Evolution to Complexity: replication, elongation and assembly in an RNA world

Tomoyuki Yamamoto[A] and Paulien Hogeweg[B]

[A,B] Bioinformatica, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands
[A] Department of Mathematics, Hokkaido University, Sapporo, 060-0810 Japan
E-mail: ty@math.sci.hokudai.ac.jp

Abstract. A spatially resolved model of RNA world is studied, where primer induced replication, concatenation and random cutting are considered. The increase of diversity of sequences and complexity of shapes are observed; A hierarchically orgnized "replication network" is formed, and evolution to long sequences and assembly of further long sequences are obtained through it. These results suggest a scenario for overcoming the error threshold and for the evolution of enzymatic activity.

1 Introduction

The RNA world is a hypothesis for the origin of life (see [1] for review) and several attempts to simulation it have been made(e.g.[2]). An important problem here is that the error threshold [3] [4] prevents long molecules to increase in the population.
In the present paper, we will investigate this problem by simulation of an RNA-replicator model on a plane. The secondary structure is included. We obtain a replication network with high diversity and an increase of the variety of shapes. Then our model appears to suggest a way to pass the information threshold.
Single-stranded RNA molecules can replicate by themselves (e.g. [5]), but the length is bounded by the error threshold. Then, we assume the concatenation of them is also possible, as is done in DNA [6].
Spatial resolution has several merits for the generation of diversity of sequences. Especially, we believe the spatial clustering supports elongation and repair of molecules via the replication network (see §3.1).
Recently, McCaskill et al. studied DNA/RNA amplification system (e.g. [7]). They have shown a formation of cooperative amplification network and reaction-diffusion like patterns on a plane. While this work is similar to ours, we concentrate on RNA: we include RNA secondary structure and different kinds of phenomena are obtained, like assembly of molecules.

2 Model

The model consists of two parts. One is the plane where RNA polymers exist, the other is the "soup" which supplies oligomers and monomers.

The plane is descretized to a grid (typically 100×100) and each gridcell can be occupied by a molecule, which satisfies the minimum length limit (typically 5). The molecules are assumed to be attached to the plane, but limited diffusion by Margolus' method [8] does occur.
The soup supplies oligomers (typically, length 2 to 5) and monomers as substrates for the replication.
The secondary structure of molecules is evaluated by the Vienna RNA package by Hofacker et. al [9].
We apply three reactions, replication, concatination and random cut.
replication We assume the primer is required to start replication and the replication proceeds only from 5' to 3' ends. Dangling 5' end is allowed only for the template (i.e. template do not elongate). Folded, double-stranded molecules cannot be replicated. Point mutations are applied on the part being replicated. The threshold value for the binding energy (typically -2.50 kcal/mol, averaging the number of bonds) is applied. When no matching primer molecule is found in the nearest neighbor, a random oligomer is taken from the soup.
concatenation The molecules can be concatenated when a molecule pairs two molecules in its neighborhood to form a hemiduplex strand, bridging over two strands. The threshold for the binding energy is applied for both matching parts.
random cut The single stranded areas in the molecules can be cut (typically 1.0×10^{-5} per nucleotide). The double-stranded region cannot be cut and a folded molecule is more stable than a single-stranded molecule.

3 Results

Both the population growth and the increase of diversity are frequently obtained. Although sometimes the plane will be ocuppied by trivial sequences (see §3.3), variety of sequences and complexity of secondary structures are obtained around 30% of the simulations (see §3.4).
To illustrate the behavior, let us see the case when the replication fidelity is 1.0. In fig.1a, the sequences of frequent molecular species, whose populations are larger than 5, are shown. In fig.1b, some of the longest sequences are shown. There are about 3000 molecules after 100000 time steps from random initial condition where we start with 100 short molecules. The secondary structure is shown by the bracket representation [10].

