Exploring Connectivity in Sequence Space of Functional RNA

Chenyu Wei[1,2], Andrew Pohorille[1,2], Milena Popovic[2,3], and Mark Ditzler[2]
[1]Department of Pharmaceutical Chemistry, UCSF; [2]Exobiology Branch, NASA Ames Research Center, Moffett Field, CA 94035; [3]Blue Marble Space Institute of Science, Seattle, WA 98145
*chenyu.wei@nasa.gov

Abstract: Emergence of replicable genetic molecules was one of the marking points in the origin of life, evolution of which can be conceptualized as a walk through the space of all possible sequences. A theoretical concept of fitness landscape helps to understand evolutionary processes through assigning a value of fitness to each genotype. Then, evolution of a phenotype is viewed as a series of consecutive, single-point mutations. Natural selection biases evolution toward peaks of high fitness and away from valleys of low fitness [1,2], whereas neutral drift occurs in the sequence space without direction as mutations are introduced at random. Large networks of neutral or near-neutral mutations on a fitness landscape, especially for sufficiently long genomes, are possible or even inevitable [1,3,4]. Their detection in experiments, however, has been elusive. Although a few near-neutral evolutionary pathways have been found [5-7], recent experimental evidence indicates landscapes consist of largely isolated islands [8.9]. The generality of these results, however, is not clear, as the genome length or the fraction of functional molecules in the genotypic space might have been insufficient for the emergence of large, neutral networks. Thorough investigation on the structure of the fitness landscape is essential to understand the mechanisms of evolution of early genomes.

RNA molecules are commonly assumed to play the pivotal role in the origin of genetic systems. They are widely believed to be early, if not the earliest, genetic and catalytic molecules, with abundant biochemical activities as aptamers and ribozymes, i.e. RNA molecules capable, respectively, to bind small molecules or catalyze chemical reactions. Here, we present results of our recent studies on the structure of the sequence space of RNA ligase ribozymes selected through in vitro evolution. Several hundred thousands of sequences active to a different degree were obtained by way of deep sequencing. Analysis of these sequences revealed several large clusters defined such that every sequence in a cluster can be reached from any other sequence in the same cluster through a series of single point mutations. Sequences in a single cluster appear to adopt more than one secondary structure. The mechanism of refolding within a single cluster was examined. To shed light on possible evolutionary paths in the space of ribozymes, the connectivity between clusters was investigated. The effect of length of RNA molecules on the structure of the fitness landscape and possible evolutionary paths was examined by way of comparing functional sequences of 20 and 80 nucleobases in length. It was found that sequences of different lengths shared secondary structure motifs that were presumed responsible for catalytic activity, with increasing complexity and global structural rearrangements emerging in longer molecules.

References: [1] Kauffman SA and Levin S (1987) *Journal of Theoretical Biology* 128:11-45. [2] Gavrilets S (1997) *Trends in Ecology & Evolution* 12:307-312. [3] Fontana W and Schuster P (1998) *Science* 280:1451-1454. [4] Gravner J et al. (2007) Journal of Theoretical Biology 248:627-645. [5] Schultes EA and Bartel DP (2000) *Science* 289:448-452. [6] Mandal M and Breaker RR (2004). *Nature Structural & Molecular Biology* 11:29-35. [7] Held DM et al. (2003). *Journal of Molecular Evolution* 57:299-308. [8] Jiménez JI et al. (2013) *Proceedings of the National Academy of Sciences USA* 110:14984-14989. [9] Petrie KL and Joyce GF (2014) *Journal of Molecular Evolution* 79:75-90.

EXPLORING CONNECTIVITY IN SEQUENCE SPACE OF FUNCTIONAL RNA

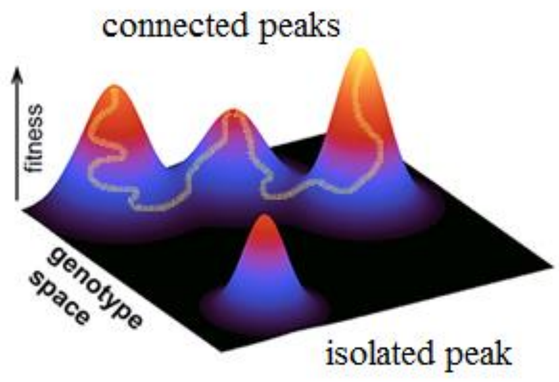
Chenyu Wei*[1,2], Milena Popović[1,3], Andrew Pohorille[1,2], Mark Ditzler[1]

