Published in final edited form as:

Q Rev Biophys. 2011 November ; 44(4): 433–466. doi:10.1017/S0033583511000059.

A new way to see RNA

Kevin S. Keating¹, Elisabeth L. Humphris¹, and Anna Marie Pyle^{1,2,*}

¹Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06511, USA

²Howard Hughes Medical Institute and Department of Chemistry, Yale University, New Haven, CT 06511, USA

Abstract

Unlike proteins, the RNA backbone has numerous degrees of freedom (eight, if one counts the sugar pucker), making RNA modeling, structure building and prediction a multidimensional problem of exceptionally high complexity. And yet RNA tertiary structures are not infinite in their structural morphology; rather, they are built from a limited set of discrete units. In order to reduce the dimensionality of the RNA backbone in a physically reasonable way, a shorthand notation was created that reduced the RNA backbone torsion angles to two (η and θ , analogous to ϕ and ψ in proteins). When these torsion angles are calculated for nucleotides in a crystallographic database and plotted against one another, one obtains a plot analogous to a Ramachandran plot (the η/θ plot), with highly populated and unpopulated regions. Nucleotides that occupy proximal positions on the plot have identical structures and are found in the same units of tertiary structure. In this review, we describe the statistical validation of the η/θ formalism and the exploration of features within the η/θ plot. We also describe the application of the η/θ formalism in RNA motif discovery, structural comparison, RNA structure building and tertiary structure prediction. More than a tool, however, the η/θ formalism has provided new insights into RNA structure itself, revealing its fundamental components and the factors underlying RNA architectural form.

1. Pseudo-torsions as reduced representations for RNA conformational space

1.1 The problem of RNA backbone complexity

The challenge of building an RNA structure is much like building a house from flexible rods that rotate in eight different places. Every time you attempt to shape the rod in one spot, it rotates in another spot, and you soon discover that it is impossible to build anything stable without reducing the dimensionality of the system and finding a new type of building material.

We first confronted this problem while attempting to build a three-dimensional model of a small hairpin-loop region (Domain 5, D5) within a self-splicing group II intron. This RNA

substructure is not highly complex, as it consists only of an extended RNA duplex that is capped by a GNRA tetraloop and interrupted in its center by an asymmetric, two-nucleotide bulge (Abramovitz *et al.* 1996). However, when we attempted to create an *ab-initio* model of D5 using the version of MC-SYM that was available at that time (Major *et al.* 1991), we obtained hundreds of distinct structural solutions (Duarte & Pyle, 1998). The program was, correctly, representing the fact that each nucleotide in an RNA structure can flex and rotate about seven individual torsion angles (α , β , γ , δ , ϵ , ζ , and χ) (Fig. 1a) and adopt one of at least two major sugar pucker configurations (C2'-*endo* and C3'-*endo*). The output from MC-SYM reflected the reality that each RNA building block has eight degrees of freedom. In a protein structure, each peptide building block has only two degrees of freedom: ϕ and ψ (phi and psi, Fig. 2a), thereby representing a significantly simpler geometry problem than the modeling of RNA.

When confronted with the numerous models for D5 structure, we decided to address the problem by visually examining each one. We found that, while each structure was slightly different, the models could actually be grouped into a few major structural categories. On visual inspection, members of these categories appeared similar, and although individual torsion angles within these groups varied, they compensated for one another resulting in similar overall shapes. This phenomenon had been previously documented on a smaller scale: early studies of RNA structure noted the 'crankshaft' effect, where compensatory rotations of torsion angles helped to maintain base-stacking interactions (Holbrook *et al.* 1978; Olson, 1982). As a result, we began to wonder whether the individual backbone torsion angles of RNA were, in fact, useful indicators or predictors of specific RNA structures.

Despite the wide ranges of standard torsions within our model structures, it was not clear as to whether this variation would actually be seen in real, crystallographically determined structures. We therefore examined whether standard torsion angles could uniquely describe the conformation of nucleotides within the most rigid form of natural RNA structure: the A-form helix. In order to approach this problem, we asked whether the empirically determined and generally accepted ranges for individual torsions that were published in the literature (Saenger, 1984) were observed in two high-resolution crystal structures of A-form RNA (Portmann *et al.* 1995), one of which was ultimately refined to 1.4 Å (Egli *et al.* 1996). Surprisingly, we observed that only a few helical nucleotides within the crystal structures fell within the standard torsion range defined previously for A-form RNA. In addition, we observed wide variation in standard torsion angles for A-form nucleotides within the crystal structures, suggesting that individual torsion angle ranges were not precise descriptors of discrete conformations (Duarte & Pyle, 1998). We recently repeated this examination using several helical RNA crystal structures solved at 1.2 Å or better, and the results matched our previous observations: A-form nucleotides contain a wide range of standard torsion values (Table 1).

Based on this experience, we sought a new way to describe RNA that reduced the dimensionality of RNA backbone configuration and did not implicitly depend on the identity of bases or pairings within a molecule. Nevertheless, we wanted the new formalism to accurately capture the discrete conformations that were empirically observed for whole

RNA structures and specific RNA motifs. Perhaps more importantly, we wanted to employ a description that was intuitively understandable (few humans can think in eight dimensions) and that computers could process rapidly. To this end, we developed the η/θ formalism for describing RNA backbone configuration.

1.2 The development of η and θ as descriptors of nucleotide conformation

Inspired by the simplicity of protein structure, we sought to determine whether the RNA backbone could be approximated by connecting sequential phosphates and sugars through a series of artificial rods (pseudo-bonds; Fig. 1*b*), of which there would be two per nucleotide. We predicted that the two torsion angles between these virtual bonds might provide a metric of conformational space, much like the function of ϕ and ψ in proteins (Fig. 2), and we hoped that the simplified, two-variable description would be comprehensible to humans and computers alike. Our approach was inspired by previous attempts to reduce the dimensionality of nucleotide conformation using tools such as modular blocks (Westhof *et al.* 1996), virtual bonds (Olson, 1975; Olson & Flory, 1972), and principal component analysis of correlations in standard torsion angles (Beckers *et al.* 1998).

When we used the phosphate and C4' atoms as anchor points for the virtual bonds, a plot of the resultant pseudo-torsion angles resulted in clusters of data (Fig. 3*a*) (Duarte & Pyle, 1998), as observed in a $\phi - \psi$ plot that is calculated from protein structures (Fig. 2*b*). Such clustering was not observed for pseudo-torsions resulting from other choices of backbone anchor points. After observing this clustering, we named these torsions η and θ , where η referred to the torsion of C4'_{i-1}, P_i, C4'_i and P_{i+1}, and θ referred to the torsion of P_i, CA'_i, P_{i+1} and C4'_{i+1} (Fig. 1*b*). The two-dimensional plot of these torsions was therefore named an η/θ plot. In 1998, when this work was done (Duarte & Pyle, 1998), the database of solved structures was small and there were few large RNA molecules that contained elements of complex tertiary motifs. Nonetheless, distinct nucleotide clusters could be identified in both the center of the η/θ plot (which turned out to represent the abundant nucleotides involved in A-form helices) and in various well-spaced regions of the plot (Fig. 3*a*, colored blocks).

Having observed that nucleotide conformers tend to cluster in η/θ space, it was important to determine whether there was any physical meaning to the co-localization of these points on the plot. A first challenge for this analysis was to define the spatial boundaries of each cluster, and to determine which points belonged inside these boundaries. We would then be able to analyze the clustered spots and determine if they represented structurally similar nucleotides in all-atom space. At the time of this analysis (Duarte & Pyle, 1998), the clusters contained an insufficient number of data points for application of a statistically rigorous clustering analysis (this was performed later, with a more mature database (Wadley *et al.* 2007), *vide infra*). We therefore chose to categorize only the most well-populated clusters and to use simple geometric shapes to visually define their boundaries (colored blocks, Fig. 3*a*). Once the clusters were designated, we then examined each spot within a clustered group, referring back to the original PDB file from which each set of coordinates was calculated (Fig. 3*b-i*). This analysis was made possible by the development of the first in a series of computational tools designed specifically to investigate RNA conformational space

using the pseudo-torsional formalism: AMIGOS (Algorithmic Method of Identifying and Grouping Overall Structure) (Duarte & Pyle, 1998).

Much to our surprise, given the radical reduction in dimensionality of the η/θ approximation, nucleotides within a given cluster shared very similar three-dimensional structures and often belonged to the same type of building block within known RNA motifs. For example, a pronounced group of co-localized points is found in the upper left region of the plot, centered at approximately η 225°, θ 30° (Fig. 3c) (Duarte & Pyle, 1998). The vast majority of nucleotides falling within this cluster are part of a sharp turn in which a new stacking axis is established on the 5'-side of the nucleotide. This is the conformation adopted by the second nucleotide within canonical U-turns and GNRA tetraloop motifs, and we therefore named this the 'stacked turn region'. A similar analysis was performed with eight different regions of the plot (Fig. 3b–i), resulting in the identification of discrete regions of η/θ space in which the vast majority of nucleotides represented the same type of RNA conformational unit. Each of these building blocks represents an obligate component of a specific RNA tertiary structural motif (Duarte & Pyle, 1998).

It was clear from this primitive analysis that the η/θ formalism represented a useful tool for analyzing and building RNA structures. It was also significant that one of the two anchor points chosen for the virtual bonds defining η and θ (the phosphorus) represents the atom that is the most easily recognized in electron density maps of RNA X-ray diffraction. Thus, we anticipated that the approach might one day be useful in building structures from crystallographic data. However, we realized that both structure building and further cluster analysis would require a considerably larger database of solved structures, which was not available at the time. In the meantime, the η/θ formalism was used as a platform for the development and use of new tools designed to explore the diversity of RNA conformational space (Beuth *et al.* 2005; Correll & Swinger, 2003; Giambasu *et al.* 2010; Huppler *et al.* 2002; Jovine *et al.* 2000; Keating *et al.* 2008; Scharpf *et al.* 2000; Sigel *et al.* 2004; Szep *et al.* 2003; Tamura & Holbrook, 2002), which was a burgeoning new problem in structural biology.

1.3 Other virtual bond systems

It is important to note that η and θ were not the first pseudo-torsional system for conceptualizing RNA conformation. Indeed, a phosphate- and sugar-atom-based virtual bond system was independently developed three separate times (Fig. 4). The first of these was published by Olson and Flory in 1972, employing a virtual bond system anchored at the phosphate and C5' atoms (Olson & Flory, 1972). Three years later, Olson published a simpler system that represented each nucleotide by a single virtual bond using only the phosphate atom (Olson, 1975). In 1980, Olson published her final virtual bond system, which employed the C4' atom in place of the C5', as this was better able to account for the effects of base-stacking (Olson, 1980) (Fig. 4a). Several months later, Malathi and Yathindra published an identical P-C4' system (Malathi & Yathindra, 1980) and conducted a number of analyses using this system throughout the early 1980s (Malathi & Yathindra, 1980, 1981, 1982, 1983, 1985) (Fig. 4b). Interestingly, one of these studies published in 1985 (Malathi & Yathindra, 1985) contained an analysis of pseudo-torsional values using an

$\omega'_v - \omega_v$ plot (Fig. 4c), which is conceptually identical to the $\eta - \theta$ plot presented above (Fig. 3a). When we first published the $\eta - \theta$ pseudo-torsions in 1998 (Duarte & Pyle, 1998), we were unaware of the Yathindra system and its application as a two-dimensional plot for analyzing tRNA structure. It is inherently significant that three separate research groups independently converged on the phosphate and C4' atoms as anchor points for a virtual bond system, as it underscores the robustness and utility of this specific methodological approach.

