DIMACS Technical Report 97-83
December 1997

RNA Based Computing: Some examples from RNA catalysis and RNA editing

by

Laura F. Landweber [1]
Department of Ecology and Evolutionary Biology
Princeton University
Princeton, NJ 08544
lfl@princeton.edu

[1] Thanks to DIMACS for stimulating this research, which was originally presented at the second DIMACS meeeting on DNA Based Computers in June, 1996. The research of the author was supported in part by a Burroughs Wellcome New Investigator Award in Molecular Parasitology and by NSF MCB-9520253. *Keywords*: in vitro selection, directed evolution, RNA editing, RNA ligase, ribozyme, SELEX

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs-Research, Bell Labs and Bellcore.

DIMACS is an NSF Science and Technology Center, funded under contract STC–91–19999; and also receives support from the New Jersey Commission on Science and Technology.

ABSTRACT

This report will focus on a combined treatment of the fields of RNA evolution and DNA computers. Using examples of in vitro selection of functional RNA molecules from large pools of random sequences, we show that these experiments can be viewed as an "RNA computer", in the language of DNA computation. By applying a strict set of criteria to a pool of heterogeneous DNA or RNA sequences, both types of experiments find sequences that contain solutions to hard molecular computational problems. These can even be nondigital problems that an electronic computer cannot solve. For example, a problem we posed to a pool of RNA molecules was one of catalysis, or the ability to ligate a small substrate RNA molecule to itself. In vitro evolution searches for the best solution or class of solutions contained within a large library of approximately 10^{15} unique sequences. The advantage of these "molecular computers" is their ability to examine and eliminate billions of possible solutions in parallel. We also describe an example of a cellular paradigm for RNA-based computing: RNA editing.

1 Molecular Selection

In vitro selection is a technique which, like the Polymerase Chain Reaction (PCR), has revolutionized molecular biology, because it allows the study and construction of novel nucleic acid molecules with specific properties or functions. Like DNA computers, in vitro selection is a seiving process of collecting desired molecules and discarding unwanted ones from large combinatorial libraries of 10^{13} to 10^{16} unique sequences (Figure 1).

Independently invented by three laboratories in 1989 and 1990 ([6], [4], [16]), in vitro selection allows amplification and isolation of rare molecules based solely on their specific functional properties and not on their primary sequence. All random molecules in molecular selection experiments are flanked by identical primer binding sites for PCR amplification; therefore it is essentially "PCR by function". The ability to pass a functional assay and to survive to the next step depends on either binding a target ligand or undergoing a self-modification reaction. This leads to differential amplification of the "fittest" molecules. In a catalytic assay for RNA enzymes, or ribozymes, the active molecules themselves must become modified by the reaction in such a way as to provide a "fishing hook" to distinguish them from inactive ones and to allow their capture from a large sequence pool. This limits the types of RNA catalysts (ribozymes) which can be retrieved by in vitro selection to those that can act in *cis*, allowing "self-tagging" of active ribozymes. Any molecule which catalyzes the reaction only in *trans*, meaning it acts on another molecule but not on itself, will be lost during the filtering process.

Through multiple rounds of selection and amplification, in vitro evolution allows amplification, mutation and isolation of single desired nucleic acid molecules from large pools of heterogeneous sequences, leading to purification of molecules present in as few as one in 10^{16} copies. Since these molecules can have novel catalytic properties and myriad applications-such as binding a specific ligand or catalyzing a reaction, the ability to harness combinatorial chemistry to isolate new ribozymes has launched a whole new industry of generating designer RNA tools. By expanding our knowledge of RNA's functional repertoire, the technique has also revealed the possibility of uncovering early catalysts in RNA evolution and even resuscitating active RNA molecules that have been "extinct" for over 3.9-billion years.

Normally the genetic carrier between DNA - informational molecule - and protein - functional molecule, RNA is an ideal unit for in vitro evolution, because it can combine both genotype and phenotype into one molecule. Therefore selection can act directly on RNA or DNA, because the primary sequence is contained within the actual molecule. In addition, the sequence is easily mutated and recovered, allowing experimental evolution on test-tubes of nucleic acids.

