Article

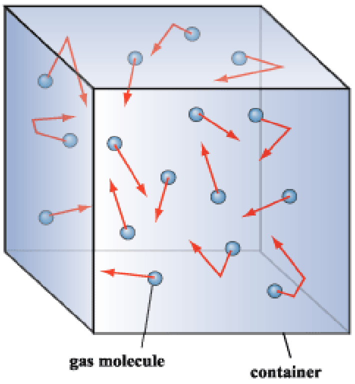
# RNA Secondary Structure as a Model Toward the Understanding of Infinite-Organized-Complexity from Simple Initial Rules

James F. Lynn
May, 2009
www.rnaparse.com

## Introduction

In a general sense, complexity is used to describe a *system* of many parts which form intricate structure or structures. I'd like to modify that definition by further defining and dividing “complexity” into two separate classes and omitting “structure” from one of the two definitions.

For purposes of this discussion, complexity is divided into two separate classes of “disorganized complexity” verses “organized complexity.” Disorganized complexity can be likened to some number of elemental atoms of a mixed gas held within the confines of a container – it possesses neither meaningful structure and contains little or no *inherent* information given the confined atoms are of a type that do not further combine to form increasingly complex molecules.



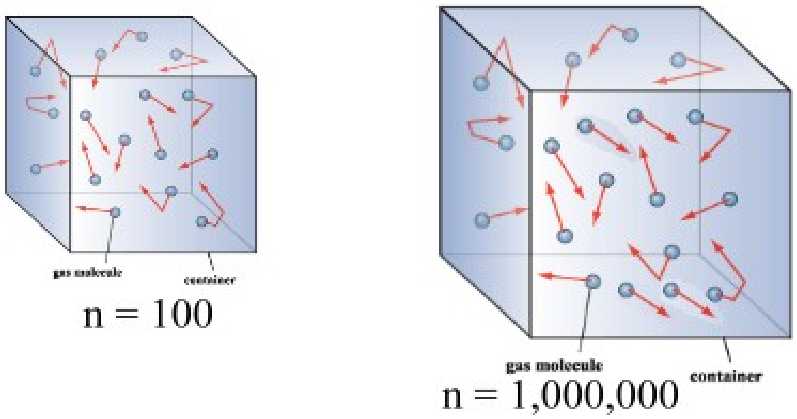
Disorganized complexity: Atoms held within a container possess no structure – their random motion is neither created from information nor creates information. Increase the number of atoms and the level of disorganized complexity increases but of great importance note the level structural information does not increase.

Disorganized complex systems lack the ability to increase its inherent information content no matter how large the system becomes by increasing the number of its parts. In that sense, a small disorganized complex system is no more complex than a large one. Exactly the opposite is true of systems of organized complexity.

## Organized Complexity

Any organized system, closed or not, must contain a minimum of two parts or objects that possess *inherent information* and act upon one another in some meaningful way. A system containing one part is uninteresting in the fact that it cannot form any degree of organized complexity beyond simple self-to-self interaction.

Remembering our container of gas above we can imagine the parts as atoms of hydrogen and oxygen: a large number of oxygen molecules quickly combine to form O2, and the hydrogen largely stays in a mono-atomic state. The inherent information possessed by our two molecules is limited to relatively simple rules of atomic bonding and the information content cannot be increased by adding more atoms of  oxygen and hydrogen. That is to say, the level of organized complexity *does not increase* as the system's size increases.



Two containers of gas, one containing 100 atoms and the other 1,000,000 atoms essentially possess the same amount of information. Because the information content of this system is limited, the degree of organized complexity is held at a constant no matter how much we increase the number of members in the system. It is thus said to be disorganized.

## RNA – Organized Complexity

Up to now we have not well defined what makes up an organized-complex system. Essentially it is this; Like our disorganized system, an organized-complex system contains “parts” or members that possess some inherent property that allows them to interact with other members of the system. Hence, the point of bifurcation at which a system becomes organized verses disorganized is directly related to three things; the number and type of members in the system and to the properties of the individual components.