More recently, while adapting the η/θ formalism for crystallographic RNA model building (Keating & Pyle, 2010) the C4' atom was replaced by the C1' atom (resulting in the η' and θ' torsions, *vide infra*). Subsequently, another research group also independently determined that C1' was superior to C4' for this type of application (Gruene & Sheldrick, 2011; Keating & Pyle, 2010). Thus, for purposes of automated building of RNA into electron density, the P and C1' atoms appear to be the optimal anchors for a virtual bond system.

2. Validating and testing the η/θ formalism

After the publication of Duarte & Pyle (1998), the number of high-resolution RNA crystal structures climbed sharply. Many of these molecules contained complex architectural elements and provided much-needed information on the diversity of tertiary structural motifs (Fig. 5). This wealth of new information can be attributed, in part, to the successful crystallization and structural analysis of ribosomes and their subunits (Ban *et al.* 2000; Ramakrishnan, 2002; Schlutzenzen *et al.* 2000; Wimberly *et al.* 2000), which contain massive rRNA components (Fig. 5c). But other molecules such as self-splicing introns (Adams *et al.* 2004; Golden *et al.* 2005; Guo *et al.* 2004; Juneau *et al.* 2001), large and small ribozymes (Ferre-D'Amare *et al.* 1998; Kazantsev *et al.* 2005; Rupert & Ferre-D'Amare, 2001; Serganov *et al.* 2005; Torres-Larios *et al.* 2005) and riboswitch RNAs (Batey *et al.* 2004; Montange & Batey, 2006; Thore *et al.* 2006) were also structurally characterized at this time (Fig. 5a,b). These RNAs are particularly rich in complex tertiary contacts and their structures greatly expanded our understanding of conformational diversity in RNA molecules. By the beginning of 2006, these new RNA structures resulted in a greatly expanded database for analysis of η/θ space and they set the stage for rigorous evaluation of pseudo-torsional formalisms.

The new database not only provided more structures for calculating η/θ coordinates, it provided a larger set of higher-resolution structures. This allowed us to apply more stringent filtering criteria and to select the highest-quality files for inclusion in the database. As in our previous work, we avoided sample redundancy (e.g. only one tRNA^{phe} is included, despite the numerous examples in the protein data bank) and we now applied a resolution cut-off of 3.0 Å. For structures in this category, the positions of P and C4' can typically be reasonably well determined. Further selection criteria were imposed on individual nucleotides within the database, resulting in a final set of 7407 nucleotides (Wadley *et al.* 2007).

The η/θ coordinates were calculated for all nucleotides within the final dataset and results were displayed on a scatter plot (Fig. 6a). As predicted, clusters of nucleotides were more densely packed than we had observed previously (Duarte & Pyle, 1998), and additional clusters appeared (Wadley *et al.* 2007). However, there were still a number of points in this

scatter plot that did not fall into any clearly defined cluster. Additionally, despite the maturation of the structural database, we recognized that additional clusters were likely to exist. As a result, we sought a global method for analyzing structural similarity between any arbitrary set of proximal nucleotides on the η/θ plot.

2.1 A global correlation between η/θ and conformation

In order to examine the global correlation between η/θ coordinate co-localization and structural similarity as defined by RMSD (root mean square deviation), we chose random pairs of nucleotides from the dataset and plotted their distance from one another in η/θ space (in degrees) *versus* the RMSD of their backbone position (Fig. 7a). A striking linear relationship is observed (correlation coefficient $R^2 = 0.80$, $P \ll 0.001$), indicating that, irrespective of relative plot location, co-localized nucleotides share similar structures (Wadley *et al.* 2007). Surprisingly, the tight correlation is maintained even when the position of base atoms are included in the RMSD calculation (Fig. 7b). It is particularly illustrative to compare these results with a similar plot that is calculated using standard torsion angles. In the latter case, there is no correlation between RMSD and proximity of nucleotides in standard torsion space for all but the most similar of nucleotides (Fig. 7c,d; $R^2=0.50$). This analysis confirms that the η/θ pseudo-torsions, but not the individual backbone torsion angles, can serve as global descriptors of discrete RNA conformations.

2.2 Identifying and characterizing high-density regions of the η/θ plot

After confirming the global relationship between the pseudo-torsions and nucleotide conformation, we further examined the clustering in the η/θ plot. While the high-density regions of the plot were visually evident, they were not defined or circumscribed using standard statistical techniques. In order to identify the statistically significant regions of the plot and to establish their boundaries, we experimented with a series of window functions that convert the η/θ plot into a density function, thereby differentiating signal from noise (Wadley, 2006). The Blackman window function (a kernel smoothing technique commonly employed in astronomy (Kolb, 1980)) provided the best balance between efficiency and accuracy (Harris, 1978), allowing us to readily visualize and categorize the most highly populated regions of the plot (Wadley *et al.* 2007). The scatter plot was thereby transformed into a topological map of η/θ space, and each region of high density appeared as a proportionally sized peak on the plot (Fig. 6b).

Having delineated densely populated regions of the η/θ plot by purely statistical methods, we then sought to determine the extent to which nucleotides within each region were structurally similar at atomic resolution. The best way to make this determination was to conduct a direct RMSD superposition of nucleotides within each region, comparing the nucleotides with one another and with a prototype that was most representative of this region η/θ space. We wrote a script that enabled us to communicate directly between each pair of η/θ coordinates (e.g. a point on the η/θ plot) and the original location of a nucleotide within a specific PDB file, allowing us to call up individual sections of original structures and compare them to one another by superposition. From this analysis, we calculated a 'regional score', defined as the fraction of nucleotides that superimpose with an RMSD lower than 0.95 Å (Wadley *et al.* 2007).

2.3 The importance of sugar pucker as an additional variable

Examination of these scores revealed that a few regions of the η/θ plot represented overlapping sectors that were composed of two structurally distinct populations. In order to increase dispersion of the plot and differentiate these populations graphically, we set out to determine the structural feature that distinguished these conformers. In order to address the overlap issue, we filtered the constituents of each region by C2' and C3' *endo* sugar pucker based on the pseudo-phase angle of the furanose ring (Saenger, 1984) (later, we also applied other metrics to discriminate sugar pucker, *vide infra*). This analysis revealed that members of the overlapping regions differed in their sugar conformation, and were easily separated by constructing different η/θ plots for the two different major types of sugar pucker (Fig. 8a) (Wadley *et al.* 2007). With the regional scoring system established through RMSD comparison, and the sugar pucker differentiation in place, we were then able to define bona-fide structural clusters of nucleotides in η/θ space by setting a lower limit on the regional score (initially 70%).

This resulted in 11 spatially and structurally distinct clusters of nucleotides in η/θ space (Fig. 8b,c) (Wadley *et al.* 2007). In each cluster, the degree of structural identity increases as one moves to the center of each region, resulting in contour plots that resemble a geographic topographical map, where the center of the region approaches > 95%. The majority of data points fall within these spatially and structurally defined clusters, underscoring the finding that RNA conformation is highly discrete and that η/θ space is a reasonable proxy for all-atom structural space (Wadley *et al.* 2007).

We were able to assign the plot clusters to specific nucleotide conformations and to associate these with specific RNA tertiary structural motifs. These analyses allowed us to begin analyzing and utilizing RNA conformational space in new ways. For example, at this stage we began using sets of allowed η/θ coordinates to build and model RNA molecules (*vide infra*) (Wadley *et al.* 2007). In addition, we noticed that the base-planes of nucleotides that share η/θ space are almost identical, despite of the fact that χ and other descriptors of base location are not explicitly included in the η/θ formalism (Wadley *et al.* 2007). This led to us to investigate the role of backbone conformation in setting base orientation (*vide infra*).

3. The η/θ formalism as platform for innovation in RNA structural biology

Immediately after the first pseudo-torsional validation paper was published in 1998, the η/θ formalism was applied as a tool by structural biologists. Like the PROCHECK program for protein crystal structures (Laskowski *et al.* 1993), AMIGOS quickly became useful for building η/θ plots of new structures and scanning them for substructures that appear in unusual or potentially 'disallowed' regions. This provides the crystallographer with a method for double-checking a model and determining whether a section of RNA should be built differently (Beuth *et al.* 2005; Huppler *et al.* 2002; Jovine *et al.* 2000; Scharpf *et al.* 2000). However, the pseudotorsions are capable of significantly more than simply classifying individual RNA nucleotides. In the past decade, η and θ have been used in a wide variety of applications.

3.1 PRJMOS: an η/θ adaptation for RNA motif comparison and identification

When the nucleotides comprising a specific RNA structural motif (e.g. a GNRA tetraloop, A-platform, etc.) are diagrammed on an η/θ plot, it becomes apparent that the path between these points contains valuable information. The path is a 'signature' for a motif that is uniquely described by a string η/θ coordinates (Fig. 9a) (Duarte *et al.* 2003). While this is clearly a useful formalism, it becomes unintelligible once a motif or RNA substructure exceeds a certain size. This problem can be circumvented, and the information content within the path preserved, by incorporating nucleotide position as a third dimension of the plot (Fig. 9b) (Duarte *et al.* 2003). This serves to pull the string of η/θ coordinates out of the plane, connecting them along the sequence axis (Fig. 9b). The result is a computationally searchable roadmap of an RNA structure that contains all of the relevant information about constituent structural motifs. Also called an 'RNA worm', these maps can be created in seconds for any RNA structure, regardless of size, and they can be used to align and compare the structural features of RNA molecules.

Once an RNA substructure has been solved through crystallography or NMR (nuclear magnetic resonance), PRIMOS (Probing RNA Structures to Identify Motifs and Overall Structural changes), a computational tool designed to perform structural comparisons and motif searches using the pseudo-torsional formalism) can be used to create a characteristic 'worm' for the element, which serves as a unique identifier (Figs 9b and 10d). The worm is then used to search the entire database of RNA structures for identical or related motifs. In this way, a database of computationally searchable fingerprints for known tertiary structural motifs was created and used for structural analysis. For example, worms of the GNRA tetraloop and other common RNA structural elements were created and then used as 'bait' for screening the library of solved structures, which have themselves been converted to a linearized 'worm' formalism. The 'bait' worm is strung along the worm of an intact target structure, and their degree of conformational overlap at any given interval is continuously scored, thereby revealing structural matches between bait and target (Duarte *et al.* 2003). This approach allowed investigators to rapidly examine the database to determine the composition of motifs within their structures, to observe motifs in multiple structural contexts and to investigate how individual motifs contribute to larger substructures. In one such example, the hook-turn (Fig. 10c, initially identified from a crystallographic structural study of a ribosomal RNA loop E sequence motif) was demonstrated to be a recurrent motif by using its worm as 'bait' for screening the library of solved ribosome structures (Fig. 10d) (Szep *et al.* 2003). This exercise revealed the presence of four other hook-turn examples, which facilitated phylogenetic analysis of the family members and established the hook-turn as a building block within RNA tertiary structures (Szep *et al.* 2003).

