Computational Analysis of Conserved RNA Secondary Structure in Transcriptomes and Genomes

Sean R. Eddy

Howard Hughes Medical Institute Janelia Farm Research Campus, Ashburn, Virginia 20147; email: eddys@janelia.hhmi.org

Annu. Rev. Biophys. 2014. 43:433–56

The *Annual Review of Biophysics* is online at biophys.annualreviews.org

This article's doi: 10.1146/annurev-biophys-051013-022950

Keywords

noncoding RNA, lncRNA, probing, SHAPE, statistical inference

Abstract

Transcriptomics experiments and computational predictions both enable systematic discovery of new functional RNAs. However, many putative noncoding transcripts arise instead from artifacts and biological noise, and current computational prediction methods have high false positive rates. I discuss prospects for improving computational methods for analyzing and identifying functional RNAs, with a focus on detecting signatures of conserved RNA secondary structure. An interesting new front is the application of chemical and enzymatic experiments that probe RNA structure on a transcriptome-wide scale. I review several proposed approaches for incorporating structure probing data into the computational prediction of RNA secondary structure. Using probabilistic inference formalisms, I show how all these approaches can be unified in a well-principled framework, which in turn allows RNA probing data to be easily integrated into a wide range of analyses that depend on RNA secondary structure inference. Such analyses include homology search and genome-wide detection of new structural RNAs.

Contents

INTRODUCTION

Some of the most important and controversial questions in molecular biology and genomics today are about the biological functions of RNA (9, 36, 76, 107–109). Advances in sequencing technology have made it possible to survey RNA transcript populations comprehensively using cDNA sequencing (68), tiled microarrays (35), and now RNA-seq (62). As technology has become more sensitive, a large number of putatively noncoding RNA (ncRNA) species have been detected, and the apparent complexity of RNA transcript populations has grown (16, 20, 30).

Noncoding RNA (ncRNA): RNA that does not code for protein, which, depending on context, may include mRNA untranslated regions (UTRs) and/or nonfunctional RNA

Transcriptional noise: RNA species produced by background error rates of other normal RNA biogenesis processes

There are two fundamentally opposed views of this growing complexity. One view is that it indicates a vast unappreciated repertoire of functional noncoding RNAs (9, 15). Another view is that many supposed noncoding transcripts are the result of experimental artifacts, analysis errors, and transcriptional noise (4, 99, 108). On the one hand, the repertoire of functions for RNA certainly continues to expand for noncoding RNA transcripts (2, 39, 40, 83, 118), *cis*-regulatory RNA sequences in messenger RNAs (32, 42, 44, 80, 84, 92), and catalytic RNAs (74, 90). On the other hand, studies have shown that high-throughput experiments and computational analysis pipelines suffer from serious systematic artifacts (65, 108, 132), and RNA biogenesis, like any biochemical process, must have some background level of infidelity (75, 99). Thus, the question is not whether or not all newly discovered RNA transcripts are functional; rather, the question is, for any one of them, how to tell the difference.

Computational methods of RNA sequence analysis are at the crux of addressing these questions, if only because the data sets are large. Historically, these methods assume that the RNA to be analyzed is already known to be functional and that a secondary structure is involved. RNA secondary structure prediction (67, 134, 135), structure-guided sequence alignment (18, 54, 120, 124), and database similarity searching with RNA sequence/structure consensus models (8, 18, 49, 89) are examples of these classic computational problems. Now, it has also become important to be able to judge whether or not an RNA sequence is likely to have a biological function, as well as whether or not the RNA has a secondary structure that plays a role in its function. For example, signatures of evolutionary sequence conservation help distinguish functional RNAs from transcriptional noise, and signatures of RNA secondary structure conservation help distinguish RNAs that function via primary sequence alone from those that depend on a more complex structure.

A class of computational RNA analysis methods has been developed that seeks to identify novel structural RNAs in genome sequences (6, 11, 13, 26, 72, 86, 102, 106, 113, 115, 123, 129). Structural RNA detection methods work by looking for evolutionarily conserved RNA secondary structure using comparative analysis of patterns of covariation in homologous genome sequence alignments. As a result, these techniques detect structural RNAs, including both structural noncoding RNA genes and *cis*-regulatory RNA structures, but they do not detect functional RNAs that act as linear sequences.

Rather than helping to clarify the results coming from systematic transcriptomics, computational methods for structural RNA detection have sown a parallel line of confusion. These methods have been used to predict hundreds, thousands, or even millions of novel structural noncoding RNAs (ncRNAs), especially in large mammalian genomes (71, 72, 79, 91, 95, 114, 117, 123). The large number of candidate ncRNAs produced via computational predictions and experimental transcriptomics has sometimes been seen as independent confirmation of the existence of a vast hidden complexity of functional RNA, but the computational approaches are subject to their own list of potential artifacts.

The problem with computational RNA structure detection approaches is that they are unreliable (5). Their signal-to-noise ratios are poor, and they are being used at a perilously ragged edge of statistical significance. Because of difficulties in establishing appropriate negative controls, such as adequately realistic homologous multiple genome alignments that are known not to be functional structural RNA, there are large uncertainties in calculating statistical significance. Small errors that are well within these uncertainties could erase the majority of the predictions. Thus, these methods need to be improved.

The identification of new data sources that could be incorporated into genome-wide sequence analyses to increase the detectable signal for structural RNAs might dramatically improve the methods discussed above. One such data source has begun to look feasible. There is renewed interest in using chemical modification and enzymatic cleavage experiments to probe RNA secondary structure, using both well-established reagents (such as RNases and dimethyl sulfate) (10, 38, 60, 77) and, especially, a powerful new class of reagents used in a technique called selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) chemistry (59, 119). RNA structure probing experiments have been coupled to high-throughput sequencing readouts, allowing these approaches to be applied at scale to probe many RNAs in parallel, including transcriptome-wide structure probing (37, 43, 48, 105, 133).

Structure probing data are noisy and statistical in nature. They provide an informative but ambiguous signal of RNA structure. Several computational methods have been proposed already for incorporating probing data into single-sequence RNA secondary structure prediction methods

RNA secondary structure: an essentially two-dimensional representation of an RNA in terms of its intramolecular nested base pairing interactions that form stems and loops

Transcriptomics: systematic discovery, quantitation, and cataloging of individual RNA transcripts

Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): a structure probing chemistry with favorable sequence-independent properties

RNA structure probing: using enzymatic or chemical modification and/or cleavage experiments to differentiate base-paired from single-stranded nucleotides in an RNA

(12, 29, 52, 53, 69, 77, 116, 131). As yet, it remains unclear which of these approaches is most powerful, most principled, or most generalizable to more complex problems in comparison with single sequence structure prediction.

In what follows, I expand on the above themes, and I conclude by showing how all the existing approaches for incorporating RNA structure probing data into RNA structure prediction can be viewed from a unified statistical inference perspective. This perspective suggests ways of naturally incorporating RNA structure probing data into all other classes of computational RNA analysis methods that depend on RNA secondary structure inference. Such methods include de novo genome-wide structure detection and homology search.

HETEROGENEITY OF RNA FUNCTION AND BIOGENESIS

It is necessary to appreciate the extreme heterogeneity of RNA functions in order to understand the limitations of functional RNA discovery and analysis methods. RNAs can fold into complex three-dimensional structures. They can present sequence or structural motifs for binding regulatory macromolecules. They can use complementary base-pairing of linear sequence to recognize other nucleic acid sequences with exquisite specificity and efficiency. They can use complementarity to template nucleic acid synthesis. The act of transcription itself may have a function, rather than the RNA that it produces (55).

Different functional RNAs combine and deploy these modalities in a variety of different ways (28, 82). RNAs can serve as informational messages, as in protein-coding messenger RNAs. RNAs can act as structural and catalytic machines, much as protein enzymes and protein complexes do, as in ribosomal RNAs. RNAs can act as scaffolds, deploying a set of protein binding motifs (either linear or structural) to facilitate assembly of a multiprotein complex, as in signal recognition particle RNA (73) or telomerase RNA (130). RNAs can act as templates for complementary RNA or DNA synthesis, as in the core of a telomerase RNA. RNAs can act as complementary guides, targeting a shared protein machine to several different specific nucleic acid targets, such as the small nucleolar guide RNAs that direct specific 2′-O-ribose methylations and pseudouridylations of other RNAs (81). *Cis*-regulatory RNA motifs act as posttranscriptional signals and switches (78), with roles in essentially every imaginable step of RNA biogenesis and trafficking. Regulatory RNA motifs may simply be small linear sequence targets of an RNA binding protein (80, 93) or may be complex RNA machines, such as riboswitches (92).

RNA biogenesis is also heterogeneous (51). Noncoding RNA genes may be transcribed by RNA polymerase I, II, or III, often as larger precursor transcripts that undergo trimming and processing. Some functional RNAs are generated by processing of pre-messenger RNAs (pre-mRNAs), including many intron-encoded microRNAs (miRNAs) and small nucleolar RNAs (snoRNAs) (104, 121). Unlike messenger RNAs, functional noncoding RNA transcripts are often not 5′ capped and 3′ polyA+; instead, they exhibit a variety of 5′ and 3′ ends, including circular RNAs with no ends at all (58). Functional RNAs can range in size from very small (a 4–10-nucleotide protein binding *cis*-regulatory site or a 20–25-nucleotide miRNA transcript) to very large (a large RNA catalyst or a scaffold of several thousand nucleotides).

There is no such thing as an unbiased screen for functional RNAs. No one characteristic signal discriminates functional RNA from other sequences. Functional RNAs may have conserved secondary and tertiary structure, but many do not. A linear RNA sequence can function as a message, a guide, a scaffold, a template, or a signal. Functional RNAs may be independent transcripts arising from noncoding RNA genes, discoverable by transcriptomics, but the roles of *cis*-regulatory RNA sequences in posttranscriptional gene regulation are at least as important (and still relatively understudied relative to transcriptional regulatory signals).

