The RNA structure alignment ontology

JAMES W. BROWN,[1] AMANDA BIRMINGHAM,[2] PAUL E. GRIFFITHS,[3] FABRICE JOSSINET,[4] RYM KACHOURI-LAFOND,[4] ROB KNIGHT,[5] B. FRANZ LANG,[6] NEOCLES LEONTIS,[7] GERHARD STEGER,[8] JESSE STOMBAUGH,[5] and ERIC WESTHOF[4]

[1]Department of Microbiology, North Carolina State University, Raleigh, North Carolina 27695, USA
[2]Thermo Fisher Scientific, Lafayette, Colorado 80026, USA
[3]Department of Philosophy and Centre for the Foundations of Science, University of Sydney, NSW 2006, Australia
[4]Architecture et réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, Strasbourg 67084, France
[5]Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder, Colorado 80309 USA
[6]Centre Robert Cedergren, Département de Biochimie, Université de Montréal, Montréal, Québec, H3T 1J4, Canada
[7]Department of Chemistry and Center for Biomolecular Sciences, Bowling Green State University, Bowling Green, Ohio 43403 USA
[8]Institut für Physikalische Biologie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

ABSTRACT

Multiple sequence alignments are powerful tools for understanding the structures, functions, and evolutionary histories of linear biological macromolecules (DNA, RNA, and proteins), and for finding homologs in sequence databases. We address several ontological issues related to RNA sequence alignments that are informed by structure. Multiple sequence alignments are usually shown as two-dimensional (2D) matrices, with rows representing individual sequences, and columns identifying nucleotides from different sequences that correspond structurally, functionally, and/or evolutionarily. However, the requirement that sequences and structures correspond nucleotide-by-nucleotide is unrealistic and hinders representation of important biological relationships. High-throughput sequencing efforts are also rapidly making 2D alignments unmanageable because of vertical and horizontal expansion as more sequences are added. Solving the shortcomings of traditional RNA sequence alignments requires explicit annotation of the meaning of each relationship within the alignment. We introduce the notion of "correspondence," which is an equivalence relation between RNA elements in sets of sequences as the basis of an RNA alignment ontology. The purpose of this ontology is twofold: first, to enable the development of new representations of RNA data and of software tools that resolve the expansion problems with current RNA sequence alignments, and second, to facilitate the integration of sequence data with secondary and three-dimensional structural information, as well as other experimental information, to create simultaneously more accurate and more exploitable RNA alignments.

Keywords: databases; ontology; sequences; structural bioinformatics

INTRODUCTION TO MULTIPLE SEQUENCE ALIGNMENTS

Alignments of RNA sequences allow us to identify functionally important regions and to trace the evolutionary history of related molecules by placing equivalent parts of different sequences at equivalent positions for ease of comparison. Alignments are usually represented as two-dimensional (2D) matrices. Rows in a sequence alignment represent individual sequences, and columns represent individual residues from different sequences that are thought to be related. Gap symbols indicate positions where a sequence lacks a residue that is present at corresponding positions of other sequences (either because of an insertion or deletion, or because only part of the sequence is available). All sequence alignments thus represent a series of implicit assertions: that the residues found in each column all correspond to one another in each of the different RNA sequences. The meaning of this correspondence relation can be that these residues are believed to occupy equivalent positions in the three-dimensional (3D) structure of the molecule, or that they are believed to be related by sequence homology (i.e., that the sequences have a common ancestor), or typically, both. We propose that these assertions of correspondence should instead be made explicitly and discriminately, and that the assignment of correspondence be made between blocks of residues and elements of higher order structure as well as individual residues, as appropriate

Reprint requests to: Eric Westhof, Architecture et réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, Strasbourg 67084, France; e-mail: e.westhof@ibmc.u-strasbg.fr; fax: 33-388-602218.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.1601409.

in each context. We demonstrate how these conceptual advances can improve the construction, interpretation, and usefulness of RNA alignments.

HOW RNA SEQUENCES AND STRUCTURES ARE ALIGNED IN PRACTICE

The practice of manually aligning diverse RNA sequences differs substantially from the "matrix of nucleotides" alignment paradigm, and can be enhanced by alternative methods of representing alignments. RNA sequences can be aligned on the basis of sequence similarity (i.e., primary structure), on the basis of shared patterns of secondary structure, by incorporating additional constraints imposed by the 3D architecture, or by some combination of these. For highly similar sequences, e.g., 5S rRNA (Pavesi et al. 1997; Gardner et al. 2005), an alignment based solely on sequence similarity will also correctly align higher-order structural features. However, because there are only four bases, the ability to produce good alignments by sequence similarity diminishes rapidly as sequence conservation decreases (Gardner et al. 2005). The underlying secondary structure then becomes an essential guide to alignment, as in the signal recognition particle (SRP) RNAs (Larsen and Zwieb 1991). Here, one aligns two columns simultaneously using covariation information, for example, to allow A–U and G–C Watson–Crick pairs to substitute for one another, while avoiding mismatches. Elements of the secondary structure that are shared by aligned molecules can thus serve as landmarks for alignment even in the absence of conserved sequences or similarity in the sequences as a whole, and can allow the alignment of more distantly related sequences, because the secondary structure evolves more slowly than the primary sequence. Rigorous alignments of distantly related RNA sequences typically require consideration of both sequence and secondary structure, and are best performed manually.