There is no limit to the size of an RNA structure that can be analyzed by PRIMOS, which readily detects subtle structural changes that occur on ligand binding to RNA molecules even as large as the ribosome. Ligand-induced conformational changes are detected by creating worms of the bound and free RNA structures and then computationally passing them through one another using a scoring algorithm (Fig. 10a,b) (Duarte, 2002; Duarte *et al.* 2003). Conformational rearrangements that occur in response to ligand binding are

identified as sharp signals (Fig. 11), even in cases where direct structural superposition has failed.

For example, there are multiple structures of ribosomal subunits bound to diverse antibiotics, which have been reported to induce local structural changes at the sites of ligand binding (Brodersen *et al.* 2000; Klein *et al.* 2001; Nissen *et al.* 2000; Ogle *et al.* 2001; Ramakrishnan, 2002; Schmeing *et al.* 2002; Wimberly *et al.* 2000). These studies utilized conventional methods such as visual inspection and RMSD superposition, rendering it difficult to detect all of the structural changes induced by the antibiotics (Wadley, 2006). However, PRIMOS can rapidly assess the structural features within ribosomal RNAs in the presence and absence of antibiotics and readily detect ligand-induced conformational changes throughout the molecule, even at positions far from the ligand-binding site (Fig. 11) (Duarte *et al.* 2003). In one case, PRIMOS flagged nucleotides that act as a hinge between sections of 30S ribosomal RNA.

PRIMOS may have been successful because it is unbiased by the user input required for anchoring sites of superposition. Unlike superposition, subtle structural changes are not rendered statistically irrelevant by the vast majority of nucleotides that remain static. One might expect that this type of approach would be computationally expensive, but due to the linearized form of the data, it is not. A PRIMOS comparison of two ribosomal RNAs (each thousands of nucleotides in size) takes less than a minute.

In another application, PRIMOS provides a useful way to determine if a structural element is totally novel and represents the first discovery of an RNA building block. New structural features within the catalytic cores of group I and group II introns were identified in this way (Adams *et al.* 2004; Keating *et al.* 2008, 2010; Sigel *et al.* 2004; Strobel *et al.* 2004). PRIMOS has also been useful in classifying and differentiating motifs, as demonstrated by the discovery that there are two types of S-turns (Fig. 12): one class is superimposable and relatively common, while another class has an altered backbone architecture that distinguishes it from other examples (the S2 motif) (Duarte *et al.* 2003).

3.2 COM PADRES: an automated approach to motif discovery

While PRIMOS can be used to discover new motifs and substructures, its implementation requires an *a priori* search probe, in the form of an RNA worm. We wondered whether the η/θ formalism could be used to develop a completely automated approach for motif discovery that is unbiased by prior knowledge. It seemed plausible that the entire database of high-quality solved RNA structures could be sifted with worms representing known substructures, allowing novel units of architecture to be separated and characterized.

In order to develop this methodology, we took advantage of the fact that a worm of any size can be rapidly searched for content. Using a computational approach called COMPADRES (**Com**parative **Al**gorithm to **Dis**cover **Re**curring **E**lements of **S**tructure), we generated a set of massive worms by cutting up small, overlapping units of RNA structure and stringing them together in a linear series (Wadley & Pyle, 2004). We then created a comprehensive motif library using PRIMOS (Duarte *et al.* 2003). Members of this motif library were then used to scan the worms for matches, thereby establishing the known and unknown sectors of

worm architecture. Novel sectors were flagged using this approach, and they all represented previously undiscovered elements of RNA tertiary structure (Wadley & Pyle, 2004). These elements represent the first RNA building blocks ever to be identified without human input or by heuristic classification; rather, they were found using a completely automated, mathematical approach for structural discovery.

Although these results were inherently exciting, it was nonetheless essential to determine whether any of these new substructures qualified as a bona-fide motif. The generally accepted definition of a motif is a recurrent element of RNA structure, which must be observed more than once in the database of solved structures. In order to determine whether any new elements represented actual motifs, each novel sector was used as the search probe for a PRIMOS analysis (Wadley & Pyle, 2004). This allowed us to parse each of the novel sectors into one of two categories: totally unique substructures and recurrent motifs. The completely unique elements were simply classified and reported (Wadley, 2006), as they may be useful for structure prediction and modeling in the future. However, many of the novel substructures were common and recurrent, and thereby qualified as bona-fide motifs despite the fact that humans had never identified them by visual inspection (Wadley & Pyle, 2004). This is particularly remarkable given the striking structural complexity of the novel motifs, and the fact that all of them are found at sites that are critical for the function of their parent macromolecules. The success of the approach underscored the serious limitations imposed by human perception during structural analysis (Wadley, 2006), and the need to employ computational approaches for evaluating macromolecular features.

The motifs discovered by COMPADRES are generally new types of turns, although one appears to be a metal-binding motif (Wadley & Pyle, 2004). Perhaps the most structurally interesting motif identified in the study is the π -turn (Fig. 13), which induces an unusually tight kink in the RNA that forms a binding site for proteins (Wadley & Pyle, 2004). This motif has been observed in a duplicated form, resulting in a vase-like substructure that is decorated with extra-helical bases extending away from the backbone scaffold (Fig. 13d). Like the individual π -motifs, the duplicated form appears ideally suited for molecular recognition of amino acid side chains and other ligands.

Having discovered a library of new RNA motifs, it was of interest to evaluate the contexts in which they were found. It was expected that RNA substructures with similar form would be embedded in similar superstructures. However, a striking finding of this study was that a given motif can be identified within vastly different primary sequence and secondary structural contexts. The π -turn is a particularly striking example, as this element can form within single-stranded, double-stranded or junction RNA composed of diverse sequences (Fig. 13c). A similar conclusion was reached by Krasilnikov *et al.* who showed that similar architectural elements within RNase P RNA are derived from secondary structural regions that appear completely different without knowledge of tertiary structure (Krasilnikov *et al.* 2004). Taken together, these findings underscore the challenges that are inherent in RNA tertiary structural prediction: sequence and secondary structure alone do not yet provide sufficient information for predicting tertiary architectural forms. While progress is being made (*vide infra*), additional RNA tertiary structures are needed to fully dissect the components of individual RNA motifs and their interaction partners.

One might expect, particularly given their diverse secondary structural contexts, that motifs identified through COMPADRES do not contain members with identical base plane orientation. After all, η and θ do not contain explicit information about base location. Nonetheless, members of each motif family superimpose with an RMSD of ~ 1 Å, even when nucleobases are included in superimposition (Fig. 13b) (Wadley & Pyle, 2004). Visual inspection confirms that the bases (regardless of sequence) have almost identical locations within a motif family. This is consistent with previous observations that nucleotides with similar η and θ values share similar base plane location, underscoring the role of the RNA backbone in directing nucleobase location.

Methods such as COMPADRES provide a mechanism for learning as much as possible about RNA structural diversity, and for comprehensively mining the structural database for information about RNA architectural building blocks. But scientifically, the implications of the work are more significant than motif discovery: In the COMPADRES study, a computer method discovered a wealth of information about RNA without any human guidance or human perceptual intervention. This means that one can design tools that are capable of discovering new knowledge on their own, without carrying along the biases of the investigator, and that these tools can facilitate innovation in structural biology. In computer science, this is referred to as unsupervised learning. While the application of this technique to biological data is certainly not unique to our research (Golub *et al.* 1999; Quackenbush, 2001), its application to RNA structural data demonstrates its great potential.

3.3 Other tools for structural analysis using the η/θ formalism

In addition to the tools and techniques developed in our lab, η and θ have been key components of algorithms developed elsewhere. For example, the pseudo-torsions are used extensively in the Vfold model (Fig. 14), which has been used to study RNA folding and stability (Cao & Chen, 2005, 2006; Cao *et al.* 2010; Tan & Chen, 2008). This model represents a RNA molecule using a highly simplified system: all helices are modeled using idealized η and θ values, and loops are represented as P-C4' and C4'-P bonds on a diamond lattice, where each phosphate and C4' atom must fall exactly on a lattice point (Fig. 14a,b). The model can be used to accurately estimate entropy, which can then be used to predict RNA secondary structure and melting temperatures (Cao & Chen, 2005). Additionally, the Vfold model has been used to study the folding of loops (Cao & Chen, 2005) and pseudo-knots (Cao & Chen, 2006; Cao *et al.* 2010) (Fig. 14c,d), as well as the effect of salt on hairpin stability (Tan & Chen, 2008).

The pseudo-torsions also form the basis of the iPARTS server, which uses η and θ to align RNA tertiary structures (Wang *et al.* 2010). Traditionally, structures are aligned using RMSD; that is, by minimizing the distance between corresponding atoms in the two structures. However, this requires pre-determining which atoms are considered 'corresponding', which is not always straightforward with dissimilar structures. The iPARTS algorithm uses η and θ to identify corresponding regions of two structures (Wang *et al.* 2010). In order to accomplish this, the η/θ plot was divided into 23 different clusters, and a unique letter was assigned to each cluster. A structure can then be represented as a string of letters by assigning each nucleotide to a cluster based on its $\eta - \theta$ coordinates. In this

way, the pseudo-torsions are used to reduce a three-dimensional structure to a one-dimensional sequence of letters. Two such sequences can then be aligned using a variety of one-dimensional alignment techniques, which can easily find matching regions that may have been missed when examining the full three-dimensional structure. These alignments can then be used to assess structural and functional similarity.

3.4 RCrane: automated building of RNA structural models for crystallography

The above applications use η and θ to analyze existing structures by identifying structural motifs, predicting thermodynamic properties or performing structural alignments. However, the pseudo-torsions can also be used as a tool in crystallographic model building by helping to build RNA structure into electron density maps, which are the result of an X-ray crystallography experiment. One of the current challenges of RNA structural studies is the low resolution typical of RNA crystallography (Fig. 15), which leads to difficulty and errors in modeling the RNA backbone. This problem is exacerbated by the flexibility of the RNA backbone and the lack of computational tools for RNA modeling. In order to meet this challenge, we developed a technique for building the RNA backbone into low-resolution electron density maps using the RNA pseudo-torsions (Fig. 16b) (Keating & Pyle, 2010).

This technique uses slightly modified versions of the pseudo-torsions (Fig. 16a). While η and θ are defined using the phosphate and C4' atoms, this study employed an alternative set of pseudo-torsions, η' and θ' , which use the phosphate and C1' atoms (η' is the torsion about C1'_{i-1}–P_i–C1'_i–P_{i+1} and θ' is the torsion about P_i–C1'_i–P_{i+1}–C1'_{i+1}) (Fig. 16a). The η' and θ' torsions are more suitable when interpreting crystallographic density because the C1' atom is covalently bound to the nucleoside base and therefore can be more easily and accurately located within a low-resolution map (Gruene & Sheldrick, 2011; Keating & Pyle, 2010).

The building technique combines these modified pseudo-torsions with the consensus backbone conformer library (Richardson *et al.* 2008), which enumerates a limited number of allowed all-atom configurations for the RNA backbone (Fig. 17). This library therefore provides a set of discrete choices for fitting structure into electron density. It should be noted, however, that the conformers are defined using the suite division of the backbone rather than the traditional nucleotide division. While a nucleotide is centered about the ribose sugar and spans two phosphates, a suite is instead centered about the phosphate and spans two sugars (Figs. 1a and 16a) (Murray *et al.* 2003). Thus, a suite is equivalent to the first half of one nucleotide combined with the latter half of the previous nucleotide. As a result of this, this study interpreted the pseudo-torsions in (θ' , η') pairs that span a suite rather than (η' , θ') pairs that span a nucleotide (Fig. 16a).

