STATISTICAL ANALYSIS OF RNA BACKBONE

By

Eli Hershkovitz

Guillermo Sapiro

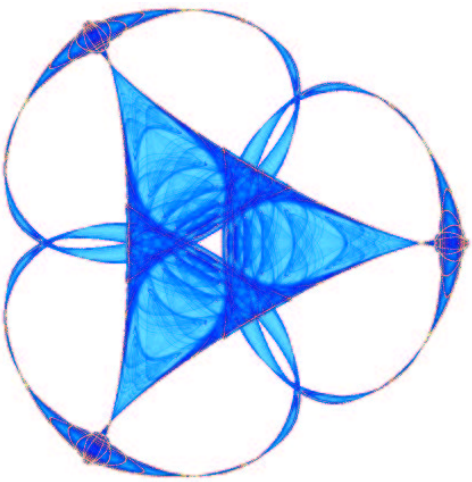
Allen Tannenbaum

and

Loren Dean Williams

IMA Preprint Series # 1981

(June 2004)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA

514 Vincent Hall
206 Church Street S.E.
Minneapolis, Minnesota 55455–0436

Phone: 612/624-6066 Fax: 612/626-7370
URL: http://www.ima.umn.edu

Statistical Analysis of RNA Backbone*

Eli Hershkovitz†
Guillermo Sapiro‡
Allen Tannenbaum§
Loren Dean Williams¶

Abstract

Local conformation is an important determinant of RNA catalysis and binding. The analysis of RNA conformation is particularly difficult due to the large number of degrees of freedom (torsion angles) per residue. Proteins, by comparison, have many few degrees of freedom per residue. In this work we use and extend classical tools from statistics and signal processing to search for clusters in RNA conformational space. Results are reported both for scalar analysis, where each torsion angle is separately studied, and for vectorial analysis, where several angles are simultaneously clustered. Adapting techniques from optimal quantization to the RNA structure, we find torsion angle clusters and RNA conformational motifs. We validate the technique using well known conformational motifs and show how the simultaneous study of the total torsion angle space leads to the discovery of new motifs. The proposed technique is fully automatic and based on well-defined optimality criteria.

1 Introduction

Nucleic acid polymers play important roles in the storage and transmission of information. In particular, RNA can both encode genetic information and catalyze chemical reactions [9]. As the only biological macromolecule capable of such diverse activities, it has been proposed that RNA preceded DNA and protein in early evolution [2]. Over the past 15 years, the database of RNA conformation and interaction (the NDB [22]) has evolved rapidly, or to be more accurate, has exploded, in both size and complexity. The database has been transformed from tRNA and RNA oligonucleotides to moderately sized globular RNAs to massive complexes containing multiple large RNA molecules, many proteins, ions, water molecules, etc. These large complexes are a rich source of new information, but do not surrender to traditional methods of analysis. These complexes are of sufficient size that one can gather and analyze statistics that were not previously available. The development of techniques for discovering statistical rules governing RNA conformation and interaction will help answer fundamental biological and biochemical questions including those related to non-protein enzymology and the origins of life. The goal here is to discover repetitive elements of interaction and conformation (motifs).

*Work supported by ONR, DARPA, NSF, ARO, AFOSR, and NIH.

†Schools of Electrical & Computer Engineering and Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250, eli@theor.chemistry.gatech.edu

‡Electrical and Computer Engineering and Digital Technology Center, University of Minnesota, Minneapolis, MN 55455, guille@ece.umn.edu

§Schools of Electrical & Computer Engineering and Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250, tannenba@ece.gatech.edu.

¶School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332, loren.williams@chemistry.gatech.edu.

The analysis of RNA conformation presents particular problems. For proteins, for each amino acid residue there are two conformational degrees of freedom: torsional angles θ and ϕ. The observed protein conformation is generally confined to limited regions of this two dimensional space (*Ramachandran plot*) [24, 25]. For RNA the dimensionality is much greater. For each nucleotide residue there are **seven** degrees of freedom; see Figure 1 and [27]. Each RNA residue has six backbone torsional angles and one angle χ that describes the rotation of the base relative to the sugar. Differences in dimensionality are a distinguishing characteristic of RNA conformational analysis in comparison to protein conformational analysis.

To deal with the high dimensionality or RNA conformation, several approaches have been explored. A reduced set of two pseudo-torsional angles per residue was proposed in [5]. This reduction in dimensionality from seven to two simplifies the analysis, but sacrifices information. Alternatively work in [8, 18, 26] attempts to retain information from the full conformational space. The approach in [8] gives a structural alphabet based on the discretization of the conformation distribution function via binning the torsion angles *taking one angle at a time*. The binning method uses observed minima of the torsion angle frequency distributions to define boundaries between conformational classes.

The approaches of [18, 26] decompose the seven dimensional space into various subspaces of three dimensions. It is possible to locate centers of frequency clusters in torsional subspaces. The restriction to three-dimensional subspaces arises from requirements for manual (visual) detection of the frequency clusters. In addition, a filtering stage is described in [18] to remove conformations that are suspected to arise from measurement error. Various elemental units can be parsed during conformational analysis of an RNA polymer. Richardson and coworkers [18] suggest a base-to-base unit (a "suite") instead of the chemically inspired, and more conventional phosphate-to-phosphate unit (a residue); see Figure 1. (A comparison based on *mutual information* of these two parsing approaches is given in Appendix A.) The work of [26] utilizes a dinucleotide building block to attempt to include the correlations between neighboring residues.

The primary drawback of low dimensional methods is that some clusters might avoid detection and defy description. Several distinct clusters at full dimensionality can be compressed into a single cluster at low-dimensionality. A limitation to three dimensional subspaces is arbitrary and might inaccurately characterize some regions of RNA conformational space.

Here all seven dimensions of RNA conformation are analyzed simultaneously with methods from classical signal processing. The methods used here can be automatic and virtually parameter-free. We use high-dimensional clustering, mainly *vector quantization*.[1] We extend the method by imposing well-characterized conformations, such as A-conformation, onto the clustering. The work here continues the research line of [8] (see also [19, 20]), attempting to resolve some limitations there. Vector quantization gives well-defined distortion and quality metrics. It does not involve visual inspection. The VQ approach is validated by comparison of the output with that of previously reported methods, as well as with the structural motifs library (SCOR) [12]. The VQ method allows us to describe motifs which were not found in [8].

The remainder of this paper is organized as follows. In Section 2, we provide the basic background on vector quantization. In Section 3, we begin with a particular case of vector quantization, the scalar case, which permits us not only to introduce the basic concepts but also to show the results reported in [8], and be replicated, refined and fully automated. In Section 4, we use the full power of vector quantization to analyze sets of four and seven torsion angles simultaneously, extending some of the results reported previously in such works as [8, 18]. We moreover present a modification of the basic vector quantization algorithm, namely *cluster merging*, which is fully motivated by RNA properties and is needed to adapt this classical signal processing technique to the study of RNA structure. Section 5 presents the motifs that were found by our method and compares our findings with known structural motifs. Finally, in Section 6, we summarize our methods as well as describe some key research directions. We have also included three appendices. Appendices A and B give some very preliminary results on the use of other techniques from statistical

[1] Previously, vector quantization was used in the context of protein structure; e.g., [10].

signal processing, mutual information and principal component analysis, for the study of RNA motifs. In Appendix C, we summarize some of the key results of the binning method [8] for the convenience of the reader.

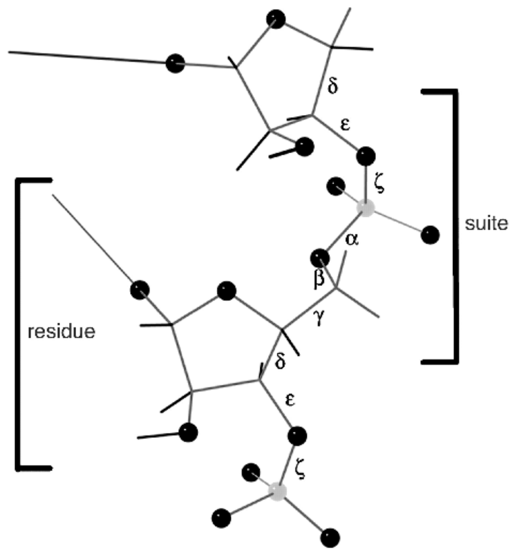


Figure 1: *RNA backbone with six torsion angles labeled on the central bond of the four atoms defining each dihedral. The two alternative ways of parsing out a repeat are indicated: A conventional nucleotide residue spans the atoms from a 3 phosphorous atom and a 5 oxygen atom, (changing residue number between O5' and P), whereas an RNA suite goes from sugar to sugar (or base to base). This image was obtained from [18].*

2 Background on Scalar and Vector Quantization

Vector quantization (VQ) is a clustering technique originally developed for lossy data compression. In 1980, Linde *et al.*, [15], proposed a practical VQ design algorithm based on a training sequence. The use of a training sequence by-passes the need for multi-dimensional integration, thereby making VQ a practical technique, implemented in many scientific computation packages such as Matlab (www.mathworks.com).

