On the importance of cotranscriptional RNA structure formation

DANIEL LAI,1 JEFF R. PROCTOR,1 and IRMTRAUD M. MEYER2
Centre for High-Throughput Biology, Department of Computer Science and Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4

ABSTRACT
The expression of genes, both coding and noncoding, can be significantly influenced by RNA structural features of their corresponding transcripts. There is by now mounting experimental and some theoretical evidence that structure formation in vivo starts during transcription and that this cotranscriptional folding determines the functional RNA structural features that are being formed. Several decades of research in bioinformatics have resulted in a wide range of computational methods for predicting RNA secondary structures. Almost all state-of-the-art methods in terms of prediction accuracy, however, completely ignore the process of structure formation and focus exclusively on the final RNA structure. This review hopes to bridge this gap. We summarize the existing evidence for cotranscriptional folding and then review the different, currently used strategies for RNA secondary-structure prediction. Finally, we propose a range of ideas on how state-of-the-art methods could be potentially improved by explicitly capturing the process of cotranscriptional structure formation.

Keywords: RNA secondary-structure prediction; RNA structure formation in vivo; cotranscriptional RNA folding

INTRODUCTION
The primary products of all DNA genomes are RNA transcripts consisting of linear sequences of four different types of ribonucleic acids (abbreviated A, C, G, and U and chemically different from the similarly abbreviated DNA building blocks, A, C, G, and T). When a gene of the genome is activated, a corresponding transcript is synthesized in a linear fashion with its 5′ end emerging first and its 3′ end emerging last. Primary transcripts vary greatly in length from a few nucleotides (nt) to $10^4$ nt and longer. They may be processed in a number of ways, e.g., splicing and RNA editing, which may happen while the transcript is being made. The functional role of some transcripts is exerted by RNA structure that is formed when pairs of complementary nucleotides of the RNA sequence (C-G, A-U, G-U) form base pairs. In contrast to proteins, where we typically need to know its three-dimensional (3D) structure in order to study a protein’s potential functional roles, it often suffices to only know the RNA secondary structure in order to investigate its potential functional role(s). This RNA secondary structure is defined by the pairs of base-paired sequence positions in the RNA. RNA structure can either be global, i.e., span most of the transcript, or more local, i.e., be confined to a subsequence of the transcript. During its life in the cell, a single transcript may assume more than one functionally relevant RNA structure, e.g., riboswitches, which can assume two mutually exclusive structures that are both functional.

Many computational methods for RNA structure prediction, in particular, earlier and noncomparative methods, implicitly focus on predicting global RNA structures only. They are typically applied to analyze the noncoding portion of a given transcriptome because this is where globally structured RNA genes are suspected. RNA structural features, however, are also known to play important functional roles in regulating protein-coding transcripts (e.g., splicing, localization, degradation, translation initiation), yet this typically involves only local RNA structures, which only some of the computational methods for RNA secondary-structure prediction can adequately model (Pedersen et al. 2004a,b).

Recent advances in nucleotide sequencing technologies have enabled the routine sequencing of entire transcriptomes, with methods such as strand-specific RNA-seq, enabling the discovery of novel transcripts en masse. Experimental methods for RNA structure determination such as X-ray crystallography and NMR can provide atomic-resolution 3D solutions, but remain relatively costly and comparatively slow. Computational methods for predicting RNA secondary
structures based on RNA sequence information alone are therefore key to assigning potential functional roles to the transcriptome and identifying worthwhile targets for experimental validation.

When available, computational structure prediction can be aided by results from RNA footprinting experiments. Such experiments can estimate the pairing status of individual nucleotide positions in a single sequence with chemical probes, but cannot identify the pairing partner involved in a base pair. Such methods, when paired with next-generation sequencing technologies, in protocols such as Frag-seq, PARS, and SHAPE-seq, show great promise in generating high-throughput RNA secondary structure probe maps (Wan et al. 2011). Nonetheless, footprinting results still require algorithms to derive the overall most likely solution, again emphasizing the need for reliable and efficient computational methods.

There exists by now ample experimental evidence that RNA structure formation starts cotranscriptionally, i.e., while the RNA is transcribed from the genome. The process of cotranscriptional structure formation is key to determining the resulting functional RNA structure(s) in vivo and that this process can be influenced by a range of intrinsic as well as extrinsic factors. Yet, nearly all state-of-the-art methods for computational RNA secondary structure prediction ignore the structure formation process and focus exclusively on the end result, i.e., a single, final RNA structure. There already exist a few computational methods that aim to explicitly simulate the cotranscriptional folding pathway by capturing key features of the folding environment in vivo. Because their prediction accuracy has so far been evaluated on only a few select sequences of typically short length, however, they are currently viewed as folding-pathway prediction methods rather than RNA secondary-structure prediction methods.

We argue that ignoring the formation process often yields decent structure predictions, especially for short and globally structured transcripts (<200 nt), but that in order to increase the prediction accuracy for longer transcripts and to reach a conceptually better understanding, we ought to aim to take some effects of cotranscriptional folding into account.