As with previous studies of the C4'-based pseudo-torsions, we examined η' and θ' in the context of sugar pucker. As pucker is exceedingly difficult to determine directly from low-resolution electron density, the building technique determines pucker using the base–phosphate perpendicular distance (Davis *et al.* 2007; Murray, 2007), which requires only the phosphate and glycosidic bond coordinates to accurately assess pucker. When a θ'/η' plot is divided by pucker, the consensus backbone conformers show remarkably tight clustering (Fig. 16c) (Keating & Pyle, 2010). As a result of this correlation, the C1'-based pseudo-

torsions and the base–phosphate perpendicular distance can be used to predict backbone conformers, which can then be built into electron density.

The accuracy of these predictions was tested using jackknife validation with the RNA05 dataset (Richardson *et al.* 2008). This showed that the first (i.e. most likely) predicted conformer was correct 80% of the time for non-helical suites and 84% of the time for helical suites, and that one of the first three conformers was correct 97% of the time for non-helical suites and 98% of the time for helical suites (Keating & Pyle, 2010) (Fig. 18a). As a further test of this technique, two high-resolution crystal structures were rebuilt using only the published phosphate and base coordinates. The sarcin/ricin domain (PDB code: 1Q9A (Correll *et al.* 2003)) and guanine ribo-switch (PDB code: 2EES (Gilbert *et al.* 2007)) were both accurately rebuilt (Fig. 18 b,c). Between the two structures, 76 of the 88 suites with assigned conformers were correctly predicted and built using only the first conformer prediction, and in the majority of the remaining 12 suites, the mis-prediction caused only imperceptible changes in the rebuilt structure (Keating & Pyle, 2010). Additionally, for two suites of the guanine riboswitch, the rebuilt structure showed a noticeably better match to the electron density than did the original coordinates (Keating & Pyle, 2010). Thus, this technique can produce an accurate all-atom representation of the molecule even when starting from an imprecise low-resolution density map.

Independent of our research, Gruene & Sheldrick (2011) also used the C1 ‘-based pseudo-torsions to develop a promising approach toward automated building of RNA into electron density. They demonstrated a technique for automatically locating phosphates and bases within density maps, and they showed that η' and θ' could be used to assign order and connectivity to these bases and phosphates. The technique identifies phosphates by their strong, tetrahedral-shaped electron density, and finds bases by searching for large, planar ‘blobs’ of density (Gruene & Sheldrick, 2011). Thus, the combination of this method with our crystallographic model building technique (Keating & Pyle, 2010) holds great promise for fully automated building of RNA structure into electron density.

In addition to the current applications of these methodologies to X-ray crystallography, there is great potential for using η' and θ' to interpret cryo-electron microscopy results, which also produce electron density maps, albeit at lower resolution. Additionally, it would be interesting to investigate the applications of the pseudo-torsions toward the modeling of RNA in NMR. While NMR experiments do not produce electron density maps, the reduced dimensionality of the pseudo-torsions could prove useful in modeling RNA to match experimental spectra.

3.5 RNA structure prediction: building RNA *in-silico* using pseudo-torsions

A second application of the η and θ formalism in RNA structural modeling is the prediction and building of RNA structures *de novo*. The modeling of RNA in three dimensions is a challenging problem, complicated by its highly charged backbone and inherent backbone flexibility. Approaches to RNA modeling can typically be classified as all-atom (Das & Baker, 2007; Parisien & Major, 2008) or coarse grained (Ding *et al.* 2008; Flores *et al.* 2010; Jonikas *et al.* 2009b), depending on whether models of RNA are built in full atomic detail or whether each nucleotide is modeled in a simplified representation using pseudo-

atoms (see Shapiro *et al.* (2007) for a review). All-atom approaches have been successful in the modeling of small RNAs, but tend to scale poorly with structure size, making modeling of structures larger than tRNA difficult. In contrast, coarse-grained approaches can model significantly larger RNA structures, but result in much lower-resolution detail than seen with all-atom modeling. While advances have been made in bridging the two scales of RNA modeling (Jonikas *et al.* 2009a), approaches that combine the detail and accuracy of all-atom modeling of RNA with the speed and freedom from size constraints afforded by coarse-grained modeling are still needed.

The η/θ formalism may be especially well suited for combining the benefits of a reduced representation of RNA and all-atom detail. The η/θ notation is simplistic, involving only two atoms per nucleotide (C4' and P), but nevertheless can identify RNA structural features and motifs in 3D, a task that had previously required direct structural superimposition. The fact that the simplified pseudo-torsions can identify and classify 3D structure highlights an important finding on a nucleotide by nucleotide level: the closer any two arbitrary RNA nucleotides are in pseudo-torsional space, the closer their structural similarity, as measured by RMSD (*vide supra*) (Wadley *et al.* 2007). This is true not only for the highly populated regions of the η/θ plot (Wadley *et al.* 2007) but also in the less dense regions of the plot containing highly unusual, non-A form conformations. Surprisingly, this structural similarity often extends to base positioning as well, such that nucleotides close in pseudo-torsional space are also found to have virtually identical base orientations. Taken together, this suggests that nucleotides with similar η and θ values may be, to some degree, structurally interchangeable and that $\eta - \theta$ values might be useful as a shorthand proxy for specific conformational states of RNA.

In order to test these ideas, we asked whether realistic strands of RNA could be built *in silico* using only the pseudo-torsional formalism. A random selection of 500 strands, each 10-nucleotides in length, were chosen from solved crystallographic structures and η/θ values were calculated for each nucleotide in the strand. The 500 RNA strands were then created *in-silico* by computationally joining nucleotides with similar η/θ values from other solved structures (Fig. 19) (Wadley *et al.* 2007). Impressively, using pseudo-torsional information alone, the average pair-wise RMSD between the 500 models and starting 10-nucleotide strands was 1.7 ± 1.0 Å. In contrast, the RMSD observed when nucleotides with random η/θ values were selected and joined *in-silico* was 5.5 ± 2.5 Å. High model-building accuracy was observed even when the η/θ formalism was used to build specific RNA structural motifs such as a GNRA tetraloop (Fig. 19a) and a RNA bulge region (Fig. 19b) (Wadley *et al.* 2007). Importantly, these modeled motifs were built entirely from nucleotides that had not been found in naturally occurring contiguous instances of the specific motif being built.

These initial attempts at building RNA strands *in silico* allowed substitution of any nucleotide within a structural dataset for any other, based solely on their respective distance in pseudo-torsional space. However, these results could be extended more generally in the future by creating a library of discrete RNA conformations by systematically varying η and θ values (stepping by 10 or 20 degrees at a time, for example) and selecting a single conformational representative from all structurally solved instances. Such a process should

generate a set of discrete RNA conformations spanning the full diversity of nucleotide conformational accessibility for use in RNA structural modeling.

4. What we learn about RNA by looking through the η/θ lens

In addition to the direct applications of the RNA pseudo-torsions, such as motif finding and structure building, the studies of η and θ have revealed a great deal about RNA structure itself.

4.1 RNA conformation is not dictated by sterics alone

Our initial observation that models containing vastly different standard torsions could be grouped into similar structural categories suggests that bigger forces are at work in constraining RNA structure than simple sterics and geometry. We also observed that the distribution of nucleotides on an η/θ plot could not be explained purely by steric constraints. For example, Wadley calculated nucleotide structures using only steric constraints and plotted these nucleotides in η/θ space (Wadley, 2006). The distribution of nucleotides on this η/θ plot was different from the distribution of nucleotides taken from published crystal structures. A number of η/θ regions that were allowed sterically were almost entirely unpopulated by real nucleotides, indicating that RNA structural forms are dictated by many additional factors beyond sterics. A similar phenomenon is observed when sterically allowed and disallowed regions are calculated using the standard torsions: clusters of observed nucleotides do not always overlap directly with the regions predicted to be sterically optimal (Murthy *et al.* 1999).

This is completely different from what is seen in protein structure. For a Ramachandran plot (Ramachandran *et al.* 1963), (ϕ , ψ) coordinates can be divided into allowed and disallowed regions on the basis of steric constraints alone (Ramakrishnan & Ramachandran, 1965). Disallowed regions represent backbone structure that would certainly contain steric clashes for non-glycine and non-proline amino acids. For nucleotide structures, however, structural constraints must include other aspects of the folded RNA environment, such as base-stacking and base-pairing, electrostatic terms, hydration and other energetic constraints that even now have not been adequately captured computationally (Wadley *et al.* 2007).

4.2 There are limited number of basic conformational units in RNA structure

The large number of backbone torsions (Fig. 1a) would seem to indicate that the RNA backbone is infinitely flexible; however, this is clearly not the case. In η/θ space, there are only 11 highly populated clusters within the existing database (Fig. 8c) (Wadley *et al.* 2007). Additionally, the consensus backbone conformer library (Fig. 17) contains only 46 allowed backbone suite configurations (Richardson *et al.* 2008). Thus, at two very different levels of detail, RNA backbone structure is sharply constrained. While the RNA backbone allows more freedom than the protein main chain, RNA tertiary structure must still exist within the constraints imposed by the backbone.

4.3 The link between backbone and base

The pseudo-torsions clearly reveal a strong link between backbone conformation and base location. For example, η and θ are highly accurate predictors of RMSD (i.e. structural similarity) even when base atoms are considered (Wadley *et al.* 2007) (Fig. 7b). This relationship between backbone and base is not readily apparent when the standard torsions are examined (Fig. 7d). Furthermore, most of the clusters in η/θ space still show strong structural similarity even when base atoms are considered. Thus, even though these clusters were defined entirely on the basis of backbone conformation, the clustering still reveals information about base location (Wadley *et al.* 2007).

This link is further confirmed by the ability to rebuild realistic structures, with accurate nucleobase positioning, using only the pseudo-torsions (*vide supra*) (Wadley *et al.* 2007). In addition, a biopolymer chain elasticity (BCE) model has also shown accurate backbone structures of simple RNA hairpins can be reproduced without consideration of base-base interactions (Pakleza & Cognet, 2003; Santini *et al.* 2003). The BCE model assumes that the nucleic acid backbone behaves as a flexible thin rod and demonstrates that the structural constraints imposed by the end of the stem are sufficient for determining loop structure. This ability to accurately build RNA structure using only backbone information, whether it be pseudo-torsions or end conditions shows that RNA conformation is not driven solely by interactions between bases.

4.4 The importance of quality filters when studying the RNA backbone

In any scientific study, assuring the quality of the data is an obvious concern, but this task is especially important and especially difficult when examining the RNA backbone. Most large RNA structures are solved using X-ray crystallography. However, due to the low resolutions typical of RNA crystallography and the lack of computational tools, backbone modeling errors are unavoidable. Similarly, NMR studies of RNA frequently provide incomplete information about the specifics of backbone structure (Furtig *et al.* 2003). As a result, stringent quality filtering is commonly applied to studies of all-atom backbone structure (Murray *et al.* 2003; Richardson *et al.* 2008). However, these filters are still important when considering η and θ , even though the pseudo-torsions examine the backbone at a lower level of detail. As shown in Fig. 20, the differences in η/θ plots constructed with different filtering stringencies are readily apparent, with the stronger filtering resulting in a plot with sharper distinctions between favorable and unfavorable regions.