Secondary structure can be added to an RNA alignment using a base-pairing mask (a row containing matched pairs of parentheses to designate which columns are Watson–Crick base paired) (Fig. 1). We refer to sequence alignments containing the secondary structure as "secondary structure sequence alignments." The RNA secondary structure contains all pseudoknots and is a superset of the RNA 2D structure (the 2D structure is the nested set of Watson–Crick base pairs, excluding pseudoknots) (Haas et al. 1994; Massire et al. 1998). In order to annotate pseudoknots, matched-pair symbols other than normal parentheses must be introduced. The 3D architecture results from the assembly of the 2D structure elements (helices, hairpins, single-stranded regions) through tertiary interactions, and thus the resulting secondary structure represents all 3D helices present in the final architectural fold. In an ideal secondary structure sequence alignment, there is a precise one-to-one correspondence between pairs of columns (X,Y): if the residue in column X pairs with the residue in column Y in any one sequence in the alignment, then the residue in column X should pair with the residue in column Y in all sequences in the alignment (see Figs. 1, 2). Similar considerations apply to hairpin loops matching the specifications for a GNRA tetraloop and many other RNA structural features. These types of correspondence are a prerequisite to detailed phylogenetic or comparative structural analysis, and are also essential for inferring structures directly from alignments.

Thus, alignments are constructed by identifying sequence or structural elements that are common to some subset of the sequences, aligning the regions that clearly correspond to one another, aligning the resulting subalignments to one another, and identifying new features that are revealed as shared by the new alignment. Figure 3 illustrates some of these types of correspondences, and highlights examples of them in two distantly related RNase P sequences. This procedure differs radically from the automated procedure, as implemented in Clustal and related programs, of aligning pairs of sequences based on similarity in the primary sequence, building a matrix of pairwise distances between the sequences, and then building a multiple alignment by aligning the sequences and/or subalignments to one another.

This structural view of an RNA alignment also differs conceptually from the traditional sequence alignment based on a matrix of nucleotides. In this view, it is not just nucleotides that are being aligned, but also regions of nucleotides, base pairs, helices, and any other elements of

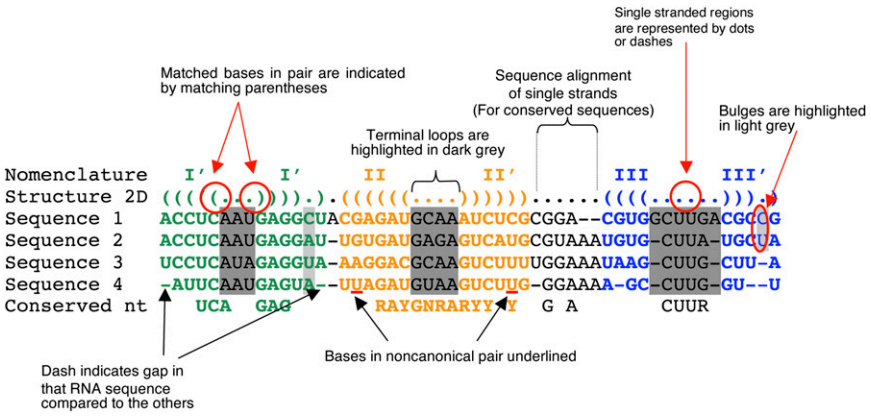

FIGURE 1. Abstract example of an RNA sequence alignment showing typical features. This simplified diagram shows many features common in sequence alignments, including representation of paired and unpaired regions, gaps, kinds of loops, etc. Some features can be conveniently represented using existing software. Others, such as noncanonical bases, cannot.

```
thousand 000-0000000-----------------0000-----------------0--00-00000000
hundred  000-0000000-----------------0000-----------------0--00-00000000
tens     112-2222222-----------------2233-----------------3--33-33333444
ones     890-1234567-----------------8901-----------------2--34-56789012
helices  J2/3<---------P3-5'------->..L3..<----------P3-3'---------->J3/4
pairing  ----((((((((((((((----(((((------)))))-----)))))))--))-))))----
pairing  ----ABCDEFGHIJKLMN----OPQRS------SRQPO-----NMLKJIHG--FE-DCBA----
Ssolfat  UAA-CGGGG-----------------CAAA----------------C-CCUGAGGA
Sacidoc  UUA-CGGGA-----------------AUA-----------------U-CCUGAGGA
Msedula  CCA-CGG-------------------GAAA-------------------CUGGGGA
Apernix  CCA-CGGCCCCCC-------------AGCCA---------------GGG--GG-GCUGAGGA
Pfurios  UGC-CGGGC-----------------UUUAU---------------G-CCCGAGGA
Tlittor  CCU-CGGGU-----------------AUUUG---------------A-CCCGAGGA
Mthermo  UGA-CGGUCCC---------------UCAA----------------G--GG-GCUGAGGA
MthMarb  UGA-CGGCCCA---------------UUUU----------------U--GG-GCUGAGGA
Mformic  UAC-CGGUUUCUAUAGAU---------UUAAU----------GUCUGUAGUUAA-ACUGAGGA
Tvolcan  UGA-CGCC-------------------GUAA-----------------GGUGAGGA
Mbarker  UGA-CGGGCC----------------UUCG-----------------GG-UCUGAGGA
Hcutiru  UGCCCGUGCC----------------GUGA-----------------GG-CAUGAGGA
Hvolcan  UCC-CGUGCCCG--------------AGA----------------CG--GG-CAUGAGGA
Hmorrhu  CAC-CGCGGCGUACC---GACAGGCAC-ACAC-GUGCCAGCG----GGUAC--GCACGCGAGGA
Ngregor  UGC-CGCGGGCGUC-------------GUGC--------------GACG--CG-CGCGAGGA
```