In the following, we first review the variety of mechanisms that have been shown to influence cotranscriptional folding in vivo. This summarizes primarily experimental, but also some theoretical evidence for cotranscriptional folding. We then provide an overview of the currently existing methods for RNA secondary-structure prediction. This part of the review is not aimed at providing a detailed description of every existing method for RNA secondary-structure prediction, but rather at highlighting the different underlying concepts used by these methods. At this point, we also cover methods for predicting RNA folding pathways that already capture some effects of cotranscriptional folding. To conclude, we propose a range of ideas on how cotranscriptional folding could be captured in computational methods for RNA secondary-structure prediction in order to further improve their prediction accuracy.

**EXPERIMENTAL AND THEORETICAL EVIDENCE FOR COTRANSCRIPTIONAL FOLDING**

**Directionality of transcription**

One of the most obvious differences between the in vivo and the typical in vitro setting is that RNA transcripts in vivo emerge sequentially starting with the 5′ end, whereas in vitro experiments start with an already synthesized molecule. The directionality of the molecule’s synthesis in vivo may thus lead to structural asymmetries during its cotranscriptional folding that may, in turn, influence the resulting functional RNA structure(s).

**Transcription, transcription speed, and variations thereof**

Whether or not folding can happen during synthesis depends, among other things, on how the timescale of RNA synthesis compares with that of RNA structure formation. The speed of transcription not only depends on the underlying organism, but also on the polymerase responsible for generating the transcript in question. It ranges from 200 nucleotides per second (nt/sec) in phages, to 20–80 nt/sec in bacteria and 10–20 nt/sec for human polymerase II (Pan and Sosnick 2006). On the other hand, RNA folding is known to occur on a wide range of time scales; some RNAs fold in 10–100 msec (Al-Hashimi and Walter 2008), whereas kinetically trapped conformations can persist for minutes or hours (Sosnick and Pan 2003; Thirumalai and Hyeon 2005; Al-Hashimi and Walter 2008). Experiments in the early 1980s have shown that RNA structure formation can happen during transcription (Boyle et al. 1980; Kramer and Mills 1981), i.e., cotranscriptionally, and that folding in vivo can happen on the same timescale as RNA synthesis (Brehm and Cech 1983). The latter was first shown for the cotranscriptional and structure-dependent self-splicing of the *Tetrahymena* group I intron (Brehm and Cech 1983).

Since then, several in vitro experiments have confirmed that RNA folding can happen cotranscriptionally and that the speed of transcription not only affects the overall folding rate, but also transient structures as well as the final structure (Pan et al. 1999; Helimann-Miller and Woodson 2003a,b). Lewicki et al. (1993) and Chao et al. (1995) showed that altering the natural speed of transcription can yield misfolded and functionally inactive transcripts. Experimental studies of the *Tetrahymena* self-splicing intron are consistent with the view that a set of identical RNA molecules partitions into an active and an inactive pool, and that this partitioning is highly influenced by the cotranscriptional folding environment, including the RNA transcription rate (Koduvayur and Woodson 2004).

For a given transcript, the speed of transcription is not necessarily constant. Transcriptional pausing can serve as an additional mechanism for fine-tuning cotranscriptional folding (Toulme et al. 2005; Wickiser et al. 2005; Wong et al. 2007). This pausing happens at specific transcript positions and for...
well-defined time intervals (ranging from $10^{-6}$ sec to 10 sec). In bacteria, pausing can be due to interactions between the emerging RNA and the polymerase and/or polymerase-associated protein factors (Liu et al. 1996; Landick 1997; Mooney et al. 1998). The flavin mononucleotide (FMN)-dependent riboswitch in *Bacillus subtilis* (Wickiser et al. 2005) is a beautiful example of how these features can be combined into a cotranscriptional feedback loop in which the binding of a metabolite selects one of two possible cotranscriptional folding pathways whose resulting RNA structure determines whether transcription is terminated or not.

**Self-interactions including transient RNA structures**

One of the key features of any RNA sequence is that it can interact with itself via base pairs between complementary nucleotides to form RNA structure. During cotranscriptional folding, already formed structures can unpair and yield to other structures, in which case, we refer to them as “transient structures.” In other cases, it is energetically unfavorable for an existing structure to yield to a new conformation, thereby forming a kinetic trap. Transient structural features thus have the potential to significantly influence the cotranscriptional folding pathway and the resulting functional RNA structure(s) (see Fig. 2, below). Most of our current knowledge of transient structures, which we also refer to as *cis* RNA–RNA interactions, stems from dedicated experimental studies of select folding pathways that explore how RNA structure changes as a function of time.

Folding pathways of RNA transcripts in vitro have been the subject of intense study for a long time. Initial experiments primarily studied how already synthesized and fully denatured RNA molecules fold, whereas more recent studies examine cotranscriptional folding pathways in vitro and, most recently, also in vivo (Adilakashmi et al. 2009; Woodson 2010). Because any of these experiments are technically sophisticated, our current view derives from a few well-studied test cases such as the hairpin ribozyme (Donahue et al. 2000; Fedor 2002, 2009; Mahen et al. 2005, 2010) and the *Tetrahymena* intron (Koduvayur and Woodson 2004; Jackson et al. 2006). These ribozymes are comparatively easy to study in vivo because their cleavage relies on distinct structural features whose products are easier to detect than the corresponding functional structures.