TRANSCRIPTOMICS APPROACHES TO SYSTEMATIC DISCOVERY OF NONCODING RNAs

Powerful experimental transcriptomics approaches (35, 62, 68) have resulted in the description of large numbers of putative noncoding RNA transcripts, especially what are called long noncoding RNAs (lncRNAs). lncRNAs are loosely defined as apparently noncoding mRNA-like transcripts (5′ capped, 3′ polyA+ or polyA−, and transcribed by RNA polymerase II) that are at least 200 nucleotides long.

Long noncoding RNA (lncRNA): noncoding RNA transcripts longer than (i.e., other than) the abundant classes of small RNAs such as microRNAs, snoRNAs, snRNAs, and tRNAs

Pervasive transcription: especially in mammalian genomes, the observation that most of a genome is transcribed at a detectable level

A small number of lncRNAs have known functions. One of the best studied is Xist, a very large (19-kb) ncRNA transcript that triggers the heterochromatization (Barr body formation) of one of the two X chromosomes in females. Among recent lncRNA discoveries, one of the best studied is HOTAIR, an ∼2-kb RNA in the HOXC cluster that apparently regulates the transcription of HOXD loci in *trans* via a mechanism having to do with chromatin modifications and association with both the PRC2 histone H3K27 methylation complex and the coREST complex (83, 103). Similar to Xist and HOTAIR, there is evidence that several other lncRNAs are involved in chromatin modification. A few other lncRNAs have been proposed to have other functions.

For the most part, well-studied lncRNAs such as Xist, HOTAIR, and others including MALAT1 (125) provide substantial circumstantial evidence for their functionality, even leaving aside the detailed experimental studies that have focused on them individually. These molecules are highly expressed and localized to the nucleus. They contain evolutionarily conserved sequence regions. Their sequences are unique and are devoid or highly depleted of the transposable element (TE) remnants that are so abundant elsewhere in the human genome.

In contrast, most of the RNAs in the large catalogs of lncRNAs that have yet to be experimentally characterized, lack the aforementioned expected characteristics of functional RNAs. For example, in one recent meta-analysis of 127 human RNA-seq libraries—notable for the thoroughness of its data availability, which allowed me to reanalyze the work in considerable depth—Hangauer et al. (30) identified 53,864 lncRNA loci expressed above a chosen threshold. I have replotted two key observations from their paper in **Figure 1** in order to make two points.

First, we must distinguish genome coverage from expression level when discussing pervasive transcription (9, 20, 108, 109) and lncRNAs. **Figure 1*a*** illustrates how most (here 67%) of the genome is detectable in cellular RNA (9), but only if we look at very low expression levels relative to mRNA transcripts. At expression levels that are more typical of known and annotated genes (coding and ncRNA both), only a small fraction of the genome is covered (108). For example, in **Figure 1*a***, 85% of the coding mRNA exons are covered at a read depth of at least 100, whereas only 5% of the genome and only 3% of the lncRNAs are covered at the same threshold. Functionally characterized lncRNAs are unlike the bulk of the lncRNA distribution because they tend to be expressed at levels that are comparable to or even higher than those of coding mRNAs (see, for example, H19, NEAT1, and MALAT1).

Second, **Figure 1*b*** shows how most cataloged lncRNAs show sequence divergence in comparison to repeat sequences that are assumed to be nonconserved and neutrally evolving. Only a few show segments of sequence conservation, including most of the well-characterized functional lncRNAs. For example, only 7% of RefSeq exons fall below a threshold conservation score of 1 in **Figure 1*b***, whereas 99.9% of repeats and 91% of lncRNAs are below this threshold. GAS5 is an exception that proves the rule: GAS5 is an inside-out snoRNA carrier gene, whose function is to have conserved introns processed into snoRNAs (94). Moreover, according to my analysis with RepeatMasker, 56% of the sequence of these 53,864 lncRNAs consists of TE remnants, essentially indistinguishable from genomic background (53%).

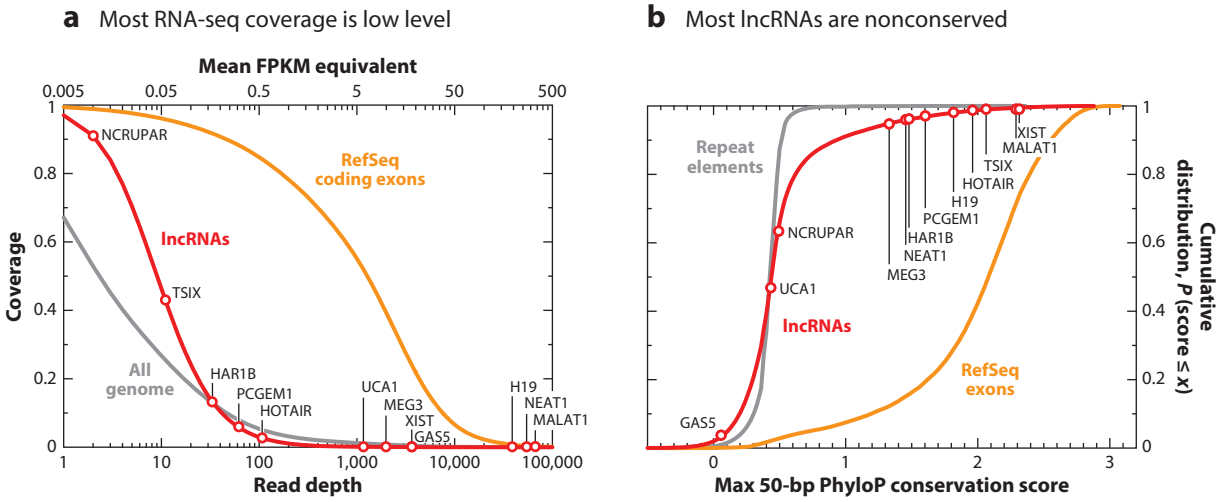

Figure 1

Two key observations about pervasive transcription and long noncoding RNA (lncRNA) catalogs (30). (*a*) Gray line: fraction of the uniquely mappable genome (2,570 Mb; assembly version hg18) covered at different thresholds of minimum read depth per genome position. Red line: coverage of the sequence of 53,864 lncRNAs (38 Mb). Orange line: coverage of the sequence of 364,265 coding exons of 34,978 RefSeq coding genes (34 Mb). Red circles placed along the lncRNA distribution mark the median read depth over 12 functionally characterized lncRNAs (*y*-axis position for these points has no meaning). Read depth units from Reference 30 are roughly convertible to mean FPKM (fragments per kilobase per million mapped reads) units; means are calculated over 127 RNA-seq libraries. FPKM = 1,000 * read depth/(read length per fragment)/(millions of fragments), where the aggregate data set has 3.39 billion fragments with a mean read length of 60 nucleotides per fragment (top *x*-axis label). Data reanalyzed and replotted with permission from Hangauer and colleagues (supplementary figure 1 and data set 8 from Reference 30, plus a BED file of read depth coverage per genome position provided by M. Hangauer). (*b*) Cumulative distribution of sequence conservation (maximum in 50-bp windows, as measured by PhyloP in a placental mammalian genome alignment) for human repeat elements (*gray line*; presumed to be neutrally evolving), exons of 31,204 RefSeq coding genes (*orange line*), and 53,864 lncRNAs defined by Reference 30 (*red line*). Conservation values for 12 characterized functional lncRNAs are marked with red circles placed along the lncRNA cumulative distribution (their *y*-axis positions have no meaning). Figure redrawn from the same data used in figure 3C of Reference 30, with permission.

To be functional, an RNA need not necessarily be expressed at levels comparable to those of known mRNAs, nor evolutionarily conserved, nor devoid of TE remnants (15). However, other more likely explanations exist for low-level nonconserved transcripts with TE content similar to genomic background.

One source of lncRNAs is transcriptional noise (75, 99). Some authors (76) have taken transcriptional noise to mean random transcription, a uniform haze across the genome, implying that the observation that a lncRNA is expressed in a tissue-specific manner is evidence of functionality (30). However, the neutral expectation is that cryptic RNA transcription and processing are driven by randomly occurring (specific and discrete, but cryptic) short binding sites for regulatory proteins. Expression patterns of these discrete cryptic transcripts will follow the specific spatiotemporal expression patterns of the regulatory proteins that activate them (17).

Other sources of lncRNAs are computational analysis errors, including failures to recognize predictable experimental artifacts. Half of the so-called noncoding RNAs in a pioneering paper on lncRNAs from the Functional Annotation of the Mouse 3 (FANTOM3) project (68) are cloning artifacts that arose by internal priming on polyA tracts in pre-mRNA (65). False transcribed regions

are created by cross-hybridization artifacts on genome tiling arrays (108) and by mismapping of RNA-seq reads (even uniquely mapped reads) (132).

Even defining a transcript as noncoding is surprisingly difficult (34). Many real proteins are shorter than the typical open reading frame (ORF) length cutoff of ≥100 amino acids used for defining ncRNA (31). More powerful methods for recognizing coding genes using comparative sequence analysis are often used (45, 112), but they are often trained and their accuracy evaluated on complete sequences of normal proteins rather than on mRNAs expected to contaminate a lncRNA catalog (these are not average mRNAs but are instead an extreme tail of the coding mRNA distribution that is enriched for short coding genes and partial transcript sequences). Even basic rules for defining ncRNA have proven inexplicably difficult to apply. The FANTOM3 bioinformatics pipeline failed to recognize that 27% of the so-called ncRNAs that they identified in fact do contain an ORF of ≥100 amino acids, and 25% of them have a BLASTP similarity to the protein database of $E \leq 10^{-10}$ (65), even though these criteria were among those used in the FANTOM3 analysis of coding potential.

Putative lncRNAs need to be treated as heterogeneous, not as a class. Only some are likely to be functional RNAs, and these are likely to have a variety of functions. Careful computational analyses can help prioritize and sort putative lncRNAs into different categories. Improved computational tools of all sorts will help these analyses. Such tools include read mappers with lower false positive mapping rates, spliced transcript assemblers that assemble longer and more complete RNA transcripts from RNA-seq data, more quantitative measurements of sequence conservation and evolutionary constraint, more powerful methods for detecting small coding regions, and better methods for detecting homology between RNA sequences.