FIGURE 2. Example RNA sequence alignment. This example is helix P3 and the adjacent joining regions in RNase P RNA from representative Archaea. The first seven rows are annotations. Rows *1–4* are standard numbering, relative to the *Methanothermobacter thermoautotrophicus* RNA. Row *5* contains human-readable secondary structure labels. Columns are indicated in the second and third rows. Row *6* is the machine-readable base-pairing mask. Row *7* is a human-readable guide to the pairings specified in the previous row; column "A" pairs with "A," "B" pairs with "B," etc. The remaining rows are individual sequences; data taken from the RNase P Database (Brown 1999).

structure in the RNA. In this view, the nucleotides need not be considered (although they usually would be); it is the structures and the building blocks forming those structures that are being aligned.

LIMITATIONS OF CURRENT RNA ALIGNMENTS

The simple 2D matrix paradigm of sequence alignments has proven enormously useful, but is insufficient for today's massive sequence databases. We need large-scale integration of information regarding sequence, function, evolution, and structure in human- and machine-readable formats that facilitate reuse of data and knowledge. Organizational schemes are urgently needed for denoting correspondences between elements larger than individual residues (so that meaningful vertical slices of an alignment can be chosen for display) and for denoting relationships among the sequences themselves (so that meaningful horizontal slices of an alignment can be chosen for display—these slices might be discontiguous, such as to allow both halves of a putative helix to be displayed simultaneously). These issues are summarized in Table 1.

Indiscriminate assignment of correspondence, but only between residues, leading to horizontal expansion

In a traditional sequence alignment, every nucleotide in any column of the alignment is implicitly considered to correspond to all of the nucleotides in other sequences in that column. In regions of good sequence and structural conservation this is reasonable, but in regions of sequence or structural variation, the traditional alignment implies unreasonable nucleotide-to-nucleotide correspondence between all sequences. The proper approach to representing these regions in a traditional alignment is to use runs of consecutive gaps to isolate regions in which correspondence between sequences is not clear. However, this quickly results in unmanageable alignments dominated by gaps (in alignments of many RNAs, such as RNase P RNA, tmRNA, and MRP-RNA; these gaps make up the bulk of the alignment) (Schmitt et al. 1993; Brown 1999; Andersen et al. 2006). In RNase P RNA (see Figs. 2, 3), helices P3 and P12 are highly variable in both sequence and length, and although generally alignable between closely related species, the alignment of the individual nucleotides of these elements between more evolutionarily distant groups is probably not meaningful. In addition, there are numerous elements that are present in only some examples of these RNAs (e.g., P13, P14, P19), as well as alternative elements that have different structure but reside in the same region of the RNA (e.g., P6 versus P5.1). However, other parts of the alignment of these homologous sequences are meaningful at the primary sequence level, and we need to be able to capture and display this information.

Meaningful alignments also often cannot be assigned to the nucleotides in regions that vary in length, even if the corresponding regions are easily defined. For example, in RNase P (Figs. 2, 3), the loop L3 varies somewhat in length. Although it might be argued that the region of nucleotides that form the loop correspond in these different cases, it will usually be neither possible nor meaningful to specify structural correspondence with nucleotide-by-nucleotide resolution. Similarly, it is seldom clear which base pairs in a helix correspond across different sequences when the length (i.e., number of base pairs) of a helix varies. Nonetheless, the traditional alignment forces the user (whether human or machine algorithm) to assign correspondence on a per-nucleotide or per-base-pair basis.

These issues can be avoided by adopting an alignment approach in which correspondence between nucleotides can be assigned specifically where appropriate, and otherwise left undefined. It should also be possible to assign correspondence between regions of nucleotides, leaving the nucleotide-for-nucleotide correspondences unspecified.

Vertical expansion and organization

RNA alignments expand vertically due to the rapid growth in the number of sequences produced by high-throughput sequencing. When there are more than a small number of sequences, not all can be displayed at the same time nor be