The basic concept of in vitro selection is the differential amplification of molecules with desired properties from pools of random sequence (Figure 1). Amplification of selected molecules increases their frequency in subsequent rounds, or generations, leading to survival of the fittest molecules. Multiple rounds of in vitro selection plus evolution allow rapid amplification, mutation, and isolation of a single desired nucleic acid molecule that can bind a novel target or perform a specific function, from a pool of as many as 10^{16} heterogeneous

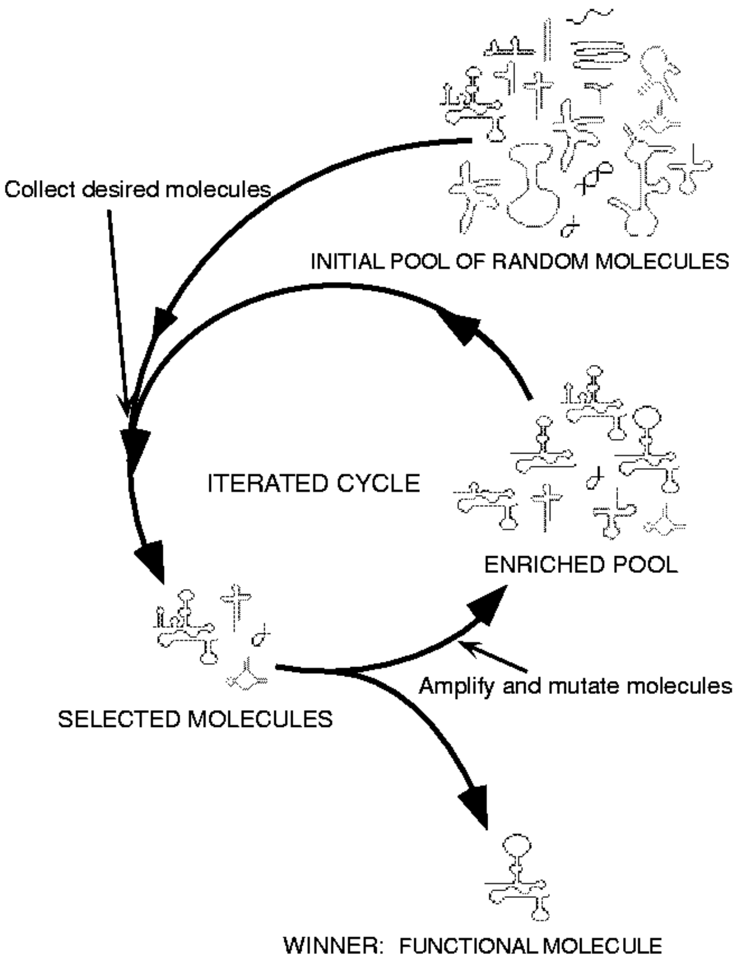

Figure 1: Outline of in vitro evolution. The "winning" molecule or class of molecules at the end are the most efficient at surviving the selection cycle [11].