This filtering is not without caveats, though. Even with the enormous increase in RNA structural data over the past decade, the quantity of data is still far more limited than what is available for protein structure. The application of these quality filters further decreases the amount of data. For example, confining datasets to structures approaching atomic resolutions is a laudable goal. Indeed, a popular protein side-chain rotamer library (Lovell *et al.* 2000, 2003) was constructed using structures of 1.7 Å resolution or better. However, applying a similar criterion to RNA structure would result in an unusably small dataset. Thus, a careful balance must be struck between data quantity and quality.

4.5 The importance of multiple descriptors of the RNA backbone

The pseudo-torsions and the consensus backbone conformers (Richardson *et al.* 2008) describe RNA structure at two very different levels of detail. The pseudo-torsions offer information about the general path of the backbone, while the backbone conformers describe structure at an all-atom level of detail. However, these two systems provide a complementary pair of descriptors. The existence of the backbone conformer library does not reduce the utility of the pseudo-torsions, as the lower level of detail provided by the pseudo-torsions proves useful in a number of situations such as motif searching or interpreting low-resolution maps. In these cases, additional details frequently obscure the useful information. Additionally, the pseudo-bonds used to calculate η and θ span the backbone between three neighboring bases. Therefore, the pseudo-torsions describe the local 'context' of a nucleotide and include information about backbone orientation on the 5' and 3' sides (Wadley *et al.* 2007). In particular, this contextual information helped in the development of the consensus conformer library, as θ/η were used to confirm differences between similar conformers (Keating & Pyle, 2010; Richardson *et al.* 2008).

Conversely, the simplicity of the pseudo-torsions does not reduce the utility of the consensus backbone conformers. The all-atom detail contained within conformers is necessary when examining individual hydrogen bonds and other interactions. This type of information is frequently necessary to understand the structural details of a specific motif. Perhaps more critically, this level of detail is necessary for understanding the chemistry carried out by ribozymes. Because of this all-atom detail, accurate backbone conformers are the ultimate goal of any *de novo* modeling or crystallographic building technique.

4.6 The complementarity of backbone and base descriptors

While descriptors of the backbone are crucial to the study of RNA, the details of base-pairing and -stacking are also critical for RNA folding, structure and function. Thus, descriptors of nucleobase structure are an obvious complement to the η/θ and consensus conformer systems. For example, the Leontis–Westhof nomenclature for base-pairing is invaluable when examining RNA secondary and tertiary structure (Leontis & Westhof, 2001; Lescoute & Westhof, 2006). By combining descriptions of both the backbone and the base, it is possible to give a thorough characterization of the structure of a region of RNA. Such characterizations can provide a complete picture of RNA molecules and motifs.

5. Analogous approaches in the protein world

The pseudo-torsions η and θ allow the RNA backbone to be described in a way that is highly analogous to the backbone torsions ϕ and ψ in proteins. Thus, it is not surprising that the development η/θ tools to analyze, classify and build RNA structure have parallels from four decades of research in the protein world. Uses of the η/θ formalism have included structure quality evaluation (AMIGOS) (Duarte & Pyle, 1998; Wadley *et al.* 2007), motif identification (PRIMOS) (Duarte *et al.* 2003), motif discovery (COMPADRES) (Wadley & Pyle, 2004) and structural modeling building (RCrane) (Keating and Pyle 2010). Analogous applications based on ϕ and ψ have been developed for each of these tasks for proteins as well.

The program PROCHECK (Laskowski *et al.* 1993) can be used to flag amino acids within newly determined protein structures with abnormal ϕ and ψ torsions that might have been incorrectly refined. Protein backbone ϕ and ψ torsional ranges can be used to identify amino acids likely to be found in differing conformational states, such as alpha-helices and beta sheets (Ramachandran *et al.* 1963). Based on these torsional ranges, several programs were subsequently developed to identify and classify loops (Oliva *et al.* 1997; Venkatachalam, 1968; Wintjens *et al.* 1996) as well as larger motifs in proteins (Hutchinson & Thornton, 1996; Kato & Takahashi, 1997). Armed with the knowledge that the protein backbone torsions ϕ and ψ could be directly related to the basic building blocks of protein structure, the protein field also began searching for and classifying new and existing protein substructures, protein domains and protein families. Several tools were developed to aid in searching protein databases and discovering new substructures and the relationships between them (Holm & Sander, 1994; Orengo *et al.* 1994). Finally, the protein torsions ϕ and ψ have been useful in developing libraries of discrete, backbone-dependent rotameric conformations of individual amino acids (Dunbrack & Karplus, 1993; Lovell *et al.* 2000, 2003; Ponder & Richards, 1987), the use of which has led to great successes in the protein modeling world (Dahiyat & Mayo, 1997; Kortemme *et al.* 1998; Kuhlman *et al.* 2003). While comparable successes have yet to be achieved in the modeling of RNA, the development of crystallographic and *de novo* modeling applications using the η/θ formalism is currently an active area of research, both in our lab and others.

Despite the many similarities in the application and development of tools using the η/θ pseudo-torsions for RNA and ϕ and ψ torsions for proteins, it is worth pointing out an important difference between the $\eta - \theta$ plot for RNA and the comparable $\phi - \psi$ Ramachandran plot for proteins. Ramachandran plots were initially determined, in part, by modeling the conformations theoretically allowed by the steric constraints of polyalanine and shown to correlate with energetically allowed states of the protein backbone (Ramachandran *et al.* 1963). Such a clear link between sterically allowed regions of the $\eta - \theta$ plot and empirically observed RNA conformations has not been established (*vide supra*) (Wadley, 2006) and, to date, the relationship between force-field conformational energies and the $\eta - \theta$ torsions has not been determined.

6. Tool availability

Starting with our initial publication on pseudo-torsional space in 1998, all our computational tools have been freely available for other investigators to use. The simplicity of our initial AMIGOS scripts, written at the time in perl4, helped the η/θ formalism catch on and be adapted into the complete set of computational tools available for analysis today.

Our current set of η/θ tools are all freely available at <http://pylelab.org/software/>. The AMIGOS II program is the most comprehensive tool for structure analysis. It includes all the capabilities of AMIGOS and PRIMOS and presents them with a graphical interface. RCrane is also available, which is a plugin for Coot that implements the techniques described in Section 3.4 and helps crystallographers in building new crystal structures.

Instructions for accessing the Vfold package are available online at <http://vfold.missouri.edu/chen-software02.html> and iPARTS is available online at <http://bioalgorithm.life.nctu.edu.tw/iPARTS/>.

7. Conclusions

By allowing us to perceive and evaluate RNA molecules in a different manner, the η/θ formalism has made it possible to capture RNA conformational features in new ways that are important for the study of RNA structure, RNA folding and the interaction of RNA with ligands. The biophysics and computational biology of RNA molecules is a young field that is rapidly exploding with biological significance. It is therefore important to tailor our intellectual frameworks and methods to meet the specific needs of the RNA research community and to reflect the unique characteristics of RNA molecules. While the η/θ formalism is one such attempt, we anticipate that there are many creative ways to understand macromolecular form and function. In reading about the development of the η/θ formalism, we hope that other researchers will be encouraged to develop entirely new approaches for thinking about RNA and protein structure.

Acknowledgements

We are grateful for the insights gained through discussion with Jane and David Richardson, and members of their laboratory, particularly Laura Murray and Gary Kapral. In addition, we want to thank Chuck Duarte and Leven Wadley, whose innovations and imagination provided the foundation for this work. We want to thank Eric Westhof, Neocles Leontis and Bohdan Schneider for helpful discussions and for their role in establishing the RNA Ontology Consortium (ROC), which served as an early forum for this type of research. A. M. P. would like to thank Helen Berman for her early encouragement to develop this project and Wilma Olson for helpful discussions. And perhaps most of all, we want to thank the many RNA researchers, hopefully all cited here, who began implementing the η/θ formalism to study RNA structure and to develop new algorithms based on the approach. Their feedback and commentary has been invaluable to our research program. This work was supported, in part, by an NIH training grant T15 LM07056 to K. S. K. and by NIH Grant GM50313 to A. M. P. Anna Marie Pyle is an Investigator of the Howard Hughes Medical Institute.

References

- Abramovitz DL, Friedman RA, Pyle AM. Catalytic role of 2'-hydroxyl groups within a group II intron active site. *Science*. 1996; 271:1410–1413. [PubMed: 8596912]
- Adams PL, Stahley MR, Kosek AB, Wang J, Strobel SA. Crystal structure of a self-splicing group I intron with both exons. *Nature*. 2004; 430:45–50. [PubMed: 15175762]
- Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*. 2000; 289:905–920. [PubMed: 10937989]
- Batey RT, Gilbert SD, Montange RK. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature*. 2004; 432:411–15. [PubMed: 15549109]
- Beckers ML, Melssen WJ, Buydens LM. Predicting nucleic acid torsion angle values using artificial neural networks. *Journal of Computer-Aided Molecular Design*. 1998; 12:53–61. [PubMed: 9570089]
- Berman HM, Olson WK, Beveridge DL, Westbrook DL, Gelbin A, Demeny T, Hsieh SH, Srinivasan AR, Schneider B. The Nucleic Acid Database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical Journal*. 1992; 63:751–759. [PubMed: 1384741]
- Beuth B, Pennell S, Arnvig KB, Martin SR, Taylor IA. Structure of a *Mycobacterium tuberculosis* NusA–RNA complex. *EMBO Journal*. 2005; 24:3576–3587. [PubMed: 16193062]