A VQ is analogous to an approximator. The technique is similar to "rounding-off" (say to the nearest integer). An example of a 1-dimensional VQ is shown in Figure 2. Here, every number less than -2 is approximated by -3. Every number between -2 and 0 is approximated by -1. Every number between 0 and 2 is approximated by +1. Every number greater than 2 is approximated by +3. Figure 2 also presents a two-dimensional example. Here, every pair of numbers falling in a particular region is approximated by the red star associated with that region.

The general VQ design problem can be stated as follows. Given a vector source with known statistical properties, a distortion measure, and number of desired codevectors, find a codebook (the set of all red stars) and a partition (the set of blue lines) that result in the smallest average distortion.

We assume that there is a training sequence (e.g., the measured torsion angles in RNA backbone) consisting of M source vectors of the form $T = \{x_1, x_2, ..., x_M\}$. We assume that the source vectors are k-dimensional, e.g., $x_m = \{x_{m,1}, x_{m,2}, ..., x_{m,k}\}$, for $1 \leq m \leq M$. Let N be the number of desired codevectors and let $C = \{c_1, c_2, ..., c_N\}$ be the codebook, where each c_n, $1 \leq n \leq N$, is of course k-dimensional as well. Let S_n be the cell associated with the codevector c_n and let $P = \{S_1, S_2, ..., S_N\}$ be the corresponding partition of the k-dimensional space. If the source vector x_m is in the encoding region S_n, then its approximated by c_n, and let us denote by $Q(x_m) = c_n$ (if $x_m \in S_n$) the corresponding map

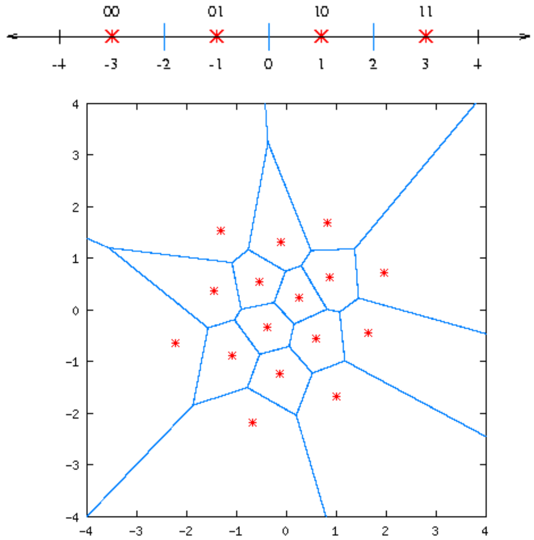


Figure 2: *One (top) and two (bottom) dimensional examples of clustering via (vector) quantization. All the points in a given interval (in one-dimension) or a given cell (two-dimensions) are represented by the red marked "center." This image was obtained from [4]. (This is a color figure.)*

(each vector is simply associated to the closest center from C). Then, assuming for example a squared error distortion measure, the average distortion is given by

$$D := \frac{1}{Mk} \sum_{m=1}^{M} \| x_m - Q(x_m) \|^2, \quad \text{where} \quad \| e \|^2 := e_1^2 + e_2^2 + ... + e_k^2.$$

The design problem then becomes the following: Given the training data set T and the number of desired codebooks (or clusters) N, find the cluster centers C and the space partition P such that the distortion D is minimized. This problem can be efficiently solved with the LBG algorithm [7, 15], and as mentioned above, its implementation can be found in popular scientific computing programs. Additional details on the technique can be found, e.g., in [4], from which we have prepared this summary. In future work we plan to use more advanced techniques such as those reported in [21].[2]

3 Scalar Quantization: Automatic Binning of Single Torsion Angles

To provide an accessible introduction to VQ, a brief discussion of scalar quantization (SQ) is provided here. SC is a natural extension of our previous work, and is extensible to VQ. With SQ one can fully automate the previous binning method described in [8], where torsion angles are treated individually. In [8], conformational space is partitioned into boxes each containing one conformational state, i.e., *rotamer* or a subset of conformational states; see also [18]. The box boundaries were set by visual inspection, using minima of torsion angle frequency distributions as guides.

As was known from [23, 28], four torsional angles $(\alpha, \gamma, \delta, \zeta)$ (which we call the *identifier angles*) are sufficient for this classification. Here the results of that work are reproduced with SC, in a fully automatic fashion. In Appendix C more details are described from the work in [8], reproducing key tables for the convenience of the reader.

[2] Vector quantization is often also known in the literature as k-means clustering.

We argued in [8] that the frequency histograms of the four identifier torsion angles have a clear multi-peak structure, see Figure 3 and details below. Since the peak structure is the cornerstone for our proposed classification method, we describe here these results for a larger set of RNA structures than those reported in [8]. In particular, two data sets are used. One follows the work reported in [8], and is for a single RNA with 2914 residues (HM LSU 23S rRNA, RR0033), while the second one follows work reported in [18], and is for a collection of 132 RNAs,[3] giving a total of 10463 residues (redundancies have not been eliminated). Here, as in the rest of this work, residues with undefined or unknown torsion angles were omitted. Coordinates were obtained from the *Nucleic Acid Database* [22]. We have not performed the filtering of [18]. That method may indeed improve the results. As mentioned above, in the SQ, we limit the analysis to the torsion angles (α, γ, δ, ζ) (see Figure 1), since the others are either dependent on these angles or have distributions which are almost unimodal [23, 28]. There is no intrinsic limitation which restricts one to this reduced set of angles, and indeed being fully automatic, the process can be easily applied to larger sets.

Figure 3 shows the distributions for the four angles from the large and small datasets. The two datasets of histogram features have a strong resemblance, suggesting the generality of the cluster classification method for analysis of RNA conformation.

One potential problem with visually-based classification methods such as the binning in [8] and the technique presented in [18], in addition to being limited to ad-hoc observations of three or less angles at a time (see more on this below), is that the resolution (and amount of data) may not be sufficiently fine, which may make it difficult to distinguish distinct features in the data, and clusters can be confused and merged.

This issue is demonstrated in the behavior of the torsional angle ζ. It is known that minima are predominantly at staggered conformations [27], at gauss+, gauss-, and trans. The histogram of this angle gives a clear peak only at the gauss- peak. The other two peaks are flat and wide enough to merge into a large plateau. As demonstrated below, this degeneration causes a loss of important structural information. In [18], the degeneracy is removed by filtering. VQ retains these details without the need for filtering (shown below).

[3]With NDB and PDB codes: ar0001, 02, 04, 05, 06, 07, 08, 09, 11, 12, 13, 20, 21, 22, 23, 24, 27, 28, 30, 32, 36, 38, 40, 44; arb002, 3, 4, 5; arf0108; arh064, 74; arl037, 48, 62; arn035; dr0005, 08, 10; drb002, 03, 05, 07, 08, 18; drd004; pd0345; pr0005, 06, 07, 08, 09, 10, 11, 15, 17, 18, 19, 20, 21, 22, 26, 30, 32, 33, 34, 36, 37, 40, 46, 47, 51, 53, 55, 57, 60, 62, 63, 65, 67, 69, 71, 73, 75, 78, 79, 80, 81, 83, 85, 90, 91; prv001, 04, 10, 20, 21; pte003; ptr004, 16; rr0005, 10, 16, 19, 33; tr0001; trna12; uh0001; uhx026; ur0001, 04, 05, 07, 09, 12, 14, 15, 19, 20, 22, 26; urb003, 08, 16; urc002; urf042; url029, 50; urt068; and urx053, 59, 63, 75.

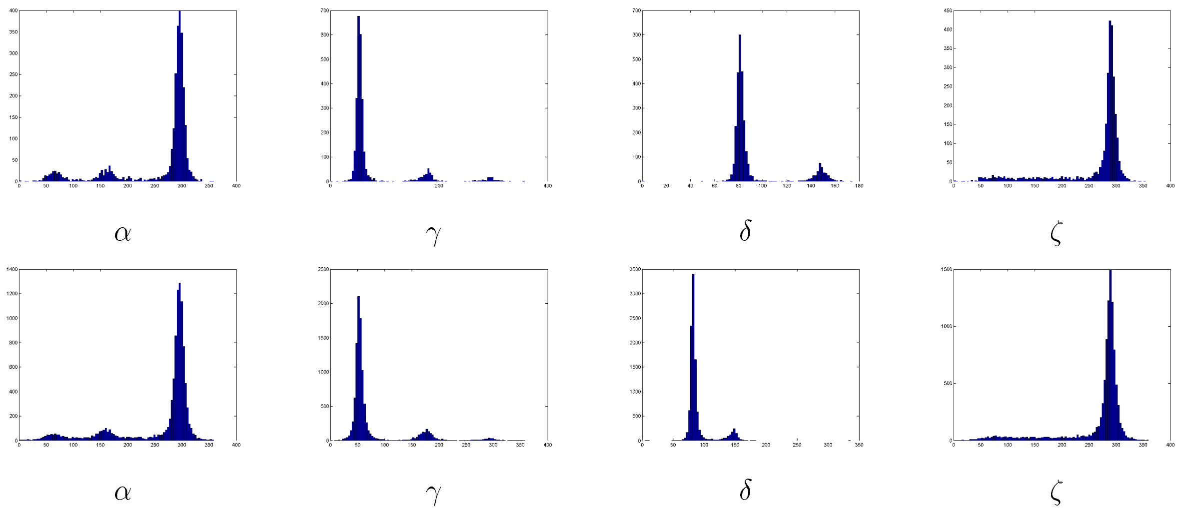


Figure 3: *Distributions of the torsion angles α, γ, δ, and ζ for the single RNA (first row) and the collection of RNAs (second row). We observe the similitude among the distributions, marking the presence of "rotamers" not only for a given RNA but also across RNAs. We also observe clear modes, which are automatically detected by the proposed clustering technique. In addition, note that the ζ torsion angle has a large tail not present in the other distributions.*