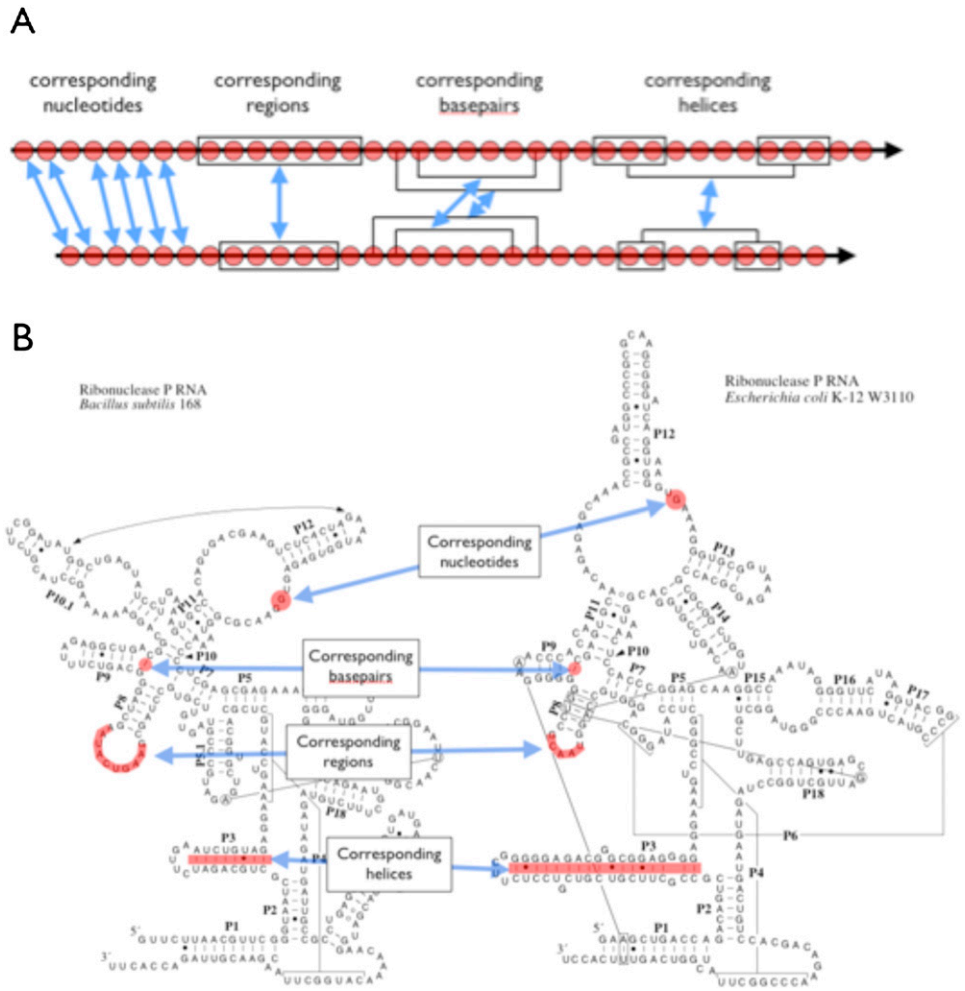

FIGURE 3. Example bacterial RNase P RNA secondary structures and correspondences. (*A*) The correspondence relationship between two conceptual RNA sequences; corresponding nucleotides (all that is possible in a traditional sequence alignment), corresponding regions, corresponding base pairs, and corresponding helices. (*B*) These types of relationships in the context of the secondary structure of RNase P RNA. Type B RNase P RNA is represented by that of *Bacillus subtilus* strain 168, and type A RNase P RNA is represented by that of *Escherichia coli* strain K12 W3110. Helices are numbered P1–P19 according to Haas et al. (1994). Taken from the RNase P Database (Brown 1999).

managed by the human user. The ability to scroll around in large virtual windows in current alignment editors such as BioEdit (Hall 1999) only partially alleviates the difficulty in visualizing all of the relevant data simultaneously to facilitate editing an alignment. Nor does the user typically want to display all of the sequences in an alignment. In order to selectively display relevant sequences, these would need to be organized hierarchically into groups—a taxonomy. In some cases, this taxonomy could be phylogenetic (e.g., rRNAs); in others, it could be structural (e.g., self-splicing introns). The user could then specify within each group whether to display all sequences, or only representative sequences, at whatever level desired. A key part of this functionality would be allowing the user to reassign sequences to new groups as the alignment and taxonomy are improved; this is especially true in cases where the groups are nonphylogenetic, but horizontal gene transfer can also make it essential to move sequences in ways that conflict with the organismal phylogeny.

Inability to include alignable, nonsequence information in alignments

Current RNA sequence/structure alignments cannot consistently annotate additional alignable, but nonsequence information in the alignment. This is information that belongs to specific regions of the alignment (i.e., sets of corresponding residues or groups of residues) and includes residue numbers, non-Watson–Crick pairing types and base-pairing partners, stacking interactions, backbone conformation, and other structural or statistical annotations such as helix designations, phylogenetic "weights," and consensus data. Another notable example of information that cannot be easily included is 3D architecture.

TABLE 1. Desired features and requirements for an RNA structure alignment ontology

Desired feature	Prerequisite
The ability to be specific about the assignment of correspondence relations	Definitions of the objects that can correspond and of the types of correspondence relationships that should be captured in the ontology
The ability to collapse the alignment horizontally	A robust annotation system for sets of corresponding elements
The ability to include alignable nonsequence information	Specifications for how nonsequence information should be attached to the alignment
The ability to collapse the alignment vertically	A method to organize and group sequences
Distinctions between different types of gaps	A reformulation of the notion of gaps, e.g., distinct types of gaps for indels and absent data

Currently, there are no accepted standards for attaching such annotations to an alignment; they are instead included in alignments as lines of nonsequence data in an ad hoc fashion (see Figs. 1, 2). Consequently, this information or its meaning is not available for reuse, because it is generally lost when the alignment is stored in one of the standard file formats currently defined. Developing methods to capture, store, and transmit all relevant information for reuse is thus a high priority, especially for integrating sequence and 3D data.

Ambiguity about the meaning of gap characters

A problematic aspect of gaps in traditional alignments is that missing data (e.g., from partial sequences or from regions of crystal structures with poor resolution) is often not distinguished from real insertions or deletions. Some alignments use alternative gap characters, such as periods or tildes, but the meaning of the characters is typically implicit and not transferable between programs. The solution is to dispense with the generic "gap," replacing it with distinct notions of "outside the range of available data" and "not present in the sequence."