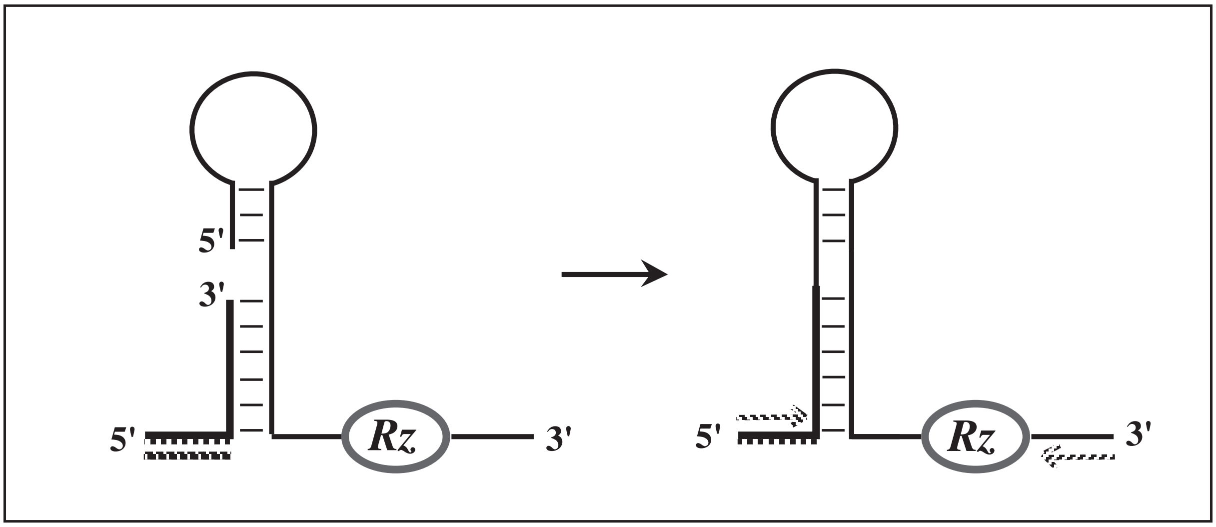

Figure 2: Selection for a ligase ribozyme (Rz). Arrows (right) indicate specific PCR primers that enable recovery of novel ligase ribozymes from large pools of random sequences [1]. This design allows differential amplification of molecules which ligate the substrate RNA (bold) onto their 5' end, thus incorporating a new DNA sequence tag (dotted) which provides both an affinity tag (shown complexed with the complementary strand on the left) and a selective PCR primer site [11].

sequences [11].

Each iterated cycle mimics several aspects of evolution. Variation is supplied in the initial pool by randomization during oligonucleotide synthesis, and additional variation can be introduced in each cycle by mutagenic PCR, which is amplification with reduced fidelity of replication [1]. This allows optimization of the selected molecules. Selection is imposed by the stringency of the protocol used to capture or "trap" desired molecules - usually a binding or a functional assay. Most binding approaches rely on standard molecular isolation steps such as affinity chromatography (in which the specific RNA sticks to a column loaded with target), filter binding (in which the RNA does not pass through a filter if it is complexed with target), gel mobility shift (in which the RNA bound to target migrates as a larger complex), or immunoprecipitation (in which an antibody recognizes the target plus RNA) ([4], [16]). An example of a direct functional assay is ligation to an oligonucleotide containing a unique sequence tag (Figure 2) or biotin. The sequence tag provides a recognition sequence for affinity chromatography, with the complementary strand serving as "bait", and biotin allows capture of ligated molecules to a streptavidin support; either strategy effectively converts the functional assay into a binding one. In addition, only those molecules that become covalently linked to the substrate RNA can be amplified by selective PCR using the sequence tag as a primer [2], [1].

In a landmark experiment [1], Bartel and Szostak used in vitro selection to isolate new

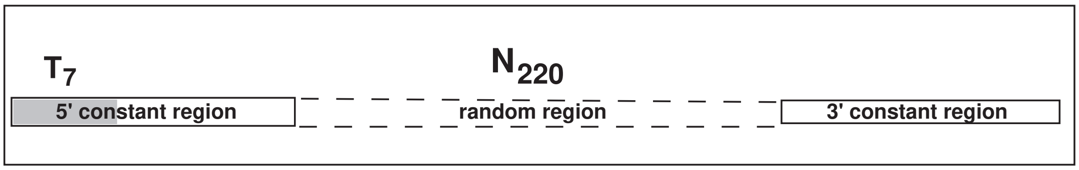

Figure 3: Construction of a random pool for in vitro selection.

RNA ligase ribozymes from a large pool of random-RNA sequences (Figure 2). Their discovery proves that one can retrieve molecules present in as few as one in 10^{15} copies which have novel enzymatic properties.