Cotranscriptional folding—whether in vitro or in vivo—tends to happen sequentially (Mahen et al. 2005, 2010) because base pairs at the 5′ end of the RNA can form first, whereas base pairs involving the 3′ end can only form once transcription is complete. This folding often involves transient RNA structure elements, i.e., structural features that are only present for a specific time span (Kramer and Mills 1981; Repsilber et al. 1999). These can direct the structure formation via one or several folding pathways toward the desired structural configuration(s). These transient features may also play distinct functional roles. They may, for example, be required for template activity during (+)-strand synthesis in some viruses (Repsilber et al. 1999) or may serve as protein-binding sites during transcription (Ro-Choi and Choi 2003). These examples once again illustrate that any given RNA transcript may have more than a single functionally relevant RNA structure during its lifetime in the cell.

Cotranscriptional folding and other reaction rates in vivo typically differ from those in vitro with folding rates in vivo being typically (Mahen et al. 2005, 2010), but not always (Donahue et al. 2000), higher than in vitro. One example is the cotranscriptional folding of the *Tetrahymena* ribozyme in vitro, which is twice as fast as the refolding of the fully synthesized and denatured molecule, but slower than the cotranscriptional folding in vivo (Heilmann-Miller and Woodson 2003a). Cotranscriptional folding pathways in vivo need not be unique (Jackson et al. 2006), and tertiary interactions can determine which of several possible folding pathways is chosen (Chauhan and Woodson 2008). Factors such as transcription speed and flanking sequences can also influence which pathway dominates (Koduvayur and Woodson 2004).

One of the few existing in vivo studies of cotranscriptional folding pathways (Sclavi et al. 1998) indirectly examined the structural folding intermediates of the *Tetrahymena* ribozyme at $10^{-5}$ sec time resolution using X-ray synchrotron radiation and chemical accessibility probing and found folding intermediates that are similar to those in vitro.

The tryptophan (*trp*) operon is a group of genes found in bacteria that act in the biosynthesis pathway of the amino acid tryptophan. The *trp* operon leader encodes a short peptide that is rich in tryptophan codons near the 5′ end of the RNA (Yanofsky 1981). Regulation of the *trp* operon is carried out in part by the *trp* operon leader through a mechanism that relies on the simultaneous transcription of a DNA gene and translation of the resulting RNA in bacteria. The *trp* operon leader assumes two mutually exclusive structural configurations that form cotranscriptionally: the attenuator, which prevents further transcription of the *trp* operon; and the anti-terminator, which permits transcription (Yanofsky 1981). When tryptophan levels are high, the ribosome proceeds rapidly through the operon leader and interferes with the anti-terminator hairpin. When tryptophan levels are low, the ribosome stalls while translating the leader peptide and allows the anti-terminator hairpin to form, and thus the *trp* operon is activated.

In addition to these experimental results, the bioinformatics community has conducted a range of computational studies to investigate cotranscriptional structure formation. Computational simulations of cotranscriptional folding pathways, e.g., Isambert and Siggia (2000), show that the basic features of cotranscriptional folding and their beneficial effects on RNA structure formation can be investigated in silico. Using a kinetic Monte Carlo Markov Chain (MCMC) to study the folding of the hepatitis delta virus ribozyme (87 nt in length), Isambert and Siggia (2000) show that cotranscriptional folding at the natural transcript speed of 50 nt/sec
is significantly more efficient than when starting with a fully denatured sequence or when using the increased transcript speed of 1000 nt/sec that is typically used in in vitro experiments. By combining computational simulations of RNA folding pathways with phylogenetic structure analyses, Schoemaker and Gultyaev (2006) investigated the effect of sRNA binding on ribosomal RNA (rRNA) structure formation during cotranscriptional folding and find that it significantly facilitates structure formation.

A bioinformatics analysis of 361 structural RNA genes (Meyer and Miklós 2004) showed that these genes not only encode information on their known functional structure, but also on transient features of their respective cotranscriptional RNA folding pathways. For this, Meyer and Miklós (2004) examined helices (defined as contiguous stretches of adjacent base pairs) that could potentially out-compete helices of the known structure. They found statistically significant 5′-to-3′ asymmetries between these competing helices and the respective helices of the known structure. More specifically, they identified two different types of transient structures: those that can yield to the functional structure and help its cotranscriptional formation and those that are more likely to act as kinetic traps during cotranscriptional folding. They showed that the former are preferentially encoded in the underlying RNA sequences, whereas the latter are suppressed.

More recently, Zhu et al. (2013) conducted a computational study of six RNA families with known transient and alternative structures in order to test whether evolutionarily related sequences not only assume similar final structures, but also share common transient structures during their respective cotranscriptional folding pathways. They find that some transient structures have been evolutionarily conserved on a level that is similar to those of the final structure. Moreover, they find that evolutionarily related sequences encounter similar transient structure features during their respective, predicted cotranscriptional folding pathways and that these features often coincide with known transient features.

To conclude, naturally occurring transcripts not only encode their functional RNA structure, but also information on how to get there via transient features that help define the corresponding cotranscriptional folding pathway.