[1]NASA Ames Research Center; [2]Department of Pharmaceutical Chemistry, UCSF;

[3]Blue Marble Space Institute of Science

Introduction

Emergence of replicable genetic molecules was one of the marking points in the origin of life, evolution of which can be conceptualized as a walk through the space of all possible sequences. A theoretical concept of fitness landscape was proposed by Wright [1,2] to help understand the evolution process, on which each genotype has a value of fitness, and the evolution of a phenotype to reach a local fitness peak is viewed through consecutive mutations. Natural selection biases evolution toward peaks of high fitness and away from valleys of low fitness, while neutral drift occurs in the sequence space without direction as mutations are introduced at random [3,4].



Evolution path and connectivity between peaks on fitness landscape

- Isolated peaks: no path-ways between peaks consisting of consecutive, viable genotypes that differ by a single mutation → *evolutionary optimization possible only through genetic recombination or alterations to the landscape*

- Connected peaks: networks of neutral or near-neutral mutations → *large volumes of genotypic space crossed without marked effect on fitness, eventually chancing upon a new fitness peak.*

Evolution of RNA genomes

- Advantages
  - early genetic and catalytic molecules,
  - abundant biochemical activities: aptamers and ribozymes.
- System in this study: *ligase ribozymes selected through in vitro evolution.*

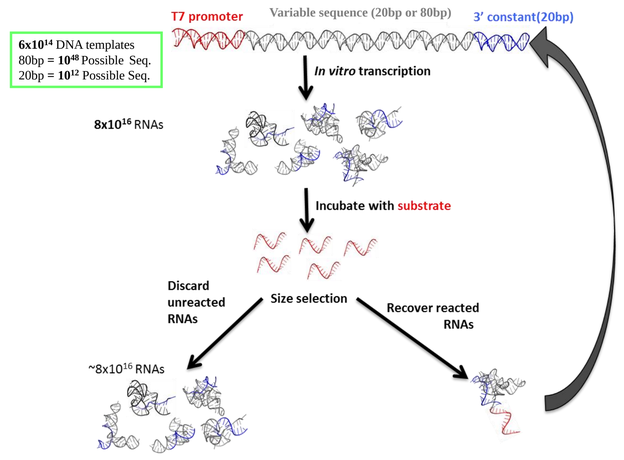
Motivation for the study

- Determine connectivity in the sequence space for small RNA ribozymes.
- Investigate length effects on the activity and connectivity to shed light on evolution process with increased length of RNA ribozyme.

Method

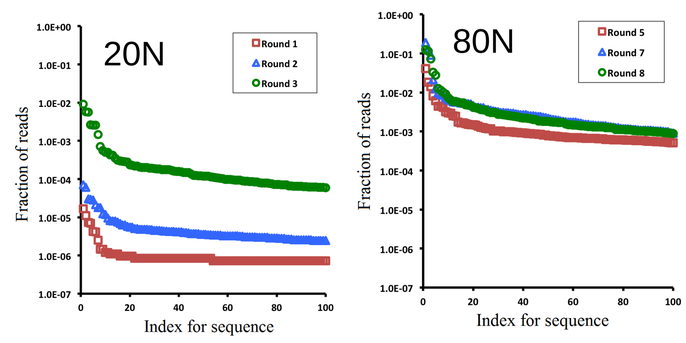
- In vitro evolution and high throughput sequencing.
- Computational methods of sequencing analysis and secondary structure prediction and comparison.

Ligase evolution with random region



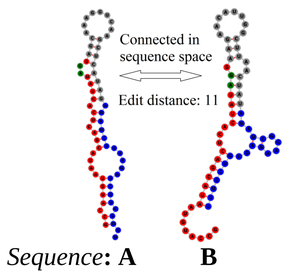
- Initial DNA library (>10^{14} unique sequences)
- High-throughput sequencing
  (> 3 x 10^6 reads per population)
- Random length variation: 20N & 80N