The pools of nucleic acid molecules used in molecular selection experiments are generally constructed with 5' and 3' constant regions that contain an RNA polymerase promoter (to allow RNA synthesis) and primer binding sites for reverse transcription and PCR (to recover and produce more copies of the selected sequences) [11] (Figure 3). The constant regions therefore allow PCR amplification of the pool, in addition to the design of the specific selection experiment. The pool in Bartel and Szostak's experiment contained 220 nucleotides of random sequence (N220), flanked by a 5' constant region which could anneal to an oligonucleotide substrate RNA, forming a small hairpin loop which positioned the 3' end of the substrate RNA adjacent to the 5' end of the pool RNA by base-pairing (Figure 2). Ligated molecules become specifically tagged with the new sequence at the 5' end of the substrate RNA. This sequence then provides both an affinity tag for chromatography and a primer for "selective PCR"-amplifying only molecules with a covalently attached 5' primer. Bartel and Szostak were able to select for RNA molecules that could catalyze ligation of the substrate RNA to the pool RNA after just four cycles of enrichment by affinity chromatography and selective amplification. We performed a similar experiment which led to the discovery of a minimal RNA ligase motif (Landweber, unpubl.), and experiments in other laboratories have led to the discovery of many classes of novel nucleic acid catalysts, including DNA enzymes isolated from pools of single-stranded DNA [11].

Nucleic acid-based computers and selection both search through a test-tube of molecules representing some subset of "sequence space" to find a molecule or class of molecules which satisfies the requirements of a strict selective filter. DNA computing mimics the field of in vitro selection by applying a strict set of functional criteria to a pool of heterogeneous DNA sequences; however, a DNA computer generally has to perform an exhaustive search of all possible solutions. In vitro selection, on the other hand, empirically finds the best possible solution in a limited sampling of sequence space. For example, a library containing 10^{15} sequences with a 100 nucleotide random region represents only a very negligible fraction of all possible sequences of length 100 ($4^{100} > 10^{60}$, which is approximately the number of protons in the most massive star in the universe!). The success of in vitro selection experiments using pools of fewer than 10^{16} unique sequences already suggests an enormous wealth of functional molecules distributed in sequence space, many of which contain "solutions" to hard molecular computational problems.

The purification of desired sequences to homogeneity usually requires ten or more rounds of in vitro selection. The functional molecules which survive a strict selective filter ultimately become the dominant species in the pool. Thus in vitro evolution reduces the pool complexity to one (or more) functional class(es) of sequences, much like searching for solutions to hard bio-computational problems [12]. The technique has quickly enjoyed popularity, mainly because of its power to examine in parallel billions of possible solutions to important biochemical problems such as catalysis or binding.

A number of parameters in molecular selection experiments greatly affect the nature of the evolutionary process being simulated and therefore the likelihood of finding a "winning" molecule within a large collection of sequences. These parameters include the stringency of the selection, the amount of initial variation in the pool, the amount of variation introduced at each successive round of amplification, the choice of salt concentrations and buffers, and numerous considerations about the design of the pool for the individual experiment [11].

The RNA pools can be one of two types. The first is completely random; all bases have an approximately equal probability (i.e. 25% each A, C, G & U) in a region of defined length, such as 220 bases (Figure 3). The pools are either synthesized as long DNA oligonucleotides or assembled from smaller ones. They are then made doubled stranded and amplified by large-scale PCR [1] and transcribed into RNA by T7 RNA polymerase. The sequence composition of such pools must be carefully checked to avoid biases introduced during DNA synthesis. Random pools are typically used to isolate novel ribozymes or aptamers-molecules that bind specific ligands. The second type of pool is a "doped," or variant, pool in which an existing sequence is mutated at a defined frequency (e.g. each consensus nucleotide with probability 85% and each non-consensus nucleotide with probability 5%). Doped pools are often used to select variants of existing ribozymes with improved efficiency or to obtain molecules that function with different substrates or in different conditions [11]. They can be synthesized either by mutagenic PCR [2] or by controlled oligonucleotide synthesis [4].