Interactions with other molecules

One key difference between the in vivo and in vitro settings is that the cellular environment typically contains a wealth of additional molecules. In vivo, these may interact with the RNA transcript and thereby influence its structure formation and the resulting RNA structure (see Fig. 2C, below). These molecules may comprise of proteins, RNA transcripts, metabolites, ligands, and different types of ions. Any intermolecular interaction between two distinct RNA molecules, i.e., any trans RNA–RNA interaction, has the potential to prevent the thus bound RNA nucleotides from engaging in other interactions including RNA structure (i.e., cis RNA–RNA interactions). This may either stabilize or destabilize existing RNA structure features, which may, in turn, influence the cotranscriptional folding pathway and the resulting RNA structures.

Due to the methodological challenges of investigating RNA folding in vivo and in real time, we currently have only limited insight into folding pathways in vivo (Sclavi et al. 1998; Heilmann-Miller and Woodson 2003a; Jackson et al. 2006; Chauhan and Woodson 2008). Numerous recent in vitro experiments that replicate specific aspects of the complex in vivo environment and rapid progress regarding in vivo methodologies (Adilakshmi et al. 2009; Alexander et al. 2011) are likely to change this.

So, which interactions between RNA transcripts and other molecules have been experimentally confirmed to be functionally important for RNA structure formation?

Ligand–RNA interactions

One of the most-obvious examples in which RNA structure formation is influenced by trans interactions is so-called riboswitches. The change of one distinct RNA structure to another is usually triggered by the binding of a metabolite or ion, but can also be induced by a temperature change, at least in bacteria (thermoswitches) (Johansson et al. 2002; Giuliodori et al. 2010; Narberhaus 2010). The two distinct structural conformations of a riboswitch are typically located in the 5′ UTRs of messenger RNAs (mRNAs) and are mutually exclusive because they engage two overlapping subsequences. The structural change triggers a change of the gene’s expression by altering either its transcription, translation, or splicing (Serganov 2009; Roth and Breaker 2010). Nechooshtan et al. (2009) identified a pH-responsive riboregulator upstream of the alx open reading frame (ORF). For a high pH, the translationally active RNA structure is formed during transcription, which involves two well-defined transcriptional pausing sites. Frieda and Bock (2012) succeeded in directly observing the cotranscriptional folding of the puBuE adenine riboswitch. Using an optical assay that allowed them to monitor folding transitions in individual transcripts in real time, they showed that the transcriptional outcome of the riboswitch is kinetically controlled. Perdrizet et al. (2012) present strong evidence that the btuB riboswitch in Escherichia coli depends on the precise transcriptional pausing of its polymerase to guide its folding into its native structure (Hopkins et al. 2011).

Protein–RNA interactions

In order for many large RNAs to fold in vitro into their functional structure without any other trans-acting molecules (apart from water), it is necessary to raise the concentration of metal ions (e.g., of Mg2+) significantly above normal levels in vivo (Gregan et al. 2001; Fedorova et al. 2002). Several in vitro experiments have shown that the ion concentration can be lowered if specific proteins are added that stabilize the RNA structure (Gampel and Cech 1991; Caprara et al. 1996; Matsuura et al. 1997; Weeks 1997; Osterseer et al.
and that there exist dedicated cellular mechanisms for dealing with misfolded RNA structures, e.g., by sequestering and degrading them as shown for the *Tetrahymena* intron (Jackson et al. 2006). Most RNA chaperones identified so far are proteins that resolve misfolded RNA structures by binding stretches of double-stranded RNA with low affinity and in a sequence-unspecific way. Other RNA chaperones bind single-stranded RNA and facilitate the transition from the incorrect to the correct structural conformation by lowering specific kinetic barriers (Herschlag 1995).

Chaperone-assisted folding has been extensively studied for proteins, whereas comparatively little is known about the extent and mechanisms underlying chaperone-assisted RNA folding. What we know is that most of these proteins play a wide range of other functional roles in addition to being RNA chaperones and that they share no obvious similarities in terms of sequence and structure motifs (Woodson 2010). Additionally, unlike protein chaperones, RNA chaperones typically do not require any ATP to encourage refolding (Herschlag 1995; Weeks 1997; Schroeder et al. 2004; Rajkowitsch et al. 2007).

**Trans RNA–RNA interactions, i.e., interactions with other transcripts**

*Trans* RNA–RNA interactions, i.e., interactions with other transcripts, involve the same elementary building blocks as RNA structure or *cis* RNA–RNA interactions, namely, base pairs between pairs of complementary nucleotides. This implies that *trans* RNA–RNA interactions involve two single-stranded stretches of RNAs. They differ in that regard from protein–RNA interactions, which may involve single-stranded or double-stranded RNA (and may happen in a sequence-specific or unspecific way).

If a single-stranded stretch of RNA sequence is to be bound in a sequence-specific way, it should be much easier in terms of evolution to come up with a corresponding, near-complementary RNA sequence than to devise an RNA-binding protein that would bind in an equally sequence-specific way. One would therefore expect *trans* RNA–RNA interactions to be much more abundant than sequence-specific protein–RNA interactions with single-stranded RNAs (Smit et al. 2007; Meyer 2008).