Understanding the peak shape of each cluster is crucial for probabilistic RNA design and for understanding local dynamics of folding. The peak shapes of the clusters contain important information on RNA dynamics, but might also be influenced by coordinate error. It appears that better fitting for the major clusters (see below for the limits of these clusters), is obtained using exponential distributions, and not Gaussian distributions as argued for example in [8]. For the first dataset, the kurtosis[4] for the main peak is 5.3 for α and 4.6 for ζ, clearly indicating a significant deviation from Gaussian distributions (whose kurtosis is 3). The log-likelihood while fitting an exponential function improves by 24% with respect to fitting a Gaussian for the α torsion angle and by 23% for the ζ torsion angle. Similar behavior is observed for the other dataset, although in some cases the improvement is more moderate (e.g., for the first mode of α in the first dataset, the improvement is about 16%).

Using the automatic and optimal quantization technique described in the previous section, and requesting the number N of codevectors following [8] (or just from visual inspection for now, this will be made automatic later), we found the codevectors or centers of the clusters $C = \{c_1, ..., c_N\}$ given in Table 1. Later on, for the classification, we enumerate the clusters in each coordinate by increasing values. For example, a residue whose torsional angles are in the third peak (center) in α, the first in γ and δ, and the third in ζ will be enumerated as 3113; see Table 1.

The results are similar for the two data sets. For γ, two of the centers are very close to each other, and will be merged during clustering. This demonstrates a possible problem of over-clustering by scalar quantization (or any other automatic clustering technique). In the next section, a simple merging algorithm is proposed to treat this difficulty. Once again, although the number of clusters is pre-defined, this could be accomplished as part of the automatic process; see Section 4.

Regarding ζ, if additional clusters are requested, e.g., 3 clusters for the first dataset, these clusters are automatically found at 85.86 (1), 188.25 (2), and 289.27 (3), thereby splitting the large tail (following the description in [18], but in an automatic fashion). These additional centers will also appear when considering

[4]The degree of peakedness of a distribution, defined as a normalized form of the fourth central moment of a distribution, $\mu_4/(\mu_2)^2$, where μ_i denotes the i-th central moment.

	Dataset 1
α	68.3 (1), 169.7 (2), 294.3 (3)
γ	50.4 (1), 60.0 (1), 175.8 (2), 292.3 (3)
δ	81.7 (1), 147.8 (2)
ζ	118.0 (2), 286.7 (1)
	Dataset 2
α	68.6 (1), 167.8 (2), 294.0 (3)
γ	50.1 (1), 65.0 (1), 174.4 (2), 290.2 (3)
δ	82.7 (1), 144.4 (2)
ζ	116.4 (2), 286.0 (1)

Table 1: *Cluster centers automatically computed by our technique. Numbers in parentheses are used for cluster identification.*

torsion angles in vectorial form in the next section, and will be used to search for motifs. Further increasing the number of clusters does not produce in general a significant change in the distortion D, an indication that the selected number of clusters is sufficient; see Section 4.

The clustering (binning) method that results from scalar quantization as described so far has one major difference with the one described in [8]. For scalar quantization no bins are classified as "other." In the scalar quantization case, every bin is populated. Every residue is associated with a specific set of four centroids (by simple proximity via the map Q defined in Section 2), each one corresponding to one of the four torsion angles $(\alpha, \gamma, \delta, \zeta)$. In Table 2 we give the corners of the boxes that define these bins. We could of course easily and automatically add the "other" class if so desired, by simply forcing the torsion angles not to be "too far" from the center of the bin. This can be quantified for example by the standard deviation of each bin.

bin index	1	2	3
α	$[0 - 115]$	$[115 - 220]$	$[220 - 360]$
γ	$[0 - 120]$	$[120 - 220]$	$[220 - 360]$
δ	$[50 - 118]$	$[118 - 170]$	
ζ	$[10 - 130]$	$[130 - 220]$	$[220 - 360]$

Table 2: *Enumeration of the bins obtained by scalar quantization and their boundaries.*

The scalar quantization method was used to automatically cluster the four identifier torsion angles. The fundamental difference between the binning method in [8] and the scalar quantization method is that bin boundaries were established manually by inspection of frequency histograms, while the clusters borders were accomplished automatically computed via a distortion minimization criterion. The four identified torsion angles of all the residues in RR0033 were classified by scalar quantization, with the three clusters in ζ described above. The tetraloop motif, [1, 11, 16, 30], was used to compare the results here to our previous work [8] and to other methods such as SCOR. A SQ tetraloop is defined by four consecutive residues, with residues 1,3,4 in state 3113 (in the A conformation). Residue 2 is in state 2113. In other words, all torsion angles except α of residue 2 are clustered around their most populated peaks. Sixteen SQ tetraloops are observed, initiating at residues

$$253, 469, 577, 691, 805, 1327, 1469, 1500, 1596, 1629, 1863, 2249, 2412, 2630, 2696, 2877.$$

All of the SQ tetraloops except 1500 and 1596 are similarly classified as tetraloops by SCOR. For these two cases the values of the angles that cause these classification differences are very close to the cluster border. These results are an indication of the utility of the automatic clustering technique derived from VQ, and its value will be further demonstrated below.

In summary, the results of automatic classification by scalar quantization are very similar to the manual binning method of [8], except for an extra refinement (obtained automatically) in the ζ coordinate. As mentioned above, it can be shown that any increase in the number of clusters in the four coordinates will not reduce the distortion D. Indeed, it seems that any attempt to increase the refinement will only worsen the results.

4 Vector Quantization: Automatic and Simultaneous Binning of Multiple Angles

Information can be lost in scalar quantization. This loss occurs because each angle is considered in separation from the others. Scalar clustering is a one dimensional projection that can merge clusters that are distinct in projections of higher dimension. For a schematic illustration of this problem, see Figure 4.

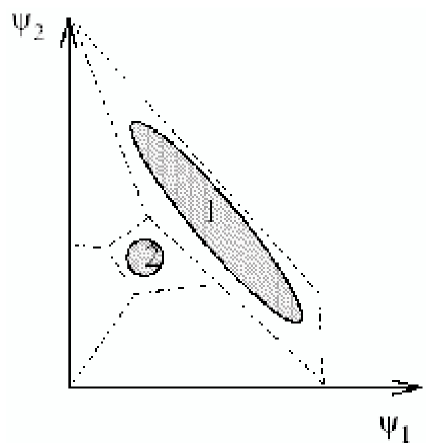


Figure 4: *In this example, the conformations space is projected onto two torsional angles ψ_1 and ψ_2. There are clearly two population clusters 1 and 2. The individual torsion angle histograms will give only one peak with a negligible tail. The automatic vectorial binning via VQ will map the two clusters into separate bins.*

VQ analysis overcomes this problem. For an illustration of the methodology, consider VQ analysis of two angles (with $k = 2$).[5] For example, requesting $N = 6$ clusters for the pair (α, ζ) we obtain the centers

$$C = \{(69.1, 284.2), (291.0, 165.6), (287.4, 79.0), (167.6, 284.6), (287.7, 280.0), (105.3, 109.8)\}.$$

The α component of the automatically detected centers is as in the case of scalar quantization, while the ζ component includes terms that appear both when we request 2 and 3 bins for ζ in the scalar case. That is, VQ for $k = 2$ finds additional relevant clusters in ζ when considered as a vector in conjunction with α. In Figure 5, the torsion angles are plotted (blue dots) together with the cluster centers (red stars). Repeating

[5]To further demonstrate the importance of the simultaneous study of torsion angles and to make the figures simpler, and since this exercise is for the moment for illustrative purposes only, we exclude the residues of RR0033 in A-conformation, which constitutes over 60% of the RNA. A-conformation is characterized by the angles $(\alpha, \gamma, \delta, \zeta)$ each in the modes corresponding to their respective major peaks.

this exercise for $N = 9$ clusters for (α, ζ), Gives the centers

$$\begin{aligned} C = \quad \{ \quad & (292.2, 68.3), (68.3, 283.8), (176.5, 122.6), (157.5, 284.9), \\ (\quad & 66.7, 102.4), (213.1, 287.0), (293.4, 284.0), (295.3, 188.0), (293.5, 132.0)\}. \end{aligned}$$

Figure 6, contains plots of the torsion angles (blue dots) together with the cluster centers (red stars), showing that while the main cluster centers are closely located to those when only 6 centers were considered, the three additional centers split the broad distributions (lower left region, where one center became two) as well as to split the very popular conformations (e.g., additional center at the main α pick, right of the figure).