In vitro selection allows the exploration of the sequence space of catalytic RNA molecules, sometimes referred to as "catalytic task space" [7]. We and others have selected RNA molecules with desired properties from combinatorial libraries containing over 10^{15} unique sequences. The steps always involve an iterative procedure of artificial selection and PCR amplification of the rare molecules. There are several opportunities for interaction between computational molecular biology and these experiments. For example, one can use the data gathered from in vitro selection experiments to model in vitro evolution of the pool of RNA molecules. The patterns of gene-frequency changes in such a system resemble patterns in natural populations under strong directional selection. In addition, in vitro selection experiments allow one to map out both the underlying ruggedness of the adaptive landscape as well as the distribution of functional RNA molecules in sequence space. These are problems which have hitherto been treated theoretically by Eigen, Schuster, Kauffman and others (e.g. [14]); however in vitro selection now provides a means to test numerous predictions and also to refine the theory. Indeed, the molecular "solutions" to computational problems imposed by in vitro selection can be treated as satisfiability problems, within the framework of nucleic acid-based computers.

```
DNA    G    G     GTTTTGG   AGA       G ATTTGG  A
RNA   uGuuuuGuuuuuGUUUUGGuuuAGAuuuuuuuGuAU  GGuuAuuu
```

Figure 4: Comparison of a representative region of the *H. mariadeanei* COIII mRNA sequences with the corresponding region of the mitochondrial DNA. DNA sequences in upper case; uridines in mRNA that are added by RNA editing in lower case; encoded thymidines deleted from the mRNA indicated by gaps.

2 RNA Editing

RNA editing is the modification of RNA sequences by addition, deletion, or substitution of bases. Found in a wide variety of eukaryotes, ranging from parasitic protozoa to humans, this remarkable process alters the sequence of messenger RNA before translation into protein such that the resulting protein sequence often differs dramatically from the original gene sequence, which sometimes does not encode a protein at all. For example, in *Trypanosoma brucei*, RNA editing by addition and deletion of literally hundreds of uridine residues creates initiation and termination codons, alters the structural features of transcripts, and creates over 90% of the amino-acid codons [9]. The main features of RNA editing are identified by differences between RNA and genomic DNA sequences.

Figure 4 illustrates an extreme example of RNA editing of the mitochondrial cytochrome oxidase subunit III (COIII) transcript in *Herpetomonas mariadeanei*. 91% of the 288 amino acid codons in *T. brucei* [5] and *Herpetomonas* [9] are created by editing, with this effect restricted only to uridines (Figure 4). The other bases - A, C, and G - are completely conserved between the DNA and the RNA sequence.

The edited proteins also have a fast molecular clock, which accumulates amino acid substitutions at least twice as fast as the unedited versions. This is due to frameshift mutations that are introduced by fixed changes in the number and position of U's inserted or deleted by editing and then compensated by editing at another site that restores the reading frame (Figure 5). Editing therefore allows combinatorially diverse protein products to be generated from a single gene, either over evolutionary time [9] or within an individual cell [15]. The surprising discovery that RNA editing is a source of sequence mutation, rather than a proper "editing" or correcting mechanism, underscores the importance of the question of why it is still used by some organisms to generate a sequence encoding a single conserved protein [10].