Functionally important *trans* RNA–RNA interactions include the well-known class of microRNA–mRNA interactions, which alter gene expression on the mRNA level (Lagos-Quintana et al. 2001), interactions between snoRNAs and ribosomal RNAs, which edit rRNAs before ribosome assembly (Bachellerie et al. 2002); and snRNA–mRNA interactions, which are key during mRNA splicing (Horowitz 2012). Both mRNA splicing and ribosome assembly can occur cotranscriptionally.

Large-scale transcriptome studies of higher organisms such as mouse and human show that a large fraction of the transcriptome does not encode any proteins, e.g., Carninci...
(2010). These noncoding transcripts are diverse with regard to length, expression patterns and levels, and functional roles, if known. This has given rise to a wealth of different names for these transcripts, which we shall simply call noncoding RNAs (ncRNAs) in the following.

One well-studied example is the short DsrA ncRNA in E. coli, which alters the structure of the rpoS mRNA upon binding, thereby enabling its translation. In order for this trans RNA–RNA interaction to happen, the structure of the ncRNA DsrA first needs to be destabilized by binding the Sm-like protein Hfq (Mikulecky et al. 2004; Soper and Woodson 2008; Soper et al. 2010; Hopkins et al. 2011). Several other examples of structure-mediated translation regulation via trans RNA–RNA interactions between a short ncRNA and an mRNA have been found, primarily in bacteria (Geissmann et al. 2010; Lioliou et al. 2010). The short ncRNA is often an anti-sense transcript of the corresponding mRNA, the trans RNA–RNA interaction typically involves a short stretch of near-complementarity, and a protein is often required as a third ingredient for the regulatory mechanism to be functional. Yet another example of a functionally relevant trans RNA–RNA interaction is the formation of the 30S ribosomal subunit in bacteria, which requires the transient interaction with the leader sequence of the rRNA operons (Balzer and Wagner 1998).

Another well-studied example is the hok/sok toxin–antitoxin system in E. coli, which provides a mechanism for preservation of the R1 plasmid after cell division (see Fig. 1; Steif and Meyer 2012). This system consists of three overlapping genes. The host-killing hok gene induces cell death upon translation of its protein. The mok (modulation of killing) gene overlaps hok on the same mRNA transcript, and translation of the mok reading frame must occur in order for translation of hok to occur. The sok (suppression of killing) gene encodes a short anti-sense RNA that binds and prevents translation of mok and thus, indirectly, also the translation of hok. In cells that possess the R1 plasmid, the unstable sok RNA is produced in high quantities and prevents cell death caused by the longer-lived hok RNAs. Following mitosis, the sok RNA is rapidly degraded in any daughter cells that lack the R1 plasmid, allowing the hok gene to induce cell death. The mechanism of the hok/sok system depends on several structural features of the hok mRNA. Alternative structural configurations reduce the degradation rate of the hok mRNA, and several transient hairpins at the 5′ end prevent binding of sok RNA during transcription (Steif and Meyer 2012).

Summary

The overall view that emerges is that the cotranscriptional folding pathways are determined both by intrinsic features encoded in the RNA sequence itself such as transient and final structural features, and by extrinsic features such as the speed of the transcribing polymerase, and trans-interaction partners (e.g., proteins, ligands, RNA transcripts, and other trans-interaction partners). In vivo, both types of features are combined in the appropriate cellular context and determine the functional RNA structure(s) being formed.

A range of experimental evidence supports the notion of fairly well-defined cofolding pathways in vivo. These pathways are, on the one hand, robust enough to guide the formation of the correct functional RNA structure under typical cellular conditions, but, on the other hand, are—if required—flexible enough to yield different structural and functional outcomes, if the cellular environment significantly changes (Wickiser et al. 2005).

CAPTURING COTRANSCRIPTIONAL FOLDING IN METHODS FOR RNA SECONDARY-STRUCTURE PREDICTION

Existing methods for RNA secondary-structure prediction

A wide variety of computational methods already exist for predicting RNA structural features. Most RNA structure prediction...
methods that can technically handle long, naturally occurring transcripts such as rRNAs only aim to capture the RNA secondary structure rather than its tertiary structure. Fortunately, many functional features can already be studied on this level of abstraction. In the following, we therefore focus on methods for RNA secondary-structure prediction (rather than also covering methods for predicting tertiary RNA structure, which are currently limited to sequences of ~100 nt length).

Existing methods for predicting RNA secondary structure can be broadly grouped into two categories: those that take a single RNA sequence as input and those that work in a comparative way by taking a set of homologous RNA sequences as input. There also exists a different class of prediction methods that explicitly predict cotranscriptional folding pathways in terms of RNA secondary-structure changes over time. They aim to capture the structure formation process in vivo and are typically limited to analyzing transcripts of a few hundred nucleotides in length. These methods are currently viewed as folding-pathway prediction methods rather than RNA secondary-structure prediction methods.

Comparative methods for RNA secondary-structure prediction currently provide the state-of-art in terms of prediction accuracy, in particular, for long RNA sequences. Apart from one recently introduced new method, CoFold (Proctor and Meyer 2013), none of the currently existing noncomparative or comparative methods for predicting RNA secondary structures, however, explicitly capture cotranscriptional folding or its overall effects.

In the following, we review the existing methods and propose ways of capturing some effects of cotranscriptional folding explicitly in order to further improve their prediction accuracy.

