I was delighted when Tim Nilsen asked me to contribute an article for this special 20th anniversary issue. I’ve been very fortunate to witness and participate in the transformation of our understanding of RNA and the biology that is emerging because of it. My happenstance and serendipitous entry into RNA research provided opportunities for me to explore and meander, and benefit from the nucleic acid sequencing and computer technologies that were in their infancy when I was starting my scientific career. These then cottage industries have blossomed into major academic and biotechnology centers that have changed the manner scientific research is practiced, and provides for an alternative and complementary means to decipher complex systems.

Past to present

Introduction

As an undergraduate in the mid to late 1970s my broad interests were molecular biology with specific interests in molecular evolution and a growing interest in computers and mathematical abstractions to understand the system. At that time as an undergraduate at the University of California at San Diego (UCSD), a major question in molecular evolution was—“what came first, DNA or proteins.” RNA was mostly presented as that labile molecule, either as the mRNA intermediate between the stable DNA and the protein as the final product, or as the static structure that is attached to amino acids (tRNA) and bound by proteins in the ribosome (rRNA). I don’t recall discussion about the early origin and importance of RNA in San Diego. Although in hindsight I wonder if I did (and forgot) or should have. In the late 1960s Carl Woese (on sabbatical in San Diego, hosted by Orgel), Leslie Orgel (at the Salk Institute in San Diego), and Francis Crick (colleague and personal friends with Orgel) published the first articles about the importance of RNA in the earliest forms of life. And (3) both Harry and Carl said that RNA came before proteins and DNA, long before others appreciated this. Precious lessons, especially at the beginning of a career.

RNA folding

Less than a year after starting my graduate studies the first complete 16S rRNA sequence was determined. My task, my thesis project, given my interests in computers, mathematics, and evolution was to help determine its secondary structure from the analysis of only that one sequence. However we did not fully understand at the onset that for the E. coli 16S rRNA, the number of possible helices outnumbered the correct helices 1000 to 1. And while we expected the correct secondary structure would contain the most stable helices, we did not know, a priori, that the vast majority of the most stable helices would not be in the correct secondary structure, while many of the helices that are putatively significantly less stable are. How to proceed with my thesis project? In addition to our efforts to solve this RNA secondary structure with first principles, I was attempting to use the electron microscope to determine long range interactions in the rRNA. Little to no success on this latter project. Two strikes and I felt like I was in the batter’s box with Sandy Koufax on the mound. This challenge, the determination of the correct helices in the secondary and three-dimensional structure, from a single class. However my formal introduction to the importance of RNA would come as soon as I started graduate school at the University of California, Santa Cruz (UCSC). However this second opportunity almost slipped away. I had reservations about attending graduate school at UCSC. The presentation of RNA and the ribosome in Watson’s Molecular Biology of the Gene did not excite me. Fortunately Bill Loomis, my Molecular Biology professor at UCSD convinced me to go, saying—“Work for Harry Noller. He is young and creative. He will be famous”.

My first three lessons as a graduate student in Santa Cruz, starting in the fall of 1977: (1) Harry Noller said that everything you know is wrong. (2) Carl Woese (and George Fox) discovered the third form of life, the Archaeabacteria (later renamed Archaea) and called rRNA the magic molecule. And (3) both Harry and Carl said that RNA came before proteins and DNA, long before others appreciated this. Precious lessons, especially at the beginning of a career.
sequence and a collection of principles and rules was and continues to be one of the grand challenges in biology. A solution will be more than an accurate prediction, it will reveal dynamics that transcend physics, chemistry, and biology. However, this challenge, as defined, requires knowledge, rules, and mechanisms that we lacked (and continue to under appreciate).