A NEW VIEW OF ALIGNMENTS

An ontological perspective is required to resolve the problems discussed above and to open the way for truly integrative approaches to displaying, storing, and manipulating RNA sequence and structure data. This requires more than ontological definitions of traditional alignments, although this is useful and is underway (Thompson et al. 2005); instead, we suggest an entirely new view of the "alignment." This view provides the solution to both horizontal and vertical expansion by explicitly encoding the information that allows the user to selectively hide less important information and to determine the relative importance of various components of the data. The data must thus be annotated in detail in both the horizontal (sequence-specific) and vertical (position-specific) dimensions, perhaps with multiple annotations in each dimension.

The "correspondence" relationship

The purpose of an alignment is to designate elements in different molecules that correspond to one another, i.e., the designation of a relationship ("corresponds to") between various parts in two or more macromolecules, as defined in Table 2. In a traditional sequence alignment, these are the one-to-one correspondences between residues of different sequences implied by the fact that they are in the same column of the matrix. Our new view of an alignment defines an alignment as a set of correspondence relations, not necessarily between individual residues. Formally, a region of an RNA sequence can consist of a single nucleotide or of a set of nucleotides. Two regions correspond if they are annotated with the same correspondence relation (defined below). A set of regions corresponds if all pairs of regions in the set correspond with the same correspondence relation. Any given region always corresponds with itself. Correspondence relations are thus reflexive, symmetric, and transitive; they constitute equivalence relations that can partition a set into disjoint subsets or equivalence classes.

The most contentious aspect of this definition of an alignment is usually the choice of the term used to describe what we call "correspondences." Homology is an obvious possibility. The term "homology" was originally introduced as a rigorous way to express the observation that the same structure exists in modified forms in different species: "the same organ in different animals under every variation of form and function" (Owen 1843). With the general acceptance of the theory of evolution, "homology" has primarily been used to denote structures with a shared evolutionary ancestry (Table 2). As such, however, homology is something inferred rather than directly observable. More problematic are multiple appearances of "the same" recurrent motif within a single RNA molecule, where these instances may or may not be related to one another through duplication, and cases where "the same" motif has arisen independently. For example, structurally similar kink-turn motifs appear six times in the large subunit ribosomal RNA of *Haloarcula marismortui* (Klein et al. 2001), and the hammerhead ribozyme has evolved at least

TABLE 2. Definition of terms

Terminology	Definition
Correspondence	A relation between regions of an RNA alignment that can occur between molecules or within a molecule. These relations are reflexive, symmetric, and transitive.
Region	Consists of a single RNA nucleotide or a set of RNA nucleotides. Regions can be continuous spans of nucleotides or discontinuous collections of contiguous spans. Single base pairs, terminal loops, junctions, etc., are all examples of regions.
Homology	A correspondence that implies descent from a common ancestor with evolutionary continuity.
Similarity	A correspondence that can be defined in terms of a quantitative measurement, typically, at some structural level.
Sequence similarity	A similarity defined at the primary sequence level, e.g., 95% sequence identity.
Secondary structure similarity	A similarity defined at the secondary structure level, e.g., 50% of base pairs in common.
3D structure similarity	A similarity defined at the 3D structure level, e.g., 3 Å RMSD.
Base pairing	A relation between two RNA nucleotides, defined by base–base hydrogen-bonding interactions.
Function	The properties of a biological entity for which it is maintained by evolutionary selection.

three times: at least once in nature and at least once each from random-sequence pools in the Breaker and Szostak laboratories (Tang and Breaker 2000; Salehi-Ashtiani and Szostak 2001; Hammann and Westhof 2007). However, for some purposes (such as to define a sequence profile for matching the motif), we would want to align these kink-turn motifs or these hammerhead ribozymes based on shared structure and function, despite the fact that they share no common ancestor. Thus, calling all interesting correspondence in alignments ''homology'' would be misleading.

An alternative to homology is similarity, describing commonality that can arise either by common descent (homology) or convergence (analogy) (Table 2). Similarity is a useful term because it is directly observable (once the similarity metric, e.g., pairwise sequence identity or some other scoring scheme for sequence alignments, or a method of measuring distances among atomic coordinates or geometric features such as base planes, is defined) and meaningful for molecules that do not share ancestry, such as SELEX products or convergently evolved structures. However, objects resemble or differ from one another in indefinitely many ways and have no determinate degree of similarity, unless a specific similarity metric is chosen. The choice of a similarity metric must be justified by assumptions about which points of resemblance are relevant given the theoretical context. For example, the use of ''phenetic'' approaches in taxonomy, which were intended to free taxonomy from theoretical assumptions by grouping organisms based on raw similarity, failed because specific kinds of similarity are most useful for relating organisms to one another and because generic statistical measures of similarity tend not to converge on any underlying truth as more features are considered (Mickevich 1978; Panchen 1992; Griffiths 2007). Moreover, the term ''similarity'' suggests placing sequences on a continuum, whereas an alignment involves using similarity metrics to identify elements from different sequences as ''the same'' (e.g., placing them in equivalence classes). For both reasons, the term ''correspondence'' seems preferable.