**Noncomparative, MFE methods for RNA secondary-structure prediction**

Historically, noncomparative methods that take a single RNA sequence as input came first. These use the so-called minimum-free energy (MFE) approach, which aims to identify the (usually pseudoknot-free) RNA secondary structure that minimizes the overall free Gibbs energy of the transcript. They include well-known methods such as MFold, RNAfold, and related programs (Zuker and Stiegler 1981; Hofacker et al. 1994; Mathews et al. 1999; Zuker 2003). These methods mirror the in vitro setting, where a fully synthesized RNA has infinite time to settle into its thermodynamically most favorable configuration. They implicitly assume that the functionally relevant secondary structure is the thermodynamically most stable one. Predictions are generated by efficiently searching the search space of all possible (usually, pseudoknot-free) RNA secondary structure for the structure with the lowest overall MFE. This is typically done using a dynamic programming algorithm.

Several methods based on the suboptimal folding algorithm introduced by Wuchty et al. (1999) have been developed that explicitly consider an ensemble of RNA secondary structures close to the minimum free energy. RNAsubopt, a program included in the ViennaRNA package (Hofacker et al. 1994; Hofacker 2003), provides a list of low-energy secondary structures above a user-defined energy cutoff above the minimum free energy. Sfold (Ding and Lawrence 2003; Ding et al. 2004; Chan et al. 2005) uses a statistical approach to sample RNA secondary structures from the ensemble of RNA secondary structures at thermodynamic equilibrium, where the probability that the algorithm picks a particular structure is proportional to the structure’s probability in the structural ensemble. While these methods consider structures that differ from the MFE configuration, they still assume that the RNAs are in thermodynamic equilibrium. Moreover, they ignore the kinetic nature of cotranscriptional formation and the effect it may have on the resulting structure or ensemble of structures.

In 1996, Morgan and Higgs (1996) investigated a set of long RNAs (comprising 16S rRNAs, 23S rRNAs, and RNase P) and found significant discrepancies between the evolutionarily conserved RNA structure features and the respective predicted MFE structures. They concluded that these differences “cannot simply be put down to errors in the free energy parameters used in the model” (Morgan and Higgs 1996) and hypothesized that these may be due to effects of kinetic folding in vivo.

To test this hypothesis, Proctor and Meyer (2013) recently introduced the new RNA secondary-structure prediction method called CoFold, which is the first to combine thermodynamic with kinetic considerations. They incorporate one overall effect of kinetic folding into a minimum free-energy prediction method: the reachability of potential pairing partners during cotranscriptional folding. CoFold demonstrates a significant performance improvement over minimum free-energy methods alone, particularly for longer RNA sequences of >1000 nt for which one usually observes a marked decrease in prediction accuracy. Capturing this overall effect of cotranscriptional folding yields RNA secondary structures with similar, but slightly higher free energies compared with the MFE structure. These results promise that there may be great value in accounting for other effects of cotranscriptional folding to improve noncomparative methods for RNA secondary-structure prediction.

**Comparative methods for RNA secondary-structure prediction**

Rapidly increasing amounts of genome sequencing data for a variety of organisms have given rise to a conceptually new approach to RNA secondary-structure prediction that takes as input a set of homologous RNA sequences rather than a single RNA sequence of interest (e.g., Knudsen and Hein 1999, 2003; Hofacker et al. 2002; Mathews and Turner 2002; Perriquet et al. 2003; Ji et al. 2004; Pedersen et al. 2004a,b; Ruan et al. 2004; Touzet and Perriquet 2004; Witwer et al. 2004; Ruan et al. 2004; Touzet and Perriquet 2004; Witwer et al. 2004).
formation as a function of time. All of these methods aimed to explicitly simulate cotranscriptional structure prediction. In parallel to the development of the RNA secondary-structure prediction currently provide the state-of-art in terms of prediction accuracy. They tend to significantly outperform noncomparative methods (Gardner and Gigerich 2004), but typically require a high-quality input alignment provided by the user to reach their optimal performance (e.g., Perriquet et al. 2003; Ji et al. 2004; Touzet and Perriquet 2004; Holmes 2005; Meyer and Miklós 2007 for methods that do not require a fixed input alignment).

Overall, comparative methods for RNA secondary-structure prediction currently provide the state-of-art in terms of prediction accuracy. They tend to significantly outperform noncomparative methods (Gardner and Gigerich 2004), but typically require a high-quality input alignment provided by the user to reach their optimal performance (e.g., Perriquet et al. 2003; Ji et al. 2004; Touzet and Perriquet 2004; Holmes 2005; Meyer and Miklós 2007 for methods that do not require a fixed input alignment).

All of these methods generate predictions by first identifying pairs of covarying alignment columns to detect conserved base pairs and then combining these into a single (and, usually, global) consensus RNA secondary structure. For this, they use (1) a modified MFE framework that also accounts for conservation of base pairs and aims for overall energy minimization; (2) a probabilistic framework such as stochastic context-free grammars (SCFGs) combined with likelihood maximization; (3) a nondeterministic, yet probabilistic approach such as Bayesian Markov Chain Monte Carlos (MCMCs) that samples from a posterior distribution that is subsequently combined with a post-processing step to extract a consensus structure; or (4) a combination of heuristic, ad hoc procedures.

