Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2011 ; 8(6): 1604–1619. doi:10.1109/TCBB.2010.128.

Nonparametric Clustering for Studying RNA Conformations

Xavier Le Faucheur,

School of Electrical and Computer Engineering, Georgia Institute of Technology, UA Whitaker Building, 313 Ferst Drive, Atlanta, GA 30332-0535. xavier@gatech.edu

Eli HersHKovits,

School of Electrical and Computer Engineering, Georgia Institute of Technology, UA Whitaker Building, 313 Ferst Drive, Atlanta, GA 30332-0535. eli@bme.gatech.edu.

Rina Tannenbaum, and

School of Materials Science and Engineering, Georgia Institute of Technology, Love Building, Room 274, 771 Ferst Drive, NW, Atlanta, GA 30332-0245. rina.tannenbaum@mse.gatech.edu.

Allen Tannenbaum

School of Electrical and Computer Engineering, Georgia Institute of Technology, UA Whitaker Building, Room 4201, 313 Ferst Drive, Atlanta, GA 30332-0535. tannenba@ece.gatech.edu.

Abstract

The local conformation of RNA molecules is an important factor in determining their catalytic and binding properties. The analysis of such conformations is particularly difficult due to the large number of degrees of freedom, such as the measured torsion angles per residue and the interatomic distances among interacting residues. In this work, we use a nearest-neighbor search method based on the statistical mechanical Potts model to find clusters in the RNA conformational space. The proposed technique is mostly automatic and may be applied to problems, where there is no prior knowledge on the structure of the data space in contrast to many other clustering techniques. Results are reported for both single residue conformations, where the parameter set of the data space includes four to seven torsional angles, and base pair geometries, where the data space is reduced to two dimensions. Moreover, new results are reported for base stacking geometries. For the first two cases, i.e., single residue conformations and base pair geometries, we get a very good match between the results of the proposed clustering method and the known classifications with only few exceptions. For the case of base stacking geometries, we validate our classification with respect to geometrical constraints and describe the content, and the geometry of the new clusters.

Index Terms

RNA conformation; clustering; potts model; statistical mechanics

1 Introduction

In this work, we apply a clustering algorithm based on the Potts model [1] to the problem of data mining of RNA structures. RNA is a biopolymer, that is, involved in a multitude of activities in the cell [2], and hence, possesses a high-degree of structural and functional variability.

The functional diversity of the RNA molecule depends on the ability of the RNA polymer to fold into a large number of precisely defined spatial forms. Over the last few years, the data bank of RNA structures has considerably grown due to major efforts made by experimentalists leading to great advances in crystal growth protocols. The Nucleic Acid Database (NDB) [3] and the protein data bank [4] accommodate the structural information of numerous RNA molecules from small RNA structures with only a few nucleotides to very large ones with thousands of nucleotides, e.g., ribosomal RNA. The structural resolution achieved in these RNA molecules (obtained by the use of various characterization techniques) allows one to give an estimate for the location of individual atoms [5].

One of the main challenges of bioinformatics is to develop data mining tools for such documented RNA structures, in order to establish a clearer understanding of the structure/function relationships in these molecules. In most cases, this problem is too complicated to be solved computationally [6]. To date, efforts in this respect have focused on finding repetitive smaller substructures, i.e., *structural motifs* [7]. If the functionality of a specific substructure from a given structural motif is known, then the functionality of other substructures with a similar three-dimensions form can be assumed to be similar. Therefore, the main task in the classification of these structural motifs is to define a similarity measure for substructures and to cluster motifs accordingly [8], [9], [10], [11]. In this work, we will limit our analysis to the clustering of the most basic building units of the RNA, namely, the single nucleotide and the nucleotide doublets.

RNA nucleotides (residues) are comprised of two distinct moieties: a flexible backbone consisting of ribose rings bridged by phosphate groups and rigid bases consisting of either purines or pyrimidines. Most of the nucleotide interactions in an RNA molecule are due to interactions between bases. Given the differences between the flexible backbone and the rigid bases in RNA residues, the three-dimensions structure can be described by two complementary representations (see Fig. 1): the backbone conformations [8], [9], [12], [13], [14], [15] of a single residue and the geometries of the base interactions [16].

The building block for the backbone consists of either the residue or the base-to-base suite [10] (see Fig. 1a). In the representation of the flexible backbone, residues are well-described by a set of six torsional angles, whereas suites necessitate considering seven torsional angles.

The representation of base interactions depends on six parameters, which describe the relative translation and rotation that are needed to align one base with the other. In this type of conformation, the co-ordinate system is composed of three rotation angles and a three-dimensions vector representing the base-to-base distance. Note that the representation is not unique and depends on the choice of origin for the transformations.

Whereas the distances and angles are continuous parameters, differentiation of substructures and structural classification in both representations requires discrete criteria. For example, base pair geometries may be organized into 12 classes with respect to the interacting edges of the bases [17]. Single nucleotide conformations can be classified into groups of rotamers [10].

For both representations, the recognition and definition of the classes are formulated as a segmentation problem, which deals with partitioning of the continuous data space into a finite collection of well-defined subspaces. This segmentation is done by recognizing the underlying clusters in the data space.

There are numerous different clustering methods, which can be classified into parametric (for example, k-means) and nonparametric methods such as hierarchical graph methods [18]

[19]. Parametric methods are characterized by the assumption that the number of clusters in the data is known and nonparametric methods are based on a prior knowledge of the distribution of data points within the clusters. These classical methods are not very accurate when the underlying distribution of the data points cannot be well approximated. The deficiencies of clustering algorithms are especially evident in the case of RNA, where data is hard to get and resolution is poor.

To date attempts to perform structural data mining on RNA have been done using either qualitative observations of the data space [8], [11], or parametric clustering analyses [9], [20], [21]. The qualitative classification technique is tedious and inaccurate since the dimensionality of the data space can be larger than three and because the distribution of data points can be fuzzy. The parametric clustering techniques are complicated by the need for prior knowledge [9], [20]. Therefore, in this work, we introduce the use of nonparametric clustering methods. A previous attempt in this direction has proposed a hierarchical graphical clustering approach for the case of base pair interactions [21]. The main drawback of this method was that the choice of the hierarchy level at which the proper clustering was defined could not be determined.