Comparative analysis

Carl Woese provided the fix, plan B if you will, and as Woese described it—“Using one of Mother Nature’s oldest tricks” and assuring me when I commented that I didn’t have any experimental results to show for my earlier projects—“Don’t worry, Mother Nature have done the experiments for you”. Plan B is predicated on a simple Many (sequences) to One (structure) concept, and our task was to identify that one RNA secondary structure that is formed from the many different 16S rRNA sequences. This approach has itself evolved in its quest to identify and even attempt to define structural elements that are “conserved” during the evolution of the RNA. Different names have been tagged to this method—comparative, phylogenetic, covariation and others. I prefer “comparative” for the broad search for similar structural elements, and “covariation” for the specific application of comparative analysis that identifies positions in a sequence alignment that have similar patterns of variation. My PhD thesis, completed some 11 years before the RNA journal started, described my contributions to the comparative secondary structure of 16S rRNA. Additional refinements to the 16S rRNA, and the 23S rRNA comparative secondary structure models were subsequently published over the next 15 or so years, culminating in about 2000. These improvements in the comparative structure models included the removal of base pairs with an increasing amount of uncoordinated variation at the two paired positions (i.e., variation but no covariation), identification of new canonical (e.g., A:U ↔ G:C) and non-canonical (e.g., U:U ↔ C:C) base pairs that occur within regular “canonical” helices and non-canonical helices (e.g., pseudoknots, helices with a single base pair, tertiary interactions). A few years after the RNA journal started, the crystal structures for the 30S and 50S ribosomal subunits were solved at high-resolution. We were pleased, but not surprised when nearly 100% of the base pairs in the covariation-based structure model, including all of the irregular tertiary and tertiary-like base pairs were in the Thermus thermophiles 16S and Haloarcula marismortui 23S rRNA crystal structures. For some the problem was now solved, and comparative analysis had served its purpose. Others even said that I should now retire (What?). Yes, comparative analysis has determined the correct secondary structure. However comparative analysis has much more to reveal about RNA structure and evolution, and plan A, aka the RNA folding problem—the accurate determination of an RNA secondary structure with first principles—remains to be solved.

Present to future

Data and databases

Mother Nature’s successful experiments are inscribed in the nucleic acid sequences. And comparative analysis have benefitted from these sequences. Many lessons and biological principles can be deciphered from these linear array of nucleotides. Now enter the parallel development of the nucleic acid sequencing and computer technologies. They, in conjunction with the internet and the concomitant “networked Science,” are completely transforming the process of scientific discovery. Jim Gray, recipient of the Turing award for his seminal work on database systems eloquently discussed the exaflood of data and the need to develop new computational systems to retrieve, organize, distribute, visualize, and analyze this deluge of information. These data issues are most applicable for our broad and unfettered understanding of RNA. These methods and concepts are part of the new discipline—Data Science, and benefits from the massive amounts of “BIG DATA” that is overwhelming us.

While current computer technology make it possible to transform significant amounts of different types of data into knowledge, the quality and knowledge about the data is as essential as the amount of data. While Noller and Brosius provided highly accurate sequence data for the first complete 16S and 23S rRNAs, the quality of the current 16S rRNA sequences are of concern. For example papers now report that at least one in 20 16S rRNA sequences contain substantial errors (I suspect the number and ratio of bad RNA sequences is even higher and this rate will continue to increase). Others are concerned that microbiome RNA sequence databases that are filled with a sufficiently large percentage of “contaminants” should be regenerated from scratch with only high quality sequences and annotation. We must wonder to what extent the analysis and interpretation of microbiomes are adversely affected by bad data.

And while the scaling of the increased amount of data to the current problems in RNA research is important, it is also important to organize and annotate the data for new types of analysis. At the end of his career Carl Woese wrote about biological systems that are more than “a mere collection of parts” and that “All of these problems are different aspects of one of the great problems in all of science, namely, the nature of (complex) organization.” Freeman Dyson expanded upon these ideas—“The big problems … cannot be understood by reducing them to elementary particles and molecules. New ways of thinking and new ways of organizing large databases will be needed.” And most recently Keith Yamamoto at UCSF said “If we just aggregate the data, we’d just have a pile of data. The challenge is to devise ways to analyze it, mine it and make sense of it.” These words from the wise should be taken seriously. Our understanding of RNA will benefit, especially when the traditional RNA scientists (who publish their articles in the RNA journal)
collaborate with scientists who publish in journals such as *GigaScience* and *Nature’s Scientific Data*.