Often these computational analyses are subtractive. They look for positive signals of something else (e.g., coding regions, experimental artifacts) to winnow down a set of candidate lncRNAs and enrich for functional RNAs. One of the more interesting areas to me is the development of methods for detecting evolutionarily conserved RNA secondary structure. Conserved RNA secondary structure is one of the few affirmative signals we can look for in a functional RNA.

COMPUTATIONAL DETECTION OF CONSERVED RNA STRUCTURE

An evolutionarily conserved RNA secondary structure might be the most general feature shared by many functional RNAs. Obviously, a drawback of using conserved structure as a detection strategy is that this strategy will miss functional RNAs that act primarily by their linear sequences. Even in the best-studied lncRNAs, it remains somewhat unclear whether there is much conserved RNA structure. RNA secondary structures have been proposed for parts of HOTAIR (37, 103), parts of Xist (50), and for other lncRNAs (66). In addition, MALAT1 clearly has a fascinating transfer RNA (tRNA)-like structure at its 3′ end (125, 126). But lncRNAs that act as scaffolds, for example, for chromatin modification complexes, could well bind those complexes via single-stranded RNA sequence motifs. Nonetheless, computational detection of conserved secondary structure is a useful signal to positively identify at least a subset of functional RNAs against a background of other less interesting explanations, and this technique has an advantage (over transcriptomics, for example) in that it can also detect *cis*-regulatory structures in mRNAs.

Development of Computational Methods for RNA Structure Detection

The first attempts to develop a general genome-wide approach to detecting RNA structures looked for regions of a single sequence predicted to fold into RNA structures that are more thermodynamically stable than expected (41). However, random RNA sequences fold into secondary

structures with predicted thermodynamic stabilities that are similar to those of functional RNAs, so this approach was deemed insufficiently powerful for genome-wide screens (85). At best, ~30% of structural RNAs could be detected with an estimated false positive rate of ~10 per megabase of genome screened (85).

Attention moved to exploiting the evolutionary conservation of RNA secondary structure in homologous sequence alignments as an additional source of signal to discriminate real functional RNAs from background using pairwise (86) or multiple alignments (11, 13). The 2001 Rivas QRNA algorithm was estimated to detect ~80% of structural RNAs at an estimated false positive rate of ~20 per megabase, allowing for a screen of the small *Escherichia coli* genome (87).

The general idea of detecting conserved RNA structure in multiple sequence (or multiple genome) alignments has now been extended and implemented in many ways, such as in RNAz (115), EvoFold (72), CMfinder (129), FOLDALIGN (102), and other approaches (6, 26, 106, 113, 123). These programs have been used to predict regions of structural RNA in large eukaryotic genomes, especially the human genome (71, 72, 79, 91, 95, 114, 117, 123). In one recent screen of the human genome, for example Smith et al. (95) predicted 4.1 million structural RNAs in the human genome, at an estimated sensitivity of ~30% and an estimated false positive rate of ~170 per megabase. The authors described this false positive rate as historically low.

In fact, the stringency demanded from these approaches has declined while the ambition to screen large mammalian genomes has increased. Moreover, there is substantial uncertainty in how false positive rates are estimated, either by shuffling or simulating negative multiple alignments. My laboratory (98) abandoned attempts to extend QRNA screens to large genomes when we found that our rate of experimental confirmation of the expression of predicted intergenic RNA loci was far lower than the computationally predicted false positive rate. The Hughes laboratory (4) reached the same conclusion in their experimental follow-up of a QRNA screen of the mouse genome. The current false positive rates from this class of methods remain too high to justify their use on large genomes (5).

Current Methods Remain Insufficiently Reliable

Consider the recent computational screen by Smith et al. (95) as a specific example of the poor reliability of current methods. These authors applied two different approaches, RNAz 2.0 (27), and a new method called SISSIz (24), to the human genome using a comparative analysis of a multiple alignment of 35 mammalian genomes. RNAz and SISSIz, like all methods in this class, work by scoring one small alignment window at a time (here 200 nucleotides) under a model that looks for RNA structure conservation and by classifying that window as a structural RNA prediction if it passes a chosen score threshold. The whole genome alignment is scored in overlapping windows (in this case, 200-nucleotide windows overlapped by 100 nucleotides, both forward and reverse complement); Smith et al. (95) scored 50 million alignment windows.

The false positive rate—the fraction of windows incorrectly scored as structural RNAs—is a critical number to estimate. To estimate a false positive rate, we have to devise a negative control, namely, a way to obtain windows that are known not to contain a structural RNA yet are matched controls for all other background properties of genomic alignment windows (e.g., sequence conservation, GC% composition, indel pattern). This is a hard problem. Two main approaches have been used to address it. One approach is to shuffle alignments by columns, preserving properties such as nucleotide composition in the window and primary sequence conservation in each column (1). Another approach is to simulate synthetic alignments according to a phylogenetic model (6, 23, 24).

It is easy to create poor negative controls. Naive shuffling of an alignment shatters indel patterns into many single-base insertions and deletions, disrupts background dinucleotide composition

[which tends to have a strong effect on RNA structure calculations (127)], and homogenizes conservation and GC% composition across a window that might encompass a local region of high GC% or high conservation that already tends to score highly (85). Real genome alignments may tend to score well only because of these confounding background effects, not because they contain RNA structures. However, the more a shuffling procedure tries to preserve more realistic background effects, the more it tends to preserve the original alignment. For example, a shuffling procedure used in Reference 5 altered the order of only 53% of the alignment columns, on average, and was thus probably inadequate to destroy all the signal of a true RNA structure. Different methods produce very different predicted false positive rates. For example, Smith et al. (95) show tenfold differences in the false positive rates measured by simulations with SISSIz (24) versus shuffles with Multiperm (1).

Smith et al. (95) calibrated their score thresholds to allow 1% false positive predictions per 200-nucleotide alignment window, using both shuffled and simulated negatives. Therefore, they expected to see ~500,000 false positives when scoring a total of 50 million windows. Their screen actually detects 4.1 million positive windows. Because we estimate that 500,000 of these windows are false, all of the excess detections (3.6 million) should be true. This gives us a so-called empirical false discovery rate (FDR) of 12% [500,000/(4.1×10^6) = 0.12]. [Smith et al. (95) varied how they generated negative control windows and reported an FDR of 5–22%.]

False discovery rate (FDR): the fraction of a set of predictions that are statistically expected to be false positives

Expectation value (E-value): the number of false positives expected at or above some score threshold

Empirical False Discovery Rates are Vulnerable to the Choice of Negative Control

Empirical FDRs are only as good as the estimate of the number of expected false positives under the assumed null hypothesis. If we underestimated the number of false positives by just tenfold, a 12% FDR might really be 100%. Essentially all of our statistically significant candidates could be false.

Could the estimated FDR be off by tenfold? Yes, easily. Consider the more familiar task of a BLASTN DNA similarity search. We typically do not trust BLASTN expectation values (E-values) to be more accurate than within a few orders of magnitude. BLASTN's estimated false positive rate, though quite good, is confounded by many nonrandom biases that occur in real genome sequences (e.g., composition bias, repetitive sequence) that generate false positives at a higher rate than randomized expectation predicts. For example, if we found 100 BLASTN hits in a database search at an E-value threshold of 10, we would not assume that 90 hits were true, but an empirical FDR calculation does make such assumptions.

The null hypothesis for detection of conserved RNA structure is much more complex than that for BLASTN because the former requires matching of an even more complicated set of relevant properties of non-RNA genomic alignment windows. Thus, it would be prudent to have less confidence in the accuracy of these false positive estimates than in BLASTN-based estimates, as background biological signals that could easily confound a structural RNA detector may not be taken into account in current shuffled or simulated negative controls. Short and long inverted DNA repeats are one example. These elements are abundant in genomes, partly because of the activity of DNA transposons, and can look like RNA hairpins in a genome sequence analysis, even if they are never expressed as RNA.

Thus, the application of these methods to large genomes seems premature and perilous, although the fundamental idea is sound. The discriminatory power of these methods needs to be increased substantially. One way to do this is by incorporating additional sources of information to increase the signal-to-noise ratio. The use of deeper multiple sequence alignments analyzed with more powerful phylogenetic models of sequence covariation patterns constrained by RNA structure is currently the main path forward in the field. However, dramatic improvements in

the use of chemical and enzymatic RNA structure probing experiments are opening up another interesting direction.

Partition function: a sum over an ensemble; normalization factor for converting free energies of individual structures to probabilities in the ensemble

Ensemble: the set of all possible secondary structures for an RNA sequence; often in the context of assigning each structure a probability

PROBING-DIRECTED RNA STRUCTURE PREDICTION

Chemical and enzymatic modification experiments have long been used to probe RNA structure. Various reagents differentially attack paired versus unpaired nucleotides and generate cleavages or base modifications that can be assayed by sequencing (7, 19, 33, 61). Historically, interpreting chemical or enzymatic modification patterns has been something of a black art, and the experiments have been done on one RNA at a time. Recently, better reagents have been developed, including those involved in SHAPE chemistry (described below), and several genome-scale methods have coupled RNA structure probing to high-throughput sequencing (37, 43, 48, 105, 111, 133). Thus, it has become feasible to probe the structure of every RNA in a transcriptome simultaneously. However, it remains unclear how structure probing data should best be incorporated into RNA structure analysis algorithms, even for the simplest problem of single sequence RNA structure prediction.

In a landmark study, Deigan et al. (12) proposed a method for incorporating SHAPE probing data as soft constraints into single-sequence RNA structure prediction. Their paper is a touchstone for understanding a growing body of work from several laboratories. Since its publication, several papers have introduced alternative methods (69, 77, 116, 131); at least one paper has extended the method proposed by Deigan et al. (12) to another chemical probe, DMS (dimethyl sulfate) (10), and another has extended it to the prediction of pseudoknotted RNA structure (29). To describe this work, first it helps to give some background on single sequence RNA secondary structure prediction, as well as on SHAPE chemistry.