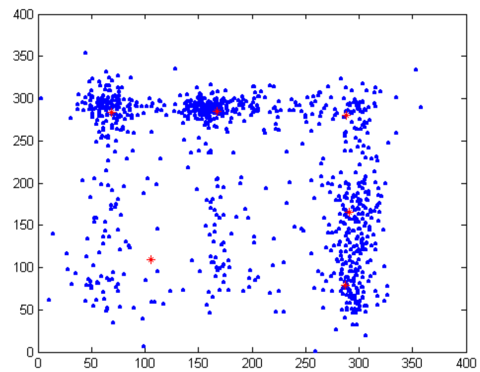


Figure 5: *Torsion angles for the pair* (α, ζ) *(blue dots) together with the 6 cluster centers (red stars).*

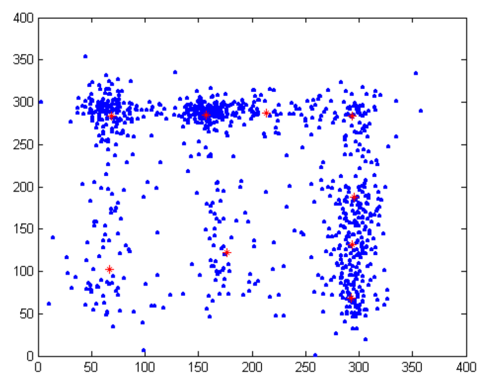


Figure 6: *Torsion angles for the pair* (α, ζ) *(blue dots) together with the 9 cluster centers (red stars).*

It is clear from the illustrations above that high dimensional clustering is necessary. All torsion angles should be considered simultaneously. The framework described in Section 2 permits that. Analysis of one dimension (one torsion angle at a time), four dimensional $(\alpha, \gamma, \delta, \zeta)$, or the full seven dimensional torsion angle space is of equal complexity with automatic VQ methods. Of course, due to the "curse of dimensionality," more data is needed at higher dimensions. However, the work here is not limited by quantity of data. The dispersion within the clusters (i.e., the peak shape) might be used to infer energy potentials and dynamical processes.

The first test of the vector quantization method used four dimensions ($k = 4$), the four identifier angles $(\alpha, \gamma, \delta, \zeta)$ of RR0033. To cluster these four angles, one must determine the optimum number of clusters

(N). False clusters arise if N is too large (over-partitioning). Distinct clusters are merged if N is too small (under-partitioning). Several metrics are used here to optimize N. The relationship between the distortion, D, defined in Section 2, and the number of clusters N is useful for optimizing N. In addition, the observation of overlapping clusters, indicates over-partitioning.

Figure 7 shows a plot of D as a function of the number of clusters N. The distortion reaches a plateau value for $N \approx 50$. Vector quantization was performed for $N = 40, 50, 60$. $N = 60$ gave all the populated bins defined in [8]. All three cases however, appear to be over-partitioned. This over-partitioning is especially pronounced in the A-conformation region. Most neighboring clusters in this region are overlapping. This overlap is not surprising, since these clusters are so highly-populated (over 60% of this RNA) that any distortion minimization approach will tend to invest a lot of resources (i.e., centers) there. This phenomenon emphasizes the need to impose structural definitions onto the clustering process, as described below.

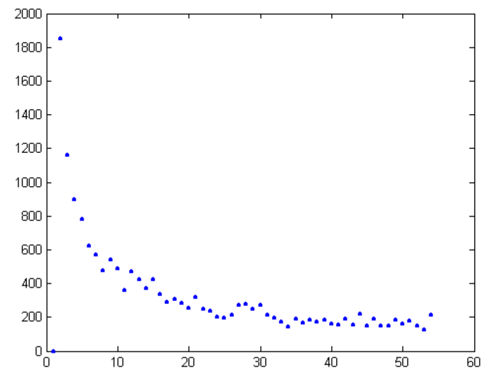


Figure 7: *Error as a function of the number of clusters for the vector* $(\alpha, \gamma, \delta, \zeta)$.

The full quantization of the conformation space, based on all seven torsional angles was also performed. The algorithm is fast enough to perform a full quantization of the 2800 residues of RR0033 to 60 classes in a few seconds. The distortion D plateaus at about 60 classes; see Figure 8. $N = 60$ gives the represent the most populated 15 bins from [8], and is in good correspondence to the results of the four dimensional quantization. Additional partitioning of up to $N = 100$ reveals very sparsely populated new classes, see discussion section. Here a "new class" is "far" from previously found classes. Classes are here considered "close" (or overlapping) when their centroids are in the same bin (as derived from the SQ, see Table 2), and "far" otherwise.

4.1 Merging

"Closeness" is the first component of a merging criteria. Specifically, we require that two clusters with centroids that reside within the same bins are merged into a unique cluster, subject to conditions mentioned below. [6]

Note that binning, whether by observation as in [8] or automatically done via SQ as described above, gives a partition of the torsion angles space into multi-dimensional boxes. There is no *a priori* reason to believe that the basin of attraction of the specific energy minimum that defines a native conformer will have such a shape. Using vector quantization with merging can shift and change the basin and its boundaries. An

[6] Another possible merging criteria is to merge clusters as long as they do not change the total distortion D above a given threshold.

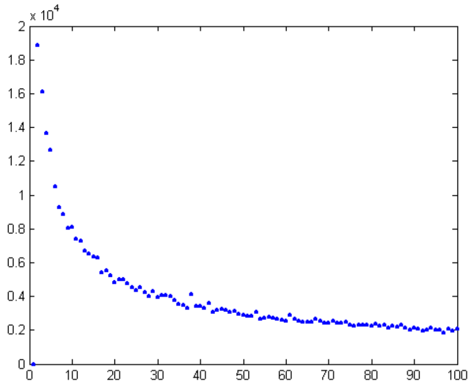


Figure 8: *Error as a function of the number of clusters for the vector with all 7 torsion angles.*

additional advantage of this method is that, as mentioned before, vector quantization naturally partitions the entire torsion space.

In order to possibly discover new motifs, we added some natural conditions to the merging in the full dimensional case. First, we merge two clusters only if the angles (β, ϵ, χ) have the coordinates of their centroids within the same peak. These peaks may be quite small and very difficult to observe via simple histogramming.

The second additional merging condition is a structural one: We define a "tagged cluster" as a cluster that corresponds to a well-established conformation such as A-helical or tetraloop RNA. Tagged clusters are protected, and are not merged with other clusters. Although there are relatively few of these clusters, they represent a large fraction of the RNA. Approximately 60% of globular RNA is found in the A-conformation.

The results of the proposed algorithm (VQ followed by merging of non-tagged clusters, or *modified vector quantization*) are presented in Table 3. Each row contains the ASCII code of the bin that matches the coding method of [8][7] (see Table 14 in Appendix C), and the enumeration of the peaks (numbers as obtained from the scalar quantization).

In addition to being fully automatic and capable of handling all the torsion angles at once, a clear advantage of the VQ method as compared to manual binning is the smaller numbers of classes that are needed to classify the structure. Vectorial binning is based on 26 clusters versus 38 bins in the usual binning method [8]. The main reason for this reduction in the number of classes is that the clustering algorithm does not recognize the "transition states" bins, or the bins classified as "other" from [8]. These are regions of conformation states that are very sparsely populated and which probably include energy bottlenecks between the low energy conformations. The result is that conformations that may be measurement error are included in the structural analysis; [18].

5 Automatically Finding Motifs: Validation

Most motifs that are already known have highly conserved three dimensional structures. They are also known to have higher stability than random loops. Finding those with the modified vector quantization method proposed above can be used as a validity measure of the algorithm, and this is the goal of the present

[7] In this code, the most popular residues are given the most popular letter of the alphabet. Classifying and labeling every residue with an ASCII letter allows one to used well-developed methods of searching and analysis of text to analyze RNA conformation. Reading text, establishing words and their relationships can allows unique insights into the three-dimensional structure that is encoded. See Appendix C for additional details.

Number	ASCII Code	Associated 4D Box	Remarks
1	a	3113	More than half of the cluster centroids are in this box
2	A	3113	$\beta \in [125-155]$ Appears in mismatch motifs
3	e	3112	$\epsilon \in [170-240], \zeta \in [180-230]$, Kink-turn motif.
4	J	3112	
5	E	3111	Takes part in the E-loop motif.
6	U	3213	$\beta = 94^0$ Takes part in the E-loop
7	u	3213	
8	o	2113	$\alpha_{center} \in [140-180]$, $\chi_{center} = 160^o$
9	O	2113	$\alpha_{center} \in [180-220]$, $\chi_{center} = 200^o$ Takes part in the GNRA tetraloop
10	n	3213	
11	r	3122	$\beta \in [140-200]$,Hyper-Twist motif
12	q	3122	$\beta \in [200-260]$
13	R	3121	$\zeta \in [40-100]$, Kink-Turn motif
14	Q	3121	$\zeta \in [100-160]$
15	h	3221	
16	d	1322	
17	z	3212	
18	s	2121	
19	t	1113	Starting conformation for an α stack
20	f	1112	
21	v	3323	Starting conformation for an α stack
22	c	1123	Takes part in kink-turn motif
23	i	2213	Crank shaft of A-form RNA
24	g	2123	
25	y	1312	Another crank shaft from A-form RNA
26	l	1213	

Table 3: *Results of the modified VQ on individual residues, see text for details.*

section. In particular, we compare the sites of different known motifs with search algorithms based on: 1) manual binning following [8]; 2) 4D vector quantization with the angles $(\alpha, \gamma, \delta, \zeta)$; and 3) 7D vector quantization with the whole torsion angles set.