Our relation of ''correspondence'' captures the fact that several different measures of similarity are relevant to an alignment. Each form of correspondence recognizes a kind of similarity which, at the appropriate level of focus, is relevant to the purposes for which alignments are constructed (e.g., investigating structure and function, reconstructing homology, etc.). These forms of correspondence are arranged hierarchically, so that portions of two sequences can be recognized as corresponding, while leaving open whether the parts that compose them correspond. Correspondence can either occur between molecules or within a molecule. Repetitions within a molecule, or ''serial correspondence,'' can either be due to duplication and divergence from a common ancestor (such as the ''serial homology'' attributed to paralogous genes or, at higher levels of biological organization, the repetitions of a developmental process such as repeated segments in an arthropod), or can be independently evolved (in the case of simpler motifs such as tetraloops). One key challenge in dealing with small RNA motifs is that convergent evolution to the same state (homoplasy) is common, and it may be impossible to determine in principle whether a particular correspondence is due to homology or convergent evolution because of insufficient statistical power.

The use of the term ''corresponds to'' retains the distinct notions of homology and different kinds of structural similarity (e.g., in the nucleotides and base pairs that make up the core hammerhead motif) as different types of correspondence. In many cases, both will apply; much of an alignment of ribosomal RNAs, for example, would represent both historical (homology) and morphological (structural_similarity) correspondences. In many cases, however, an alignment might contain distinct correspondences of each type.

Elements of RNA structure that can correspond

In order to be useful, the relationship "corresponds to" must be linked to objects—in this case, RNA elements that can "correspond to" one another in different instances of the RNA. In a traditional sequence alignment, the implicit object of this correspondence is nucleotides (or even gaps). In an RNA structure alignment, the elements involved would include nucleotides, but also should include other types of structural elements (Figs. 3, 4). This requires at least some ontology of RNA structure, which might usefully begin with a rudimentary ontology of RNA secondary structure.

In addition to nucleotides, this ontology should include regions, i.e., contiguous spans of nucleotides or discontiguous collections of these contiguous spans. Examples of such regions would be the "joining regions" between helices in a secondary structure, the 5′ and 3′ strands of these helices, and the hairpin loops capping helices. The nucleotides within corresponding regions may or may not be assigned correspondences individually, and nucleotide–nucleotide correspondences may be assignable between some RNAs and not others (even in cases where the regions correspond).

An RNA structure alignment also requires correspondence relationships between base pairs (including non-canonical base pairs) (Leontis and Westhof 2001), not just the nucleotides that comprise them, as defined in Table 2. The canonical base pairing of two regions of an RNA create a helix; like correspondences between regions, correspondence relations can be applied to helices whether or not the underlying base-pair correspondences can be assigned, and whether or not the helix is uniformly base paired. Note that a region can consist of a single nucleotide, and a helix can consist of a single base pair.

Types of correspondence

As mentioned above, the most common types of correspondence will be "homology" and "structural_similarity," each of which can involve a single base or base pair, a region or set of regions, a helix or set of helices, or any other collection of structural elements. In general, correspondence relations may be named; for example, in the RNase P sequences shown in Figure 3, a stem capped by a hairpin loop in both sequences is called "P12" and is related by both structural_similarity and homology (although individual bases in the loop and base pairs within the helix are not necessarily related to one another by either relationship). Within P12, we have loop and helix regions, illustrating the general principle that regions of correspondence can contain other regions of correspondence. Homology correspondences need not maintain structural relationships; for example, two sequences that are very similar and related evolutionarily might fold into different structures (an artificial example in which neighboring sequences fold into different structures comes from Schultes and Bartel 2000).

Structural similarity and homology are two important correspondences, and have received most attention thus far because they are two features that alignments are widely used to represent. However, as with any notion of similarity in science, correspondence relations rely on an underpinning theory about which features are important and which can be disregarded. For example, multiple sequence alignments are widely used to describe the interaction between an miRNA and its target, yet this relationship implies neither homology nor structural similarity and breaks several of the rules for correspondence relations (e.g., it is asymmetric and intransitive). However, the set of miRNAs that target the same mRNA site can meaningfully be considered to correspond to each other in the same way that base-pairing partners in a specified helix of different examples of a homologous RNA correspond to each other. In the case of miRNAs and their targets, the base pairing is in *trans* rather than *cis*, (i.e., the base pair is composed of nucleotides on separate RNA strands), which also occurs in many other RNA:RNA interactions. If there are multiple miRNAs that target the same region of an mRNA, these correspondence relations might be treated either as distinct pairings (different instances of a structure) or as alternatives in the same way that the pairings in an RNA with more than one alternative secondary structure would be treated (i.e., as different structures). A careful choice is thus required about which relationships are to be modeled by the correspondence relation, involving a trade-off between generality and convenience in the common cases.

THE ISSUE OF "GAPS"

In the view of a structure alignment presented here, the correspondence relations are separate from the form in

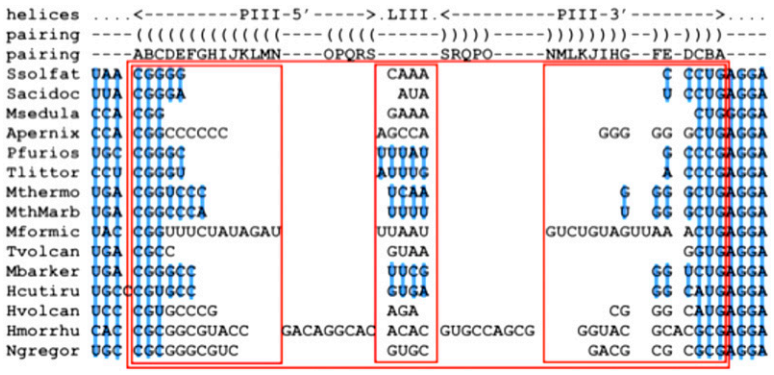

FIGURE 4. Example RNA sequence/structure alignment. This is the same alignment as shown in Figure 2 with explicit correspondence between nucleotides shown in blue and explicit correspondence between regions shown with red boxes. Correspondence relations between base pairs and helices are not displayed here. Note that indels (gaps) are not required.