Among the most promising alternatives for the clustering of such data is a method based on the statistical-mechanical Potts model [1]. In this approach, a spin parameter is assigned to each data point and an interaction parameter is attached to each pair of neighboring data points. The closer two points are, the stronger the interaction between them, and the more likely they will belong to the same cluster. The partition of the data space into clusters can be found by employing spin-spin correlation functions. The merit of this technique is that no prior knowledge of the data point distribution is needed in order to devise the different parameters. Indeed, the Potts model is a nonparametric hierarchical clustering approach, which gives intrinsic and objective criteria for defining the correct level of the hierarchy for producing clusters that optimally match the underlying physical model. This is done by linking the data space to the space of an alternative physical problem, where one tries to find paramagnetic regions in a potential induced arrangement of magnetic particles.

Susceptibility and *temperature* turn out to be natural parameters and variables for describing the clustering problem. *Temperature* plays the role of hierarchy level or *depth* in the iterative subdivision of the data into clusters. The *susceptibility* graph gives a simple criterion for selecting the temperature at which clusters are determined [1]. The values of other parameters intrinsic to the Potts model can be optimized by the use of some straightforward numerical criteria [1], [22]. The Potts model is advantageous relative to standard parametric methods by virtue of the fact that no prior knowledge about the number of clusters is needed and is also much simpler to employ than most other nonparametric algorithms.

In the present work, we describe the results of applying the Potts clustering method to the RNA structure problem. We have validated known clustering results and found clustering for the base stacking case, which to the best of our knowledge had not been solved in the past. We have validated this new clustering case by applying certain geometrical constraints emerging from the RNA Watson-Crick (W-C) double helix structure.

The remainder of this paper is organized as follows: In the next two sections, we will describe and discuss the Potts model that we have employed for clustering. In Sections 4 and 5, we describe specific applications to the single and double nucleotide classification problems and compare these results to some known classification. Then, we report the results of the clustering method for the base-stacking problem. Finally in Section 6, we summarize our results and draw some conclusions.

2 Background on Data Clustering Using the Potts Model

As a priori information on the number and the size of conformational classes may not be available for a given data set, a nonparametric clustering method fits our RNA structure classification problem. Plus, such methods are more suitable to find new elements in the classification. The method, that is, presented here has been proposed by Blatt et al. [1] and is based on a Potts spin model, which was developed for the analogous problem of the physical properties of an inhomogeneous ferromagnet. In this model, clusters consist of magnetic islands containing sites with similar Potts states.

2.1 Description of the Model

We now give some specific details on the Potts model. We refer to the N points of the given data set as *magnetic sites*. Each site is assigned a Potts spin denoted by s . Spin values are taken from a set of q distinct integers, where q is a parameter to be set. The allocation of a spin value to every magnetic site results in a unique spin configuration S that entirely defines the state of the system. One can then define q^N different spin configurations. Moreover, the spins s_i and s_j of two sites i and j are said to be “aligned” if they have the same value (i.e., if $s_i = s_j$).

Also, two spins s_i and s_j interact with each other with a strength J_{ij} . For computational reasons one assumes that a given spin will have a significant interaction with some of its closest neighbors only [1]. The method for choosing neighbors is described in Section 3.1.

The energy of a spin configuration S is evaluated by the following Hamiltonian:

$$H(S) = \sum_{\langle i, j \rangle} J_{ij} \cdot (1 - \delta_{s_i s_j}), \quad (1)$$

where $\delta_{s_i s_j} = 1$ if $s_i = s_j$ and $\delta_{s_i s_j} = 0$ otherwise. J_{ij} is the interaction between two

neighboring sites i, j . Here, we choose $J_{ij} = \frac{1}{K} \exp(-\frac{d_{ij}^2}{2a^2})$, where d_{ij} is the euclidean distance between i and j , a is a normalization constant, and K the average number of neighbors for a given site. We assume that no interaction exists between non-neighboring sites. According to the energy in (1), two sites with high-mutual interaction will pay a high-energy price if they are not aligned.

The probability to find the system in a given spin configuration S depends upon a *temperature* parameter T , and accordingly, one defines the probability density P_T :

$$P_T(S) = \frac{1}{Z} \exp(-\frac{H(S)}{T}), \quad (2)$$

where Z is a normalization constant. For a given temperature, one can compute the thermodynamic average of any quantity by estimating the weighted average of this quantity over all possible spin configurations with respect to $P_T(S)$.

2.2 Key Quantities and Metrics for Clustering

Within this framework, we now describe the details of the clustering process.

2.2.1 Spin-Spin Correlation and Clustering—At a given temperature T , clusters are formed by grouping sites that are most likely aligned, with respect to the corresponding probability distribution P_T . The key element in this clustering process is the introduction of

the spin-spin correlation G_{ij} defined between two sites i and j , which represents the probability for two spins s_i and s_j to be aligned. Two neighboring sites i and j are most likely aligned if their spin-spin correlation value G_{ij} is high (typically greater than 0.5). In such case, a link is set between these two sites, and then, the two sites are taken to belong to the same cluster. By applying this rule to all pairs of neighboring sites one can thus easily build connected graphs. A cluster referred to as a *magnetic grain* in the Potts model is then defined as one of these connected graphs.

2.2.2 Order Parameter and Thermodynamic Phases—Clusters evolve with temperature. At lower temperatures, larger clusters are formed, whereas higher temperatures allow for more disorganization and less clustering. Some of the temperatures will correspond to major structural changes in the cluster organization and will delimit specific thermodynamic phases. In order to concretely exhibit these different phases, one needs to consider the average magnetization of the system, $\langle m \rangle_T$, which measures the degree of ordering of the system at each temperature. Details on the computation of the average magnetization are described in [1]. By considering the variations of the degree of ordering as the temperature changes, one can distinguish among different thermodynamic phases. At low temperatures, the *ferromagnetic* phase is characterized by a well-ordered system and the presence of only one major magnetic cluster. As the temperature increases, we move to the *superparamagnetic* phase in which clusters successively break down into distinct magnetic grains. Finally, at very high temperatures, the system gets totally disordered; this is the *paramagnetic* phase.

