e. shapes of 1-mutants.

- **RNAs insensitive to thermal noise are also insensitive to mutations.**

Figure from (Ancel&Fontana,2000)

Survival of the flattest

- How mutation rates (rapidity of mutations) shape evolution?
- Under low mutation rates, fitness considerations dictate dynamics.
- Under high mutation rates, the breadth of the neutral network can be as more important as the fitness: the survival of the flattest.
- Simulations showed that populations having evolved under low mutation rates have a better adaptation potential than populations having always evolved under a high mutation rate (Wilke et al.,2001).
- Genotypes located in flatter regions are more robust to mutations.

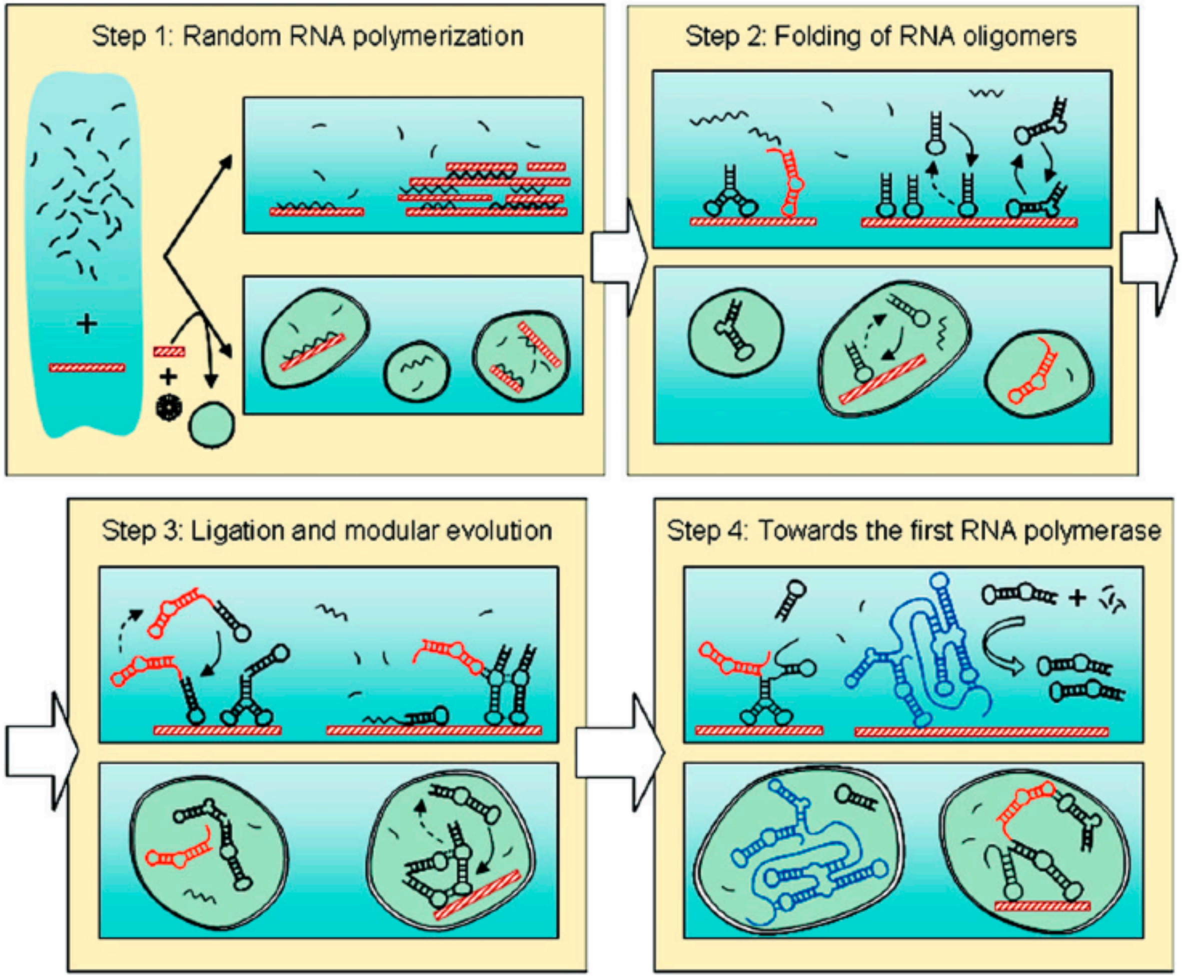
Local mutational structure

- Theory and computational experiments differ on the distribution of beneficial mutations. While the beneficial effect of mutations is predicted to be exponentially distributed, in-silico experiments showed an overabundance of small-effect mutations.

- Although they tend to be eliminated, at high mutation rates deleterious mutations (mutations changing radically the structure) are fixed through compensatory evolution. In other words evolution tends to “repair” the damages… sometimes even before.

- Epistasis regulates the effect of mutations.

Complexity through ligation


Step 1: Random RNA polymerization
Step 2: Folding of RNA oligomers
Step 3: Ligation and modular evolution
Step 4: Towards the first RNA polymerase

(Briones et al., 2009)