- Brodersen DE, Clemons WM Jr, Carter AP, Morgan-Warren RJ, Wimberly BT, Ramakrishnan V. The structural basis for the action of the antibiotics tetracycline, pactamycin, and hygromycin B on the 30S ribosomal subunit. *Cell*. 2000; 103:1143–1154. [PubMed: 11163189]
- Cao S, Chen SJ. Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA*. 2005; 11:1884–1897. [PubMed: 16251382]
- Cao S, Chen SJ. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Research*. 2006; 34:2634–2652. [PubMed: 16709732]
- Cao S, Giedroc DP, Chen SJ. Predicting loop-helix tertiary structural contacts in RNA pseudo-knots. *RNA*. 2010; 16:538–552. [PubMed: 20100813]
- Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA. Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science*. 1996; 273:1678–1685. [PubMed: 8781224]
- Chang KY, Tinoco I Jr. The structure of an RNA ‘kissing’ hairpin complex of the HIV TAR hairpin loop and its complement. *Journal of Molecular Biology*. 1997; 269:52–66. [PubMed: 9193000]
- Correll CC, Beneken J, Plantinga MJ, Lubbers M, Chan YL. The common and the distinctive features of the bulged-G motif based on a 1.04 angstrom resolution RNA structure. *Nucleic Acids Research*. 2003; 31:6806–6818. [PubMed: 14627814]
- Correll CC, Swinger K. Common and distinctive features of GNRA tetraloops based on a GUAA tetraloop structure at 1.4 Å resolution. *RNA*. 2003; 9:355–363. [PubMed: 12592009]
- Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. *Science*. 1997; 278:82–87. [PubMed: 9311930]
- Das R, Baker D. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences of the United States of America*. 2007; 104:14664–14669. [PubMed: 17726102]
- Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB III, Snoeyink J, Richardson JS, Richardson DC. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research*. 2007; (35(Web Server issue)):W375–W383. [PubMed: 17452350]
- Ding F, Sharma S, Chalasani P, Demidov VV, Broude NE, Dokholyan NV. *Ab initio* RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*. 2008; 14:1164–1173. [PubMed: 18456842]
- Duarte CM. Computational approaches to the analysis and prediction of RNA structure. PhD Dissertation thesis. New York, NY: Columbia University; 2002.
- Duarte CM, Pyle AM. Stepping through an RNA structure: A novel approach to conformational analysis. *Journal of Molecular Biology*. 1998; 284:1465–1478. [PubMed: 9878364]
- Duarte CM, Wadley LM, Pyle AM. RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Research*. 2003; 31:4755–761. [PubMed: 12907716]
- Dunbrack RL Jr, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *Journal of Molecular Biology*. 1993; 230:543–574. [PubMed: 8464064]
- Egli M, Portmann S, Usman N. RNA hydration: a detailed look. *Biochemistry*. 1996; 35:8489–8494. [PubMed: 8679609]
- Ferre-D’Amare AR, Zhou K, Doudna JA. Crystal structure of a hepatitis delta virus ribozyme. *Nature*. 1998; 395:567–574. [PubMed: 9783582]
- Flores SC, Wan Y, Russell R, Altman RB. Predicting RNA structure by multiple template homology modeling. *Pacific Symposium on Biocomputing*. 2010; 15:216–227. [PubMed: 19908374]
- Furtig B, Richter C, Wohnert J, Schwalbe H. NMR spectroscopy of RNA. *Chembiochem*. 2003; 4:936–962. [PubMed: 14523911]
- Giambasu GM, Lee TS, Sosa CP, Robertson MP, Scott WG, York DM. Identification of dynamical hinge points of the L1 ligase molecular switch. *RNA*. 2010; 16:769–780. [PubMed: 20167653]
- Gilbert SD, Love CE, Edwards AL, Batey RT. Mutational analysis of the purine riboswitch aptamer domain. *Biochemistry*. 2007; 46:13297–13309. [PubMed: 17960911]

- Golden BL, Km H, Chase E. Crystal structure of a phage Twort group I ribozyme-product complex. *Nature Structural and Molecular Biology*. 2005; 12:82–89.
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999; 286:531–537. [PubMed: 10521349]
- Gruene T, Sheldrick GM. Geometric properties of nucleic acids with potential for auto-building. *Acta Crystallographica Section A*. 2011; 67:1–8.
- Guo F, Gooding AR, Cech TR. Structure of the tetrahymena ribozyme: base triple sandwich and metal ion at the active site. *Molecular Cell*. 2004; 16:351–362. [PubMed: 15525509]
- Harris FJ. Use of windows for harmonic-analysis with discrete Fourier-transform. *Proceedings of the IEEE*. 1978; 66:51–83.
- Holbrook SR, Sussman JL, Warrant RW, Klm SH. Crystal structure of yeast phenylalanine transfer RNA. II. Structural features and functional implications. *Journal of Molecular Biology*. 1978; 123:631–660. [PubMed: 357743]
- Holm L, Sander C. Searching protein structure databases has come of age. *Proteins*. 1994; 19:165–173. [PubMed: 7937731]
- Huppler A, Nikstad LJ, Allmann AM, Brow DA, Butcher SE. Metal binding and base ionization in the U6 RNA intramolecular stem-loop structure. *Nature Structural Biology*. 2002; 9:431–435.
- Hutchinson EG, Thornton JM. PROMOTIF – a program to identify and analyze structural motifs in proteins. *Protein Science*. 1996; 5:212–220. [PubMed: 8745398]
- Jonikas MA, Radmer RJ, Altman RB. Knowledge-based instantiation of full atomic detail into coarse-grain RNA 3D structural models. *Bioinformatics*. 2009a; 25:3259–3266. [PubMed: 19812110]
- Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D, Altman RB. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*. 2009b; 15:189–199. [PubMed: 19144906]
- Jovine L, HaInzl T, Oubridge C, Scott WG, Li J, Sixma TK, Wonacott A, Skarzynski T, Nagai K. Crystal structure of the ffh and EF-G binding sites in the conserved domain IV of Escherichia coli 4.5S RNA. *Structure*. 2000; 8:527–540. [PubMed: 10801497]
- Jucker FM, Heus HA, Yip PF, Moors EH, Pardi A. A network of heterogeneous hydrogen bonds in GNRA tetraloops. *Journal of Molecular Biology*. 1996; 264:968–980. [PubMed: 9000624]
- Juneau K, Podell E, Harrington DJ, Cech TR. Structural basis of the enhanced stability of a mutant ribozyme domain and a detailed view of RNA – solvent interactions. *Structure*. 2001; 9:221–231. [PubMed: 11286889]
- Kang HS, Tinoco I. A mutant RNA pseudoknot that promotes ribosomal frameshifting in mouse mammary tumor virus. *Nucleic Acids Research*. 1997; 25:1943–1949. [PubMed: 9115361]
- Kato H, Takahashi Y. SS3D-P2: a three dimensional substructure search program for protein motifs based on secondary structure elements. *Computational and Applied Bioscience*. 1997; 13:593–600.
- Kazantsev AV, Krivenko AA, Harrington DJ, Holbrook SR, Adams PD, Pace NR. Crystal structure of a bacterial ribonuclease P RNA. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102:13392–13397. [PubMed: 16157868]
- Keating KS, Pyle AM. Semiautomated model building for RNA crystallography using a directed rotameric approach. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107:8177–8182. [PubMed: 20404211]
- Keating KS, Toor N, Perlman PS, Pyle AM. A structural analysis of the group II intron active site and implications for the spliceosome. *RNA*. 2010; 16:1–9. [PubMed: 19948765]
- Keating KS, Toor N, Pyle AM. The GANC tetraloop: a novel motif in the group IIC intron structure. *Journal of Molecular Biology*. 2008; 383:475–81. [PubMed: 18773908]
- Klelm DJ, Moore PB, Steitz TA. The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *Journal of Molecular Biology*. 2004; 340:141–177. [PubMed: 15184028]
- Klelm DJ, Schmelng TM, Moore PB, Steitz TA. The kink-turn: a new RNA secondary structure motif. *EMBO journal*. 2001; 20:4214–4221. [PubMed: 11483524]

- Kolb, EWTMS. The Early Universe. New York: Addison-Wesley; 1980.
- Kortemme T, Ramirez-Alvarado M, Serrano L. Design of a 20-amino acid, three-stranded beta-sheet protein. *Science*. 1998; 281:253–256. [PubMed: 9657719]
- Krasltnikov AS, Xiao Y, Pan T, Mondragon A. Basis for structural diversity in homologous RNAs. *Science*. 2004; 306:104–107. [PubMed: 15459389]
- Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. *Science*. 2003; 302:1364–1368. [PubMed: 14631033]
- Laskowski RA, Macarthur MW, Moss DS, Thornton JM. Procheck - a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*. 1993; 26:283–291.
- Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. *RNA*. 2001; 7:499–512. [PubMed: 11345429]
- Lescoute A, Westhof E. The interaction networks of structured RNAs. *Nucleic Acids Research*. 2006; 34:6587–6604. [PubMed: 17135184]
- Lovell SC, Davis IW, Arendall WB III, De Barker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by Calpha geometry: phi, psi and Cbeta deviation. *Proteins*. 2003; 50:437–450. [PubMed: 12557186]
- Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. *Proteins – Structure, Function and Genetics*. 2000; 40:389–408.
- Major F, Turcotte M, Gautheret D, Lapalme G, Fillion E, Cedergren R. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science*. 1991; 253:1255–1260. [PubMed: 1716375]
- Malathi R, Yathindra N. A novel virtual bond scheme to probe ordered and random coil conformations of nucleic-acids – conformational statistics of polynucleotide chains. *Current Science*. 1980; 49:803–807.
- Malathi R, Yathindra N. Virtual bond probe to study ordered and random coil conformations of nucleic-acids. *International Journal of Quantum Chemistry*. 1981; 20:241–257.
- Malathi R, Yathindra N. Secondary and tertiary structural foldings in tRNA. A diagonal plot analysis using the blocked nucleotide scheme. *Biochemical Journal*. 1982; 205:457–160. [PubMed: 7138511]
- Malathi R, Yathindra N. The heminucleotide scheme: an effective probe in the analysis and description of ordered polynucleotide structures. *Biopolymers*. 1983; 22:2961–2976. [PubMed: 6626696]
- Malathi R, Yathindra N. Backbone conformation in nucleic acids: an analysis of local helicity through heminucleotide scheme and a proposal for a unified conformational plot. *Journal of Biomolecular and Structural Dynamics*. 1985; 3:127–144.
- Montange RK, Batey RT. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature*. 2006; 441:1172–1175. [PubMed: 16810258]
- Murray LJW, Arendall WB, Richardson DC, Richardson JS. RNA backbone is rotameric. *Proceedings of the National Academy of Sciences of the United States of America*. 2003; 100:13904–13909. [PubMed: 14612579]
- Murray, LW. RNA Backbone Rotamers and Chiropraxis. PhD Dissertation thesis. Durham, NC: Duke University; 2007.
- Murthy VL, Srinivasan R, Draper DE, Rose GD. A complete conformational map for RNA. *Journal of Molecular Biology*. 1999; 291:313–327. [PubMed: 10438623]
- Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. The structural basis of ribosome activity in peptide bond synthesis. *Science*. 2000; 289:920–930. [PubMed: 10937990]
- Ogle JM, Brodersen DE, Clemons WM Jr, Tarry MJ, Carter AP, Ramakrishnan V. Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science*. 2001; 292:897–902. [PubMed: 11340196]
- Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJ. An automated classification of the structure of protein loops. *Journal of Molecular Biology*. 1997; 266:814–830. [PubMed: 9102471]

