Informatics challenges in structured RNA

Alain Laederach

Submitted: 31st March 2007; Received (in revised form): 2nd June 2007

Abstract

The world of regulatory RNAs is fast expanding into mainstream molecular biology as both a subject of intense mechanistic study and as a tool for functional characterization. The RNA world is one of complex structures that carry out catalysis, sense metabolites and synthesize proteins. The dynamic and structural nature of RNAs presents a whole new set of informatics challenges to the computational community. The ability to relate structure and dynamics to function will be key to understanding this complex world. I review several important classes of structured RNAs that present our community with a series of biologically novel informatics challenges. I also review available informatics tools that have been recently developed in the field.

Keywords: RNA; folding; informatics; riboswitch; ribosome; RNAi

Central to many of the regulatory networks in the cell, RNA’s role in regulating both transcription and translation is garnering much attention from both wet and dry labs [1–3]. The fact that RNA generally exists as a single-stranded molecule, and the additional conformational flexibility of the sugar-phosphate backbone, allows RNA to adopt stunningly complex three-dimensional structures [4]. These structures confer function to the RNA including catalysis, metabolite sensing and transcriptional control [5–7]. The unique molecular architecture of RNA, combined with the wide range in size of functional RNAs, presents our community with a series of novel opportunities in RNA informatics [8].

Much like proteins, RNAs have elements of primary, secondary and tertiary structure [9]. The linear nature of the RNA biopolymer (containing Guanine, Cytosine, Uracil and Adenine bases) and the ability of bases to form Watson–Crick (or canonical) base-pairs (G with C and A with U) makes the prediction of RNA secondary structure from primary sequence a contemporary challenge in RNA informatics. Significant advances in the field of RNA secondary structure have been made in the last 2 years (for a review see [8] as well as [10]), but are not the main focus of this review. I will rather focus on emerging informatics challenges in the area of structured RNAs, and their dynamic behavior. Specifically this review addresses challenges and opportunities related to RNA molecules for which the three-dimensional structure has been determined experimentally. Furthermore, I will review recent RNA/Protein (RNP) complexes of particular significance in the emerging field of RNA interference (RNAi) [6, 11].

FUNCTIONALLY IMPORTANT STRUCTURED RNAs

Figure 1 illustrates the difference in size of three functionally important classes of structured RNAs. Riboswitches are small (100 nucleotides, Figure 1A) single stranded RNAs that act as metabolite sensors in Bacteria [5]. They are generally found upstream of the Gene they control (in this case Thiamine synthesis) and undergo a conformational change in the presence of the metabolite (Thiamine) [12]. When the metabolite is bound to the Riboswitch, it generally inhibits translation thereby down-regulating the production of the corresponding protein product. Self-splicing introns are transcribed...
with the mRNA and fold into a catalytically active conformation (Figure 1B) that auto-splices itself out of the mRNA effectively processing the message for translation [2]. The Ribosomal RNAs (rRNAs) are molecular machines (Figure 1C illustrates the 30S rRNA subunit) that perform the translation reaction in the cell (synthesis of protein). RNAs therefore come in many shapes and sizes that each have specialized functions within the cell.

Common to all three RNAs illustrated in Figure 1 are dynamic conformational changes that enable and regulate the function of the biopolymer. Riboswitches undergo significant conformational changes upon ligand binding, while catalytic RNAs and rRNAs must fold into active structures. The dynamics of these assembly reactions are critical to the regulatory role of the molecule, as correct sequencing of the reaction directly impacts function. In RNA, the kinetics of folding and assembly are particularly important, as most RNAs that have been studied to date exhibit complex kinetic behavior during their folding reaction.

**STRUCTURAL ELEMENTS OF RNA**

As mentioned previously, RNA structures have elements of primary, secondary and tertiary structure. The primary sequence of RNA is comprised of four types of bases (C, G, A and U, Figure 2A). Figure 2B illustrates the secondary structure diagram of the P4P6 subdomain of the L-21 *Tetrahymena thermophila* group I intron [13]. This diagram differs somewhat from the manually laid out secondary structure diagrams often published with RNAs (see for example Figure 2 in [14] for a particularly aesthetic diagram). The diagram was generated automatically using the RNAMLVIEW software [15], which is now incorporated in S2S [16] (http://bioinformatics.org/S2S/). The layout attempts to best represent the three-dimensional reality (Figure 2C) of the
molecule in two dimensions. Helical segments of the molecule are drawn as sequential pairs of stacked bases. Base pairs are indicated with lines, and long-range tertiary contacts are identified by red dotted lines.