Transitions between phases can be investigated by evaluating the susceptibility χ , which tells us about the variance of the magnetization

$$\chi = \frac{N}{T} (\langle m^2 \rangle_T - \langle m \rangle_T^2). \quad (3)$$

Large fluctuations in the susceptibility characterize successive subdivisions of the magnetic grains. Hence, one wants to detect major peaks in the susceptibility, and therefore, determine at which temperatures major subdivisions occur in the cluster decomposition process. Each major peak corresponds to either cluster splitting or cluster disaggregation (i.e., the cluster melts away). Thus, the system defines a metastable state configuration over an interval of temperatures, that is, delimited by two successive peaks. However, as we observe the configuration of the system over such an interval of temperatures, we notice that some of the data points (usually located at the fringe of their clusters) will tend to successively dissociate from the clusters as temperature increases, without creating any major effect on the whole clustering configuration. Accordingly, the temperatures that immediately follow a peak in the susceptibility graph are those temperatures at which we will perform and analyze clustering. More specifically, temperatures that are chosen for clustering will be taken at local minima that immediately follow a peak in the susceptibility graph. Finally, at higher temperatures, an abrupt decrease in χ characterizes the transition to the paramagnetic phase.

2.3 Monte Carlo Simulation

The computations of the average magnetization and the spin-spin correlation both involve the notion of thermodynamic averaging. As proposed in [1], a Monte Carlo simulation allows us to generate M configuration samples at each temperature, with respect to $P_T(S)$. The method is based on a Markov chain process generated by the implementation of the Swendsen-Wang Monte Carlo algorithm [23]. This Monte Carlo approach turns out to be computationally efficient by enabling us to flip a whole set of spins in one iteration, instead of changing the configuration one site at a time.

3 Discussion on the Method

Even though this Potts model-based clustering is considered to be a nonparametric algorithm, nevertheless many parameters need to be adjusted. Those parameters usually allow for flexibility in the process of forming clusters. We now detail the analysis of the most significant parameters.

3.1 Mutual Neighbors

As previously mentioned, non-neighboring sites have no interaction. Several different options exist to define the concept of neighborhood [1]. Here, we characterize neighbors using standard mutual neighborhood conditions. Two sites are *mutual neighbors with value K* if each is a K -nearest neighbor of the other.

The chosen value K can vary from 0 to $N - 1$, and this choice will have a potential impact on the outcome of the algorithm. Instead of employing a homogeneity parameter to find K as in [22], we propose a novel approach using coordination numbers. Indeed, for our data sets, the type of method described in [22] turns out to provide very large values of K . However, very large values of K not only make the algorithm computationally expensive but also tend to shrink the thermodynamic region of interest for clustering. We, therefore, define and use another criterion based on sphere packing theory, in order to efficiently choose the parameter K . Sphere packing theories study the arrangements of spheres in a given volume and the resulting densities. In this manner, they naturally connect to the notion of coordination number.

Specifically, in this work, we model our sites as equal-radius spheres and assume a maximal density. The number of nearest neighbors for this case is also known as the *kissing number problem*. A kissing number is the maximal number of spheres that can touch a single sphere in d dimensions. In d dimensions for $d = 2$ or three [24], any given sphere has $6(d - 1)$ closest neighbors and so for any given site, we pick its $6(d - 1)$ closest neighbors, and consider only these as potential mutual neighbors. This method allows us to have a standard and straightforward way to choose K , while providing quite reasonable results.

3.2 Advantages of the Potts Model and Comparison to Classical Methods

The Potts model clustering carries advantages over other nonparametric clustering algorithms. Hierarchical linkage methods [18], [19] are commonly used methods that successively decompose the data into clusters according to a chosen linkage criterion. This criterion defines the way distances between clusters are calculated and the order in which clusters are successively formed, split, and merged. Hence, intermediate clustering configurations obtained through this iterative decomposition can be represented graphically by a solution tree, usually referred as *dendrogram*.

However, these linkage clustering methodologies present several undesirable characteristics.

First, no clear and unambiguous indication is given about the depth at which one should explore the tree in order to obtain reasonable clustering. The major steps of the iterative cluster construction are usually not easily detectable. In [25], a criterion is defined that tries to provide a way to find appropriate levels in the decomposition at which clustering should be performed. However, this criterion does not have any real intuitive meaning and is not necessarily straightforward to use. Therefore, even with such a criterion, the optimal clustering configuration becomes more difficult to estimate for a nonexpert analyst, since a large number of potential configurations are available along the tree. The Potts model, reduces the amount of uncertainty by limiting the decision making process. Indeed, very few

peaks are usually observed over the superparamagnetic phase, and one only needs to choose among those few peaks to obtain a final clustering.

Second, an issue of concern with hierarchical linkage methods is the legitimacy of clustering in certain cases. Indeed, when no real inhomogeneity exists in the data set, clustering should not occur. However, with classical methods, the decomposition nevertheless proceeds and generates a nontrivial tree. When the data set is homogeneous, the Potts model does not generate any intermediate clustering, since no superparamagnetic phase is observed, i.e., the configuration jumps from one large cluster to no cluster without any transition. To illustrate this, we applied the Potts model clustering to points sets with no real inhomogeneity in the distribution. Thus, we generated ten data sets of two-dimensional random variables, uniformly distributed over $[0; 1]^2$, and observed the susceptibility graphs for each of these (see Fig. 2). Note that each of these graphs exhibits only one peak and that none of them cases shows a plateau after the peak, meaning that no superparamagnetic phase exists and no clustering configuration is formed. Not so, when we applied an hierarchical clustering method on the same random sets, we always got clear clusters.