- Olson WK. Configurational statistics of polynucleotide chains. A single virtual bond treatment. *Macromolecules*. 1975; 8:272–275. [PubMed: 1152526]
- Olson WK. Configurational statistics of polynucleotide chains – an updated virtual bond model to treat effects of base stacking. *Macromolecules*. 1980; 13:721–728.
- Olson WK. Computational studies of polynucleotide flexibility. *Nucleic Acids Research*. 1982; 10:777–787. [PubMed: 7063420]
- Olson WK, Flory PJ. Spatial configurations of polynucleotide chains. I. Steric interactions in polyribonucleotides: a virtual bond model. *Biopolymers*. 1972; 11:1–23. [PubMed: 5008180]
- Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. *Nature*. 1994; 372:631–634. [PubMed: 7990952]
- Pakleza C, Cognet JAH. Biopolymer Chain Elasticity: a novel concept and a least deformation energy principle predicts backbone and overall folding of DNA TTT hairpins in agreement with NMR distances. *Nucleic Acids Research*. 2003; 31:1075–1085. [PubMed: 12560506]
- Parisien M, Major F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*. 2008; 452:51–55. [PubMed: 18322526]
- Ponder JW, Richards FM. Internal packing and protein structural classes. *Cold Spring Harbor Symposium on Quantitative Biology*. 1987; 52:421–28.
- Portmann S, Usman N, Egli M. The crystal structure of r(CCCCGGGG) in two distinct lattices. *Biochemistry*. 1995; 34:7569–7575. [PubMed: 7779802]
- Quackenbush J. Computational analysis of microarray data. *Nature Reviews Genetics*. 2001; 2:418–27.
- Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*. 1963; 7:95. [PubMed: 13990617]
- Ramakrishnan C, Ramachandran GN. Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophysical Journal*. 1965; 5:909–933. [PubMed: 5884016]
- Ramakrishnan V. Ribosome structure and the mechanism of translation. *Cell*. 2002; 108:557–572. [PubMed: 11909526]
- Richardson JS, Schneider B, Murray LW, Kapral GJ, Immormlmo RM, Headd JJ, Richardson DC, Ham D, Hershkoviets E, Williams LD, Keating KS, Pyle AM, Micallef D, Westbrook J, Berman HM. RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA*. 2008; 14:465–481. [PubMed: 18192612]
- Rupert PB, Ferre-D'amaré AR. Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. *Nature*. 2001; 410:780–786. [PubMed: 11298439]
- Saenger, W. *Principles of Nucleic Acid Structure*. New York: Springer-Verlag; 1984.
- Santlmi GPH, Pakleza C, Cognet JAH. DNA tri- and tetra-loops and RNA tetra-loops hairpins fold as elastic biopolymer chains in agreement with PDB coordinates. *Nucleic Acids Research*. 2003; 31:1086–1096. [PubMed: 12560507]
- Scharpf M, Sticht H, Schwelmer K, Boehm M, Hoffmann S, Rosch P. Antitermination in bacteriophage lambda. The structure of the N36 peptide-boxB RNA complex. *European Journal of Biochemistry*. 2000; 267:2397–2408. [PubMed: 10759866]
- Schlutzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, Janell D, Bashan A, Bartels H, Agmon I, Franceschi F, Yonath A. Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell*. 2000; 102:615–623. [PubMed: 11007480]
- Schmeing TM, Sella AC, Hansen JL, Freeborn B, Soukup JK, Scaringe SA, Strobel SA, Moore PB, Steitz TA. A pre-translocational intermediate in protein synthesis observed in crystals of enzymatically active 50S subunits. *Nature Structural Biology*. 2002; 9:225–230.
- Scott WG, Murray JB, Arnold JR, Stoddard BL, Klug A. Capturing the structure of a catalytic RNA intermediate: the hammerhead ribozyme. *Science*. 1996; 274:2065–2069. [PubMed: 8953035]
- Selmer M, Dunham CM, Murphy FV, Welxlbaumer A, Petry S, Kelley AC, Weir JR, Ramakrishnan V. Structure of the 70S ribosome complexed with mRNA and tRNA. *Science*. 2006; 313:1935–1942. [PubMed: 16959973]

- Serganov A, Keiper S, Malinina L, Tereshko V, Skripkin E, Hobartner C, Polonskaia A, Phan AT, Wombacher R, Micura R, Dauter Z, Jaschke A, Patel DJ. Structural basis for Diels–Alder ribozyme-catalyzed carbon-carbon bond formation. *Nature Structural and Molecular Biology*. 2005; 12:218–224.
- Shapiro BA, Yingling YG, Kasprzak W, Bindewald E. Bridging the gap in RNA structure prediction. *Current Opinion in Structural Biology*. 2007; 17:157–165. [PubMed: 17383172]
- Sigel RK, Sashital DG, Abramovitz DL, Palmer AG, Butcher SE, Pyle AM. Solution structure of domain 5 of a group II intron ribozyme reveals a new RNA motif. *Nature Structural and Molecular Biology*. 2004; 11:187–192.
- Strobel SA, Adams PL, Stahley MR, Wang J. RNA kink turns to the left and to the right. *RNA*. 2004; 10:1852–1854. [PubMed: 15547133]
- Szep S, Wang J, Moore PB. The crystal structure of a 26-nucleotide RNA containing a hook-turn. *RNA*. 2003; 9:44–51. [PubMed: 12554875]
- Tamura M, Holbrook SR. Sequence and structural conservation in RNA ribose zippers. *Journal of Molecular Biology*. 2002; 320:455–474. [PubMed: 12096903]
- Tan ZJ, Chen SJ. Salt dependence of nucleic acid hairpin stability. *Biophysical Journal*. 2008; 95:738–752. [PubMed: 18424500]
- Thore S, Leibundgut M, Ban NN. Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. *Science*. 2006; 312:1208–1211. [PubMed: 16675665]
- Torres-Larios A, Swenger KK, Krasilnikov AS, Pan T, Mondragon A. Crystal structure of the RNA component of bacterial ribonuclease P. *Nature*. 2005; 437:584–587. [PubMed: 16113684]
- Venkatachalam CM. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers*. 1968; 6:1425–1436. [PubMed: 5685102]
- Wadley LM. A reduced representation coordinate system yields insights into RNA structure. PhD Dissertation thesis. New York, NY: Columbia University; 2006.
- Wadley LM, Keating KS, Duarte CM, Pyle AM. Evaluating and learning from RNA pseudotorsional space: quantitative validation of a reduced representation for RNA structure. *Journal of Molecular Biology*. 2007; 372:942–957. [PubMed: 17707400]
- Wadley LM, Pyle AM. The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic Acids Research*. 2004; 32:6650–6659. [PubMed: 15608296]
- Wang CW, Chen KT, Lu CL. iPARTS: an improved tool of pairwise alignment of RNA tertiary structures. *Nucleic Acids Research*. 2010; 38(Suppl.):W340–W347. [PubMed: 20507908]
- Westhof E, Masquida B, Jaeger L. RNA tectonics: towards RNA design. *Fold Design*. 1996; 1:R78–R88.
- Westhof E, Sundaralingam M. Restrained refinement of the monoclinic form of yeast phenylalanine transfer RNA. Temperature factors and dynamics, coordinated waters, and base-pair propeller twist angles. *Biochemistry*. 1986; 25:4868–4878. [PubMed: 3533142]
- Wimberly BT, Brodersen DE, Clemons WM Jr, Morgan-Warren RJ, Carter AP, Vornheim C, Hartsch T, Ramakrishnan V. Structure of the 30S ribosomal subunit. *Nature*. 2000; 407:327–339. [PubMed: 11014182]
- Wintjens RT, Rooman MJ, Wodak SJ. Automatic classification and analysis of alpha alpha-turn motifs in proteins. *Journal of Molecular Biology*. 1996; 255:235–253. [PubMed: 8568871]
- Word JM, Lovell SC, Labean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *Journal of Molecular Biology*. 1999; 285:1711–1733. [PubMed: 9917407]

Fig. 1.

The RNA backbone. (a) Diagram of a nucleotide showing the six standard backbone torsion angles (α , β , γ , δ , ϵ , and ζ). The nucleotide and suite divisions of the backbone are indicated. A nucleotide is centered about the ribose sugar and spans two phosphates, while a suite is centered about the phosphate and spans two sugars. (b) Diagram depicting the definitions of the pseudo-torsions, η and θ . The red lines indicate the pseudo-bonds that connect successive P and C4' atoms. The portion of the backbone shown affects a single pair of η and θ values, as the pseudo-torsions extend into the previous and next nucleotide. Figure modified from Wadley *et al.* (2007) with permission.

Fig. 2. The peptide backbone. (a) Diagram of a peptide showing the two variable backbone torsions (ϕ and ψ). (b) A Ramachandran plot of ϕ versus ψ showing approximately 81 000 non-glycine, non-proline and non-pre-proline residues from a high-resolution database, along with validation contours for favored and allowed regions. Figure reprinted from Lovell *et al.* (2003) with permission.

Fig. 3.

Features of the $\eta - \theta$ plot, (a) An $\eta - \theta$ plot published in 1998 shows all nucleotides (from a database of 53 RNA structures). Gray bars represent areas of the plot where either η or θ is in the same range as nucleotides in the helical region. Colored areas are regions of the plot that contain nucleotides that share similar structural features. Note that $\eta - \theta$ plots from later analyses are shown in Figs 6, 8 and 20. (b)–(i) Representative nucleotides from the regions of the plot indicated. (b) The helical region: the intersection of the two gray bars include nucleotides from the crystal structure of an A-form duplex (from PDB file 1rxs; Portmann *et al.* 1995). (c) Stacked turn region; exemplified by the second nucleotide of a GNRA loop (PDB file 1zif, Ade 5; Jucker *et al.* 1996). (d) The χ -switch region: includes the nucleotide 5' to the cleavage site of the hammerhead ribozyme (PDB file 300d, Cyt B170; Scott *et al.* 1996). (e) Flip-turn region; exemplified by APK A27 G pseudo-knot nucleotide G9 (PDB file 1kpd, Gua 9; Kang & Tinoco, 1997). (f) The C2'-bend region, includes tRNA^{phe} tertiary contact nucleotide G18 (PDB file 1tra, GUA 18; Westhof & Sundaralingam, 1986). (g) The stack switching region, exemplified by P456 domain pivot nucleotides (PDB file 1gid, nucleotides Ade A122, Ade A123; Cate *et al.* 1996). (h) The base twist region: includes the

last stem nucleotide of a kissing hairpin (PDB file 1kis, Ura 21; Chang & Tinoco, 1997). (i) The cross-strand stack region: includes all 5' nucleotides in sheared tandem R-R pairs (PDB file 1gid, Ade A113, Ade A206; Cate *et al.* 1996). Figure reprinted from Duarte & Pyle (1998) with permission.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 4.

A phosphate- and C4'-based virtual bond system has been independently developed three separate times. (a) The first such system was published in July of 1980 by Olson. Figure reprinted from Olson (1980) with permission. (b) Several months later – November, 1980 – Malathi and Yathindra independently published an identical virtual bond system (Malathi & Yathindra, 1980). Figure reprinted from Malathi & Yathindra (1982) with permission. (c) An $\omega'_v - \omega_v$ plot published in 1985 by Malathi and Yathindra (Malathi & Yathindra, 1985). The ω'_v angle is the torsion about the C4'-P virtual bond and is identical to θ , and the ω_v

angle is the torsion about the P-C4' virtual bond and is identical to η . (Note that the ω'_v and ω_v axes are reversed from those in the $\eta - \theta$ plots shown in Figs 3, 6, 8 and 20). The points on this plot represent nucleotides from oligonucleotide and yeast tRNA^{phe} crystal structures. Figure reprinted from Malathi & Yathindra (1985) with permission.

Fig. 5.

The late 1990s and early 2000s saw the publication of a wealth of complex RNA tertiary structures, including (a) the thiamine riboswitch (PDB ID: 1CKY; Thore *et al.* 2006), (b) the group I intron (PDB ID: 1U6B; Adams *et al.* 2004) and numerous ribosomal structures, (c) Shown here is the 16S ribosomal RNA from the bacterial 70S ribosome (PDB ID: 2JOO; Selmer *et al.* 2006). All structures are shown to scale. In each, the backbone is shown as an orange ribbon and the bases are shown in green.

Fig. 6.

The effect of windowing an $\eta - \theta$ plot using a Blackman window function. (a) An $\eta - \theta$ scatter plot of all nucleotides from the (Wadley *et al.* 2007) dataset. Each point shows the η and θ values of an individual nucleotide. (b) The result of applying the Blackman window to the dataset, colored from low to high density: blue, green, yellow and red. An upper cut-off has been applied to allow for better discernment of the peaks surrounding the helical region. Figure modified from Wadley *et al.* (2007) with permission.