Canonical base-pairs (known also as Watson–Crick pairs) in RNA will lead to an A-form helix. RNA nucleotides, however, can form non-canonical base-pairs in which different edges of the base form complimentary hydrogen bonds. As a result, stabilizing interactions can form between any two nucleotides in the RNA. A seminal paper by Leontis and Westhof describes all 12 possible base-pairing interactions based on geometrical considerations for each base type [17]. The different symbols connecting the bases in the secondary structure diagram (Figure 2B) represent different types of base-pairing as defined by Leontis and Westhof [17, 18]. S2S and RNAMLVIEW conveniently generate XML annotations of RNA structures (including base-pairing type), greatly reducing the need for specialized code bases to analyze RNA structure [16, 18]. A complete Documentation Type Definition (DTD) for RNA structure has also been developed for the markup language RNAML [19].

The ability of RNA to form 12 different types of base-pairs significantly complicates the prediction of RNA tertiary structure from primary sequence.

**Figure 2:** Structural elements of RNA. (A) Primary sequence of RNA containing the four nucleotides G, C, A and U for the P4P6 subdomain of the L-21 *Thermoplasma* group I intron [13]. (B) Secondary structure of the P4P6 subdomain as illustrated by RNAMLView [15]. The layout of the structure is automated trying to best represent the three-dimensional reality of the molecule. Long range tertiary contacts are indicated as red dotted lines, while base-pair types are as defined by Leontis and Westhof [17]. (C) Three dimensional cartoon representation (similar scheme as Figure 1) of the P4P6 subdomain (PDB ID 1GID).
Nonetheless, remarkably accurate models of large structured RNAs have been generated and later confirmed by x-ray crystallography [20]. These structural models are built in a semi-automated way and rely on the user’s intuition and knowledge of RNA structure. Fully automated structure prediction of RNA remains a contemporary challenge in the field with only limited success thus far [21]. The large number of RNA structures that have been solved by crystallography should make knowledge-based structural modeling approaches amenable in the near future. The nucleic acid database curates all new RNA and DNA structures available [22], and SCOR maintains a structural classification [23–26]. Both of these resources offer convenient entry points into the world of structured RNA. Furthermore, global statistical analyses of RNA structure are beginning to identify emergent orientations of RNA helices which can be used to improve automated structure prediction [27].

**RNA MOTIFS AND POTENTIAL FOR HOMOLOGY MODELING**

Careful analysis of RNA structures has revealed a surprising number of recurring structural motifs across species and function [15, 17, 18, 28–31]. These are three-dimensional orientations of bases that have nearly identical base-pairing and stacking interactions and are believed to constitute building blocks for complex RNA structures [18]. Work in the Levitt lab has shown that it is possible to reconstruct entire RNA structures with a library of clustered di-nucleotides [32]. Alternatively the analysis of the backbone conformation of structured RNA also allows the identification of similar motifs [33–36]. Whether the motifs are automatically or manually identified, there is a general consensus in the field that there most likely exists a set of core RNA building blocks larger than single nucleotides from which all RNA structures can be reconstructed. The identification of these motifs remains a contemporary challenge in the field. It is likely that only consideration of both backbone and base orientations will lead to a complete description of RNA structure [37].

A classic example of an RNA structural motif is illustrated in Figure 3. It is a Sarcin (also known as Sarcin-Ricin) motif, and the three overlaid structures occur in different positions within the same structure. The Sarcin motif is clearly a structural motif as the sequences of each strand differ significantly and would not be identified by sequence comparison alone [29]. Approximately 40 such motifs have been identified to date and every year new motifs are discovered. Geometrical definitions for three-dimensional RNA motifs have yet to be finalized as these require consensus building within the community. This has prompted the creation of an RNA Ontology Consortium to develop a structured, community accepted vocabulary for describing RNA structural motifs [38]. The goals of the consortium include providing the informatics community with curated structural annotation databases to facilitate informatics research in RNA (http://roc.bgsu.edu/).