The genetic information for editing is stored in the form of "guide" RNA molecules (gRNAs), very small (50-70 nucleotides) transcripts that mediate editing by base-pairing with specific regions of the edited transcript, allowing some G:U base-pairs [3]. Each gRNA contains the sequence information to edit approximately 30 nucleotides of edited RNA [9] and pairs more efficiently with the final product than with the pre-edited substrate. For every inserted U in the messenger RNA sequence, there is a corresponding A or G in the gRNA which pairs with the fully edited product (Figure 6). Complete editing proceeds 3' to 5' on

```
Hsa DNA   G AA      A        G G   GG   AGA      G  G     G G     G
Hsa RNA uuGuAAuuuuuuAuuuuuuu GuGuuuGGuuuAGAuuuuuuGuuGu  uuGuGuuuuuGu
Hme DNA   G AA         ATTTT G G   GG   AGA      G  GT    G A     G
Hme RNA uuGuAAuuuuuuuuuAUUUU GuGuuuGGuuuAGAuuuuuuGuuGU  uuGuAuuuuuGu
Hmm DNA   G AA         ATTTTTG G   GG   AGA      G  GT    G A     G
Hmm RNA uuGuAAuuuuuuuuuAUUUU GuGuuuGGuuuAGAuuuuuuGuuGU  uuGuAuuuuuGu
Hma DNA    GAA    G    G      GTTTTGG   AGA       G ATTTGG  A   G
Hma RNA uuuGAAuuuuGuuuuGuuuu uGUUUUGGuuuAGAuuuuuuuGuAU  GGuuAuuuGuuu
Tbr DNA  G  AA       G    G  GTTTTTGG   AGG       G  GTTTTG  G    G
Tbr RNA uGuuAAuuuuuuuGuuuuGu GUUUUUGGuuuAGGuuuuuuuGuuG  UUGuuGuuuuGu
```

Figure 5: An example of fixed frameshifts between species with editing. Hsa, *H. samuelpessoai*; Hme, *H. megaseliae*; Hmm, *H. muscarum muscarum*; Hma, *H. mariadeanei* [9]; Tb, *T. brucei* [5]. DNA sequences in upper case; uridines in mRNA that are added by RNA editing in lower case; encoded thymidines deleted from the mRNA indicated by gaps.

gRNA

```
 3'-UaaUUaUagaaCaUagagaCUgUaaaUaaAUAAACAACCAAAUAUA-5'
    :|||||:|:||||||:|:|||::|||||||||||||||||||||:|
5'...GuuAAuGuuuuGuAuuuuuGAuGuuuAuuuAuuuGuuGGuuuAuGu...3'
```

mRNA

Figure 6: Representative pairing between a guide RNA (top) and the edited mRNA (bottom). Lowercase u's are added by editing. Boldface u's are edited by base pairing to lowercase a's and g's in this gRNA [13].

the mRNA and requires a full set of overlapping guide RNAs. Editing by each guide RNA creates an anchor sequence for binding the next guide RNA (Figure 7; [3], [13] leading to an ordered cascade of editing events. It is also a very slow but faithful mechanism, considering that one can find hundreds of partially edited and even incorrectly edited intermediates at steady state [9] although these presumably lead to correct product formation. Thus RNA editing is a surprisingly inefficient cellular computing mechanism for producing messenger RNA sequences, although directed editing 3' to 5' ensures that only fully edited transcripts are translated, since the AUG or UUG [9] translation initiation codon is edited last.

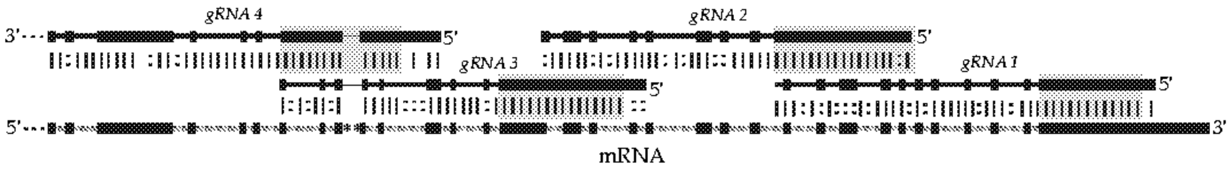

Figure 7: Editing by four overlapping gRNAs. Thick lines in the mRNA are encoded in the mitochondrial DNA. Thin shaded lines are inserted U's; the two asterisks are deleted U's [13]. Thin lines in the gRNA's are guide nucleotides (A or G) that pair with inserted U's. Vertical lines indicate Watson-Crick base pairs; colons indicate G:U wobble base pairs, illustrating formation of well-paired "anchors" between the 5' ends of gRNA's shaded boxes and the corresponding region of the mRNA.