3.1 Replication network

In fig.1a, some subsequences are shared by some of the freqent sequences. Such molecules may be orderd, as is in fig.2. In this figure, starting from the longest reverse-complementary sequence pairs (at the top), subsequences whose tail part are lost are ordered on either row, with their reverse-complementary sequences.
In some pairs, the population sizes are asymmetric (e.g. there are 2 CCUCCGGCCCG and 26 CGGGCCGGAGG). This implies an "elongation mechanism". CCUCCGGCCCG is replicated 217 times and made

a)

```
GGAGG ..... 40        CCACCGG ....... 10        GGGCCGGAGG .......... 11
CGGAG ..... 10        CCUCCGG ....... 11        CGGGCCGGAGG ........... 26
GGUGG ..... 7         GCCGGUGG ........ 9        GGCGCCGGUGG ........... 8
CGGUG ..... 8         GCCGGAGG ........ 26       GCGGGCCGGAGG ............ 48
CGGAGG ...... 39      CCGGAGGG ((...)). 6       CGGCGCCGGUGG ............ 9
CGGUGG ...... 14      GGCCCGCG ........ 9        CCACCGGCGCCG ............ 6
CCGGAG ...... 19      GGCGGGCCG ((....)). 12    CCUCCGGCCCGCG ............. 16
CCUCCG ...... 51      CGGCCCGCA ......... 6      CGCGGGCCGGAGG ............. 10
CCACCG ...... 25      GGCCGGAGG ......... 12     CCUCCGGCCCGCA ............. 7
GCCCGC ...... 6       CGGCCCGCG ......... 6      GGCCCGCGGGCCGGAGG
CCGGUG ...... 9       CGCCGGUGG ......... 7      ((((....)))).....  7
CCGGAGG ....... 44    GCGCCGGUGG .......... 13   CCUCCGGCCCGCGGGCCGGAGG
CCGGUGG ....... 31    CCGGCCCGCG .......... 7    ((((((((((....)))))))))) 15
```

b)

```
GGCCGGAGGCGGGCCUCCGGCGCGCGCGCCUGCGGGCCGGAGGGG .(((...))).((((((((.(((.....))).)))))))..
GGCCCGCGGGCCGGAGGGGCGCCGGGGGCGGGCCGGAGGGCGCGCCG .((((.((((((.....))).))).)))).((((((.....)).))).
CGGCGGGCGCCCCGCGCGCGGGUGGCGGCCCGCGGGCCGGCGCCGGUGG ..((.((((((.((.(((((((....)))))))..)).)))))).))..
AGGCGGCCCUCCGGCCCGCGCCGGGCCUCCGGCCCGCUUGCGGGCUGGGCCU
....((((((...(((((((((.((((((...))))))..)))))))))))))).
```

Fig. 1. a: A list of frequent sequences from a snaphot of a simulation. On each line, a sequence, its secondary structure and its population size are shown. b: The longest sequences, from the same snapshot. Some of them have 2 forks.

by random cut 108 times. 297 of them are used as primers for replicating longer sequences. Although this mechanism is limited to single-stranded molecules, we did obtain population growth of some long sequences by assembly (see §3.2). Note that the reverse-complimentary sequence, CGGGCCGGAGG cannot elongate, because no dangling 3' end when it matches CCUCCGGCCCGCG.

If we focus on the relationship between a long sequence and its subsequences with loss of tails, above elongation mechamism is regarded as a "repair mechamism". While long sequences are more fragile than short sequences, this mechamism supports them and the diversity of the whole system is maintained.

The replication network is an important driving force for the "evolution to complexity". As described above, elongation and repair processes maintains the variety of sequences and long sequences are formed and maintained. While we have not investigated spatial pattern formation, we think spatial clustering supports above elongation and repair. Since the reaction is local, the formation of cluster is trivial. Then, if a molecule elongate, its reverse-complementary sequence is expected to found in the neighborhood and whole cluster can be elongated. Also, the repair can be done in a similar manner.

3.2 Assembly

The long molecules (more than 20) usually have secondary structure and they cannot replicate but are formed via concatenation. However, in fig.1a, some of folded sequences are reproduced. Such long sequences are "assembled", by frequent sequences. For example, CCUCCGGCCCGCGGGCCGGAGG is a combination of CCUCCGGCCCGC and

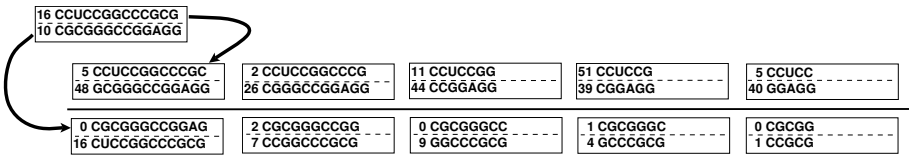

Fig. 2. An example of the replication network. Each box shows reverse-complementary pair. From the longest pair (the box at the top left), their subsequences are shown on the either row (separated by the line) , orderd by the length. The number on the left of the sequence is the population size.