5.1 GNRA Tetraloop

As described above, a *tetraloop* has been defined a four residue loop element that caps an A-helix [17]. The most abundant of the tetraloops has the GNRA sequence. The "G" forms a non-Watson/Crick base pair with the "A" that bends the loop to a "U" shape at the "N" site. This structure was found to be stable and ubiquitous in different RNA elements. The commonality and the ease of recognition of both the structure and the sequence make this motif a very good test case for any algorithm. In Table 4 we give the results of our search for the motif with the structure "aoaa." Note that "a" is the A form and "o" occurs when rotation of the α torsional angle from the g- to the trans-orientation. This forms the U-turn. The first column gives the locations of the motif, the second one the corresponding sequence of the four residues, the third one gives the conventional structural representation from the SCOR site, the fourth one shows the binning of the structure from [8], and the fifth and sixth columns show the results for the four and seven dimensional vector quantization, respectively.

We can find here a very good agreement among the methods. There is one structure, starting at 2062, that was classified as a tetraloop with the seven dimensional vector quantization only (while the four dimensional

First Residue	Sequence	Conventional Definition	Binning structure	4D VQ	7D VQ
253	UCAC	Tetraloop with GNRA fold	aoaa	aoaa	aoaa
469	GUGA	Tetraloop with GNRA fold	aoaa	aoaa	aoaa
577	GCGA	Tetraloop with GNRA fold	aoaa	aoaa	aoaa
691	GAAA	Tetraloop with GNRA fold	aoaa	aoaa	aoaa
805	GAAA	Tetraloop with GNRA fold	aoaa	aoaa	aoaa
1327	GAAA	Tetraloop with GNRA fold	aoaa	aoaa	aoaa
1469	CAAC	Pentaloop with GNRA fold	aoaa	aoaa	aoaa
1500	UAAU	Octaloop	aoaa	aoaa	aoan
1596	UAAU	Pentaloop	aoaa	aoaa	aoaa
1629	GAAA	Tetraloop with GNRA fold	aoaa	aoaa	aoaa
1863	GCAA	Tetraloop with GNRA fold	aoaa	aoaa	aoaa
2062	AUUC	Not defined	asaa	aqaa	aoaa
2249	GGGA	Tetraloop with GNRA fold	aoaa	aoaa	aoaa
2412	GAAA	Tetraloop with GNRA fold	aoaa	aoaa	aoaa
2630	GUGA	Tetraloop with GNRA fold	aoaa	aoaa	aoaa
2696	GAGA	Tetraloop with GNRA fold	aoaa	aoaa	aoaa
2877	GUAA	Tetraloop with GNRA fold	aoaa	aoaa	aoaa

Table 4: *Results for the GNRA tetraloop.*

VQ actually classifies it as a class very close by). There is another structure, starting at 1500, that was defined as a tetraloop by SCOR and the four dimensional quantization, but not by the full seven dimensional one. It should be noted that some of the GNRA structures are embedded within larger loops (those starting at $1469, 1500, 1596$).

5.2 E-Motif

Another motif with a well-defined structure is the *E-motif* [14]. This is a double-stranded motif with a pair mismatch that causes a bulging out of the $+$ strand. We were able to identify 6 such structures in the manual binning method in [8]. Here there are two conformations of the $+$ strand that give the same bulging geometry. The $-$ strand has a unique conformation. This motif seems to have large affinity to Mg^{2+} ions as well as to certain amino acids, especially Arg. All the structures from this class are classified as “looped with a dinucleotide platform in a triplet” by the SCOR library.

First residue	Sequence	Conventional Definition	Binning Structure	4D VQ	7D VQ
172	UCAGUA	S-Turn	aeshaa	aEshaa	aEszAa
210	UUAGUA	S-Turn	aeshaa	aEshaa	aEszAa
355	CCAGUA	Loops with a dinucleotide platform in a triple	aesdaa	aEsdaa	aEsdAa
585	CCAGUA	Loops with a dinucleotide platform in a triple	aesdaa	aEsdaa	aEsdAa
1069	CAGGAC		aes-aa	aEsdaa	aEgdAa
1367	AUAGUA		aeshaa	aEshaa	aEszAa
2689	AUAGUA	Loops with a dinucleotide platform in a triple	aeshaa	aEshaa	aEszAa

Table 5: *Results for the E-motif $+$ strand.*

Referring to Table 5, where the notation is as in the previous table, there is full agreement between the binning method and the seven dimensional vector quantization results. The result in residue 1069 was found only with the four dimensional VQ algorithm. Observing this region in the secondary structure figure does not reveal the bulging helical structure. The refinement of the ζ coordinate enables us to find that all of these motifs are in a conformation with ζ in the first peak. When performing the seven dimensional vector quantization, the last residue that is in the A RNA form belongs to a cluster with a centroid that deviates at β from trans to 140^o. Because of this, we did not merge this cluster with the "a" cluster. We also observe that "h" in the four dimensional quantization and the binning is replaced by "z" in the seven dimensional VQ.

First Residue	Sequence	Conventional Definition	Binning Structure	4D VQ	7D VQ
159	GAACU	S Turn	auaaa	auaaa	aUaaa
225	GAACG	S Turn	auaaa	auaaa	aUaaa
292	GACCG	Loops with a dinucleotide platform in a triple	auaaa	auaaa	aUaaa
568	GACCG	Loops with a dinucleotide platform in a triple	auaaa	auaaa	aUaaa
-	-	-	-	-	-
2053	GAACU		auaaa	auaaa	aUaaa
2701	GAACU	Loops with a dinucleotide platform in a triple	auaaa	auaaa	aUaaa

Table 6: *Results for E-motif $-$ strand.*

The results for the $-$ strand are shown in Table 6. This strand has the RNA A-form stack with a kink in the second residue. Here we can find full agreement between the binning results and the four dimensional vector quantization. The seven dimensional quantization gives the bin "U" instead of "u". The difference between the two in Table 6 is in the non-identifier angle $\beta = 94^o$ that is outside the main envelope. Here we have an example where a non-identifier torsional angle gives extra information which is needed for correct definition of the conformation.

5.3 Kink-Turn Motif

This motif described in [17] also has the double-stranded structure. The *kink-turn* consists of a bulging $+$ strand which has a conserved structure, and a $-$ strand which has a more flexible structure. We will focus our attention here on the $+$ strand only.

First Residue	Sequence	Conventional Definition	Binning Structure	4D VQ	7D VQ
43	UGAUGA	Loops with multiple triples	aedcra	aadcRa	aedcrA
93	CGAAGA	Kink-turn	aedcra	aedcRa	aedcrA
260	CAAUGU	Kink-turn	aedcra	aadcra	aedcrA
1027	GUUUGA	Kink-turn	ae*1ra	aeuvRa	aeuvra
1147	CCUAGA	Kink-turn	aedcra	aadcra	aedcrA
1312	GAUGGA	Kink-turn	aedcra	aadcra	aedcra
1601	GCAGGA	Kink-turn	aercra	aaRcra	aahcra

Table 7: *Results for the kink-turn motif $+$ strand.*

Referring to Table 7, we see some inconsistencies between the structures detected by the different methods. Two possible places for the ambiguity in the structure from the binning method are in the second place letter "e" and the fifth with the letter "r." In both of these places the ζ angle is out of the main peak, but the binning is not finely tuned enough to recognize the precise place; see also Tables 2 and 3. With seven dimensional vector quantization, it is obvious that one can find ζ in a second peak, which emphasizes an advantage of using the full dimensional quantization technique.

5.4 Hyper-Twist Motif

The *hyper-twist* is another motif that is based on the double helix structure. Here the double strand is twisted around a purine-purine mismatch. The mismatch is usually a G-A pair. This motif typically has a symmetric structure. There is a G-A pair and an A-G pair. In Table 8 we included both the $+$ and the $-$ strand.