Single Sequence RNA Structure Prediction

The most widely used methods for RNA secondary structure prediction utilize free energy minimization. A nearest-neighbor thermodynamic model (often called the Turner rules) approximates the free energy (ΔG) of an RNA secondary structure as a sum of individual free energy terms assigned to local features in an RNA structure, particularly to each base-pair stack (i.e., neighboring base pairs, hence the name nearest-neighbor model), as well as to hairpin, internal, and bulge loop lengths and various other elemental features (22, 47, 53, 128). Given the thermodynamic model, an efficient dynamic programming algorithm (e.g., the Nussinov/Zuker algorithm) guarantees finding the RNA secondary structure with the minimum free energy (67, 134, 135). A related algorithm (the McCaskill algorithm, described in more detail below) calculates the partition function, the sum over the ensemble of all possible secondary structures weighted by their predicted likelihoods in solution, according to their estimated free energies (56). Using the McCaskill algorithm, alternative structures can be sampled from the ensemble according to their probability (14).

Although single sequence RNA secondary structure prediction has been useful, its accuracy remains unsatisfactory. Accuracy is limited by fundamental problems: The residual error in the parameters [~5% (128)] is greater than the typical free energy difference between quite different alternatives in the low free energy RNA landscape, and the model neglects the contributions of tertiary contacts and divalent cations to the overall free energy of an RNA's fold.

This tantalizing state of affairs—namely, prediction accuracy that is useful but not reliable—has motivated the search for additional information that can be used to constrain structure predictions, including data from chemical and enzymatic structure probing experiments (7, 19, 33, 61). Several past approaches for incorporating probing data in structure prediction have given uncompelling

results (52, 53), partly because probing experiments give noisy and ambiguous data (53) and partly because traditional probing reagents such as DMS have a complex dependence on sequence and local structure (19), making it difficult to parameterize an ad hoc approach. A breakthrough came from the development of a probing reagent that acts in a much less context-dependent way, enabling simple ad hoc methods to be used, as described in the following section.

Selective 2′-Hydroxyl Acylation Analyzed by Primer Extension (SHAPE) Chemistry

SHAPE, introduced in 2005, stands for selective 2′-hydroxyl acylation analyzed by primer extension (59). A SHAPE reagent acylates the 2′-hydroxyl position of the ribose sugar of a nucleotide. This acylation impedes reverse transcription, so the presence of the acylated nucleotide can be assayed by primer extension. The reaction rate depends on the local geometry of the nucleotide backbone (57). Nucleotides in Watson–Crick base pairs are constrained in an incompatible geometry, resulting in low SHAPE reactivities. Unpaired nucleotides can show high SHAPE reactivities, presumably because a flexible nucleotide backbone can frequently visit a compatible geometry. Occasionally, a nucleotide may happen to be constrained in the right geometry, making that nucleotide hyperreactive (57). Several different SHAPE reagents exist with different properties, such as reagents with fast reaction rates for probing kinetics (63) or reagents with properties that are well suited for in vivo SHAPE experiments (96).

The standard data processing protocol from a SHAPE experiment (3, 46, 70, 110) yields a single normalized unitless number for each nucleotide in the probed RNA sequence. SHAPE values range from 0 to ∼2 or sometimes more, as the upper bound is ill defined because of the ad hoc nature of the so-called normalization.

SHAPE Data Analysis from a Likelihood Ratio Perspective

Similar to other structure probing reagents, SHAPE values do not unambiguously distinguish paired bases from unpaired bases. Rather, a SHAPE experiment confers probabilistic information about RNA secondary structure because the distribution of SHAPE reactivities for base-paired residues is different than for unpaired residues. For example, **Figure 2*a*,*b*** shows empirical distributions of SHAPE values collected from *E. coli* SSU and LSU rRNA that were compiled by Sükösd et al. (101) from SHAPE experiments published by Deigan et al. (12).

Intuitively, we might imagine that probing data distributions would show distinct modes for unpaired versus paired bases: a high-reactivity peak for unpaired bases, a low-reactivity peak for paired ones. However, the modes of the distributions are low for both paired and unpaired bases. The information in SHAPE data comes from the increased variance at unpaired residues. An unpaired base is more likely to have a low SHAPE reactivity than a high reactivity, but a base with a high SHAPE accessibility is much more likely to be unpaired than paired. **Figure 2*c*** shows the paired/unpaired likelihood ratio as a function of the SHAPE value. A base with a low SHAPE value (e.g., 0) is approximately five times more likely to be paired than unpaired, and a base with a high SHAPE value (e.g., 2.0) is conversely about five times more likely to be unpaired. We want to incorporate this kind of information into SHAPE-directed structure prediction.

Deigan's Pseudoenergy Approach

Deigan et al. (12) proposed a particular pseudoenergy term for incorporating SHAPE data,

$$\Delta G'_i = m \log(\alpha_i + 1) + b,$$

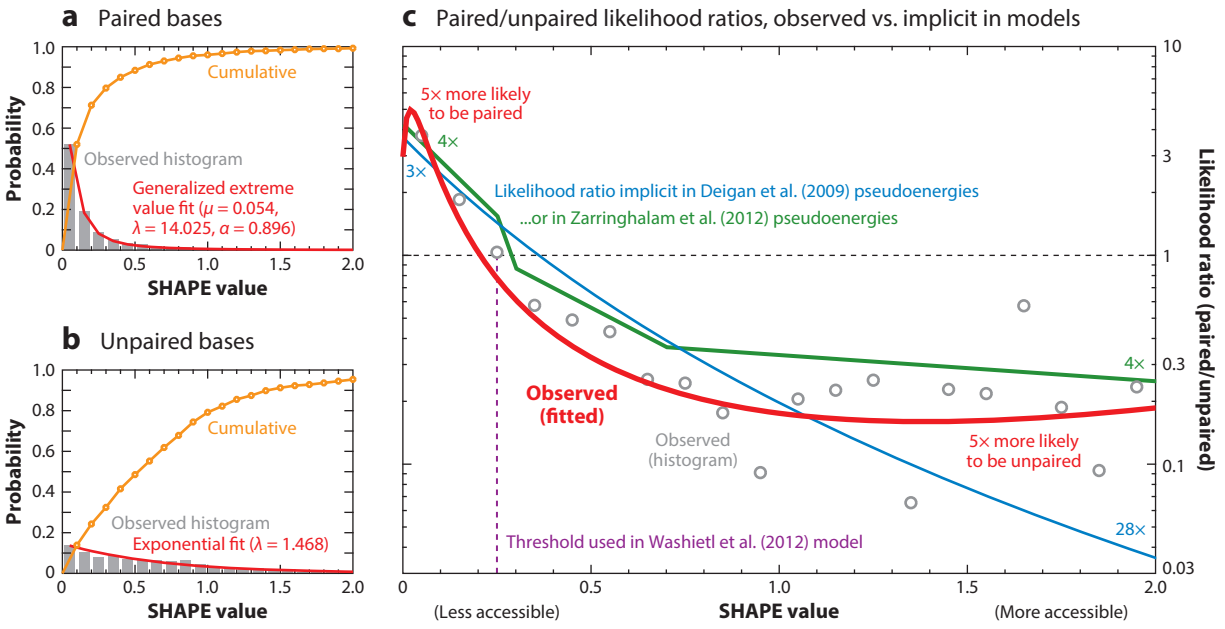

Figure 2

Observed distributions of selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) values for paired versus unpaired bases, compared to likelihood ratios implicit in different SHAPE-directed structure prediction methods. (*a*,*b*) Distributions of SHAPE values observed for (*a*) 2,531 paired nucleotides and (*b*) 1,656 unpaired nucleotides in *Escherichia coli* SSU and LSU ribosomal RNA (rRNA). Data are taken from the in vitro SHAPE experiments of Deigan et al. (12), collated by Sükösd et al. (101). Gray bars represent histograms; orange lines represent cumulative distributions. Red lines show my maximum likelihood fits to the distributions chosen by Deigan et al. (101); Panel *a* shows a generalized extreme value distribution (21), and panel *b* shows an exponential distribution. Data replotted with permission from Reference 101. Figure 1 in Reference 101 distinguished helix end pairs from base pairs internal to a stacked stem because helix ends are more flexible and therefore more accessible to SHAPE than are internal base pairs, but for simplicity I have merged all base pairs here. (*c*) Paired/unpaired likelihood ratios from the two fitted distributions in panels *a* and *b* (*red line*), from the corresponding histogram bins in panels *a* and *b* (*gray circles*), implicit in the Deigan et al. (12) pseudoenergy model (*blue line*), and implicit in the Zarringhalam (131) pseudoenergy model using their default $\beta = 0.89$ (*green line*). See text for further explanation.

where α_i is the SHAPE value for base i in the RNA; $i = 1 \ldots n$, where n is the sequence length; and m and b are free parameters with defaults set to $m = 2.6$ and $b = -0.8$ kcal mol^{-1}. This pseudoenergy term is applied to every residue i involved in a base pair (and not to unpaired bases) in the calculations in the dynamic programming recursion.

If the SHAPE reactivity is minimal (i.e., $\alpha_i = 0$), then each base in a base pair is rewarded by an additional −0.8 kcal/mol. If the reactivity is high, say $\alpha_i = 2.0$, then pairing of base i is disfavored by +2.1 kcal/mol. If the reactivity is 0.36, there is no added pseudoenergy, and the SHAPE data are considered to favor pairing or unpairing of base i equally.

Deigan et al. (12) did not justify their choice of functional form, and they set m and b empirically by grid searching a wide range of parameter settings and finding values that optimized the accuracy of *E. coli* LSU rRNA folding. However, via statistical thermodynamics, we can observe that the pseudoenergy term described above implies a likelihood model. The probability that base i is paired, given the SHAPE data, is proportional to $e^{-\Delta G'_i/\mathrm{RT}}$. Unpaired bases are implicitly assigned a pseudoenergy of 0, independent of SHAPE reactivity, so the probability that the base is unpaired is proportional to 1. The proportionality constant is the same (a simple partition function, the

sum of the two terms), so we obtain the following equation:

$$\frac{P(\pi_i = \text{paired})}{P(\pi_i = \text{unpaired})} = e^{-\Delta G'_i/\text{RT}}.$$

The notation π_i refers to the structural context of base i, which for the moment is either paired or unpaired.