Increased abundance after rounds of selections

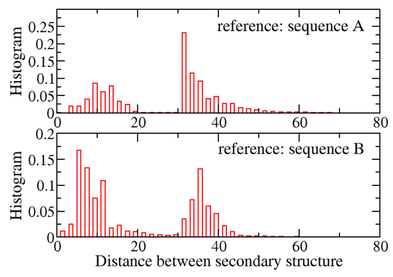


Short random region: 20N

- > 3 x10^6 sequence selected
- Distinct secondary structures: connected in sequence space
- Secondary structure comparison: two peaks with structures close to sequence "A" or "B"



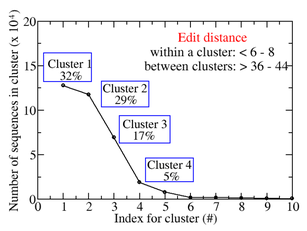
Red: sequence substrate
blue: constant region
Grey: random region; Green: ligation site



Secondary structure prediction: RNAfold, with distance comparison in tree representation

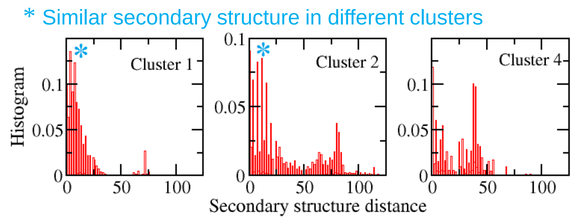
Long random region: 80N

- > 4 x 10^5 unique sequences analyzed
- Clusters in sequence space
  - connected through single mutation within a cluster
  - 4 large clusters (83% of all)

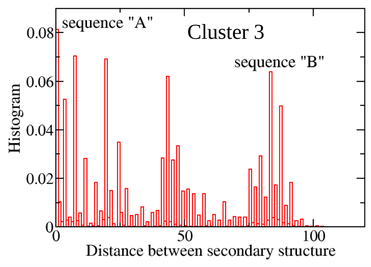


- Secondary structure comparison within clusters

* Similar secondary structure in different clusters



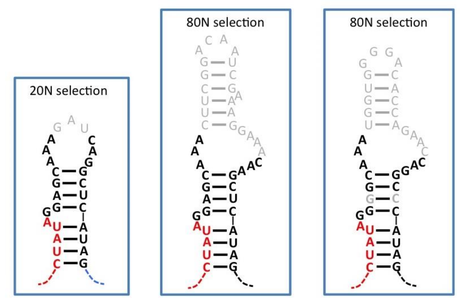
- Distinct secondary structures within a cluster



Sequence: **A** **B**

Comparison of 20N and 80N ribozymes

- Common motif identified

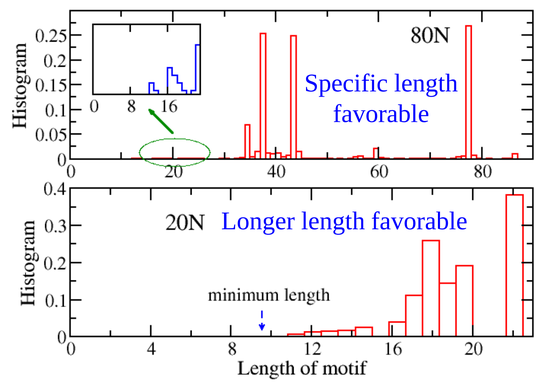


Sequence "**A**" (20N) from cluster 1 from cluster 2

sequence from peak *

Abundant in cluster 1 & 2 of 80N ribozymes

- Variation of length for the common motif: 20N & 80N



- Increased motif length → Structure complexity
- Enhanced activity in 80N

Summary

- Small RNA ligase with varied random region selected through *in vitro* evolution
  - Common motif identified at ligation site for independently evolved short and long RNA ribozymes
  - Increased structure complexity and activity in longer ribozyme
- Distinct secondary structures connected in sequence space of selected RNA
  → **connectivity on fitness landscape**

References

[1] Wright S. (1932). Proceedings of the Sixth International Congress on Genetics, 355-366.

[2] Gavrilets S. (2004). Fitness Landscapes and the Origin of Species (Princeton Univ. ress, Princeton). ISBN: 9780691119830.

[3] Kauffman S.A. and Levin S. (1987). J. Theor. Biol., 128, 11-45.

[4] Gavrilets S. (1997). Trends Ecol. Evolut., 12, 307-312.

Acknowledgements

Supported by NASA Exobiology Program

Contact information

chenyu.wei@nasa.gov