Existing methods for predicting RNA folding pathways

In parallel to the development of the RNA secondary-structure prediction methods, several methods have been developed that aim to explicitly simulate cotranscriptional structure formation as a function of time. All of these methods—e.g., RNAKinetics (Mironov et al. 1985; Mironov and Lebedev 1993; Daniilova et al. 2006), Kinfold (Flamm et al. 2000), Kinefold (Isambert and Siggia 2000; Xayaphoummine et al. 2003, 2005), and Kinwalker (Geis et al. 2008)—take as input a single RNA sequence and use a range of different statistical models, approximations, and heuristics to arrive at their predictions. Typically, they use stochastic simulation that extends the input RNA sequence at regular intervals, and simulates helix formation and disruption events over a simulated timescale. The probability that each event occurs is proportional to its theoretical chemical rate of change. They have, however, conceptual difficulties dealing with long sequences (over a few hundred nucleotides), and their performance has Until recently (Zhu et al. 2013) been only benchmarked for a few select sequences. They are thus currently viewed as folding-pathway prediction methods rather than RNA secondary-structure prediction methods.

The recent study by Zhu et al. (2013) uses three of these existing methods to show that evolutionarily related RNA sequences share common transient structural features during their predicted folding pathways, and that these features often coincide with known transient structures. The investigators propose an analysis pipeline that applies several folding-pathway prediction methods in a comparative manner by combining folding predictions across evolutionarily related RNA sequences. Moreover, this study provides solid evidence that some transient helices have been conserved during evolution.

Ideas for capturing cotranscriptional folding in methods for RNA secondary-structure prediction

The key effect of cotranscriptional folding is to make the formation of the final structure depend on its wider context, both along the sequence and in terms of time.

The key feature common to all existing noncomparative and comparative methods for RNA secondary-structure prediction is that they search the space of all possible (typically pseudoknot-free) RNA secondary structures for the optimal structure without having any notion of a folding pathway or a timescale of events (see Fig. 2). The recently introduced method CoFold (Proctor and Meyer 2013) is an exception, yet it currently only models a single overall effect, namely, the reachability of base-pairing partners during cotranscriptional folding, which effectively amounts to a weighing of different regions of the structure search space. The search of the structure space usually involves a scoring function whose overall value is being optimized during the search. The overall score for any candidate RNA structure is typically expressed as the sum or product of scores for individual structural building blocks that, taken together, cover the entire sequence. These elementary scores and the way in which they are combined by the scoring function during optimization, however, only depends on the local building blocks of the subsequence under consideration, but neither on their location within the sequence nor the RNA structure context of the surrounding sequence (see Fig. 2). Most optimization algorithms are dynamic programming algorithms that combine optimal structures for adjacent subsequences into one optimal structure for the resulting merged subsequence. The order of these steps, however, does not replicate the events during cotranscriptional folding. In particular, no region of the theoretical structural search space is marked as unlikely, if the corresponding structure feature could not readily form cotranscriptionally in vivo (see Fig. 2).

One of the intrinsic features that are known to influence the formation of RNA structure in vivo are transient...
structures as discussed above. Because these features are encoded in the RNA sequence itself, they could, in principle, be detected by any method for RNA secondary-structure prediction and subsequently used to bias the optimization process yielding the final RNA structure. Their detection could be implemented via a straightforward dynamic programming procedure that swiftly identifies all candidate helices (of some minimum length or stability) in the given input RNA sequence (Meyer and Miklós 2004). The conceptual problem is that these helices would naturally comprise both candidate transient helices as well as candidate helices of the final RNA secondary structure. These helices could be used in the optimization procedure in order to influence the local decision making (how to combine optimal structures for two subsequences into a single optimal structure for the merged subsequence). This would be one conceptual way of taking the wider structure context into account during the optimization procedure yielding the predicted final RNA structure. In the spirit of Meyer and Miklós (2004), these modifications could, for example, penalize any candidate structure that has strong competing transient helices upstream that could jeopardize its cotranscriptional formation.

Whereas the identification of candidate helices and relevant competing helices for a single sequence may be complicated due to the relatively large search space, comparative methods may generate a more accurate and smaller set of evolutionarily conserved competing helices to consider, such as those output by the conservation-based helix-finding algorithm Transat (Wiebe and Meyer 2010). If transient RNA structural features turn out to be evolutionarily conserved on a similar level to those of the final RNA structure, which is what recent results by Zhu et al. (2013) indicate, however, this may actually lower the prediction accuracy of comparative RNA secondary-structure prediction methods because they may erroneously incorporate these conserved transient helices into the predicted final RNA secondary structure. Whether or not this is the case and a cause for concern remains to be shown.

In addition to the ideas used by CoFold (Proctor and Meyer 2013) discussed above, the directionality of transcription could also be captured by rendering the scores assigned to the structural building blocks dependent on their position within the transcript, whether they are nearer to the 5’ or the 3’ end.