which they are displayed. Some version of the traditional sequence alignment (for example, that used in Fig. 4) is only one way in which the correspondence relations could be displayed, but there are others one could imagine (for example, some elaborated version of Fig. 3). The gap (indel) becomes, in this view of an alignment, an optional visualization aid rather than part of the underlying information. Gaps corresponding to unavailable data (usually in partial sequences) would become ''regions'' of unknown length and sequence, perhaps with ''N's'' specified where required to complete base pairings to specific nucleotides (whose correspondence is specified) in regions of known sequence.

PHYLOGENETIC ANALYSIS

Phylogenetic trees constructed from molecular sequences, morphological traits, or any aspect of genotype or phenotype, are based on comparison of homologous elements. The ability to specify these (as one type of ''correspondence'') explicitly and specifically rather than indiscriminately, not only to nucleotides, but higher elements of structure as well, has the potential to greatly improve the quality of phylogenetic trees based on these alignments. The use of these alignments in phylogenetic analysis will, of course, require development of appropriate evolutionary models for evolutionary changes in nonsequence information and perhaps for mixtures of sequence and nonsequence changes. At the very least, trees can be improved by including only homologous nucleotides in the analysis in the same way that inclusion ''masks'' are currently used ad hoc.

THE NEED FOR FURTHER DEVELOPMENT

In order to make the conceptual advances presented here accessible to the broader RNA community, software needs to be created that allows one to encode, interpret, and visualize knowledge about RNA sequence and structure alignments. Many software libraries that provide core functionality such as reading and writing standard file formats are available in the public domain, but, although tools for working with RNA alignments such as Arb (Ludwig et al. 2004), S2S (Jossinet and Westhof 2005), or Colorstock/Sscolor/Raton (Bendana and Holmes 2008) are very useful, some features may be missing. Some of the visualization aspects required are the ability to (1) view and annotate helical, single-stranded, and unstructured regions (with or without gaps), insertions and deletions, incomplete (partial) sequences, and numbering schemes; (2) annotate structural features of all types; (3) collapse the view of the alignment horizontally by hiding less-interesting regions of the alignment according to the user's needs; and (4) organize the alignment on the basis of structural correspondence or phylogenetic relationships so that the view of the alignment can be collapsed vertically, either by hiding groups of sequences not of immediate interest or displaying only representatives from any group of sequences. Ultimately, this functionality would be embodied in an ontology-centric RNA alignment editor facilitating convenient editing and display of correspondence relations, definition of regions and assignment to different correspondence groups, redisplay of the alignment based on different priorities for correspondences (e.g., structural similarity versus homology), etc. Reuse of existing standard file formats is essential; for example, the alignment editor might store its sequences in FASTA, its trees as a collection of Newick-format strings, and its relations as a set of labeled sets of indices into the sequences.

Key to many of the desired features of an ontology-oriented RNA structure alignment editor is the ability to annotate features in the alignment. These features can be divided into two classes: (1) those that are specific to RNAs or clusters of RNAs (rows in a traditional alignment) and (2) those that are specific to individual or clusters of corresponding elements in the RNAs (columns in a traditional alignment). The former include features such as the names of individual RNA sequences and are already incorporated to some degree in all alignment file formats and alignment editors. Annotation of features that are related to sets of corresponding elements in many sequences is not currently incorporated into alignment editors in any useful way. Examples of this type of feature would include sequence and helix numbering schemes, base-pairing specifications, structural features, names, cross-linking sites, etc.

The utility of RNA structure alignments will also depend on a robust ontology of RNA secondary and higher-order structure, because it is these descriptions of the structures of RNAs—not just the sequences—that are to be aligned. Useful ontologies already exist for nucleotides (Eilbeck et al. 2005) and base pairs (Leontis and Westhof 2001). The fundamental organizing principle of RNA structure, however, is secondary structure, and so an ontology of RNA secondary structure is the highest priority. Informal descriptions of RNA secondary structure have existed for some time (e.g., Burke et al. 1987; Wyatt et al. 1989; Hendrix et al. 2005). These will need to be adopted into a formal ontological framework. From there, formal descriptions of RNA structure motifs (both local backbone configurations and tertiary ''modules'') can be added.

CONCLUSIONS

Solving the limitations of traditional RNA sequence alignments described above requires a new view of an ''alignment,'' the ''corresponds to'' relation, and the elements of RNA structure that can correspond to one another. This work, in conjunction with the existing RNA structure ontology efforts, will ultimately lead to an alignment ontology that enables the development of new representations of

RNA data and software tools to resolve the problems with current RNA sequence alignments, and to facilitate the integration of secondary and 3D structural and other experimental information to create more accurate and useful alignments. Here, we have proposed a prototype RNA correspondence relation to initiate discussion on how best to resolve these issues. In order for the perspective on RNA structure alignments outlined above to be useful, further development should be undertaken by RNA scientists in as broad a range of specialities as possible.