GGGCCGGAGG. Further elongation is possible, if there are dangling ends. Thus we may hope a step-by-step extention of sequences.
In fig.1b, some of the longest sequences are shown. Some molecules have multiple forks. About 30% of simulations have such complex secondary structures. The formation of long, multi-fork molecules can lead to the emergence of "ribozymes". Indeed, we have obtained some interesting molecules, which are about 100 nucleotides long, and three or more forks.Also, another merit of folded molecule is stability, since the double-stranded regions do not break by the random cut.

3.3 Parasites

When the replication fidelity is high, parasitic sequences may take over. Trivial sequences, such as all-C or all-G, can easily replicate and concatnate because of high probability for the ligation of a primer. Once some all-C and all-G sequences meet and start replication, their populations grow rapidly. Such trivial sequences do not allow the formation of "ribozymes" with complicated secondary structures.

3.4 the effects of mutation

We have calculated the average ratio of parasitic invasion, maximum sequence length and maximum population size. See table.1. Those values are averaged over 100 samples. The classification is evaluated as follows. **A**: more than half out of 10 longest molecules have multiple forks. **B**: secondary structures are less seen, but not invaded by the parasites (i.e. quickly growing all-G/all-C sequences). **C**: invaded by parasites.
On the right side of the table, averaging only **A**-class runs, total population size (pop), maximum length (L_{max}), maximum population size restrincting length ≥ 10 ($P_{max}1$), maximum population size without restriction ($P_{max}2$) are shown. The averaged maximum length and population size decrease as the fidelity decreases; this is a trivial result of mutation. Also, the mutation suppresses the parasitic invasion, by giving a secondary structure to mutants of all-C and all-G sequences.
The percentage of **A**-class case have an optimum when fidelity is 0.90, due to mutatants of parasites, which have several forks. However, the

fidelity	**A**	**B**	**C**	pop	L_{max}	$P_{max}1$	$P_{max}2$
1.0	20	24	56	2483.6	67.6	20.8	44.3
0.98	32	28	40	2934.5	61.0	13.0	36.6
0.95	34	38	28	2795.1	64.3	4.3	23.1
0.90	35	61	4	2596.3	55.4	2.8	15.3
0.80	27	63	10	2544.3	48.6	2.2	9.9
0.70	19	68	13	2384.5	43.8	2.0	8.8

Table 1. Classification of simulation results and statistics. See text.

maximum population is a few at this point. Both rich diversity and population growth of a molecule are obtained around 0.95 fidelity.

4 Summary

The spatial clustering supports the complexity of replication network, where elongation and repair of sequences occur. Within the replication network, frequent sequences increase the population of assembled long sequences where the complexity increases further, beyond the boundary of error threshold. Then the enzymatic activity may expected to arise. We will study the transition to replicase dominated dynamics in the near future.

One of the authors (TY) is supported by the fellowship from Japan Society for the Promotion of Sciences.

References

1. Joyce, G.F., Orgel, L.E.: Prospects for understanding the origin of the RNA world. In Gesteland, R.F., Cech, T.R., Atkins, J.F., eds.: The RNA world, 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (1999) 49–77
2. Anderson, P.W. Proc. Natl. Acad. Sci. USA **80** (1983) 3386–3390
3. Eigen, M., Schuster, P.: The Hypercycle: A Principle of Natural Self-Organization. Springer-Verlag, Berlin (1979)
4. Swetina, J., Schuster, P. Biophysical Chemistry **16** (1982) 187–203
5. Inoue, T., Orgel, L.E. J. Am. Chem. Soc. **103** (1981) 7666–7667
6. James, K.D., Ellington, A.D. Chemistry and Biology **4** (1997) 595–605
7. Breyer, J., Ackermann, J., McCaskill, J. Artificial Life **4** (1998) 25–40
8. Toffoli, T., Margolus, N.: Cellular Automata Machines. MIT Press, Cambridge (1987)
9. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., Schuster, P. Monatshefte f. Chemie **125** (1994) 167–188
10. Konings, D., Hogeweg, P. J. Mol. Biol. **207** (1989) 597–614