3 Conclusions

The ability to isolate unique RNA molecules through in vitro selection has given rise to the field of in vitro genetics, which can be viewed either as a fantastic tool for the synthesis of novel catalysts through combinatorial chemistry or as a simple system for probing the stepwise assembly of processes involving RNA. Together, the results of in vitro selection experiments have revealed not only RNA's ability to catalyze a much wider range of reactions than previously thought but even DNA's cryptic ability to operate as an enzyme. Thus, within laws of thermodynamics, the possibilities are virtually endless for the selection of future nucleic acid catalysts, perhaps catalyzing reactions for which no known catalyst exists, such as the breakdown of toxic substances. As a tool for drug discovery, in vitro selection offers a new direction for the synthesis of potential therapeutic agents. As a tool for studying evolution, in vitro selection allows us the first opportunity to study experimentally the biological origins of catalysis and the emergence of order from random sequences. RNA editing is an example of a cellular process which uses RNA sequences as guides to convert seemingly disordered RNA sequences into functional messenger RNA: a truly RNA-based computer.

References

[1] Bartel, D.P., and J. W. Szostak. (1993). *Isolation of New Ribozymes from a Large Pool of Random Sequences*, Science **261**: 1411-1418.

[2] Beaudry, A. A., and G. F. Joyce. (1992). *Directed Evolution of an RNA Enzyme*, Science **257**: 635-641.

[3] Blum, B., Bakalara, N., and L. Simpson. (1990). *A model for RNA editing in kinetoplastid mitochondria: "Guide" RNA molecules transcribed from maxicircle DNA provide the edited information*, Cell **60**: 189-198.

[4] Ellington, A. D., and J. W. Szostak. (1990). *In vitro selection of RNA molecules that bind specific ligands*, Nature **346**: 818-822.

[5] Feagin, J.E., Abraham, J.M., and K. Stuart. (1988). *Extensive Editing of the Cytochrome c Oxidase III transcript in Trypanosoma brucei*, Cell **53**: 413-422.

[6] Joyce, G.F. (1989). *Amplification, mutation and selection of catalytic RNA*, Gene **82**: 83-87.

[7] Kauffman, S. (1993). *The origins of order: self-organization and selection in evolution*, Oxford University Press, New York, New York.

[8] Landweber, L. F., Fiks, A. G., and W. Gilbert. (1993). *The Boundaries of Partially Edited Cytochrome c Oxidase III Transcripts are Not Conserved in Kinetoplastids: Implications for the Guide RNA Model of Editing*, Proc. Natl. Acad. Sci. USA **90**: 9242-9246.

[9] Landweber, L.F. and W. Gilbert. (1993). *RNA editing as a novel source of genetic variation*, Nature **363**: 179-182.

[10] Landweber, L. F. and W. Gilbert. (1994). *Phylogenetic analysis of RNA editing: A primitive genetic phenomenon*, Proc. Natl. Acad. Sci. USA **91**: 918-921.

[11] Landweber, L. F., Simon, P. J. and T.A. Wagner. (1998). *Ribozyme Design and Early Evolution.* BioScience (in press).

[12] Lipton, R. J. (1995). *DNA solution of hard computational problems*, Science **268**: 542-545.

[13] Maslov, D.A. and L. Simpson. (1992). *The polarity of editing within a multiple gRNA-mediated domain is due to formation of anchors for upstream gRNAs by downstream editing*, Cell **70**: 459-467.

[14] Schuster, P., Fontana, W., Stadler, P. F., and I. Hofacker. (1994). *From sequences to shapes and back: a case study in RNA secondary structure*, Proceedings of the Royal Society of London **255**: 279- 284.

[15] Sommer, B., Kohler, M., Sprengel, R. and P.H. Seeburg, (1991). *RNA editing in brain controls a determinant of ion flow in glutamate-gated channels*, Cell **67**: 11-19.

[16] Tuerk, C. and L. Gold. (1990). *Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase*, Science **249**: 505-510.