First residue	Sequence	Binning Structure	4D VQ	7D VQ
$20-27$	GGUGGAUU	aaaaraaa	aaaaraaa	aaaarAaa
$516-523$	AUGAAAUC	aaraaaaa	aaraaaaa	aaraaaaa
$365-370$	GUGCGG	aaraaa	aaraaa	aaraaa
$279-281, 285-287$	CCU, AUC	iaaaaa	iaaaaa	iaaaaa
$792-799$	GAUGAAGC	aaaraaaa	aaaraaaa	aaaraaaa
$814-822$	GUGGAAGUC	aaaranzaa	aaaranHa	aaarAnHaa
$1585-1592^{*}$	CGUGGAAG	aaaarara	aaaaraRu	aaaarARu
$1602, 1605-1610$	C,GAAGCG	eraaaaa	araaaaa	araaaaa
$1881-1887$	ACUGAAU	iaaraa	iaaraaa	iaaraaa
$2015, 2016, 1771, 1847-1850$	A,U,U,AGGU	aa7aaaa	aagaaaa	aagAaaa
$2500-2505$	CGCAAG	aaaraa	aaaraa	aaarAa
$2515-2520^{*}$	CGACCG	aeaaaa	aeaaaa	aeaaaa

Table 8: *Hyper-twist motif $+$ and $-$ strands. The SCOR description of these sites is mostly of "stacked paired non-Watson/Crick double strand," or "cross strand." At $*$ the structure is considered to be a kink-turn motif. For the 7D VQ there is a clear preference for the "r" conformation to be in one of the complexes ($\zeta = 130$).*

The entries marked with a $*$ have a $-$ strand with conformation "e" instead of "r." There are some conformations which include a bulge. One of them coalesces with a kink-turn motif. It was found that all of the mismatch conformations that were marked by "r" belong to one specific cluster. We used this to unmerge this cluster from the other clusters that were merged with it before; see Table 3.

5.5 Mismatched GA-Motif

All of the above motifs are characterized by a double helix structure which may be twisted or bulged. The deformation is a result of a base pair mismatch. This is a secondary structural characteristic. We can find also a unique conformation in almost all of the above-mentioned cases. After a base pair mismatch the residue acquires the conformation marked by "A." This conformation is a single cluster. The identifier torsional angles of this conformation have the same values as that of the A-form helix. The only difference is that the β value is shifted to the shoulder of the main peak in the histogram of β. See also [26] for related results.

The α, β plot of this cluster is given in Figure 9. The binning method (even with scalar quantization) as well as 4D VQ cannot recognize this cluster, while the the full seven dimensional VQ can. A similar cluster was found with the electron density technique in [26]. This conformation appears in the following locations:

1. Hyper-Twist: 25, 818, 1590, 2504.

2. Kink-Turn: 48, 98, 265, 1152.

3. E-motif: 176, 214, 359,1073, 1371, 2693.

There is a generalization of the hyper-twist, that is a mismatched double strand that includes the "A" conformation in: $721 \in (716-726, 702-712)$, $1032 \in (1031-1041, 929-939)$, $1742 \in (1733-1744, 2035-2046)$,$1528 \in (1527-1534, 1657-1664)$, $2827 \in (2826-2830, 2910-2914)$,$2883 \in (2880-2889, 2868-2877)$. The conformation of this double strand is less conserved than the hyper-twist.

Other “bulge motifs” with pair mismatches include $442, 465, 489, 593, 2244, 2427, 2259, 2906,$ as well as the short double helix with internal loop, 2427. Some more complicated mismatch structures are

$$382, 489, 645, 1528, 1891, 1973, 2485, 2675, 2817, 2904.$$

The conformation “A” appears in three places where it cannot be associated with mismatched structures.

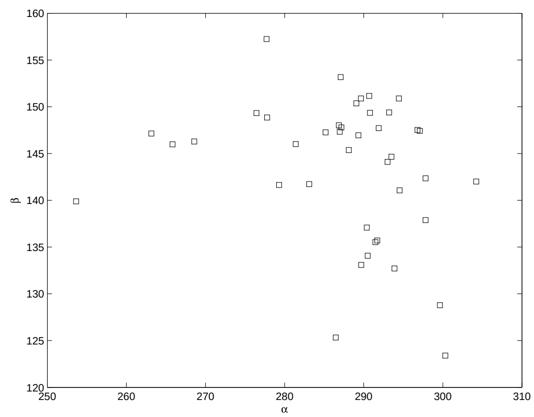


Figure 9: *The α, β torsional angles of the “A” cluster*

5.6 Helix Initiation Knee

The *helix initiation knee* is a motif that has a bend at the beginning of a helix [26]. As its name implies, such a motif is associated with a “knee” between two adjacent helices. This occurs for example in the case of the “knee” between the T stem and the acceptor stem. We defined this motif to have the binning sequence ”taaa;” see Table 9. We found this form to repeat 19 times in RR0033. Table 9 summarizes the search results for this motif.

From the 21 structures that were found to be in the desired conformation (see Table 9), only 6 were not initiating a double helix and also within these 6 cases there is a large number of tertiary interactions with other parts of the LSU. The differences among the methods are minimal and are mostly confined to the case where t (1111) is changed with the transition state p in the binning method (4111) (which includes the “other” region absent in the VQ method).

Another type of helix initiation motif has a typical conformation of “vaaa” in one of the strands. The “va” conformation was recognized in [26]. There were 13 such structures and they are summarized in Table 10.

There is full agreement between the 4D and the 7D VQ’s. Only seven of the above examples have the same binning structure. The 3-double helix is a structure where the first residue in the “v” conformation is unpaired. It seems that for this motif the binning definition of the ”v” conformation gives a more uniform motif. This will be addressed in more detail in Section 6 below.

6 Discussion

RNA conformational motifs were characterized here with statistical techniques from classical signal processing. These automatic procedures do not use visual inspection or filtering. The overriding goal is to establish fast and easily applied yet rigorous methods for analysis of RNA conformation. The simplest

First Residue	Sequence	Conventional Definition	Binning Structure	4D VQ	7D VQ
116	GAGA	Double Helix	taaa	taaa	taaa
398	UCCC	Double Helix	taaa	taaa	taaa
636	GCCA	Double Helix	taaa	taaa	taaa
762	CCCG	Double Helix	taaa	taaa	taaa
826	UAGA	Double Helix	taaa	taaa	taaa
895	ACAG	Double Helix	taan	taaa	taae
961	ACCG		taaa	taaa	taaa
1089	GAUA	Double Helix	taaa	taaa	taaa
1138	GUCU	Double Helix	taaa	taaa	taaa
1177	AGCU		paaa	taaa	taaa
1703	GGCG	Double Helix	taaa	taaa	taaa
1986	GCCG	Double Helix	taaa	taaa	taaa
2023	GAGC	Double Helix	taaa	taaa	taaa
2059	UACA		taaa	taaa	taaa
2084	CACC	Double Helix	taaa	taaa	taaa
2284	GGGC		taaa	taaa	taaa
2444	UUGA	Double Helix	taaa	taaa	taaa
2566	AGAA		taaa	taaa	taaa
2738	GAGA	Double Helix	taaa	taaa	taaa
2750	GCCG	Double Helix	paaa	taaa	taaa
2840	AAGA		paaa	taaa	taaa

Table 9: *Results for double helix initiator element.*

method used here, scalar quantization, treats each dimension in isolation of the others. SQ successfully resolves the torsion angle ζ into the three distinct clusters (three rotamers) predicted by the potential energy surface. This resolution of ζ into three was not accomplish in [8], and was found by visual inspection in [18] only after application of quality filters. This is achieved following well-defined optimality criteria, as well as the automatic analysis of multiple angles at once.

With VQ, populated clusters of RNA conformation are determined in simultaneous analysis of any dimensionality, up to the full seven dimensional torsional space. We believe this work represents the first analysis of RNA torsional space at greater than three simultaneous dimensions (i.e., of more than three torsion angles). Here four dimensional VQ was applied first to the four angles $(\alpha, \gamma, \delta, \zeta)$ that have previously been termed "identifier angles" [8], because they appear to completely specify fundamental RNA conformations. The remaining three dimensions are considered to be dependent on the four identifier angles, although are important for conformations search, see below. With a well-defined distortion measure, the number of four dimensional clusters was found by 4D VQ to be about 60. This result suggests that there are about 60 fundamentally distinct nucleotide conformational states within globular RNA. 4D VQ identified each of the populated bins reported in [8], which were obtained via manual classification. Agreement with SCOR was found as well.

We then added a merging stage to the VQ method, which is based both on cluster centroid proximity and on structural constraints, thereby adapting the generic VQ technique to the study of RNA torsion angles. For example, all clusters that meet the definition of A-helical RNA were merged into a single cluster. This initial over-population of A-helical RNA clusters was expected, since due to their popularity, VQ allocates to them a large number of resources (centroids) in order to minimize the distortion.[8]

[8]Of course, for tasks different to the one in this paper, such a merging might not be needed, and considering the different clusters can lead to a more detailed analysis, for example, of the A-helical variability in the search for "micro-conformations."

First Residue	Sequence	Conventional Definition	Binning Structure	4D VQ	7D VQ
169	AUCU		1aaa	vaaa	vaaa
331	AGGG	3-Double Helix	vaaa	vaaa	vaaa
620	ACGU	Double Helix	vaaa	vaaa	vaaa
905	CCAA	Double Helix	vaaa	vaaa	vaaa
1239	GGGA	Double Helix	vaaa	vaaa	vaaa
1438	GCUG	3-Double Helix	1aaa	vaaa	vaae
1626	AGGG	2-Double Helix	vaaa	vaaa	vaaa
1634	GUGA	Double Helix	0aaa	vaaa	vaaa
1710	AAAG	3-Double Helix	1aaa	vaaa	vaaa
1919	ACAA		*aaa	vaaa	vaaa
1996	UAGC	bulged Double Helix	1aaa	vaaa	vaaa
2526	CUUG	3-Double Helix	*aaa	vaaa	vaaa
2638	GGUC	Double Helix	vaaa	vaaa	vaaa

Table 10: *Results for double helix initiator "v" element.*

We then used this modified VQ on the full set of seven torsion angles defined by a single nucleic acid residue. This study of the full seven dimensional space led to new conformations that were not present at the one or four dimensional studies. We validated the method by comparing it with known structural motifs, as well as the SCOR classification. The minor mismatches could be a result of a too coarse clustering (different motifs merged into a single cluster). We tested adding clusters (up to 100), and found small changes that are enough to fix these discrepancies (while requiring additional merging to eliminate the not-novel clusters).