Thus, the Deigan pseudoenergy term corresponds (at 37°C) to saying that a maximally unreactive ($\alpha_i = 0$; $\Delta G'_i = -0.8$ kcal/mol) base is approximately three times more likely to be paired than unpaired, and a highly reactive base ($\alpha_i = 2.0$; $\Delta G'_i = +2.1$ kcal/mol) is approximately 28 times more likely to be unpaired than paired. **Figure 2*c*** shows a plot of the paired/unpaired ratio implied by the Deigan pseudoenergy term compared with the ratios implied by observed distributions (101).

Posterior probability: the conditional probability of a random variable of interest, given observed data, often obtained by Bayes' rule

Manhattan distance: distance between two vectors computed as the sum of absolute differences of each element, similar to walking distance on a city grid

Sample and Select Approaches

It is not obvious that the thermodynamic RNA folding model can be combined in a mathematically defensible way with structure probing data. The use of arbitrary pseudoenergy parameters looks worryingly unprincipled. For this reason, an alternative, called a sample and select approach (69, 77), aims to keep the thermodynamic folding calculation separate from the probing data constraint. The idea behind this approach is to first sample suboptimal structures from the thermodynamic ensemble (14, 56) and then to rerank these sampled alternative structures by how well they agree with the structure probing data according to some distance metric. This approach is not very powerful because it relies on being able to sample the correct structure from the thermodynamic ensemble in the first place. If the correct structure has a negligible posterior probability under the thermodynamic model, it is never sampled.

The sample and select approach requires choosing which distance should be calculated between the experimental probing data and a predicted structure. Such approaches threshold the probing data to make discrete "paired" and "unpaired" calls for each base, after which they calculate the number of discrepancies from the predicted structure (termed a Manhattan distance) (69, 77). From a statistical perspective, a better-justified measure would be the log likelihood of the observed probing data given the structure, for example, $\log P(\boldsymbol{\alpha}|\boldsymbol{\pi}) = \sum_i \log P(\alpha_i|\pi_i)$, using empirically collated $P(\alpha_i|\pi_i)$ distributions as done in Reference 101 (**Figure 2*a*,*b***).

Zarringhalam's Pseudoenergy Approach

Zarringhalam et al. (131) use a distance-based argument to criticize the Deigan approach (12) and develop a new one. They propose to optimize a distance between the SHAPE data and the structure, a Manhattan distance $\sum_i |\pi_i - a_i|$, where π_i is the predicted structure of base i and takes a value of 1 if unpaired or 0 if paired and a_i is a modified SHAPE reactivity for position i, rescaled (by an ad hoc piecewise linear transformation of the original α_i) to range 0...1. They add a pseudoenergy of $\beta|\pi_i - a_i|$ to all bases i (unpaired and paired). Impressively, Zarringhalam et al. (131) prove mathematically that this approach is guaranteed to improve (decrease) the calculated distance between the SHAPE data and the predicted structure in comparison to the thermodynamic model alone. In contrast, the Deigan approach (12) often yields higher distances for a SHAPE-constrained prediction than for an unconstrained prediction.

The argument proposed by Zarringhalam et al. (131) hinges on whether we agree that minimization of a Manhattan distance between the probing data and the predicted structure is desirable. In fact, this premise is probably not well justified, which is a shame because of the

Gibbs–Boltzmann equation: statistical thermodynamics equation relating free energy of individual states to the probability of each state in an ensemble

strong theoretical proof that followed from their premise. As is apparent in **Figure 2**, a SHAPE reactivity α_i cannot be compared directly with the probability that base i is unpaired because both paired and unpaired bases are more likely to have low α_i values.

In fact, the Zarringhalam et al. (131) and Deigan et al. (12) approaches are rather similar in terms of the paired/unpaired likelihood ratios they imply (**Figure 2**). By using a symmetrical pseudoenergy function with the same β for unpaired ($\pi_i = 1$) and paired ($\pi_i = 0$) bases, Zarringhalam et al. (131) constrain the odds ratio to be symmetrical in the sense that at minimum SHAPE values, a base is approximately four times more likely to be paired than unpaired, and vice versa for maximum SHAPE values. The ad hoc piecewise linear mapping of the SHAPE value to the range 0–1 has the effect of rather closely approximating the empirical likelihood ratio distribution (**Figure 2*c***).

Washietl's Ensemble Approach

Both approaches described above assumed that base i is either 100% paired or unpaired under SHAPE experimental conditions. This assumption is equivalent to assuming that a single RNA structure dominates in solution (even if the approach uses an ensemble calculation, as in Reference 131). What about an RNA that adopts two or more different structures in solution? In this case, the measured SHAPE data would be an ensemble-weighted average over the different structures; i.e., the data would be a function of the ensemble, rather than of a single correct structure. Could SHAPE data be used to predict not just a single optimal structure, but the whole ensemble? This is the point of an ensemble-based approach introduced by Washietl et al. (116). Their basic idea is to perturb the energy parameters by the minimal amount needed to bring the ensemble base pairing probabilities into maximal agreement with the experimental SHAPE data.

Explaining the Washietl et al. approach (116) requires further introduction to ensemble calculations. According to the Gibbs–Boltzmann equation of statistical thermodynamics, the probability that a system is in a given state i with free energy ΔG_i is proportional to $e^{-\Delta G_i/\mathrm{RT}}$, where R is the gas constant (0.001986 kcal mol^{-1} K^{-1}) and T is the absolute temperature in Kelvin. If we can enumerate the free energies of all possible states of the system, then the probability that the system will be in state i is

$$\frac{e^{-\Delta G_i/\mathrm{RT}}}{\sum_j e^{-\Delta G_j/\mathrm{RT}}}.$$

The summation over all states, $\sum_j e^{-\Delta G_j/\mathrm{RT}}$, is the partition function, often abbreviated Z, which is the quantity that the McCaskill algorithm recursively calculates over all possible RNA secondary structures for a sequence (56).

Washietl et al. (116) calculate a predicted ensemble base pairing probability $z_i(\boldsymbol{\theta}, \boldsymbol{\epsilon})$ for each residue i using a partition function calculation. This calculation uses a set of thermodynamic model parameters $\boldsymbol{\theta}$, which are perturbed by an error vector $\boldsymbol{\epsilon}$ that describes the uncertainty inherent in the parameters. Perturbing the energy parameters amounts to treating the ensemble as a random variable because the ensemble is completely determined by the energy parameters. For example, we might assume that every energy parameter θ_u for some element u of RNA structure has a normally distributed error $\epsilon_u \sim N(0, \tau_u^2)$ with variance τ_u^2. This is an attractively explicit model of the uncertainty in the Turner rules. We could then obtain variances (τ_u^2 values) corresponding to the different certainties of different parameters (for example, base-pair stacking parameters are better determined than are loop parameters). What Washietl et al. (116) actually implement, however, is an alternative in which perturbations ϵ_i are assigned to each residue i, with one position-independent variance τ^2. This choice somewhat weakens their argument that their model is more

physically grounded than a pseudoenergy model because the ϵ_i terms are now pseudoenergies, rather than an explicit error model for the energy model parameters θ_u.

Critically, Washietl et al. (116) assume that the probing data $\boldsymbol{\alpha}$ can be used to directly obtain an experimentally "observed" probability $p_i(\boldsymbol{\alpha})$ that base i is paired. They assume that this so-called experimental measurement is subject to experimental errors, and the discrepancy $z_i(\boldsymbol{\theta}, \boldsymbol{\epsilon}) - p_i(\boldsymbol{\alpha})$ is normally distributed as $N(0, \sigma_i^2)$. They further assume that this error is position independent and therefore use a single σ^2.

Under this formulation, both the energy parameters and the observed SHAPE data are assumed to be subject to unknown measurement errors, which are parameterized by variances τ^2 and σ^2, respectively. The problem can then be written as the following least-squares optimization problem:

$$\min_{\boldsymbol{\epsilon}} \sum_i \frac{\epsilon_i^2}{\tau^2} + \sum_i \frac{(z_i(\boldsymbol{\theta}, \boldsymbol{\epsilon}) - p_i(\boldsymbol{\alpha}))^2}{\sigma^2}.$$

This expression gives the maximum likelihood estimate for the perturbation vector $\boldsymbol{\epsilon}$ under the assumption that both so-called errors are normally distributed. The core of their paper then shows that this minimization can be done by gradient descent.

A big difference between the Washietl et al. (116) approach and other probing-directed structure prediction methods is that the ϵ_i pseudoenergies here are optimized for each particular RNA sequence and its SHAPE data. Each ϵ_i term is essentially a measure of how hard nucleotide i must be tweaked to agree with the SHAPE data. The ϵ_i terms constitute a position-specific profile of discrepancies between the RNA structure and the prediction of thermodynamic model. Washietl et al. (116) show an interesting example in which ϵ_i terms tend to be high for nucleotides that are modified in vivo (the thermodynamic model does not take in vivo nucleotide modifications into account).

Because the SHAPE data α do not directly report the probability that a given base i is paired, the principal weakness in this approach is in obtaining $p_i(\boldsymbol{\alpha})$. Washietl et al. (116) tried many ways of mapping α_i to an "observed" pairing probability p_i, but, in the end, they simply thresholded at 0.25, setting p_i equal to 1 (paired) for $\alpha_i < 0.25$ and equal to 0 (unpaired) for $\alpha_i > 0.25$. By discretizing p_i to 0 or 1, 100% paired or unpaired, the whole point of using an ensemble-averaged calculation is lost. Moreover, when π_i is discretized to 0 or 1, it is dubious whether the discrepancy $|z_i - p_i|$ should be treated as a normally distributed error. Rather, many p_i are just wrong.