It is less obvious how one should account for the speed of transcription, let alone variations of transcription speed and transcriptional pausing. At least for now, there is too little experimental information to hope to identify transcriptional pausing sites computationally. A change in overall transcription speed alters the ratio between the speed of transcript synthesis and the rate of structure formation. This has been experimentally shown to influence cotranscriptional folding pathways and their structural outcome. On the structure prediction side, the speed of transcription could be captured by altering the effective distances between structural features. This is exactly what the free parameter in CoFold (Proctor and Meyer 2013) is for. By changing its value, one can effectively account for different (yet constant) transcription rates and thereby optimize the program’s performance for different species. If the transcription speed is high with respect to the rate of structure formation, the emerging transcript has less time and hence fewer opportunities to explore the surrounding structure space. This has the overall effect of enlarging
effective distances, whereas a low transcription speed should have the overall effect of reducing effective distances.

A biologically diverse set of molecules can form trans interactions with transcripts in vivo. All of the existing methods for predicting RNA secondary structure including methods for folding pathway prediction assume an isolated RNA sequence as input and ignore any potential trans-interaction partners (the bulk effects of water and some ions is taken into account by most folding-pathway prediction methods). If and how these trans interactions influence the cotranscriptional structure formation not only depends on the type of interaction (RNA–RNA, RNA–protein, etc.), but also very much on the timing of the interaction with respect to the structure formation. For example, a protein that binds the emerging transcript early on and for a short time has a very different influence on structure formation from that of a protein that binds the final RNA structure only.

Early and persistent types of trans interactions could be captured in RNA secondary-structure prediction methods by preventing the bound (and either single-stranded RNA [ssRNA] or double-stranded RNA [dsRNA]) subsequence from engaging in other interactions, in particular, other RNA structural features. Technically, this is fairly easy to achieve via a slight modification of the default optimization procedure by assigning a large penalty to all structure solutions that do not keep the bound subsequence single or double stranded. This feature is already implemented by all RNA secondary-structure prediction methods that allow known RNA structural features to be taken into account (e.g., Zuker and Stiegler 1981; Knudsen and Hein 2003; Pedersen et al. 2004b). This assumes, however, that details about the interaction site (subsequence, ssRNA vs. dsRNA) are known up-front, which is often not the case.

Any trans interactions of a more transient nature, however, are hard to capture computationally by any of the existing methods for RNA secondary-structure prediction because this would require them to have some notion of time-ordered steps, which they currently do not have.

Suggestions for further improving methods for folding-pathway prediction

The existing folding-pathway prediction methods already mimic the in vivo folding as they fold the RNA sequence cotranscriptionally at a constant transcription speed (which needs to be specified by the user). This is, however, only a first approximation of the complex in vivo situation. Because these methods explicitly predict folding pathways, they already model cis RNA–RNA interactions and, in particular, transient RNA structural features. At least for now, these methods do not predict variations of transcription speed and do not capture potential trans interactions with other molecules from the in vivo environment.

If details about trans interactions are known up-front (timing, binding site, ssRNA vs. dsRNA), these could be fairly easily captured by preventing the known binding site from engaging in other interactions. This has already been done for select examples and allowed us to computationally investigate the effect of trans interactions on cotranscriptional RNA structure formation (Schoemaker and Gultyaev 2006).

SUMMARY

With 75% of the human genome being transcribed (Djebali et al. 2012), the investigation of transcriptomes and how they are regulated has never been more important. RNA structure is one important feature by which transcripts can influence their fate in the cell. There is by now ample experimental and solid theoretical evidence that RNA structure formation already starts during transcription and that events during the cotranscriptional folding determine which functional RNA structure(s) are being formed. Yet, as of now, the process of structure formation is completely ignored by almost all state-of-the-art methods for RNA secondary-structure prediction. We argue that capturing some aspects of the structure formation process in predictive models could significantly improve these methods and provide evidence for this in form of a new method (Proctor and Meyer 2013). These initial results are very encouraging because they show that a significant improvement in prediction accuracy can already be gained by modeling a single overall effect of cotranscriptional folding and without making the underlying prediction algorithm much more complex. Beyond this, we propose detailed ideas of how different aspects of cotranscriptional folding in vivo could also be captured in silico.

One of the most simple and encouraging messages from the mounting (and sometimes dauntingly complex) experimental results is certainly the realization that the transcript in the cell does not explore all of the structure search space.

ACKNOWLEDGMENTS

This project was supported by grants to I.M.M. from the Natural Sciences and Engineering Research Council (NSERC) of Canada and from the Canada Foundation for Innovation (CFI). D.L. is funded by an NSERC Postgraduate Scholarship. J.R.P. holds an Alexander Graham Bell Canada Graduate Scholarship from NSERC, with additional funding from the CIHR/MSFHR Bioinformatics Training Program at the University of British Columbia. CIHR is the Canadian Institutes of Health Research, and MSFHR is the Michael Smith Foundation for Health Research in Canada.

REFERENCES


On the importance of cotranscriptional RNA structure formation

Daniel Lai, Jeff R. Proctor and Irmtraud M. Meyer

RNA 2013 19: 1461-1473

References
This article cites 138 articles, 29 of which can be accessed free at:
http://rnajournal.cshlp.org/content/19/11/1461.full.html#ref-list-1

Open Access
Freely available online through the RNA Open Access option.

Creative Commons License
This article, published in RNA, is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported), as described at http://creativecommons.org/licenses/by-nc/3.0/.

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.