We found a conformational signature for the existence of a mismatch motif, an umbrella motif that includes the bulging or twisted double-stranded cases. We found this conformation only when we used the modified 7D VQ, showing the importance of working with the whole conformational space, and thereby the need for formal analysis technique as the one here described, that go beyond ad-hoc visualization-based approaches.

In the next step (in progress), we will seek relationship between neighboring clusters using the method of *mutual information.* As has been done for secondary structures in protein research, e.g., [6], it is important to study the dispersion within clusters. It seems likely that information on shapes of potential energy surfaces and RNA dynamics is contained within the cluster shape. Some preliminary results on mutual information for RNA are described in Appendix A. Finally, following work on proteins [6], we can perform principal component analysis (PCA) on various clusters. Preliminary results on this topic are presented in Appendix B.

To conclude, in this paper we have seen how some standard techniques from statistical signal processing are useful for the analysis of RNA structure. These techniques cover the automatic discovery of torsion angles clusters, their grouping into motifs, and the analysis of motif populations. These techniques can be augmented with novel clustering approaches being developed by the learning and signal processing communities, and investigating those, together with the search for new motifs and the extension of the preliminary results in Appendices A and B, are the subject of some of our current efforts.

Appendix A: Mutual Information Analysis of RNA Parsing

In the above sections, we discussed the use of vector quantization, a classical technique from statistical signal processing, to cluster RNA residues and find torsion angles. All the work was done at the level of residue parsing, while it can be easily repeated at the level of suite parsing [18]; see Figure 1. The motivation for the latter is the high correlation between the adjacent phosphate torsional angles ξ and α. This correlation was established for dinucleotides and short oligonucleotides [27]. Here we will extend the relation to any RNA molecule using tools from information theory. This brings yet another classical tool from statistical signal processing that can bring some light into important RNA questions.

To try to further understand the differences between the two forms of parsing the RNA backbone, as well as the general dependency of the different torsion angles among themselves, we computed the mutual information between α and ζ, both for residue parsing ($\alpha(i)$ against $\zeta(i)$) and for suite parsing ($\alpha(i)$ against $\zeta(i-1)$). This is repeated for all torsion angles. Mutual information is defined as follows [3]: Let x and y be two random variables. First, the *entropy* of x is defined as

$$H(x) := -E_x[\log(P(x)]$$

where $E_x[\cdot]$ stands for the expectation. Entropy measures (in bits) the randomness of a signal, the larger the entropy the more random the variable is. The *joint entropy* is defined as

$$H(x,y) := -E_x[E_y[\log(P(x,y))]]$$

and summarizes the degree of dependence of x on y, while the *conditional entropy* if given by

$$H(y|x) := -E_x[E_y[\log(P(y|x))]]$$

which summarizes the randomness of y given knowledge of x. We can now define the *mutual information*,

$$MI(x,y) := H(y) - H(y|x) = H(x) + H(y) - H(x,y),$$

which is a measure of the reduction of the entropy (randomness) of y given x.

In the case of residual parsing, we obtained $MI(\alpha,\zeta) = 0.85$, while for suites parsing we obtain $MI(\alpha,\zeta) = 1.17$.[9] This increase in mutual information indicates that these torsion angles are functionally more dependent with suites parsing.[10] In Table 11 we provide the mutual information for all the seven torsion angles (the values in the diagonal are the entropy) for residual parsing. When we consider suite parsing, the mutual information between a given angle $(\alpha(i), \beta(i), \gamma(i), \delta(i), \epsilon(i), \zeta(i), \xi(i)$ and $\zeta(i-1)$ is given by

$$(1.1716, 0.9435, 0.6646, 0.4623, 0.6529, 0.9067, 0.8273)$$

respectively. From Table 11, we observe that the mutual information is not a direct consequence of the proximity between the angles (otherwise, we would have monotonic functions per row and column).

To further illustrate the torsion angle dependency, we repeat the above computations when we remove the A-helix from the data, obtaining Table 12 for residual parsing and

$$(2.3465, 1.8704, 1.5771, 0.8699, 1.5474, 2.2656, 1.7937)$$

for suite parsing ($(\alpha(i), \beta(i), \gamma(i), \delta(i), \epsilon(i), \zeta(i), \xi(i)$ against $\zeta(i-1)$). We first note that the improvement in mutual information is now negligible (from 2.2989 for residue parsing to 2.3465 for suite parsing). We also observe, as expected, that the entropy increased, since we have removed the most popular conformation. Lastly, the mutual information has significantly increased in general.

[9] Both α and ζ have $H = 4.6$.

[10] For computing the MI, we quantized the α and ζ torsion angles in 100 bins. We also tested for different numbers of bins and always the mutual information increased for suite parsing.

	α	β	γ	δ	ϵ	ζ	ξ
α	4.5983	0.9956	0.8081	0.4538	0.6809	0.8473	0.8393
β	0.9956	4.4875	0.6862	0.3550	0.4955	0.7480	0.6550
γ	0.8081	0.6862	3.5670	0.3852	0.4332	0.5624	0.5683
δ	0.4538	0.3550	0.3852	2.7089	0.5566	0.5898	0.6666
ϵ	0.6809	0.4955	0.4332	0.5566	4.3744	0.8524	0.6801
ζ	0.8473	0.7480	0.5624	0.5898	0.8524	4.6199	0.8744
ξ	0.8393	0.6550	0.5683	0.6666	0.6801	0.8744	4.2753

Table 11: *Mutual information for all the torsion angles, considering a residual parsing.*

	α	β	γ	δ	ϵ	ζ	ξ
α	5.8792	1.9543	1.7290	0.9308	1.5874	2.2989	1.9515
β	1.9543	5.3180	1.4260	0.6587	1.1823	1.8622	1.4699
γ	1.7290	1.4260	4.6269	0.8218	1.0330	1.5486	1.3073
δ	0.9308	0.6587	0.8218	3.4348	0.8487	1.0948	1.0853
ϵ	1.5874	1.1823	1.0330	0.8487	4.9315	1.7319	1.3214
ζ	2.2989	1.8622	1.5486	1.0948	1.7319	5.8107	1.9958
ξ	1.9515	1.4699	1.3073	1.0853	1.3214	1.9958	5.2165

Table 12: *Mutual information for all the torsion angles, considering a residual parsing, and ignoring the A-helix conformations.*

Appendix B: Principal Component Analysis of Key Motifs

Following the work on proteins [6], we can perform principal component analysis (PCA) on the tetraloop structures.

The basic procedure is as follows. Let L denote the number of residues in the motif ($L = 4$ for tetraloops) and N the number of samples (27 for our first example). The first step in the PCA procedure is to compute the covariance matrix C, which is a square matrix of dimension $4L$ (four angles per each residue), whose elements are given by

$$C_{i,j} = \frac{1}{N-1} \sum_{m=1}^{N} (x_{mi} - < x_i >)(x_{mj} - < x_j >)$$

where $< x_i >$ is the i-th coordinate of the mean structure. We then compute the eigenvalues and eigenvectors of this matrix, λ_q and $\vec{v}_q$. The distribution of the eigenvalues will tell us the number of modes in this class.

In Figure 10, left, we clearly see 2 to 3 dominant eigenvalues for this data set, considering the 4 angles $(\alpha, \gamma, \delta, \zeta)$. In the middle, we repeat the computation for a total of 261 tetraloops,[11] considering now all the six torsion angles ($\alpha, \beta, \gamma, \delta, \epsilon, \zeta$), and defining a tetraloop as the combination

$$(3?11?3, 3?11?3, 2?11?3, 3?11?3)$$

where the symbol $?$ stands for "don't care" for those angles. We observe again the 2 (maximum 3) dominant eigenvalues (analysis of the eigenvectors will be reported elsewhere). When using the same data set, again

[11] RR0011, RR0033, RR0055, RR0043, RR0044, RR0060, RR0061, RR0077, RR0078 and RR0079; HLSU 50 from NDB.

with all the six torsion angles, but defining a tetraloop as

$$(3?11?3, 2?11?3, 3?11?3, 3?11?3)$$

we obtain 168 examples. The eigenvalues distribution is shown in the last figure on the right, with two dominant eigenvalues once again, even stronger than before.[12] Note that the first and second histograms of Table 10 refer to "tetraloops" in the sense just defined, while the third histogram refers the "tetraloops" in the standard sense [14, 30].

We have used simple (and linear) analysis in this case, while there is no reason to believe that the space of RNA motifs is flat. We plan to investigate the use of tools that consider the geometry of the space of motifs, e.g., [29], where orders of magnitude more data will be needed.