STATISTICAL INFERENCE FOR PROBING-DIRECTED STRUCTURE PREDICTION

A well-principled framework for combining the inherently statistical information from a probing experiment with the thermodynamic model of RNA folding is desirable. Such a framework might improve the accuracy of probing-directed structure and would allow more subtle information to be extracted from structure probing data than whether or not a base is paired. Reactivity depends on structural context, meaning that reactivity carries statistical information about structural context. For example, bases in helix end pairs tend to be more reactive to SHAPE probing than are bases in internally stacked stems (101). Chemical and enzymatic data from DMS modification and RNase cleavage mapping show more complex sequence dependencies than do SHAPE data, and the lack of principled approaches impedes the analysis of more complicated data.

A general approach can be outlined using probabilistic inference. The observation that a pseudoenergy term implies a particular paired/unpaired odds ratio (**Figure 2**) essentially means that the reverse is also true. We can therefore use the empirical likelihood distributions of SHAPE data values α_i to derive a principled approach in terms of probabilities.

Joint probability: the probability of two or more random variables together, as in P(A, B)

Bayes' rule: a basic equation in probability calculus for calculating a posterior probability; $P(B|A) = P(A|B)P(B)P(A)$

Conditional probability: the probability of one random variable given the value of another, as in $P(A|B)$

A strong approach to any inference problem, especially one involving integration of different sources of evidence, starts with writing a generative probability model that specifies the joint probability distribution of all of the data. This distribution should include the observed variables that are giving us information, the hidden variables that we seek to infer, and any additional hidden nuisance variables that our model needs to specify to calculate the joint probability. In the present context, we need a computable model of the joint probability $P(\boldsymbol{\alpha}, \mathbf{x}, \boldsymbol{\pi}, \theta, \psi)$ for the observed probing data $\boldsymbol{\alpha}$, the RNA sequence $\mathbf{x}$, the RNA secondary structure $\boldsymbol{\pi}$ (which we want to infer), the parameters θ of an RNA folding model, and the parameters ψ for a likelihood model of generating SHAPE values from a particular structure.

Optimal Structure Prediction and a Derivation of Pseudoenergies

Suppose we assume that a single correct RNA secondary structure dominates in solution. This critical assumption allows us to assume that the observed SHAPE data $\boldsymbol{\alpha}$ arose directly from that single structure $\boldsymbol{\pi}$. [Otherwise the observed data are an ensemble-weighted average $\langle\boldsymbol{\alpha}\rangle$, over an unknown $\boldsymbol{\alpha}(\boldsymbol{\pi})$ for each structure in solution; this scenario is discussed further below.] We can factor the joint distribution into a product of independent terms, for which the observed probing data are sampled as a function of the structure $\boldsymbol{\pi}$, and the probability of $\boldsymbol{\pi}$ is specified by the RNA folding model for sequence $\mathbf{x}$ as follows:

$$P(\boldsymbol{\alpha}, \mathbf{x}, \boldsymbol{\pi}, \psi, \theta) = P(\boldsymbol{\alpha}|\mathbf{x}, \boldsymbol{\pi}, \psi)P(\mathbf{x}, \boldsymbol{\pi}|\theta)P(\psi)P(\theta).$$

To simplify things a bit further, we can assume that we obtain fixed model parameters ψ and θ from an outside source; for example, we might obtain them by fitting our data to known example SHAPE data to obtain ψ (as in **Figure 2**) and by using the existing Turner energy model as θ. This means we can drop both terms because they equal 1.

By Bayes' rule, the posterior probability of any particular structure π is then given by

$$P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \mathbf{x}, \psi, \theta) = \frac{P(\boldsymbol{\alpha}|\mathbf{x}, \boldsymbol{\pi}, \psi)P(\mathbf{x}, \boldsymbol{\pi}|\theta)}{\sum_{\hat{\pi}} P(\boldsymbol{\alpha}|\mathbf{x}, \hat{\boldsymbol{\pi}}, \psi)P(\mathbf{x}, \hat{\boldsymbol{\pi}}|\theta)}.$$

Generative probability models for RNA structure prediction give us $P(\mathbf{x}, \boldsymbol{\pi}|\theta)$ directly (88). Indeed, Sükösd et al. (100) already introduced an inference equation much like the one above as a means of incorporating SHAPE data into a probabilistic method of RNA structure, PPFold (100). However, we need a bit more algebra for the thermodynamic folding model. Recall that the thermodynamic model gives us $P(\boldsymbol{\pi}|\mathbf{x}, \theta)$ via the Gibbs–Boltzmann equation:

$$P(\boldsymbol{\pi}|\mathbf{x}, \theta) = \frac{e^{-\Delta G(\boldsymbol{\pi})/\mathrm{RT}}}{\sum_{\hat{\pi}} e^{-\Delta G(\hat{\boldsymbol{\pi}})/\mathrm{RT}}}.$$

We need the joint probability (with $\mathbf{x}$), not the conditional probability (given $\mathbf{x}$), but we can expand $P(\mathbf{x}, \boldsymbol{\pi}|\theta)$ to $P(\boldsymbol{\pi}|\mathbf{x}, \theta)P(\mathbf{x}|\theta)$, and, because we are dealing with only a single given sequence $\mathbf{x}$, we can cancel the $P(\mathbf{x}|\theta)$ term out of the posterior probability equation. Similarly, the thermodynamic partition function $\sum_{\hat{\pi}} e^{\Delta G(\hat{\boldsymbol{\pi}})/\mathrm{RT}}$ is the same for both the numerator and denominator of the posterior probability equation, so it also cancels. This leaves us with the following posterior probability:

$$P(\boldsymbol{\pi}|\cdot) = \frac{P(\boldsymbol{\alpha}|\mathbf{x}, \boldsymbol{\pi}, \psi)\, e^{-\Delta G(\boldsymbol{\pi})/\mathrm{RT}}}{\sum_{\hat{\pi}} P(\boldsymbol{\alpha}|\mathbf{x}, \hat{\boldsymbol{\pi}}, \psi)\, e^{-\Delta G(\hat{\boldsymbol{\pi}})/\mathrm{RT}}}.$$

Because the denominator, summed over all possible structures $\hat{\boldsymbol{\pi}}$, behaves as a normalization constant with respect to any individual structure π, we can call it Z' by analogy to a partition function. Note, however that Z' differs from the thermodynamic partition function because it

includes the probability of the observed SHAPE data. We now take the logarithm of both sides because probability model calculations are generally done as sums of logarithms rather than as products of probabilities to avoid numerical underflows, yielding the following equation:

$$\log P(\boldsymbol{\pi}|\cdot) = \log P(\boldsymbol{\alpha}|\mathbf{x}, \boldsymbol{\pi}, \boldsymbol{\psi}) - \frac{\Delta G(\boldsymbol{\pi})}{\mathrm{RT}} - \log Z'.$$

Finally, if we are only interested in inferring the optimal (maximum probability) structure, we can drop the constant Z' and just maximize as follows:

$$\mathrm{argmax}_{\boldsymbol{\pi}} \log P(\boldsymbol{\pi}|\cdot) = \mathrm{argmax}_{\boldsymbol{\pi}} \left[\log P(\boldsymbol{\alpha}|\mathbf{x}, \boldsymbol{\pi}, \boldsymbol{\psi}) - \frac{\Delta G(\boldsymbol{\pi})}{\mathrm{RT}} \right].$$

Thus, to make the thermodynamic folding calculation a probing-directed calculation we only need to add the log probability of observing the probing data $\boldsymbol{\alpha}$ given the RNA structure.

As long as the parameterization $\boldsymbol{\psi}$ of the probing data likelihoods is factored such that it maps well onto the energy parameterization $\boldsymbol{\theta}$ of the folding model in terms of having similarly factored dependencies on elements of RNA structure and sequence context, then the above equation is readily implemented in the dynamic programming recursions of existing RNA structure prediction programs. For example, we might simply assume that the observed SHAPE data α_i for base i are independent of the sequence and depend only on what structural context i is in π_i. This context might be as simple as paired versus unpaired. Then, we add the appropriate $\log P(\alpha_i|\pi_i)$ term to the appropriate free energy parameter (scaled by 1/RT) at every step of the dynamic programming recursion that adds base i to a growing substructure, depending on whether i is unpaired or paired in that substructure term in the recursion.

In summary, this derivation suggests that an appropriate SHAPE pseudoenergy term for base i is $\Delta G'_i = \mathrm{RT} \log P(\alpha_i|\pi_i)$. This term should not be viewed as an energy at all; rather, it should be viewed as a log probability in a statistical inference approach. More sequence and structural context could easily be incorporated into this model as desired by relaxing any of the simplifying assumptions.

Ensemble Prediction

Deriving the ensemble-based approach proposed by Washietl et al. (116) in terms of statistical inference is also possible, but more difficult, so I only sketch the main issues here. In this case, we want to infer the posterior distribution over ensembles, as opposed to over just a single structure. Doing so is equivalent to inferring the posterior distribution $P(\boldsymbol{\theta}|\cdot)$, as opposed to $P(\boldsymbol{\pi}|\cdot)$, because, for a given sequence $\mathbf{x}$, the ensemble probabilities are completely determined as a function of the folding model parameters $\boldsymbol{\theta}$. As discussed in Reference 116, we might specify a prior distribution $P(\boldsymbol{\theta})$ by assuming that $\boldsymbol{\theta} \sim N(\hat{\boldsymbol{\theta}}, \tau^2)$, i.e., normally distributed perturbations around the standard Turner parameters $\hat{\boldsymbol{\theta}}$. The difficulty with this approach comes from the fact that the observed SHAPE data need to be treated as an ensemble average: $\langle \boldsymbol{\alpha}|\mathbf{x}, \boldsymbol{\theta} \rangle = \sum_{\pi} \boldsymbol{\alpha}(\boldsymbol{\pi}) P(\boldsymbol{\pi}|\mathbf{x}, \boldsymbol{\theta})$. Thus, the likelihood term $P(\langle \boldsymbol{\alpha} \rangle|\cdot)$ unfortunately becomes a nasty multiple integral over all possible unknown $\boldsymbol{\alpha}(\boldsymbol{\pi})$ vectors for all of the individual structures in the ensemble, subject to the constraint that their ensemble-weighted average equals $\langle \boldsymbol{\alpha} \rangle$. Under the simplifying independence assumptions that the SHAPE data α_i for each position depend on only a small number K of structural contexts for (x_i, π_i), we can obtain a tractable K-dimensional integration over those states (for example, $K = 2$ for $\pi_i =$ unpaired versus paired). All of this should be doable, resulting in a strongly grounded version of the Washietl et al. approach in which we can avoid directly comparing the SHAPE values α_i to the base-pairing probabilities z_i and instead utilize an empirical likelihood model for the observed SHAPE data.