Ribonucleic acid (RNA) can be thought of as the workhorse that allows living, organic complex systems to exist. RNA is transcribed from its counterpart, DNA, and is the stuff that “does” things within living cells. DNA is really nothing more than an information storage template from which RNA it built.

DNA is a linear molecule composed of four nucleotides {ATGC} held together by a backbone of sugar and phosphate. During the process of transcription RNA is formed and differs slightly in chemical structure from DNA to form the four nucleotides {AUGC.} Like DNA, RNA also forms linear molecules and each component or nucleotide of RNA possesses *property* that allows it to interact with other members of its linear arraignment.

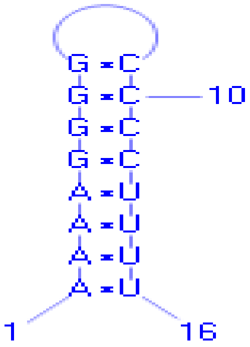
5' end -**A**-sugar/phosphate-**G**-sugar/phosphate-**C**-sugar/phosphate-**U**-sugar/phosphate-5' end

The linear structure of RNA mirrors its DNA and each strand can be millions of nucleotides long. The above strand is read by convention from the 5' (Five prime) position to the 3” position, simply as “AGCU.”

The essential property of each RNA nucleotide is quite simple with few exceptions; An RNA nucleotide possesses the ability to form weak bonds with other nucleotides in according to the following rules:

**A with U**<br>
**U with A**<br>
**G with C**<br>
**C with G**

Because of these simple properties RNA, as a linear molecule, is able to interact with other members of the system and can loop back on as if a string becoming tangled to form meaningful secondary structure as show by the simple 16 component stem-loop structure below.

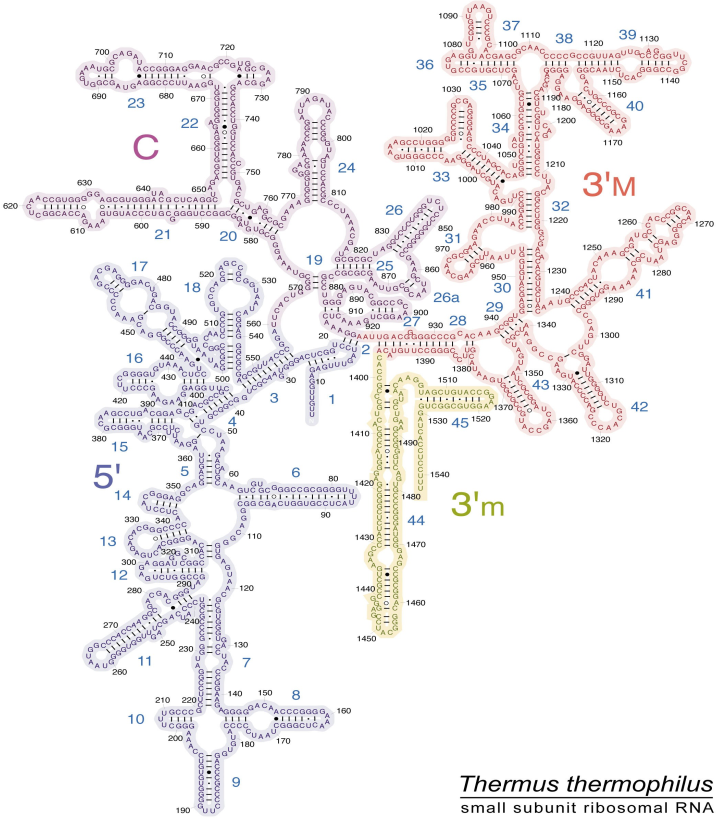


In a simple RNA stem the number of possible structures (P) is given by $n^r$ where n=4 and r= the length of the stem: A simple complement loop constructed of 4 nucleotides of length 16 as shown will theoretically form 65,536 structures. Adding just two more nucleotides raises the number to 1,048,576! Adding one of the 4 nucleotides to the loop at the top quadruples the number of possible structures.