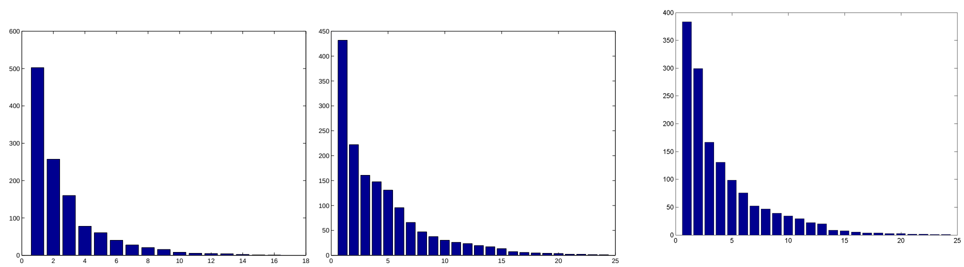


Figure 10: *Frequency plots of eigenvalues corresponding to the tetraloop PCA analysis. The first two plots use tetraloops in the sense defined in this paper while the third in the standard sense.*

Appendix C: Background on the Binning Method

In this Appendix, we briefly review the main results reported in [8]. We introduce here some minor modifications in order just to elucidate the main ideas. We finally repeat some of the key tables from [8] for easy reference on the part of the reader.

Binning as formulated in [8] is an histogram-based method for describing RNA conformation and for identifying RNA tertiary structural motifs. The conformation of each bond can be described by a small number of discrete integers. Each residue can be assigned to a distinct configuration class. Further, some of the torsion angles are dependent or highly restrained. In summary, one can reduce the full multi-dimensional torsion angle space to a set of 38 configuration classes. An ASCII code can be assigned to each configuration class. Thus the three-dimensional description of conformation is reduced to a single-dimension.

More precisely, each torsion angle of a given residue is allocated to the appropriate bin. By definition, torsion angles with single-peak distributions cannot be readily separated into distinct bins, because essentially all the angles are contained under a single envelope. Because of this the angles β, ϵ, and χ are assumed not to contribute information to the conformational description, and are ignored; see [8]. Because of their multi-peaked nature, the remaining four torsion angles and P allow a straightforward separation into distinct configuration classes. However, δ and P are correlated, both by geometric definition, and from analysis of the HM 23S rRNA data. Thus, to avoid redundancy, we eliminate P and consider only four torsion angles α, γ, δ, and ζ. The reduction in parameters led us to a four digit structural representation of the conformation of a given residue. Each residue is assigned a sequence of four integers $n_\alpha, n_\gamma, n_\delta, n_\zeta$ where each digit denotes the envelope to which a torsion angle belongs.

Binning has several important advantages:

[12]The stability of these motifs, and comparison between residue and suite parsing, is the subject of current studies.

1. It allows one to exploit the large and sophisticated pattern recognition capabilities already developed for 1-dimensional databases.

2. It allows one to combine sequence and conformational information in the same 1-dimensional representation, for example by interleaving the ASCII binning characters with sequence characters.

3. It allows one to represent conformational information along with base-pairing, tertiary interaction , etc., in simple 2-dimensional representations.

4. It can be readily tuned to a given organism, class of RNA, etc.

5. It is relatively easy to implement, and may be fully automated as indicated in this paper.

The results of the method in [8] are summarized in tables 13 and 14.

bin index	1	2	3	4
α	$[40-100]$	$[125-200]$	$[220-350]$	others
γ	$[10-110]$	$[140-210]$	$[230-350]$	others
δ	$[65-105]$	$[130-165]$	others	
ζ	$[240-350]$	others		

Table 13: *Enumeration and borders of the bins from [8].*

References

[1] S. E. Butcher, T. Dieckmann, and J. Feigon, ”Solution structure of a GAAA tetraloop receptor RNA,” *EMBO J.* **16** , pp. 7490-7499, 1997.

[2] T. Cech, “Ribozymes, the first 20 years,” *Biochem. Soc. Trans.* **30**, pp. 1162-1166, 2001.

[3] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley-Interscience, 1991.

[4] *Data Compression*, www.data-compression.com/vq.html.

[5] C. Duarte and A. Pyle, “Stepping through an RNA structure: A novel approach to conformational analysis,” *J. Mol. Biol.* **284**, pp. 1465-1478, 1998.

[6] E. Emberly, R. Mukhopadhyay, N. Wingreen, and C. Tang, “Flexibility of alpha-helices: Results of a statistical analysis of database protein structures,” *J. Mol. Biol.* **327**, pp. 229, 2003.

[7] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, January 1992.

[8] E. Hershkovitz, E. Tannenbaum, S. B. Howerton, A. Sheth, A. Tannenbaum, and L. D. Williams, “Automated identification of RNA conformational motifs: Theory and application to the HM LSU 23S rRNA,” *Nucleic Acids Res* **1**, pp. 6249-6257, 2003.

[9] S. Hecht (editor), *Bioorganic Chemistry: Nucleic Acids*, Oxford University Press, 1996.

[10] A. Hinneburg, M. Fischer, and F. Bahner, “Finding frequent substructures in 3D-protein databases,” *Data Base Support for 3D Protein Data Set Analysis – 15th International Conference on Scientific and Statistical Database Management*, pp. 161-170, 2003, Cambridge, MA.

[11] F. M. Jucker and A. Pardi, “GNRA tetraloops make a U-turn,” *RNA* **1**, pp. 219-222, 1995.

[12] P. Klosterman, M. Tamura, S. Holbrook, S. Brenner, “SCOR: a structural classification of RNA database,” *Nucleic Acids Res.* **30**, pp. 392-394, 2002.

[13] A. Leach, *Molecular Modeling: Principles and Applications (Second Edition)*, Prentice-Hall, New York, 2001.

[14] N. B. Leontis and E. Westhof, “Analysis of RNA motifs,” *Curr. Opin Struct Biol* **13**, pp. 300-308, 2003.

[15] Y. Linde, A. Buzo, and R. M. Gray, “An algorithm for vector quantizer design,” *IEEE Trans. on Comm.*, pp. 702-710, 1980.

[16] F. Michel and E. Westhof, “Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis,” *J. Mol. Biol.* **216**, pp. 585-610, 1990.

[17] J. B. Moore, “Structural motifs in RNA,” *Annual Rev. Biochemistry* **68**, pp. 287-300, 1999.

[18] L.J. W. Murray, W. B. Arendall, III, D. C. Richardson, and J. S. Richardson, “RNA backbone is rotameric,” *PNAS* **100:24**, pp. 13904-13909, 2003.

[19] V. L. Murthy, R. Srinivasan, D. E. Draper, and G. D. Rose, “A complete conformational map for RNA,” *J. Mol. Biol.* **291**, pp. 313-327, 1999.

[20] V. L. Murthy, and G. D. Rose, “RNABase: An annotated database of RNA structures,” *Nucleic Acids Res.* **31**, pp. 502-504, 2003.

[21] A. Y. Ng, M. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” *NIPS* **14**, 2002.

[22] *Nuclei Acid Database*, http://ndbserver.rutgers.edu.

[23] W. K. Olson, “Configuration statistics of polynucleotide chains. A single virtual bond treatment,” *Macromolecules* **8**, pp. 272-275, 1975.

[24] G. N. Ramachandran and V. Sasisekharan, “Conformation of polypeptides and proteins,” *Adv Protein Chem* **23**, pp. 283-438, 1968.

[25] G. N. Ramachandran and V. Sasisekharan, “Stereochemistry of polypeptide chain configurations,” *Adv. Protein Chem.* **23**, pp. 283-437 (1968).

[26] B. Schneider, Z. Moravek and H. M. Berman, “RNA conformational classes,” *Nucleic Acids Research* **32**, pp. 1666-1677, 2004.

[27] W. Saenger, *Principles of Nucleic Acid Structure*, Springer-Verlag, New York, NY, 1984.

[28] M. Sundaralingam, “Stereochemistry of nucleic acids and their constituents. Allowed and preferred conformations of nucleosides, nucleoside mono-, di-, tri, -tetraphosphates. Nucleic acids and polynucleotides,” *Biopolymers* **7**, pp. 821-860, 1969.

[29] J. B. Tenenbaum, V. De Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science* **290**, December 2000.

[30] C. R. Woese, S. Winker, and R. Gutell, “Architecture of ribosomal RNA: constraints on the sequence of ‘tetraloops’,” *Proc. National Academy of Sciences* **87**, pp. 8467-8471, 1990.

Number	ASCII Code	Associated 4D Box
1	a	3111
2	e	3112
3	r	3122
4	i	2211
5	o	2111
6	t	1111
7	n	3121
8	s	2122
9	l	1211
10	u	3211
11	c	1121
12	d	1322
13	p	4111
14	m	1122
15	h	3222
16	g	2121
17	b	4211
18	f	1112
19	y	1311
20	w	2222
21	k	4122
22	v	3311
23	x	4112
24	z	3213
25	j	2212
26	q	2112
27	1	3321
28	2	3322
29	3	1221
30	4	1321
31	5	3411
32	6	3131
33	7	4121
34	8	1212
35	9	2411
36	0	4311
37	+	3312
38	-	1222

Table 14: *ASCII code alphabet for the binning method from [8].*