ACKNOWLEDGMENTS

The RNA Ontology Consortium (ROC) has been created to foster communication addressing issues involving the community of RNA scientists (http://roc.bgsu.edu) (Leontis et al. 2006). ROC is supported by a Research Coordination Network (RCN) grant from the National Science Foundation (grant no. 0443508), and its annual general meeting takes place as part of the RNA Society meeting, where it was initiated in 2004. This work was supported in part by the Human Frontier Science Program (RGP0032/2005-C [to E.W.]). We thank Paul Gardner and Alex Bateman at Rfam for their helpful discussion and work to ensure that the concepts described in this manuscript can be incorporated into the Rfam database.

Received February 15, 2009; accepted May 26, 2009.

REFERENCES

Andersen ES, Rosenblad MA, Larsen N, Westergaard JC, Burks J, Wower IK, Wower J, Gorodkin J, Samuelsson T, Zwieb C. 2006. The tmRDB and SRPDB resources. *Nucleic Acids Res* **34:** D163–D168.

Bendana YR, Holmes IH. 2008. Colorstock, SScolor, Raton: RNA alignment visualization tools. *Bioinformatics* **24:** 579–580.

Brown JW. 1999. The Ribonuclease P Database. *Nucleic Acids Res* **27:** 314. doi: 10.1093/nar/27.1.314.

Burke JM, Belfort M, Cech TR, Davies RW, Schweyen RJ, Shub DA, Szostak JW, Tabak HF. 1987. Structural conventions for group I introns. *Nucleic Acids Res* **15:** 7217–7221.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. 2005. The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biol* **6:** R44. doi: 10.1186/gb-2005-6-5-r44.

Gardner PP, Wilm A, Washietl S. 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res* **33:** 2433–2439.

Griffiths PE. 2007. The phenomena of homology. *Biol Philos* **22:** 643–658.

Haas ES, Brown JW, Pitulle C, Pace NR. 1994. Further perspective on the catalytic core and secondary structure of ribonuclease P RNA. *Proc Natl Acad Sci* **91:** 2527–2531.

Hall TA. 1999. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* **41:** 95–98.

Hammann C, Westhof E. 2007. Searching genomes for ribozymes and riboswitches. *Genome Biol* **8:** 210. doi: 10.1186/gb-2007-8-4-210.

Hendrix DK, Brenner SE, Holbrook SR. 2005. RNA structural motifs: Building blocks of a modular biomolecule. *Q Rev Biophys* **38:** 221–243.

Jossinet F, Westhof E. 2005. Sequence to structure (S2S): Display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics* **21:** 3320–3321.

Klein DJ, Schmeing TM, Moore PB, Steitz TA. 2001. The kink-turn: A new RNA secondary structure motif. *EMBO J* **20:** 4214–4221.

Larsen N, Zwieb C. 1991. SRP-RNA sequence alignment and secondary structure. *Nucleic Acids Res* **19:** 209–215.

Leontis NB, Westhof E. 2001. Geometric nomenclature and classification of RNA base pairs. *RNA* **7:** 499–512.

Leontis NB, Altman RB, Berman HM, Brenner SE, Brown JW, Engelke DR, Harvey SC, Holbrook SR, Jossinet F, Lewis SE, et al. 2006. The RNA Ontology Consortium: An open invitation to the RNA community. *RNA* **12:** 533–541.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G, et al. 2004. ARB: A software environment for sequence data. *Nucleic Acids Res* **32:** 1363–1371.

Massire C, Jaeger L, Westhof E. 1998. Derivation of the three-dimensional architecture of bacterial ribonuclease P RNAs from comparative sequence analysis. *J Mol Biol* **279:** 773–793.

Mickevich MF. 1978. Taxonomic congruence. *Syst Zool* **27:** 143–158.

Owen R. 1843. *Hunterian lectures: Lectures on the comparative anatomy and physiology of the vertebrate animals. Part I. Fishes*, p. 374. A. Spottiswoode, London, UK.

Panchen AL. 1992. *Classification, evolution, and the nature of biology*. Cambridge University Press, New York.

Pavesi A, Percudani R, Conterio F. 1997. A novel algorithm for the search of 5S rRNA genes in DNA databases: Comparison with other methods and identification of new potential 5S rRNA genes. *DNA Seq* **7:** 165–177.

Salehi-Ashtiani K, Szostak JW. 2001. In vitro evolution suggests multiple origins for the hammerhead ribozyme. *Nature* **414:** 82–84.

Schmitt ME, Bennett JL, Dairaghi DJ, Clayton DA. 1993. Secondary structure of RNase MRP RNA as predicted by phylogenetic comparison. *FASEB J* **7:** 208–213.

Schultes EA, Bartel DP. 2000. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science* **289:** 448–452.

Tang J, Breaker RR. 2000. Structural diversity of self-cleaving ribozymes. *Proc Natl Acad Sci* **97:** 5784–5789.

Thompson JD, Holbrook SR, Katoh K, Koehl P, Moras D, Westhof E, Poch O. 2005. MAO: A multiple alignment ontology for nucleic acid and protein sequences. *Nucleic Acids Res* **33:** 4164–4171.

Wyatt JR, Puglisi JD, Tinoco I Jr. 1989. RNA folding: Pseudoknots, loops, and bulges. *Bioessays* **11:** 100–106.