Finally, with classical hierarchical clustering methods, prior knowledge of the problem is usually very helpful in order to determine appropriately the linkage criterion, that is, going to dictate the format of clustering. For example, one distinguish among *single*, *complete*, *average* types of linkage methods [18], [19]. Some will be more suitable than others according to the shape of the data set. The Potts model does not require such knowledge.

4 Backbone Structural Conformation Classification

The backbone structural conformations of RNA can be represented by either residues or suites. The structural flexibility of a residue (or suite) stems from the modes of motion of its backbone. Potential modes of motion for nucleotide backbones are restricted to N_{Tor} rotations around covalent bonds, where $N_{Tor} = 6$ for a residue or $N_{Tor} = 7$, for a suite. Accordingly, we describe the single nucleotide conformation using N_{Tor} angle parameters (see Fig. 1) Rotations of the backbone are restricted by molecular forces. Due to these restrictions, the backbone conformation distribution in the N_{Tor} dimensional torsion space is strongly nonhomogeneous. The data points are mostly restricted to lie in clusters consisting of disconnected regions of the torsional space. This clustering characteristic can be used as a similarity criterion for classification of conformations for a single residue or suite. Two conformations are considered to be similar only if they reside within the same cluster. Such clustering was performed via qualitative observations using projections of the data space onto subdimensional spaces. Thus, for residues [8], a representation of the data in six separate histograms was proposed in order to analyze the six torsional angles of the residue conformation. For suites [10], three-dimensions projections of the torsional space showed an alternative for dimensional decomposition. In both cases, classification is made difficult by the high-dimensionality of the data space.

The nonhomogeneous distribution of data points makes the problem a good candidate for the application of an automated clustering method to the original data space. In [8], a k-means based algorithm was used to find clusters in the single nucleotide (residue) conformation. The efficiency of that method is hindered by several factors: 1) there is no prior knowledge on the number of clusters; 2) k-means is based on the definition of a global metric, while actual physical forces in the RNA structure dictate a local metric, that is, unknown. The Potts algorithm seems to be a good candidate to find clusters in the conformation data space, since it does not require any prior knowledge about distribution of clusters and since it is based only on a nearest-neighbor criterion. We have performed a clustering analysis for both residue and suite representations.

4.1 Single Residue Cluster Analysis

The data that we used to examine the algorithm was composed of approximately 2,800 single nucleotide conformations from the structure of the ribosome of HM LSU 23S [26] (RR0033 in NDB). This data test case was chosen in order to compare the clustering results with previous clustering scheme [8]. This structure has a high-accuracy (resolution 2.4Å) and is often used as a test case for structural data mining of RNA. Clustering was performed on a four-dimensions data space using the four discriminating torsional angles [8], [9] α , γ , δ , and ζ . Also, as torsion angles α , γ , and ζ ($\in [0;360]$) are circular dimensions, adjustments have to be made when considering distance, neighborhoods, and interactions between points. Thus, for each dimension, our algorithm accounts for this cyclicity by allowing angles close to 360 degrees to be “neighbors” of those close to 0 degrees. Interaction between sites is similarly adapted by choosing the shortest distance between two points among all the possibilities that exist when considering all trigonometric directions.

The parameters that we used in the algorithm were carefully chosen in order to optimize and facilitate the visualization of the results. We tested different values of q , the number of possible spins, and finally used $q = 20$ for the particular application of the algorithm to backbone conformations, as well as for all other classifications that are presented in the remainder of this paper. A distance normalization parameter a equal to the average distance between mutual neighbors was found to be satisfactory. No major difference was found when scaling that parameter (we tested it for both $2a$ and $a/2$). The number of nearest neighbors K was chosen to be 18, based on an extrapolation of the explanation in Section 3.1. The calculated susceptibility diagram for backbone conformation analysis is shown in Fig. 3. Based on this graph, we can detect three main transitions in the evolution of the clusters. As previously mentioned, major transitions are represented by peaks in the susceptibility diagram. Over an interval of temperatures, that is, delimited by two successive peaks, we choose temperatures for clustering analysis as explained in Section 2.2.2: temperatures that are chosen for clustering will be taken at local minima that immediately follow a peak in the susceptibility graph. Thus, for the case of single residue conformations, we performed cluster analysis at three different temperatures. The first two temperatures were chosen to be $T_1 = 0.016$ and $T_2 = 0.061$. These two temperatures just followed the first two peaks that we observed in the susceptibility graph (see Fig. 3). The third temperature ($T_3 = 0.3126$), represented the formation of a single new cluster, which corresponded to a typical conformation of a residue in the 3' side of an A-form RNA helix. In Fig. 3, these three temperatures are marked by three vertical lines.

The analysis of the clustering may not seem straightforward because of the multippeak structure of the susceptibility diagram. To address this issue, we first performed a cluster analysis at each of the three aforementioned temperatures, from the lowest to the highest. If a cluster was found at a low-temperature and survived as a single cluster at higher temperatures, then this cluster was taken into account in the classification. If, on the contrary, a cluster was found at a low temperature but was then split into two or more significant clusters, then the new clusters were kept for analysis. Note that there can be no relevant splitting after the third peak because the fourth peak represents the final melt down of all clusters and the transition to the paramagnetic phase.

Also, comparison of our results to those from the binning method [8] allowed us to validate the pertinence of our classification. This validation was mostly limited to a decision on the “cutoff” size of the clusters (i.e., the minimal size above which a cluster is taken into consideration). Indeed, such information may be needed for the decision if a given cluster is melting or splitting.