From the above figure we can see the linear string “AAAAGGGGCCCCUUUU” bends back on itself according to rules of bonding to form an entirely new and meaningful structure, the stem-loop. Its a central axiom in RNA science that *structure* rather that specific nucleotide sequence gives an RNA molecule function and the functions of these specific secondary structures within the living cell are essential and numerous.

**Reiteration**

We have seen that disorganized systems such as free gases in a sealed container cannot form any meaningful structure beyond simple dimmers and other small molecules no matter how large the system becomes. Thus the information content is constant and static. We use the example of a linearly-arraigned system of RNA in which each component possesses of itself a basic bit of information that ultimately allow RNA to form emergent complex structures.



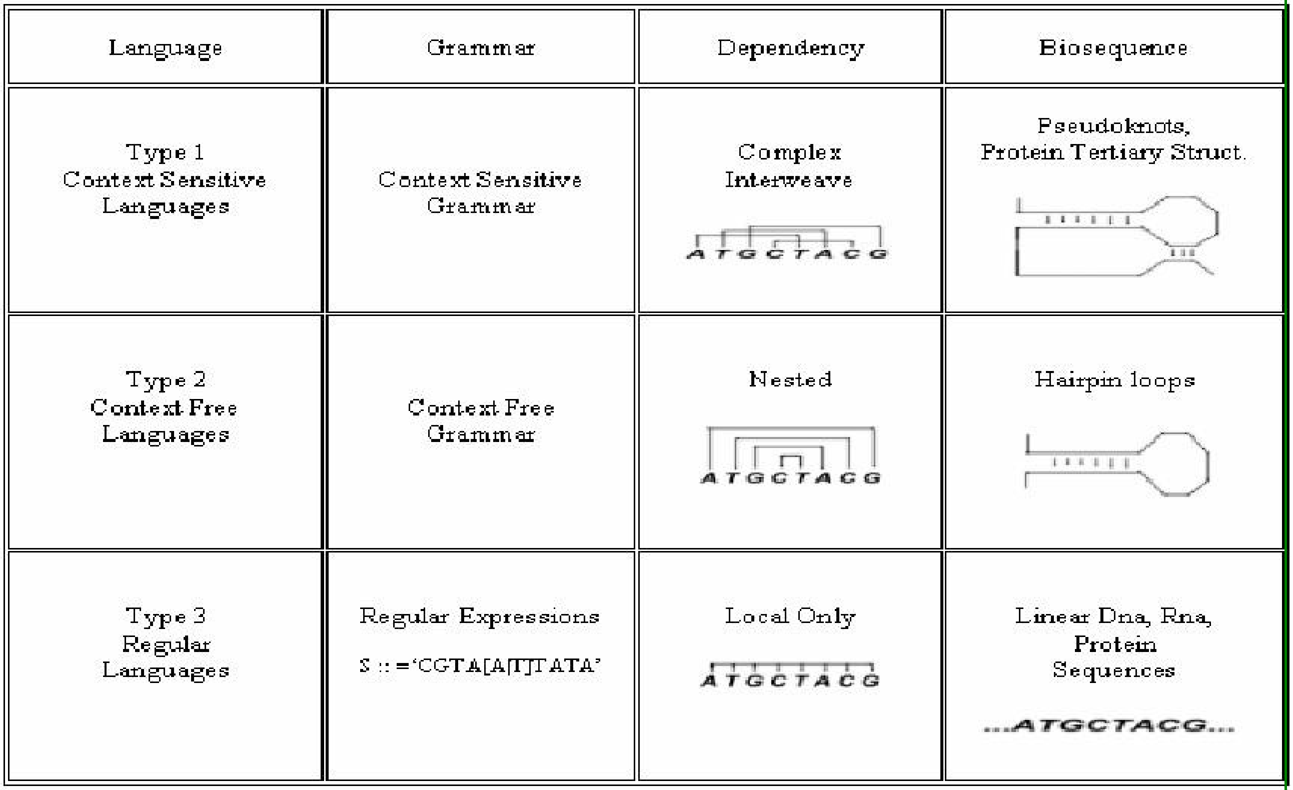
Thermus thermophilus 16S rRNA
Emergent organized complexity resulting from simple rules: RNA can form very complex meaningful secondary structures. Image credit: rna.ucsc.edu/rnacenter/ribosome_images.html