**Evolution of base pairs in RNA structure**

The two nucleotides that are base paired in the comparative structures have similar patterns of variation. This type of co-variation, or coordinated evolution is the foundation for the comparative analysis that accurately predicts the base pairs in the RNA secondary structure. However the conventional explanation for the creation and maintenance of these base pairs might not be correct. They said in the 1970s that it is easy. After a paired nucleotide—say position 500—is mutated to form a non-canonical base pair with position 545, then a random mutation is needed at 545 to recreate the canonical base pair, or a random mutation can change position 500 back to its original form. Okay. Sounds simple and logical.

How many random mutations are needed to change the pairing partner of the first initial mutation in a 1500 nucleotide RNA sequence? And how many more mutations are needed at that position to get the correct nucleotide to form the non-canonical pairing type? What is the likelihood that the random mutation needed to correct the first mutation (to recreate the canonical base pair) creates a second non-canonical base pair at another position that will also need to be corrected? And what is the likelihood that the next random mutation doesn’t correct either of these non-canonical base pairs, but instead creates yet another non-canonical base pair? And then let’s say that some non-canonical base pair types would be more deleterious and thus should be corrected ASAP! Do we increase the random mutation rate for these situations? Now let’s play out the Darwinian random mutation model for all of the invariant nucleotides and other constraints and interactions involved with this exemplar 16S rRNA. Isn’t molecular evolution fun?

This sloppy and inefficient mechanism is at odds with cellular machinery that is very refined, efficient, and even redundant to minimize errors. And since the 1970s, many new enzymatic mechanisms that alter genetic sequences have been identified and characterized. If the base pairs in the RNA structure do not evolve with a random mutation model, then what is the mechanism of this evolution? Could the answer have major ramifications for the evolution of macromolecules? Whatever the details to the answer, Orgel’s first rule is probably true—“Whenever a spontaneous process is too slow or too inefficient a protein will evolve to speed it up or make it more efficient.”

**RNA folding**

While we ultimately used Plan B, Comparative Analysis, to determine the correct secondary structures for the 16S and 23S rRNA, the challenge and need to use first principles to determine the correct RNA secondary structure has not waned. This challenge is considered by some to be one of the grand challenges in biology, and for good reason. The successful development of an algorithm, a process that determines the correct secondary and even three-dimensional structure will have a profound effect on RNA science. The most likely higher-order structures for genomic sequences can be postulated with confidence at the onset, creating a strong foundation for the subsequent experimental design, interpretation, model building, and hypothesis that follows. More accurate and meaningful interpretations of experiments will also result from the ability to understand with more confidence how mutations and other changes in the RNA sequence will influence its higher-order structure, interactions with other molecules and even function. This challenge is even grander given the daunting scale of the problem. The C/P ratio—the number of Correct helices divided by Possible helices for an RNA the size of 16S rRNA (about 1500 nucleotides)—is 1/1000, while the number of possible secondary structure models—all of the different arrangements of the possible helices—is $\sim 10^{400}$. In contrast with comparative analysis that searches for that one common structure from a set of many sequences, the RNA Folding algorithm uses knowledge (or first principles) in the form of energy values for structural elements and the means to identify and assemble these “local” elements into the “global” structure. While this knowledge was initially determined primarily from experimental methods, an increasing amount of these rules, principles, and folding pathways are derived from comparative analysis and comparative data. Thus, as stated earlier, comparative analysis not only determined the correct secondary structure for the rRNAs and other RNAs, it can be used to obtain knowledge that should be used to improve the RNA Folding programs. And while the sophistication and accuracy of the RNA Folding programs have improved, their reliability and overall accuracy are, in this opinion still woefully lacking to make an impact analogous to what the comparative analysis did for us in the 1980s. Do I believe that we can identify and refine the first principles that include energy values and folding pathways, and implement them into an RNA Folding program that nearing the accuracy of comparative methods? Yes, let the future begin.