To validate our analysis, we have quantitatively compared the Potts and binning clustering techniques. The minimal size (or “cutoff” size) of a cluster was chosen to be six. The results of the clustering analysis and the correspondence between the binning and Potts models, are presented in Table 1. The first column contains the bin index, using the alphabetical annotation as proposed in [8]. In the second column, each bin is allocated a four-digit number, where each digit encodes the spatial range of each of the four torsional angles [8] (as detailed in Table 2): α , γ , δ , and ζ . The third column gives the index of the peak after which the cluster was first identified. The fourth column contains the indices of the “corresponding” clusters found with the Potts model. The fifth column gives the number of residues that are found in both the bin and the corresponding Potts cluster as compared to the total number of residues in the bin. The sixth column compares the same number of common residues to the number of residues in the Potts cluster. The seventh column contains additional information of the typical functionalities of the residues within each cluster. As can be seen from Table 1, there is a very good agreement between the two methods, except for few cases. The first case consists of very small clusters that have split out of a larger bin (d and d' , y and y' , a and a') or very small bins that have not been recognized as clusters by the Potts method. The second and significant case consists of clusters e and a . According to the binning method, these two clusters include nucleotides that reside within the interior of the A-form RNA helical region (a) or the 3' end of it (e). The difference between these two conformations was in the angle ζ . For the bin a , ζ is in the g-conformation (see Table 2) and for e it is in the g+ or t conformation. According to the Potts method, cluster a has the same range of angles, but cluster e includes only the g+ range of ζ . The residues with ζ in the t conformation have melted away without forming any clusters. It is important to note that the cases, where residues in the e Potts clusters are those that take part in known motifs (e-loops and kink turns), while residues that are in bin e but not in cluster e do not have such affiliation.

4.2 Suite Cluster Analysis

As a second test case, we have performed cluster analysis to validate the suite structure as presented in Richardson's work [10]. First, we applied Potts clustering to the simplified case of a five-dimensions representation of the suites, where five identifier angles are taken into account: δ_{-1} , ζ_{-1} , α , γ , δ [10]. We used the RR0033 database for this suite clustering analysis in the five-dimensions conformational space. Results are attached as Supplementary Material (S1), which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2010.128>. This method was shown to give a relatively good agreement with the full seven-dimensions suite torsional space.

After validating the method on the simpler five-dimensions case, we applied the clustering algorithm on the full seven-dimensions suite representation, where the seven identifier angles are: δ_{-1} , ϵ_{-1} , ζ_{-1} ; α , β , γ , δ . For this case, we used the RNA05 database as in Richardson's work [10]. This database includes more than 9,000 suites from different RNA structures. Given the size of the data set, building a fine susceptibility diagram was found to be very time consuming. In this context, we used coarser intervals of temperatures, making the detection of local minima more difficult. In order to validate our classification, we used the same algorithm as in the single residue case but with temperature increments of 0.002. We also added a stability criterion for each cluster that was found. According to this criterion, a cluster is said to be stable if it “survives” (i.e., does not undergo major modifications) for at least two adjacent temperatures. This criterion allows one to disregard the clusters that may form around the susceptibility peaks but that actually correspond to undesired, metastable states. The analysis reveals that the majority of the clusters have been formed at the first peak. The result of the cluster analysis is presented in Table 3. This table has been built in order to mimic Table 1 in [11]. The first column of the table gives the

ASCII annotation of the bin corresponding to the given seven-dimensions cluster. When no annotation existed, we used the numerical annotation described in Section 4.1. The second column gives the total number of data points in the cluster. The third column gives the bin ASCII code of the corresponding five-dimensions representation of the cluster [11]. The fourth column gives the number of data points that are both in the Potts cluster and is associated with five-dimensions cluster bin. The fifth column gives the consensus suite cluster [11] that agrees with the new seven-dimensions Potts cluster. If no suite cluster was found to agree with the new Potts cluster, we left the entry empty. The sixth column lists, an example from the RNA05 data file of a suite with a conformation, that is, typical of the cluster. To keep this column consistent with previous results, we used the same example as in Table 1 in [11], whenever this was possible. The seven other columns give the mean dihedral angle values for each of the Potts clusters with the standard deviation in parenthesis. For most cases, there is a very good agreement between the consensus suite clusters, the five-dimensions bin suite clusters and the new Potts clusters. The most significant difference between the new Potts clusters and the consensus suite clusters is the decrease in the number of clusters. Indeed, while the consensus suite clusters includes 46 clusters, the Potts clustering algorithm generates only 32 clusters. For most cases, this decrease results from the merging of some consensus suite clusters into a large Potts cluster. For example, consensus suite clusters $1a$, $1m$, $1L$, $7a$, $9a$, and $6g$ merge all into cluster a of the Potts classification. Either these clusters were not distinguishable from each other at any temperature or they melted away from the main cluster without forming any relevant individual clusters. Some other small consensus suite clusters, such as $6j$, were not recognized as clusters by our classification. The Potts classification introduces several new clusters, among which some represent a “crankshaft” of the A-form RNA. This is cluster y , where the “crankshaft” effect is manifested by the transition α : $g_- \rightarrow g_+$ and γ_{-1} : $g_+ \rightarrow g_-$. The second “crankshaft” conformation is a variation of cluster i , where the transition is in the angle ϵ_{-1} : $g_- \rightarrow g_+$. Another new cluster is cluster u' that similarly to cluster u , includes a bend conformation in an A-form RNA single strand, but does not participate in a kink-turn motif. Cluster E is similar in conformation to cluster E' , but the angle ζ_{-1} of the Potts cluster is restricted to the g_+ orientation. Cluster F' is within very close range to cluster F , but the angle ϵ_{-1} is shifted from the g_- to the trans orientation. The new cluster does not force the bulge that appears in the main cluster F [11]. One more cluster, the 12231 cluster, seems to be a stable cluster that does not appear in other classifications.

In summary, it seems that the new technique removes some splitting between clusters as compared to the consensus suite clustering [10], but also introduces several new clusters. We should note that several of the clusters that were not present in the conformer library shown in [10] have an irregular ϵ_{-1} value. It is remarked in [10] that irregular ϵ_{-1} torsions are frequently found in RNA structures that have been improperly fit into electron density. Thus, it is likely that the clusters presented here represent certain common misfits that were intentionally excluded from the conformer library in [10]. The obvious advantage of the Potts clustering technique is that it is almost unsupervised. After finding several parameters such as the temperature range, the temperature interval of stability and the cutoff size of a cluster, the application of the technique becomes fully automated.