Hence, ever-increasing complexity from a few simple rules that govern how just four individual components of the linear strands of RNA interact with itself: RNA forms stems and loops (secondary structure) that in turn interact with other stems and loops (Tertiary structure.)

### Treating RNA Structure as a *Language*

Languages can be classified by order of complexity and of interdependency. English sentences are more than simple strings of words; they are interwoven expressions constructed from nouns, pronouns, verbs, adverbs, et cetera, that impart or convey meaning, often tied contextually to other expressions or conditions.

While all genetic biosequences are linear, one character following the next, each individual component may bond with another to form more complex structures, from simple loops through the extremely complex tertiary shapes such as globular proteins and pseudoknots.



| Language | Grammar | Dependency | Biosequence |
|---|---|---|---|
| Type 1<br>Context Sensitive<br>Languages | Context Sensitive<br>Grammar | Complex<br>Interweave<br>ATGCTACG | Pseudoknots,<br>Protein Tertiary Struct. |
| Type 2<br>Context Free<br>Languages | Context Free<br>Grammar | Nested<br>ATGCTACG | Hairpin loops |
| Type 3<br>Regular<br>Languages | Regular Expressions<br>S :: =‘CGTA[A\|T]TATA’ | Local Only<br>ATGCTACG | Linear Dna, Rna,<br>Protein<br>Sequences<br>...ATGCTACG... |

Classification of Language Type and Corresponding RNA Structures.

## Search and Match Methods

As a first step, RNA structure is matched via grammatical rules of bonding only and match independent of nucleotide sequence without using more computationally expensive methods such as multiple sequence alighnment (MSA) , lowest energy configurations and so forth.

These grammatical methods are not necessarily intended for new-motif discovery but can discover variations in known RNA motifs including multiple crossing, or pseudoknoted structure.
Since we do not try every possible combination of nucleotides but match a given structure itself, the method is linear and is able to scan entire small genomes for matches in seconds. Thus, a simple structure, ((((())))) is parsed in about the same linear, O(n) time and space as a pseudoknot (((([[[)))]]].

Grammatically derived results are inherently ambiguous but can be further filtered via any good minimal energy algorithm to weed out impossible or unlikely structures.

**Overview of Grammatical Method Used to Parse RNA Structure**

As show above, loops are relatively simple to isolate and are in essence palindromic constructs. Thus, we can design a grammar that *accepts* a certain number of nucleotides that are genetic palindromes:

```
AAAA loop TTTT
   |_______|
     stem
```

Taking things a bit further, we have discovered methods of “nesting nested structures” Such that the language becomes Type 1 or context-sensitive but remains computationally-linear.

```
AAAA GGGG TTTT CCCC
   |__________|
      |____________|
   stem1     stem2
```

These “stem and loop predicates” can be combined in any number of ways without becoming greater than computationally-linear. Any possible set of RNA secondary structure has its corresponding grammar and can be parsed by simply combining the proper stem and loop predicates.

I've heard it said that if all the protein guys would first work on RNA, then the problem of protein structure would be solved. Interestingly, loops and bulges and any number of other interactions can be written into the RNA grammars in ways that begin to resemble the complex three dimension structures of proteins – future work on RNA structure may help.