**RNA FOLDING DYNAMICS**

The ‘RNA Folding Problem’ is not limited to predicting the tertiary structure of an RNA from its...
primary sequence (Figure 2). A vast majority of biologically active RNAs have a propensity to naturally misfold, and require changes in conditions (e.g. heating) or exogenous molecules (e.g. RNA chaperones) to resolve into the active structure [7, 39, 40]. The propensity of RNAs to misfold suggests important regulatory roles related to the dynamics of the molecule and its interactions with other cellular constituents. Understanding RNA misfolding biophysically and phenomenologically is at the heart of the RNA folding problem and presents a series of novel informatics and modeling challenges.

The ability to easily synthesize RNA from DNA (using RNA polymerase) and the use of quantitative Reverse Transcriptase (RT) reactions for the analysis of RNA containing solutions has enabled the development of many RNA specific experimental modalities [41, 42]. The relative simplicity of such ‘footprinting’ experiments (compared to analogous modalities in proteins) and the development of specialized software for the analysis of the results has created a new paradigm in terms of the throughput and quality of dynamic information available for large structured RNAs [43, 44]. These new experimental modalities have in turn required the development of novel software for the analysis and modeling of RNA folding reactions [45, 46].

Hydroxyl radical footprinting experiments determine the accessible surface area of a molecule by measuring the reactivity of individual nucleotides when exposed to hydroxyl radicals. Nucleotides that are buried in the structure are less likely to react with a hydroxyl radical [41] than those on the outside [47–49]. Given the fast reaction kinetics of the hydroxyl radical with the RNA, and recent improvements in mixer technology, this technique can now resolve folding transitions with millisecond accuracy [44].

Hydroxyl radical footprinting experiments are unique in that they provide local probes of structure, as opposed to global probes (for example, the change in Radius of Gyration determined by analytical ultracentrifugation). Time-resolved hydroxyl radical footprinting experiments monitor the change in protection of each nucleotide in the RNA as a function of time. The information content of such data is sufficient to allow the determination of the underlying kinetic parameters of the system. Figure 4 illustrates the basic premise of the KinFold algorithm, which was developed to automatically model the folding reactions of large RNAs based on time-resolved hydroxyl radical footprinting experiments [45, 46]. If footprinting data were collected on the folding reaction as illustrated in Figure 4A, the time-progress curves would look like those illustrated in Figure 4B. Given that the three nucleotides indicated are all on the subdomain of the molecule that folds first, protection would initially be seen for these three nucleotides. Once the intermediate has accumulated, the other nucleotides would report protection as the concentration of folded molecule increases.

The KinFold algorithm analyzes this type of data in a two-step process. Initially, the time-progress curves are clustered using k-means clustering to establish the subdomains of the molecule that form in a concerted manner. In this case the clustering would identify the fast folding subdomain of the molecule by clustering the red, green and blue nucleotides together as illustrated in the right pane of Figure 4B. Once the time-progress curves are clustered, the data is kinetically modeled as illustrated in Figure 4C. In this case two alternative models need to be tested, as the intermediate can either be associated with the two clusters. It is important to note that a-priori the kinetic model shown in Figure 4A is not known, but that it can be automatically deduced from the analysis of the kinetic folding data illustrated in Figure 4B. This type of approach is particularly useful when studying the effects of mutation and solution conditions on the folding kinetics of RNA [45].

At the heart of the RNA folding problem is the relationship between the three-dimensional structure of the molecule and the folding dynamics. The timescales of RNA folding reactions (often on the order of seconds to minutes) makes the problem of simulating these large molecules computationally challenging at the atomic scale [50, 51]. Nonetheless, several groups have studied the structure/dynamic relationships using simplified representations of RNA (such as contact order) and come up with good correlations to experimentally observed folding rates [40, 52].