5 Base Doublet Geometry Classification

5.1 Co-Ordinate Systems

Base doublet interactions may be of different types. Thus, in this work, we applied the proposed clustering method to both base pair and base stacking interactions. For both cases, we employed two parameterization methods, which we would like to detail here.

The first co-ordinate system was proposed in [27]. It is defined by considering only translation and rotation parameters, rather than relative hydrogen bond distances. More specifically, a set of three parameters was used. Two parameters defined the projection of the glycosidic *N1/N9* (pyrimidine/purine) atom of one of the bases on the plane of the other base. The third parameter defined the rotation of one of the bases around its center of mass (COM) that was required to align it with the second base (see Fig. 4a). In the remainder of this text, we refer to this parameterization as the *COM* parameterization, and we use it as a reference case.

The alternative parameterization that has been employed in this work is based on the Pople co-ordinate system [28] with the origin in the center of the pyrimidine ring as described in Fig. 4b. The exact parameters used for this coordinate system will be different according to the nature of the base doublet interaction. Thus, base pair geometries will be represented in a two-dimensions space, whereas base-stacking geometries will be described using three parameters. Details will be given in Sections 5.2 and 5.3. In the remainder of this paper, we refer to this parameterization as the *center of pyrimidine* (COP) parameterization.

Both parameterizations present certain advantages and drawbacks. As we aim to build a method for the classification of base doublet interaction, it is important to know, which co-ordinate system to use and in what case. For this reason, we now list some of the characteristics of the COM co-ordinate system and compare it to the proposed system.

As illustrated in Fig. 4a, the COM co-ordinate system employs the angle of rotation around the COM needed to align the two bases and the (x, y) -projection of the distance between the two glycosidic nitrogen on the plane of the lower base. Note that the glycosidic atom is the closest atom to the backbone. Hence, a classification method based on the glycosidic distance is expected to give better correlation with one based on backbone classification. Also, the relative location of two glycosidic atoms is fixed for any double helix geometry. Thus, this classification should be effective for detecting any possible deviations from the double helix structure.

A drawback of this method is that the translation and the rotation are not performed within the same co-ordinate system. The rotational co-ordinate system is based on the actual COM and the location of this origin depends on the type of base. The COP parameterization employs the same co-ordinate system for both rotation and translation, whose origin is set at the location of a pseudoatom (pyrimidine ring geometrical center) rather a real atom. The pyrimidine ring seems to be involved in the majority of base pair and base stacking interactions. Hence, the COP co-ordinate system seems to be useful for base pair and base stacking classifications. However, a disadvantage of this choice of co-ordinates is the possibility of certain artifacts in the classification between the purine and pyrimidine bases. Further, because the COP origin may be far from the backbone, we may have a weaker correlation with the backbone geometry.

5.2 Base Pair Geometry

In this part, we focus on the analysis of base pair interactions. The most familiar and common case of such interactions is that of the Watson-Crick (W-C) base pairs, which is responsible for the double helical structure of polynucleic acids. Base pair geometry is flat, so that two bases lie approximately in the same plane. The base pair interaction is mediated by hydrogen bonds. Hydrogen bonds can form between an electronegative atom (electron “donor”) and a hydrogen atom bonded to another electronegative atom (electron “acceptor”), both located at specific sites in the base [29]. The relative arrangement of the donor to the acceptor is such that each base possesses three possible edges for base pair interaction [17] (see Fig. 5). This arrangement provides six different potential geometries

between interacting bases. An additional combinatorial factor of two emerges from the directionality of the strands. This gives a total of 12 classes of base pair geometries, which are referenced in data banks using the Leontis-Westhof (LW) notation [17].

An automated classification method for the base pair geometry based on an Expectation Maximization (EM) algorithm was presented in [30]. In this method, Lemieux et al., proposed an elaborate classification of the base pair geometries. They show the existence of intermediate base pair geometries, but their work mainly agrees with the LW classification. The drawback of the EM method is that it required an assumption about the structure of the underlying distribution function (N Gaussians where N had to be predefined). This disqualifies the method for our purposes, since we are interested in a minimally supervised classification method, where no prior knowledge about the underlying distribution function is needed. In another work, Sarver et al. [27] have performed a detailed by-eye classification of all existing base pair geometries. This analysis was run using the COM parameterization.

We now describe in details the parameters that were used in the COP co-ordinate system in order to describe the base pair geometry.

As mentioned above, the COP parameterization used in this work differs according to the type of the base doublet interaction (base pair or base stacking). The relative geometry of the base pair can be represented by four parameters. Three of them form the directed (vector) distance between the centers of the two hexagonal pyrimidine rings using spherical co-ordinates, i.e., distance r and polar angles θ and ϕ (see Fig. 1b). Another angle ω accounts for the rotation required to align the two pyrimidine rings. We have reduced the number of parameters to two by taking into consideration the fact that the two bases are coplanar and by eliminating the distance parameter. Hence, we use only θ and ω as the set of parameters for the clustering problem. We have run our code on the RR0082 (PDB number: 1S72) [31], i.e., the same as for the test case of the Sarver base pair classification [27].

The Potts method was applied to classify base pair geometries, using both COM and COP parameterizations.

For base pair geometry classifications, we found that both co-ordinate systems worked similarly, and so for simplicity, we will just use the COP in what follows. The main point is to highlight the Potts method.

The list of base pairs was made by choosing base doublets that fulfill the following constraints:

1. The distance between pyrimidine centers is less than 8.0 Å.
2. The minimum distance between two atoms from the two bases is less than 3.5 Å.
3. The angle ϕ is less than 115 degrees and larger than 65 degrees.
4. The normal to the two bases form an angle less than 30 degrees.