CONCLUSION

A statistical inference approach for incorporating structure probing data is easily generalized beyond single sequence RNA structure prediction. More complicated RNA structure analysis problems, including algorithms for de novo conserved structure detection and for sequence/structure homology search (64, 122), also depend on scoring schemes that require inferring an unknown secondary structure. It would be straightforward to extend the approach described here to any of these methods. Essentially, one need only include an empirical log probability term for the observed probing data at each base *i*, given the unknown structural context into which an algorithm is trying to put the base. If transcriptome-wide structure probing data become readily available in a variety of organisms (37, 105), we can imagine using these experimental data systematically across a variety of tasks in the computational analysis of RNA structure (122).

Computational RNA sequence and structure analysis is a broad topic, and I have not done justice to many areas of it in this review. In particular, I have focused mainly on functional RNA analysis and discovery in multicellular eukaryotes, especially humans because so much current controversy about pervasive transcription and lncRNAs exists in this area. Arguably, however, the richest hunting grounds for new functional RNAs are not in multicellular eukaryotes but in bacteria, where small RNAs are used extensively for posttranscriptional regulation. There is an excellent body of literature on bacterial regulatory RNAs, but I lacked the space to delve into it here (25, 97).

SUMMARY POINTS

1. Functional RNAs are heterogeneous, and no one characteristic suffices to detect them all in an unbiased fashion.
2. Putative long noncoding RNAs (lncRNAs) are likely to be a heterogeneous population that includes analysis artifacts and transcriptional noise, but a subset of lncRNAs are well expressed and evolutionarily conserved.
3. One useful positive signal that helps distinguish many functional noncoding RNAs from other explanations is the presence of an evolutionarily conserved RNA secondary structure.
4. Genome-wide computational screens for regions of conserved secondary structure are a promising means of detecting functional structural RNAs (both RNA genes and *cis*-regulatory RNA motifs), but the false positive rates of current methods are too high.
5. RNA structure probing experiments help constrain the prediction of secondary structure, and these experiments have recently been adapted to systematic transcriptome-wide measurements, offering a way to increase the signal-to-noise ratio of any computational analysis that depends on inferring RNA structure.
6. Several proposed methods for probing-directed prediction of RNA secondary structure can be unified and rationalized using principles of probabilistic inference.
7. Using similar probabilistic inference principles, structure probing data could be used to quantitatively constrain and improve other computational analyses of RNA, including homology search, alignment, and conserved structure detection.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Thanks to Matt Hangauer, Christine Heitsch, Michael McManus, Martin Smith, and Ian Vaughn for providing data; to Michael Clark, Tanja Gesell, Jan Gorodkin, Ivo Hofacker, Zsuzsanna Sükösd, Eric Westhof, and Michael Zuker for feedback and discussions; to members of my laboratory including Fred Davis, Tom Jones, Seolkyoung Jung, Eric Nawrocki, Elena Rivas, and Travis Wheeler for critical comments on the manuscript; and to the hospitality of the beautiful Centro de Ciencias de Benasque Pedro Pascual in Benasque, Spain, where most of this article was drafted.

LITERATURE CITED

1. Anandam P, Torarinsson E, Ruzzo WL. 2009. Multiperm: shuffling multiple sequence alignments while approximately preserving dinucleotide frequencies. *Bioinformatics* 25:668–69
2. Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EG, et al. 2001. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.* 11:941–50
3. Aviran S, Trapnell C, Lucks JB, Mortimer SA, Luo S, et al. 2011. Modeling and automation of sequencing-based characterization of RNA structure. *Proc. Natl. Acad. Sci. USA* 108:11069–74
4. Babak T, Blencowe BJ, Hughes TR. 2005. A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. *BMC Genomics* 6:104
5. Babak T, Blencowe BJ, Hughes TR. 2007. Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics* 8:33
6. Bradley RK, Uzilov AV, Skinner ME, Bendaña YR, Barquist L, Holmes I. 2009. Evolutionary modeling and prediction of non-coding RNAs in *Drosophila*. *PLoS ONE* 4:e6478
7. Brunel C, Romby P, Westhof E, Ehresmann C, Ehresmann B. 1991. Three-dimensional model of *Escherichia coli* ribosomal 5S RNA as deduced from structure probing in solution and computer modeling. *J. Mol. Biol.* 221:293–308
8. Bussotti G, Raineri E, Erb I, Zytnicki M, Wilm A, et al. 2011. BlastR—fast and accurate database searches for non-coding RNAs. *Nucleic Acids Res.* 39:6886–95
9. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, et al. 2011. The reality of pervasive transcription. *PLoS Biol.* 9:e1000625
10. Cordero P, Kladwang W, VanLang CC, Das R. 2012. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry* 51:7037–39
11. Coventry A, Kleitman DJ, Berger B. 2004. MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl. Acad. Sci. USA* 101:12102–7
12. **Deigan KE, Li TW, Mathews DH, Weeks KM. 2009. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. USA* 106:97–102**
13. di Bernardo D, Down T, Hubbard T. 2003. ddbRNA: detection of conserved secondary structures in multiple alignments. *Bioinformatics* 19:1606–11
14. Ding Y, Lawrence CE. 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.* 31:7280–301
15. Dinger ME, Amaral PP, Mercer TR, Mattick JS. 2009. Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. *Brief. Funct. Genomic. Proteomic.* 8:407–23
16. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, et al. 2012. Landscape of transcription in human cells. *Nature* 489:101–8
17. Eddy SR. 2013. The ENCODE project: mistakes overshadowing a success. *Curr. Biol.* 23:R259–261
18. Eddy SR, Durbin R. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Res.* 22:2079–88

12. Pioneering paper on SHAPE-directed RNA secondary structure prediction.

19. Ehresmann C, Baudin F, Mougel M, Romby P, Ebel JP, Ehresmann B. 1987. Probing the structure of RNAs in solution. *Nucleic Acids Res.* 15:9109–28
20. ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74
21. Evans M, Hastings N, Peacock B. 2000. *Statistical Distributions*. New York: Wiley. 3rd ed.
22. Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, et al. 1986. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. USA* 83:9373–77
23. Gesell T, von Haeseler A. 2006. *In silico* sequence evolution with site-specific interactions along phylogenetic trees. *Bioinformatics* 22:716–22
24. Gesell T, Washietl S. 2008. Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics* 9:248
25. Gottesman S, Storz G. 2011. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb. Perspect. Biol.* 3:a003798
26. Gruber AR, Bernhart SH, Hofacker IL, Washietl S. 2008. Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics* 9:122
27. Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF. 2010. RNAz 2.0: improved noncoding RNA detection. *Pac. Symp. Biocomput.* 15:69–79
28. Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature* 482:339–46
29. Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. USA* 110:5498–503

30. Comprehensive survey and meta-analysis of human transcriptomic data from 127 different RNA-seq libraries.

30. Hangauer MJ, Vaughn IW, McManus MT. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet.* 9:e1003569

31. Hobbs EC, Fontaine F, Yin X, Storz G. 2011. An expanding universe of small proteins. *Curr. Opin. Microbiol.* 14:167–73
32. Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. 2008. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol.* 6:e255
33. Inoue T, Cech TR. 1985. Secondary structure of the circular form of the *Tetrahymena* rRNA intervening sequence: a technique for RNA structure analysis using chemical probes and reverse transcriptase. *Proc. Natl. Acad. Sci. USA* 82:648–52
34. Kageyama Y, Kondo T, Hashimoto Y. 2011. Coding versus non-coding: translatability of short ORFs found in putative non-coding transcripts. *Biochimie* 93:1981–86
35. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, et al. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296:916–19
36. Kapranov P, Laurent GS. 2012. Dark matter RNA: existence, function, and controversy. *Front. Genet.* 3:60

37. Describes parallel analysis of RNA structure (PARS), a transcriptome-wide application of RNA structure probing.

37. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, et al. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature* 467:103–7

38. Kladwang W, VanLang CC, Cordero P, Das R. 2011. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat. Chem.* 3:954–62
39. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* 294:853–58
40. Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294:858–62
41. Le SY, Chen JH, Currey KM, Maizel JV. 1988. A program for predicting significant RNA secondary structures. *Comput. Applic. Biosci.* 4:153–59
42. Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, et al. 2007. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* 131:174–87
43. Li F, Zheng Q, Ryvkin P, Dragomir I, Desai Y, et al. 2012. Global analysis of RNA secondary structure in two metazoans. *Cell Rep.* 1:69–82

44. Li S, Breaker RR. 2013. Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. *Nucleic Acids Res.* 41:3022–31
45. Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 27:i275–82
46. Low JT, Weeks KM. 2010. SHAPE-directed RNA secondary structure prediction. *Methods* 52:150–58
47. Lu ZJ, Turner DH, Mathews DH. 2006. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res.* 34:4912–24
48. Lucks JB, Mortimer SA, Trapnell C, Luo S, Aviran S, et al. 2011. Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-seq). *Proc. Natl. Acad. Sci. USA* 108:11063–68
49. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* 29:4724–35
50. Maenner S, Blaud M, Fouillen L, Savoye A, Marchand V, et al. 2010. 2-D structure of the A region of Xist RNA and its implication for PRC2 association. *PLoS Biol.* 8:e1000276
51. Matera AG, Terns RM, Terns MP. 2007. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol.* 8:209–20
52. Mathews DH, Disney DH, Childs MD, Schroeder JL, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA* 101:7287–92
53. Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288:911–40
54. Mathews DH, Turner DH. 2002. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.* 317:191–203
55. Mazo A, Hodgson JW, Petruk S, Sedkov Y, Brock HW. 2007. Transcriptional interference: an unexpected layer of complexity in gene regulation. *J. Cell Sci.* 120:2755–61
56. McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29:1105–19
57. McGinnis JL, Dunkle JA, Cate JH, Weeks KM. 2012. The mechanisms of RNA SHAPE chemistry. *J. Am. Chem. Soc.* 134:6617–24
58. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, et al. 2013. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495:333–38
59. Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM. 2005. RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* 127:4223–31
60. Mitra S, Shcherbakova IV, Altman RB, Brenowitz M, Laederach A. 2008. High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucleic Acids Res.* 36:e63
61. Moazed D, Stern S, Noller HF. 1986. Rapid chemical probing of conformation in 16S ribosomal RNA and 30S ribosomal subunits using primer extension. *J. Mol. Biol.* 187:399–416
62. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* 5:621–28
63. Mortimer SA, Weeks KM. 2007. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.* 129:4144–45
64. Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29:2933–35
65. Nordström KJ, Mirza MA, Almén MS, Gloriam DE, Fredriksson R, Schiöth HB. 2009. Critical evaluation of the FANTOM3 non-coding RNA transcripts. *Genomics* 94:169–76
66. Novikova IV, Hennelly SP, Sanbonmatsu KY. 2012. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Res.* 40:5034–51
67. Nussinov R, Pieczenik G, Griggs JR, Kleitman DJ. 1978. Algorithms for loop matchings. *SIAM J. Appl. Math.* 35:68–82
68. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420:563–73

65. A thorough reanalysis of FANTOM3's so-called noncoding RNAs, showing that they were contaminated with several artifacts.

