COMP 564: Simulating RNA evolution

Jérôme Waldispühl
School of Computer Science
McGill University
RNA world

In a prebiotic world, RNA could have filled two distinct roles:
1. Information carrier, because it can self-replicate,
2. Catalytic role, because it can fold in complicated 3D shapes.

Over time, DNA became a molecule specialized to carry information, while proteins replaced RNA for catalyzing chemical reactions.
Outline

- Mathematical modelling
- Characterization of the evolutionary landscape
- Evolutionary dynamics
Sequence evolution

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

• Insertion
• Insertion/Deletion
• Mutation

For the sake of simplicity, we will restrict ourselves to mutation only.

Figure from (Gobel, 2000)
Genotype & Phenotype

- Use a folding program (e.g., RNAfold, RNAstructure) to compute the MFE secondary structure from the sequence (i.e., genotype).
- The secondary structure is the phenotype.

Stochastic predictions and ensemble of structures can also be used (i.e., plastic model).

Figure from (Cowperthwaite&Meyers, 2007)
Evaluating structural similarity

How to measure the similarity between phenotypes?

Hamming distance:

- The base pair distance is the minimum number of base pairs to remove/add to obtain one structure from the other.
- Structures must be of same length.

Base pair distance:
Characterization of the evolutionary landscape

Understanding the relationship between genotype and phenotype in RNA sequence-structure maps
RNA sequence-structure maps

Sequence ensemble (genotype)
- CCUCAACGAAGC
- UUUACGGCUAGC
- UAUACGGCCAGC
- UUUAAAGGCACGC
- UUUAGGGCCAGC
- UCUGAAACCGCU

Structure ensemble (phenotype)
Structural repertoire of random RNAs

Compute MFE of sequences uniformly generated at random and estimate distribution of structures.

Abundance of structures

Most abundant structures

(Stich et al., 2008)
RNA Neutral networks

- Nodes are sequences
- Edges between nodes are they differ by one mutation
- Each node is labelled with its MFE structure
- A network of nodes with the same labels is a neutral network.
Role of neutral networks

Neutral networks facilitate the exploration of the evolutionary landscape by allowing populations to explore genotype space while preserving the phenotype.

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.
- In general, neutral networks are connected.
- The fraction of neutral neighbors $\langle \lambda \rangle$ characterizes neutral networks. Theory predicts phase transition at $\lambda_c = 1 - k^{-1/(k-1)}$:
  - $\langle \lambda \rangle < \lambda_c$: many isolated parts and one giant component.
  - $\lambda_c < \langle \lambda \rangle$: 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.
Fragmentation and shape space covering

Full neutral network of GC sequence space with length=30.
\( \lambda_u \): fraction of neutral mutations in unpaired regions.
\( \lambda_p \): fraction of neutral mutations in paired regions.
Grey: fragmented networks (\( \lambda_x \) below threshold).
Red: 1-4 connected components (\( \lambda_x \) 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)
Accessible secondary structures

Exhaustive folding of sequence with given length.
Experiments conducted with five alphabets: GC, UGC, AUGC, AUG, AU.

<table>
<thead>
<tr>
<th>Chain length $n$</th>
<th>Number of sequences</th>
<th>Number of structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2^n$</td>
<td>$4^n$</td>
<td>$s_n(1)$</td>
</tr>
<tr>
<td>7</td>
<td>128</td>
<td>$1.64 \times 10^4$</td>
</tr>
<tr>
<td>8</td>
<td>256</td>
<td>$6.55 \times 10^4$</td>
</tr>
<tr>
<td>9</td>
<td>512</td>
<td>$2.62 \times 10^5$</td>
</tr>
<tr>
<td>10</td>
<td>1,024</td>
<td>$1.05 \times 10^6$</td>
</tr>
<tr>
<td>12</td>
<td>4,096</td>
<td>$1.68 \times 10^7$</td>
</tr>
<tr>
<td>14</td>
<td>$1.64 \times 10^4$</td>
<td>$2.68 \times 10^7$</td>
</tr>
<tr>
<td>16</td>
<td>$6.55 \times 10^4$</td>
<td>$4.29 \times 10^9$</td>
</tr>
<tr>
<td>18</td>
<td>$2.62 \times 10^5$</td>
<td>$6.87 \times 10^{10}$</td>
</tr>
<tr>
<td>20</td>
<td>$1.05 \times 10^6$</td>
<td>$1.10 \times 10^{12}$</td>
</tr>
<tr>
<td>25</td>
<td>$3.36 \times 10^7$</td>
<td>$1.13 \times 10^{15}$</td>
</tr>
<tr>
<td>30</td>
<td>$1.07 \times 10^9$</td>
<td>$1.15 \times 10^{18}$</td>
</tr>
</tbody>
</table>

(Schuster&Stadler,2007)
Neutral paths

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

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Alphabet</th>
<th>Degree of neutrality ($\lambda$)</th>
<th>Neutral path length $d_H(X_0, X_f)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single fold</td>
<td>GC</td>
<td>0.08</td>
<td>$\approx 45$</td>
</tr>
<tr>
<td>Single fold</td>
<td>AUGC</td>
<td>0.33</td>
<td>$&gt;95$</td>
</tr>
<tr>
<td>Cofold with one sequence</td>
<td>AUGC</td>
<td>0.32</td>
<td>75</td>
</tr>
<tr>
<td>Cofold with two sequences</td>
<td>AUGC</td>
<td>0.18</td>
<td>40</td>
</tr>
</tbody>
</table>

Data computed from 1,200 random sequences of length 100.
Note: we will introduce co-folding later.

(Schuster&Stadler,2007)
Evolutionary dynamics

Simulation of RNA populations under selective pressure
Evolutionary Dynamics

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

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

$(d_i$ is the distance between the structure and target structure S.)

- Replications includes errors (i.e., mutations).

The model helps to study natural selection and adaptation of populations to a phenotype.
Fitness Landscape

(Stich et al., 2010)
Genotype distribution of adapting populations

Optimized population (population fitted to target)

Adapting population (population not fully adapted to target yet)

Perturbed population (population adapting to novel target)

Note: On large neutral networks, the population subdivides in several subpopulations exploring different regions of the network. (Stich et al., 2010)
Evolutionary dynamics

- Short period of rapid phenotypic changes are punctuated by long period of stasis.
- Continuous transitions (nearby phenotype) at beginning, then discontinuous transitions (radical change) become predominant.

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.
Survival of the flattest

How mutation rates (frequency of mutations) shape evolution?

• Low mutation rates ⇒ fitness dictates dynamics.

• High mutation rates ⇒ the breadth of the neutral network is more important than fitness (survival of the flattest).

Simulations showed that populations having evolved under low mutation rates have a better adaptation potential those that evolved under a high mutation rate (Wilke et al., 2001).

More at (Stich et al., PLoS ONE, 2010)
Emergence of complexity?

Can complex structures emerge without target selection?

*Catalyzation from 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

Ligation to a mineral surface helps building complex structures.

*The role of GC content*

Search for stability (i.e., efficiency) and GC content is sufficient to explain emergence of multi-loops.

(Oliver et al., 2019)