In our representation, the relative geometry of base pairs is defined by the bases only and not by the strand geometry. Each base, being flat, has two faces: up and down. We have chosen the relative direction of the two bases to be the preliminary criterion for partition of the base pairs. Therefore, we work with two groups of data: a group with the same directionality for both bases (up-up or down-down) and a group of opposite orientation of the faces (up-down or down-up). For the base of a nucleotide in a helix with the glycosidic torsion angle in *anti*, the normal to the up face is pointing in the 5' direction and the normal to the down face is pointing in the 3' direction.

We have classified these two groups into clusters using a Potts model clustering method. Since the problem is two dimensions (θ and ω), we chose the number of nearest neighbors parameter to be $K = 6$.

The susceptibility graph for the up-down group exhibits one dominant peak, which appears at a very low temperature ($T < 10^{-4}$). The second peak corresponds to the melting point for all the clusters (the second one ending the phase of interest). The up-up group shows one significant peak with a slow decaying tail. Again, as for the up-down case, the separation into clusters appears almost instantaneously ($T = 0$). Therefore, we have chosen the clustering temperature to be $T = 0.001$ for both up-down and up-up cases. This seems to give good results in comparison to the reference classification by Sarver et al. [27]. For the case of the up-up group, the existence of a second peak made us consider the clustering configuration at $T = 0.019$, for which an additional cluster is detected.

Fig. 6 gives the two-dimensions scatter plot for the up-down configurations and exhibits the major clusters formed by the Potts model based classification. Some validation can be made using a well-known measure of structural similarity within and between clusters. Thus, we computed the root mean square standard deviation (RMSD) for points inside each cluster and between points of two different clusters. For this base pair two-dimensions problem, we computed the following in order to estimate the “intracluster” RMSD for each cluster C with n points:

$$\sqrt{\left(\sum_{i \in C} \sum_{j \in C} (\theta_i - \theta_j)^2 + (\omega_i - \omega_j)^2 \right) / (n(n-1)/2)}. \quad (4)$$

Then, for each pair of clusters (C_1, C_2), with n_1 and n_2 points, respectively, we computed the “intercluster” RMSD:

$$\sqrt{\left(\sum_{i \in C_1} \sum_{j \in C_2} (\theta_i - \theta_j)^2 + (\omega_i - \omega_j)^2 \right) / (n_1 \cdot n_2)}. \quad (5)$$

When comparing those quantities, we can see that, on average, the “intracluster” distances are 10 times smaller than the “intercluster” ones.

The clusters that emerged from the Potts classification are presented in Table 4 for the up-down case and in Table 5 for the up-up case.

The up-down class contains the majority of the data points (1,171 cases from the total of 1,406 base pair candidates). The first column gives the classification obtained with the Sarver et al., method [27] and the third one presents the results for the Potts classification. As observed in Table 4, the most predominant difference between our classification and the Sarver classification is that our classification splits most of the LW groups into subclusters.

Specifically, our overall framework can differentiate among *base pair contents*. By *base pair contents*, we refer to doublets ($Base_1 - Base_2$), where $Base_1$ and $Base_2$ indicate the family of the two bases involved in the base pair. The base family is either purine or pyrimidine, so that one distinguishes four different cases of base pairs: (*pyrimidine – purine*); (*purine – pyrimidine*); (*pyrimidine – pyrimidine*); (*purine – purine*). Thus, the extra splitting, that is, observed for both the *csS* and *biff* clusters [27] (see Table 4) seems to illustrate this differentiation. However, we showed that for most cases, extrasplitting is not an artifact of

the choice of coordinate system. If this were the case, the clusters would just correspond to distinct base pair contents. For example, for the *tHSLW* group, our method finds four distinct clusters. Some of these clusters exhibit the same base pair contents (e.g., *AG*) but are differentiated by the geometry.

Also, the subdivision of the *cWW* cluster to five subclusters seems to result from the differentiation between the standard Watson-Crick base pair geometry and the *GU* (or *UG*) base pair. All other divisions seem to be unrelated to the content groups. Since these groups are rather small (e.g., *cWW*(2) in Table 4, one must study a much larger data set to verify (or rule out) this subdivision. Two of the LW groups, *cWS* and *cSW*, which do appear in the classification proposed by Sarver et al. [27], cannot be recognized as clusters by our method.

When ignoring the extrasubdivisions that are obtained in the proposed classification, comparing the number of points in each one of the clusters that are also obtained by Sarver et al., seems to give an excellent match. Examining the results for the up-up group as presented in Table 5 gives very similar conclusions. An important conclusion from this section is that our methodology is able to reproduce the LW classification [27], and also to give some additional information.

In addition to this analysis, we have compared the proposed Potts clustering to a linkage-clustering scheme. Results obtained with linkage clustering for base pair geometries are presented in Supplementary Material S2, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2010.128>.

5.3 Base Stacking Geometry

The majority of bases in the RNA are arranged in stacks (see Fig. 8). This stacking behavior is typical of aromatic rings [32]. Stacking interactions constitute a major factor in the stability of RNA helices and their assembly [16]. The exact nature of stacking forces between the RNA bases is still an open question. As far as we know, the possible sources of stacking interactions are: π stacking [33], dispersion, electrostatics [34], [35], dipole-induced-dipole interactions [36] and hydrophobic forces [37], [38]. The lack of a known physical model for base stacking geometries prevents us from using prior references, such as the one we had for base pair geometries or backbone classification. However, the major role that base stacking interaction plays in the folding and stabilization of the three-dimensions RNA structure seems to be a strong incentive to start developing such a geometry classification model.

Another benefit of finding a classification for base stacking geometries is the development of some structural annotation for bases in the same manner, as for the backbones [8], [9]. To the best of our knowledge, the only attempt to perform a detailed and unsupervised classification of base stacking interaction was performed by Sykes and Levitt [20]. In their work, a clustering scheme has been developed that consists of a mixture of k-means and agglomerative classifications. This hybrid algorithm was used to classify all types of base doublet interactions in RNA structures. The main drawback of the algorithm is that no criterion is proposed for choosing the level in the data decomposition at which one observes the optimal classification. Different optional clustering configurations are presented for different levels. Whenever it was possible, we compared our results for the base stacking with this method.