69. Ouyang Z, Snyder MP, Chang HY. 2013. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res.* 23:377–87
70. Pang PS, Elazar M, Pham EA, Glenn JS. 2011. Simplified RNA secondary structure mapping by automation of SHAPE data analysis. *Nucleic Acids Res.* 39:e151
71. Parker BJ, Moltke I, Roth A, Washietl S, Wen J, et al. 2011. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Res.* 21:1929–43
72. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, et al. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* 2:e33
73. Peluso P, Herschlag D, Nock S, Freymann DM, Johnson AE, Walter P. 2000. Role of 4.5S RNA in assembly of the bacterial signal recognition particle with its receptor. *Science* 288:1640–43
74. Perreault J, Weinberg Z, Roth A, Popescu O, Chartrand P, et al. 2011. Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. *PLoS Comput. Biol.* 7:e1002031
75. Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.* 6:e1001236
76. Ponting CP, Belgard TG. 2010. Transcribed dark matter: meaning or myth? *Hum. Mol. Genet.* 19:R162–68
77. Quarrier S, Martin JS, Davis-Neulander L, Beauregard A, Laederach A. 2010. Evaluation of the information content of RNA structure mapping data for secondary structure prediction. *RNA* 16:1108–17
78. Rabani M, Kertesz M, Segal E. 2008. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc. Natl. Acad. Sci. USA* 105:14885–90
79. Rabani M, Kertesz M, Segal E. 2011. Computational prediction of RNA structural motifs involved in post-transcriptional regulatory processes. *Methods Mol. Biol.* 714:467–79
80. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499:172–77
81. Reichow SL, Hamma T, Ferré-D'Amaré AR, Varani G. 2007. The structure and function of small nucleolar ribonucleoproteins. *Nucleic Acids Res.* 35:1452–64
82. Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81:145–66
83. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129:1311–23
84. Riordan DP, Herschlag D, Brown PO. 2011. Identification of RNA recognition elements in the *Saccharomyces cerevisiae* transcriptome. *Nucleic Acids Res.* 39:1501–9
85. Rivas E, Eddy SR. 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* 6:583–605
86. Rivas E, Eddy SR. 2001. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2:8
87. Rivas E, Klein RJ, Jones TA, Eddy SR. 2001. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.* 11:1369–73
88. Rivas E, Lang R, Eddy SR. 2012. A range of complex probabilistic models for RNA secondary structure prediction that include the nearest neighbor model and more. *RNA* 18:193–212
89. Sakakibara Y, Brown M, Hughey R, Mian IS, Sjölander K, et al. 1994. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.* 22:5112–20
90. Salehi-Ashtiani K, Lupták A, Litovchick A, Szostak JW. 2006. A genomewide search for ribozymes reveals an HDV-like sequence in the human *CPEB3* gene. *Science* 313:1788–92
91. Seemann SE, Sunkin SM, Hawrylycz MJ, Ruzzo WL, Gorodkin J. 2012. Transcripts with in silico predicted RNA structure are enriched everywhere in the mouse brain. *BMC Genomics* 13:214
92. Serganov A, Nudler E. 2013. A decade of riboswitches. *Cell* 152:17–24
93. Silverman IM, Li F, Gregory BD. 2013. Genomic era analyses of RNA secondary structure and RNA-binding proteins reveal their significance to post-transcriptional regulation in plants. *Plant Sci.* 205:55–62
94. Smith CM, Steitz JA. 1998. Classification of *gas5* as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5′-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol. Cell. Biol.* 18:6897–909
95. **Smith MA, Gesell T, Stadler PF, Mattick JS. 2013. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.* 41:8220–36**

95. The most recent human genome-wide computational screen for conserved RNA structure detection.

96. Spitale RC, Crisalli P, Flynn RA, Torre EA, Kool ET, Chang HY. 2013. RNA SHAPE analysis in living cells. *Nat. Chem. Biol.* 9:18–20
97. Storz G, Vogel J, Wassarman KM. 2011. Regulation by small RNAs in bacteria: expanding frontiers. *Mol. Cell.* 43:880–91
98. Stricklin SL. 2006. *Noncoding RNA Genes in* Caenorhabditis elegans. PhD Thesis, Wash. Univ. Sch. Med.
99. Struhl K. 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* 14:103–5
100. Sükösd Z, Knudsen B, Kjems J, Pedersen CN. 2012. PPfold 3.0: fast RNA secondary structure prediction using phylogeny and auxiliary data. *Bioinformatics* 28:2691–92
101. Sükösd Z, Swenson MS, Kjems J, Heitsch CE. 2013. Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucleic Acids Res.* 41:2807–16
102. Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J. 2006. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.* 16:885–89
103. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, et al. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 329:689–93
104. Tycowski KT, Shu MD, Steitz JA. 1996. A mammalian gene with introns instead of exons generating stable RNA products. *Nature* 379:464–66
105. Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, et al. 2010. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods* 7:995–1001
106. Uzilov AV, Keegan JM, Mathews DH. 2006. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics* 7:173
107. van Bakel H, Hughes TR. 2009. Establishing legitimacy and function in the new transcriptome. *Brief. Funct. Genomic. Proteomic.* 8:424–36
108. van Bakel H, Hughes TR. 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* 8:e1000371
109. van Bakel H, Nislow C, Blencowe BJ, Hughes TR. 2011. Response to "the reality of pervasive transcription". *PLoS Biol.* 9:e1001102
110. Vasa SM, Guex N, Wilkinson KA, Weeks KM, Giddings MC. 2008. ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA* 14:1979–90
111. Wan Y, Qu K, Ouyang Z, Chang HY. 2013. Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. *Nat. Protoc.* 8:849–69
112. Washietl S, Findeiss S, Müller SA, Kalkhof S, von Bergen M, et al. 2011. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* 17:578–94
113. Washietl S, Hofacker IL. 2004. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.* 342:19–30
114. Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* 23:1383–90
115. Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA* 102:2454–59
116. Washietl S, Hofacker IL, Stadler PF, Kellis M. 2012. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res.* 40:4261–72
117. Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, et al. 2007. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.* 17:852–64
118. Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S. 2001. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.* 15:1637–51
119. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW Jr, et al. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* 460:711–16

101. One of few places where empirical distributions for SHAPE data have been shown.

105. Describes FragSeq, a transcriptome-wide application of RNA structure probing.

116. A conceptually different approach for SHAPE-directed structure prediction, focused on ensemble calculations.

120. Wei D, Alpert LV, Lawrence CE. 2011. RNAG: a new Gibbs sampler for predicting RNA secondary structure for unaligned sequences. *Bioinformatics* 27:2486–93
121. Westholm JO, Lai EC. 2011. Mirtrons: microRNA biogenesis via splicing. *Biochimie* 93:1897–904
122. Will S, Siebauer MF, Heyne S, Engelhardt J, Stadler PF, et al. 2013. LocARNAscan: incorporating thermodynamic stability in sequence and structure-based RNA homology search. *Algorithms Mol. Biol.* 8:14
123. Will S, Yu M, Berger B. 2013. Structure-based whole-genome realignment reveals many novel noncoding RNAs. *Genome Res.* 23:1018–27
124. Wilm A, Higgins DG, Notredame C. 2008. R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res.* 36:e52
125. Wilusz JE, Freier SM, Spector DL. 2008. 3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* 135:919–32
126. Wilusz JE, Spector DL. 2010. An unexpected ending: noncanonical 3′ end processing mechanisms. *RNA* 16:259–66
127. Workman C, Krogh A. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.* 27:4816–22
128. Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, et al. 1998. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry* 37:14719–35
129. Yao Z, Weinberg Z, Ruzzo WL. 2006. CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics* 22:445–52
130. Zappulla DC, Cech TR. 2004. Yeast telomerase RNA: a flexible scaffold for protein subunits. *Proc. Natl. Acad. Sci. USA* 101:10024–29
131. Zarringhalam K, Meyer MM, Dotu I, Chuang JH, Clote P. 2012. Integrating chemical footprinting data into RNA secondary structure prediction. *PLoS ONE* 7:e45160
132. Zhang Z, Huang S, Wang J, Zhang X, de Villena FPM, et al. 2013. GeneScissors: a comprehensive approach to detecting and correcting spurious transcriptome inference owing to RNA-seq reads misalignment. *Bioinformatics* 29:i291–99
133. Zheng Q, Ryvkin P, Li F, Dragomir I, Valladares O, et al. 2010. Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in *Arabidopsis*. *PLoS Genet.* 6:e1001141
134. Zuker M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* 244:48–52
135. Zuker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* 9:133–48

131. An approach for SHAPE-directed secondary structure prediction that contrasts to Reference 12.