From *SIAM News*, Volume 41, Number 9, November 2008

Sweet Structural Mysteries of Life

By Barry A. Cipra

Watson and Crick are widely celebrated for discovering "the" structure of DNA, when what they really did was to discern the details of base pairing in a rigid crystal of the stuff. The actual physical form of a nucleic acid—either DNA or its single-strand lackey, RNA—in its natural, decidedly non-crystalline state is a problem yet to be solved. In an invited presentation at the SIAM Conference on Discrete Mathematics, held June 16–19 at the University of Vermont, Anne Condon, a computer scientist at the University of British Columbia, surveyed progress made by computational biologists in elucidating the true structure of nature's masterpiece.

The caricature of DNA as complementary pairs of long, essentially linear chains of the bases A, C, G, and T (twisted, of course, into the famous double helix) is extremely useful for understanding how organisms store information about themselves and how they relate to one another in the grand evolutionary scheme of things. Indeed, the accomplishment of various genome projects amounts to spelling out this "primary" structure of DNA for organisms ranging from humans beings to the tiger blowfish. In broad outline, DNA is transcribed into RNA, which toddles off to cobble amino acids together into proteins, which do all the heavy lifting. But how is all this transcribing and cobbling accomplished? The physical chemistry, it seems, depends on the "secondary" and "tertiary" foldings of the molecules.

This has long been known to be the case for proteins, whose functions are functions of shape. The protein-folding problem is thought to amount to an enormously difficult computational search for a state of minimal free energy. An energy functional is easy enough to write down; finding its low point in a vast, high-dimensional landscape of conformations is not so easy. Much the same is true for long chains of nucleotides.

Take RNA, for example. Biologically relevant strands of RNA are made up of nucleotides, from the hundreds to the low thousands, arranged like beads on a string. As the string flops around in space, stretches made up of complementary bases (A with U, C with G, U being RNA's version of T) can join by means of energy-reducing hydrogen bonds (see Figure 1), leaving loops of unpaired bases, which typically carry out the biological job the molecule was so intelligently designed to do. As might be expected, tight loops put a strain on the structure, elevating the overall energy. Biochemists have worked out models that, assuming fixed environmental conditions, such as temperature, can be used to compute a reasonable estimate of the free energy for any secondary structure—that is, any collection of complementary substrings base-paired by hydrogen bonds. (Tackling the fully three-dimensional "tertiary" structure seems, for the moment, out of the question.)

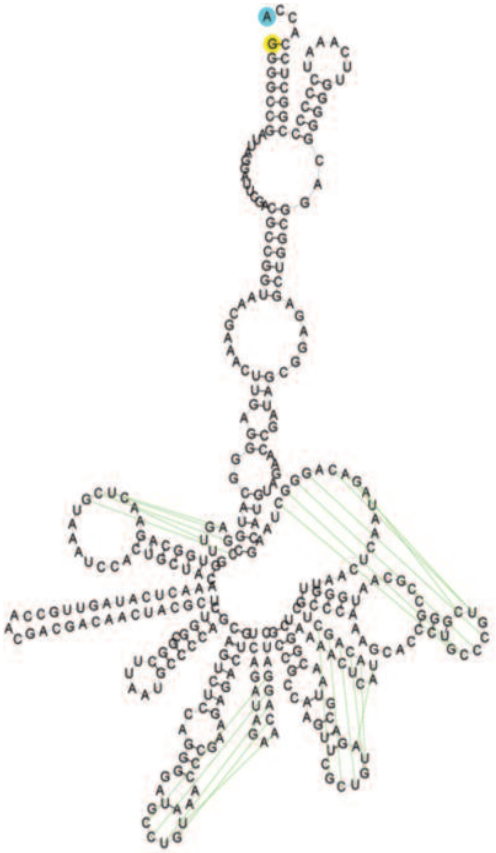

Figure 1. *The secondary structure of RNA typically has lots of hydrogen bonds. In the RNA shown here, some of the hydrogen bonds (green lines) form a computationally nasty set of "kissing hairpins." The blue and yellow A and G at the top are the ends of the strand. Image from Mirela Andronescu.*

Secondary structure is easily represented in the form of a graph: If the nucleotides are equispaced points along a line—e.g., the integers 1 to n on the number line—the hydrogen bonds beween complementary base pairs in a segment form a set of nested arcs (see Figure 2). The nesting results from the directionality of nucleic acids (in the jargon of biochemistry, a strand of DNA or RNA has a 5′ end and a 3′ end); complementary strands or segments must run in opposite directions.