Fig. 7.

Scatter plots of RMSD versus distance in the η - θ plane or standard torsional angles. For each plot, the best fit line for 10 000 random pairs of nucleotides from the dataset is shown. (a) RMSD of backbone atoms *versus* distance in the η - θ plane. The correlation coefficient is 0.80. (b) RMSD of backbone, sugar and base atoms *versus* distance in the η - θ plane. The correlation coefficient is 0.81. (c) RMSD of backbone atoms *versus* distance of standard torsional backbone angles. The correlation coefficient is 0.50. (d) RMSD of backbone, sugar and base atoms *versus* distance of the standard torsional angles (including χ). The correlation coefficient is 0.50. Figure reprinted from Wadley *et al.* (2007) with permission.

Fig. 8.

Clusters of non-helical nucleotides in the $\eta - \theta$ plot become more apparent after the dataset is divided by sugar pucker. (a) A scatter plot of the $\eta - \theta$ values of all non-helical C3'-endo (top) and C2'-endo (bottom) nucleotides. (b) A 3D view of the plot of C3'-endo (top) and C2'-endo (bottom) nucleotides with a 60° wide Blackman window function applied. (c) A contour plots resulting from analyzing the C3'-endo (top) and C2'-endo (bottom) density plots in (b). Contour levels are shown at 1σ , 2σ and 4σ levels, and scores are given in that order. These cluster scores report the percentage of nucleotides within the specified region that are superimposable with the corresponding prototype nucleotide. Contours with small populations (< 9) are not shown. The blue bars span the helical η values and the helical θ values for C3'-endo nucleotides. The pink elliptical area near the center of plot indicates the helical region that was initially excluded from the analysis. Figure modified and reprinted with permission from Wadley *et al.* (2007).

Fig. 9.

RNA motifs can be identified using η – θ worms. (a) A two-dimensional representation of the worm for the UUCG tetraloop motif. (b) A three-dimensional representation of a worm for the group II intron domain V structure (Sigel *et al.* 2004). This worm clearly reveals the location of a GAAA tetraloop and an extra-helical bulge (both of which are indicated in red on the worm and the structure). Figure reprinted from Duarte *et al.* (2003) with permission.

Fig. 10. PRIMOS analysis of the ribosome. (a) The tertiary structure of the 50S subunit of the ribosome (Ban *et al.* 2000). (b) The 50S subunit represented as an $\eta - \theta$ worm. (c) The hook turn, a motif found in the ribosome that was initially identified using PRIMOS (Szep *et al.* 2003). (d) A three-dimensional worm for the hook-turn motif. (d) Reprinted from Szep *et al.* (2003) with permission.

Fig. 11.

PRIMOS analysis can reveal changes between two related structures. Here, two 30S structures are compared: one unbound (PDB code IBL; Ogle *et al.* 2001) and one bound by paromomycin and a tRNA anticodon stem-loop (PDB code 1KQS; Schmeing *et al.* 2002). The line at 25° indicates a threshold above which nucleotides are considered to have different conformations in each complex. Some regions undergoing conformational changes between the complexes are indicated: the A site (A1492), the P site (C1397) and a site in the platform domain (C748). Figure reprinted from Duarte *et al.* (2003) with permission.

Fig. 12.

Pseudo-torsion analysis using PRIMOS revealed that there are two types of S-turn motifs, referred to as the S1 (classical S-turn) motif and the S2 motif. (a) Characteristic RNA worms for analogous portions of S1 (black) and S2 (red) motifs. (b) S1-motif structure with backbone ribbon (PDB code: 480D; Chang & Tinoco, 1997). Nucleotides for the S1 worm (U2653-U2656) are in black. (c) S2-motif structure (PDB code: 1JJ2; Klein *et al.* 2001). Nucleotides for the S2 worm (G892-A895) are in red. Figure reprinted from Duarte *et al.* (2003) with permission.

Fig. 13.

The COMPADRES technique was used to identify a number of novel motifs, including the π -turn shown here. (a) An example of an isolated π -turn (PDB file, 1JJ2 0 :A408-C12; Klein *et al.* 2001) (from the 50S ribosomal subunit of *Haemaphysalis marismortui* (H50S). The five structurally similar nucleotides (blue) are flanked by two helical strands (yellow). Numbering is from 5' to 3'. (b) A superposition of the backbones of the seven π -turns found in our dataset. (c) Locations of the four H50S π -turns (highlighted in red) in secondary structure. (d) Two of the π -turns found in the H50S occurred symmetrically opposite each

other, shown here in their helical context. Nucleotides not part of the canonical π -turns are shown in blue. Figure reprinted from Wadley & Pyle (2004) with permission.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 14.

The VFold model, which is based on the η and θ pseudo-torsions, can be used to predict RNA tertiary structure and folding (Cao *et al.* 2010). Here, it is used to predict the three-dimensional structure of a pseudo-knot. (a) The predicted pseudo-knot secondary structure. (b) The predicted virtual-bond level tertiary structure. (c) The all-atom structure constructed from the scaffold shown in (b). (d) The all-atom structure after additional refinement. Figure reprinted from Cao *et al.* (2010) with permission.

Fig. 15.

Typical RNA electron density maps for structures solved at various resolutions. Structures shown in (b–i) were retrieved from the Nucleic Acid Database (Berman *et al.* 1992), and electron density maps were calculated using observed structure factors and calculated phases, (a) Pie chart showing the resolutions of all large RNA structures (structures that contain a chain of at least 25 nucleotides), as retrieved from the Nucleic Acid Database (Berman *et al.* 1992). Numbers in parentheses are the number of structures in the specified resolution range. Note that structures in the 2.5–3.5-Å resolution range account for nearly two-thirds of all large RNA structures, whereas protein structures are typically solved at far higher resolutions. Maps are shown at (b) 1.04, (c) 1.75, (d) 2.25, (e) 2.75, (f) 3.3, (g) 3.8, (h) 4.5 and (i) 6.21 Å resolutions. Figure reprinted from Keating & Pyle (2010) with permission.

Fig. 16.

RCrane uses the pseudo-torsions for crystallographic model building. (a) The η and θ' pseudo-torsions, which use the C1' atom in place of C4'. Additionally, the suite and nucleotide divisions of the backbone are indicated. (b) The model building process. Starting with the experimental electron density (top), a crystallographer builds phosphates and bases (middle). The detailed backbone structure can then be automatically predicted and constructed (bottom). (c) A $\theta' - \eta'$ plot showing suites of the RNA05 filtered dataset (Richardson *et al.* 2008). Each color and shape combination corresponds to a specific conformer as indicated in the key. Ellipses correspond to the Gaussian functions (at the 1 σ level) used in conformer predictions. Only conformers with leading C2' endo sugar pucker and ending C3' endo sugar pucker are shown. Figure reprinted from Keating & Pyle (2010) with permission.

Fig. 17.

The consensus backbone conformers describe 46 allowable configurations for the RNA backbone (Richardson *et al.* 2008). Here, sample backbone structures of six of these conformers are shown. Note that the consensus conformers use the suite division of the backbone rather than the nucleotide division (See Figs. 1a and 16a).

Fig. 18.

The RCrane method results in highly accurate backbone structure. (a) Jackknife validation shows that conformer predictions are highly accurate. Prediction accuracy for conformers ranked as most likely, second most likely, etc., by the conformer prediction process. Standard error is <0.3% for all bars. (b, c) The sarcin/ricin domain (PDB code: 1Q9A; Correll *et al.* 2003) and guanine riboswitch (PDB code: 2EES; Gilbert *et al.* 2007) crystal structures were rebuilt using RCrane. The rebuilding used only the published phosphate and base coordinates, and was able to accurately and automatically reconstruct the backbone. Shown here are (b) the S-motif from the sarcin/ricin domain and (c) the J1/2 linker from the guanine riboswitch. The original structures are shown as green sticks and the rebuilt structures are shown in ball-and-stick representation. Atoms built within 0.5 Å of the published coordinates are shown as white spheres, and atoms built within 0.8 Å are shown in yellow. Suite numbers and conformers are labeled. Note that the rebuilt structure has not been minimized or refined against the electron density. Figure reprinted from Keating & Pyle (2010) with permission.

Fig. 19.

RNA strands can be accurately rebuilt by computationally joining nucleotides with similar η/θ values from other structures. (a) An example of an *in silico* tetraloop superimposed on the original (1S72 0:8994; Klein *et al.* 2004). The backbone RMSD of the *in silico* strand (red) to the original tetraloop (blue) is 0.78 Å, despite the fact the nucleotides used to build the strand do not belong to any naturally occurring tetraloop. (b) A bulge region (1S72 0:1391–1398; Klein *et al.* 2004) from the 50 S ribosomal subunit. The *in silico* strand (blue) superimposed on the original (red) with a backbone RMSD of 0.91 Å. Figure reprinted from Wadley *et al.* (2007) with permission.

Fig. 20.

Data filtering is important for $\eta - \theta$ plots. All plots shown here were constructed using the RNA05 dataset (Richardson *et al.* 2008) with differing filtering criteria. (a) Plots with no filtering applied. 7,372 C3'-endo (top) and 791 C2'-endo (bottom) nucleotides are shown, (b) Plots where nucleotides containing atoms with B factors > 60 have been excluded. 3733 C3'-endo (top) and 458 C2'-endo (bottom) nucleotides remain. (c) Plots with additional quality filters applied to remove nucleotides containing a steric clash (van der Waals overlap > 0.4 Å, as measured by MolProbity clashscore; Word *et al.* 1999). 1548 nucleotides C3'-endo (top) and 218 C2'-endo (bottom) nucleotides remain and are shown in the plot. Note that for all plots in a–c, only nucleotides with a well-defined sugar pucker are shown (C3'-endo: $\delta = 84 \pm 30$, pseudo-phase angle of furanose ring (Saenger, 1984) between $0^\circ - 36^\circ \pm 18^\circ$, base-phosphate perpendicular distance > 2.9 Å; C2'-endo: ($\delta = 147 \pm 30$, pseudo-phase angle of furanose ring between $144^\circ - 180^\circ \pm 18^\circ$, base-phosphate perpendicular distance > 2.9 Å). Additionally, for nucleotides with alternative conformations, only the first instance listed in the pdb file was used.

Table 1

Angle ranges for A-form helical nucleotides

Torsion	Crystallographic range ^a	Saenger range ^b
Standard torsions		
α	147°–303°	265°–310°
β	145°–193°	165°–210°
γ	45°–184°	45°–60°
δ	70°–88°	75°–95°
ϵ	197°–232°	170°–210°
ζ	269°–308°	280°–320°
Pseudo-torsions		
η	149°–185°	
θ	171°–229°	

The standard torsions alone are poor discriminators of A-form structure due to the wide range of allowed torsion values. The pseudo-torsions present far narrower ranges and can easily be used to identify A-form nucleotides.

^a Crystallographic ranges were determined using high-resolution helical structures ($\leq 1.2\text{\AA}$ resolution, PDB IDs: 1QCU, 2Q1R, 2V7R, 2VUQ, 3GVN, 434D). The first and last nucleotide of each chain were excluded due to the greater flexibility allowed at chain ends.

^b For reference, torsion ranges published in 1984 are provided (Saenger, 1984).