NUCLEIC ACID RECOGNITION IN RNA INTERFERENCE

RNAi is fundamentally changing the world of molecular biology as it is both a technology that enables specific regulation of gene expression,
and a mechanism by which cells regulate their genes [6, 11]. RNAi refers to the general mechanism by which double stranded RNAs are processed into single stranded RNAs with complimentary sequences to mRNAs. These single-stranded RNAs bind to the mRNA and down-regulate the associated gene. MicroRNAs (miRNAs) are naturally occurring hairpin structures that the cell uses for genetic regulation. Small interfering RNAs (siRNA) are double-stranded synthetic molecules that are used in molecular biology to target specific genes for down-regulation. Both miRNAs and siRNAs are processed by a series of proteins for which the three-dimensional structures have very recently been solved (a comprehensive review of RNA-induced silencing complex (RISC) structures can be found in [53]).

RNA-protein interactions are critical to the regulation of RNAi. The importance of these regulatory networks is such that even viruses have evolved proteins to interact with siRNAs. The viral proteins inhibit the silencing activity of siRNAs thereby disabling the defense mechanisms in the cell. One such example is the p19 protein virus from tombusvirus (Figure 5A) which is able to recognize double stranded RNAs of length 22 by projecting two alpha-helical ‘reading heads’ that contain terminal tryptophans that bind to the two ends of the siRNA [54].

The Argonaute class of proteins (Figure 5B) are part of RISC and provide catalytic and structural function to the complex [55–57]. The atomic reality of the recognition of the double-stranded RNA (Figure 5B, inset) by the protein illustrates the
challenges associated with predicting and understanding RNAi function. The interface between the protein and siRNA involves at least 7 nucleotides and ~30 amino acids. The fact that the protein recognizes both strands makes it unlikely that sequence analysis alone will be able to predict the specificity of the interaction. Furthermore, several domains in the protein are involved in binding to the RNA and this suggests that sequence motif type analyses alone will not be successful in predicting these types of interactions.

The most significant informatics contribution to the world of RNAi to date has been the development of algorithms to detect miRNAs in genomic sequences (see for a complete review [58]). miRNAs are generally expressed as $2 \times 22$ nucleotide long hairpins, and this sequence pattern can readily be identified in non-coding genomic sequences [55]. A more interesting informatics challenge, however, is the identification and characterization of miRNA promoters, which allows a systematic analysis of the interaction networks within RNAi controlled genes [59]. Nonetheless, only a physics-based understanding of the molecular details of the RISC complex in the context of the cellular milieu will enable a comprehensive prediction of RNAi function.

**THE FUTURE OF RNA INFORMATICS**

The linear nature of DNA has made it very amenable to sequence-based informatics analyses. Furthermore, genomic technologies have produced an overwhelming amount of genetic information ideal...
for informatics analyses. The Rfam database (RNA families) provides a comprehensive repository of structured RNAs [60, 61]. These RNAs are identified from genomic sequences based on their sequence similarity to known structured RNAs. The 500,000 structured RNAs in Rfam demonstrate the ubiquitous nature of these fascinating molecules and their biological importance.

In 1994, Stanley and co-workers identified patterns similar to language in what was then termed ‘junk DNA’ [62] using statistical and linguistics approaches [63–65]. Incredibly, in little over a decade, we now know that much of the ‘junk DNA’ is in fact transcribed into RNA, and that these molecules play critical regulatory roles. The physical reality of these RNAs is necessarily three-dimensional, and only a geometric understanding of their shape and dynamics will yield insight into their function. The future of RNA informatics will require integration of this four-dimensional reality (time and structure). The systematic and integrative approach to biology will require interdisciplinary collaborations across multiple fields of computation including the engineering and modeling communities.

Key Points
- RNA is both a messenger of genetic information and a molecule capable of carrying out reactions in the cell.
- The function of RNA is intimately related to its structure and dynamic behavior.
- RNA molecules interact with proteins in the cell to regulate gene expression.
- Structured RNAs are comprised of motifs that are found in many structures.

Funding
Funded by the NIH through NIGMS K99-GM079953, P01-GM66275 and an NIH Roadmap grant U54 GM072970 for the National Centers for Biomedical Computation and the NSF 0443508 for the RNA Ontology Consortium.

Acknowledgements
I wish to thank Jesse Stombaugh for providing the atomic coordinates for Figure 3. AL is a Damon Runyon Cancer Foundation Research Fellow.

References