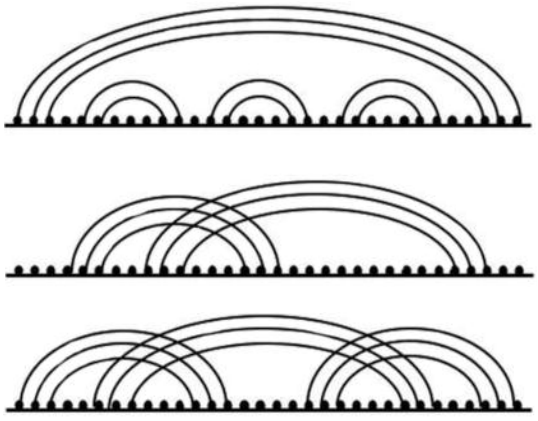

Figure 2. *RNA structures can be pseudoknot-free (top), or they can have H-type pseudoknots (middle) or kissing hairpins (bottom). Things could be worse, of course, but luckily that rarely seems to happen. Courtesy of Anne Condon.*

Many RNA structures are "pseudoknot-free," meaning that there are no intersections among the various nested sets of arcs (Figure 2, top). Others have "H-type pseudoknots" (Figure 2, middle), and still others have "kissing hairpins" (Figure 2, bottom). In principle, one can look at arbitrarily complicated secondary structures, but these three types, with the additional possibility of the nesting of one type within another, are thought to be good representatives of the known complexity found in nature. (Some biologists reserve the term "secondary structure" for the pseudoknot-free case, referring to pseudoknots as "tertiary" structures. Mathematicians don't care what labels you use, as long as you're clear and consistent.)

It may be a good thing (if it's true!) that nature doesn't cough up anything too complicated, because the general prediction problem for the loop-based free-energy minimization problem was shown by R.B. Lyngsø and C.N.S. Pedersen (1999) and T.

Akutsu (2000) to be NP-hard. Restriction to kissing hairpins brings the problem into polynomial-time feasibility, but the best known algorithm, devised by Elena Rivas and Sean Eddy of the Howard Hughes Medical Institute in 1999, has a daunting $O(n^6)$ complexity, n being the number of nucleotides. Further restriction to H-type pseudoknots is a tad better: In 1999, Yasuo Uemura and colleagues of the Doi Bio-asymmetry Project in Chiba, Japan, laid out an $O(n^5)$-time algorithm for that case. (Their algorithm handles kissing hairpins in a limited way, too.) In 2003, Robert Dirks and Niles Pierce of Caltech provided a streamlined $O(n^5)$ approach to the prediction of H-type pseudoknots.

A complexity bound doesn't automatically limit the practicality of an algorithm, as the simplex method for linear programming famously proves, but for the folding problem the estimates do currently seem to rule out pseudoknot computations with more than a few hundred nucleotides. The only structures that seem easy to compute in general are those belonging to the pseudoknot-free case, for which $O(n^3)$ algorithms have long been known. Fortunately, a great many RNA structures of interest to biologists are known—or thought—to be free of pesky pseudoknots.

Run time and generality are only two criteria for judging an algorithm, Condon notes. A third important metric is accuracy. Over the years, biologists have cultivated a small garden of RNAs for which they are reasonably confident they know the structure. (In general, these are strings that are essentially the same in large numbers of related species; if a pair of complementary substrings is conserved across species, it's a fair bet that some hydrogen bonding is present. X-ray crystallography and other emerging technologies for the study of molecular structure come in handy, too.) These strings provide a nice testbed for measuring how well a model for the free energy does at predicting structure. A convenient metric, known as the F-measure, is the harmonic mean of "precision rate" and "sensitivity." If the true structure has bonds between B base pairs, and a model predicts A such bonds but gets only a of them correct, then the precision rate is a/A, the sensitivity is a/B, and the F-measure is $2(a/A)(a/B)/((a/A) + (a/B)) = 2a/(A + B)$.

A free-energy model can bristle with hundreds of parameters, some of which have been experimentally pinned down but many of which remain free to fit the data. In 1999, Douglas Turner's group in the chemistry de-partment at the University of Rochester presented a model with a fit involving several hundred parameters. Using a testbed of 1660 known RNA structures, Con-don's student Mirela Andronescu found that the "Turner99" model weighed in with an F-measure of 60%.

Andronescu then turned to machine learning, hoping that the parameter values could be tweaked to give better results. In a recent joint paper, Andronescu, Condon, Holger H. Hoos, and Kevin P. Murphy of UBC, and David H. Mathews of the University of Rochester (who was part of the original Turner99 group) introduced an approach they call "constraint generation," which starts with Turner99 and iteratively adjusts parameters to optimize their fit on a training set of RNA data (some structural, some "thermodynamic"). Their algorithm is fast and flexible, and it ups the F-measure to 67%—still barely a passing grade, but headed in the right direction.

Two of Condon's other students, Shelly Zhao and Hosna Jabbari, wondered whether they could reduce the complexity of pseudoknotted structure computations. As described in a joint paper with Condon, they got the run time down to $O(n^3)$ by assuming that RNA folds in a hierarchical fashion: pseudoknot-free folds first, followed by pseudoknots, if any. This helpful, simplifying assumption is entirely plausible, Condon says. Indeed, the hypothesis of minimal free energy is itself a simplifying assumption in that it ignores, among other things, the pathways macromolecules take to folding. It's entirely possible that RNA depends for its raison d'être on being locked into local minima by some quirk in its creation—including, perhaps, the fact that it begins to fold while it is still being transcribed.

The computational story of nucleic acids is still taking shape. "Although knowledge of RNA secondary structure is valuable to biologists," Condon says, "it is but a small step towards the ultimate goal of predicting the three-dimensional geometry of RNA foldings."

Barry A. Cipra is a mathematician and writer based in Northfield, Minnesota.