The Potts model was applied, for the base stacking case, using two alternative data sets, corresponding to two different parameterizations. These two parameterizations have been introduced in Section 5.2: that is, the COM and the COP co-ordinate system. These two

parameterizations give different *realizations* of the base stacking, and hence, can be used for cross-validation of the stacking scheme.

The data file that we used for our analysis was the same RR0033 structure that employed previously in this work. With the COM parameterization, we used a data set of base stacking cases that was provided by Leontis and Zirbel (private communication). For the COP co-ordinate system, the criteria used to select the base stacking doublets are selected are the following:

1. The two faces are “nearly” parallel in the sense that the angle between their respective normals is less than 30 degrees.
2. The vertical distance between two bases is 3–5 Å.
3. The distance between the centers of the pyrimidine rings is less than 8 Å.
4. There are at least two atoms from the two bases within 3.5 Å from one another.

It is important to note that the criteria of the two-parameterization methods (i.e., COP and COM as explicated in [27]) are not identical. As a result, the two lists of base stacking doublets used for the Potts clustering were not identical. In fact, there was about a 10 percent difference in the content of both lists. We have applied the clustering method to both data sets rather than trying to find an identical criteria for validation purposes.

Note that a common problem of various clustering techniques is the lack of robustness [19]. Thus, a small change in a data set can cause a large change in the classification. Comparing the results for the two types of parameterizations may, therefore, give a measure for the robustness of the Potts method. The common clusters of both parameterization systems provides a strong argument to their validity.

The first co-ordinate system that we have utilized involves the COP parameterization, similarly to what has been done for the base pair analysis given above. However, contrary to base pair geometry, base stacking does not involve coplanar bases. Therefore, three parameters are needed (instead of two) to describe the base stacking geometry for the COP system. These three parameters are the projections of the distance vector between the centers of the pyrimidine rings onto the x - z plane (see Fig. 4b) and the rotation angle ω (see Fig. 1b). For the COM parameterization, the co-ordinate system is the one presented in the Section 5.2 for base pairs.

In the same manner as for the base pair set up, there is a preliminary division among base stacking geometries based on the relative orientation of the bases. There are four possible arrangements for the faces of the base doublets: up-up, down-down, up-down, and down-up. We, therefore, identify four different data sets.

The up-up and the down-down stacking geometries have the same type of interaction. The interacting faces are the upward face of the first nucleotide with the downward face of the second nucleotide in the up-up group, and vice versa for the down-down case. The difference between the two cases is that the up-up group is represented inside a double helix while the down-down group is represented outside of a Watson Crick helix. Hence, the down-down base stacking geometries are free of packaging constraints as well as stacking co-operative effects, and show a different distribution (i.e., more dispersed distribution) in the configuration space. This reason led us to separate between the two groups.

Upon initial visual inspection, none of the four data spaces actually shows any obvious clustering structure. For example, examination of the three-dimensions data space of the up-up case (see Fig. 7) shows a very vague cluster structure probably due to the less restrictive

nature of stacking forces compared to forces that determine base pair and backbone orientation. Inhomogeneity is not obvious here and clustering becomes more challenging.

We have performed cluster analysis using the Potts algorithm on all the four cases: up-up, down-down, up-down, and down-up. Given the two parameterizations COM and COP, we needed to run the Potts algorithm on eight different cases. For our analysis, we chose the scale parameter a to be equal to the average distance between neighbors and according to the criterion defined in Section 3.1, $K = 12$. Also, since the dimensions along which the stacking geometries are analyzed involve both euclidean distances and angles, we normalized every variable has been normalized between 0 and 1 before calculating distances d_{ij} and applying clustering. Susceptibility diagrams are presented in Supplementary Material S3, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2010.128>. The criterion for choosing the clustering temperature stays the same as in Section 5.2. For six of the eight runs, a single temperature was found adequate to span all the clusters. Increasing the temperature beyond this reference either shrinks or melts clusters, while decreasing the temperature prompts merging of clusters. The only case where more than one temperature was required was for the up-up and up-down face groups with the COP parameterization. These cases will be explained separately in the text.

5.3.1 Validation of Results—Given the lack of prior knowledge about the clusters appearing in the stacking geometry, we used the Watson-Crick double helical structure as the basis for validation of our analysis. Thus, we first validate our clustering results on the only data points that exhibit this structure. We define a Watson-Crick double helical structure to be a double strand consisting of a contingent of two or more base pairs with a Watson-Crick geometry [32].

The double helical arrangement is the most abundant motif in the RNA structure. About half of base stacking interactions that are candidates for classification lie within the Watson-Crick double helix definition. A typical Watson-Crick helix is shown in (see Fig. 8), demonstrating that the geometry of the intrastrand stacking doublets depends upon stacked nucleotide pairs.

We distinguish among three different classes: pyrimidine-purine (Class I), purine-purine or pyrimidine-pyrimidine (Class II), and purine-pyrimidine (Class III). We will refer to these as “content classes.” It is important to note that there is no sequence symmetry in RNA, and that the numerical order (5′-to-3′ and 3′-to-5′) of bases is important. This ordering stems from the fact that each base has two different faces. The most common intrastrand face arrangement in the Watson-Crick double helix is up-up, hence we have chosen this class to be the major validation case for the clustering application.

The validation of the content classes can be done only with the COP parameterization, since the COM technique was constructed to group all stacking cases of a Watson-Crick double helix into one class. Indeed, performing clustering with the COM parameterization only reveals one cluster for all the A-form RNA double helix stacking conformations. On the other hand, the COP parameterization reveals three clusters that clearly characterize the three content classes.

For the up-up group, Table 6 gives the comparison between the three content classes and the three major clusters obtained with the Potts algorithm at $T = 0.13$. We use the term “majority cluster” to name the cluster that overlaps the most with the considered content class. As can be seen from Table 6, about 90 percent of the class I group, that is, in a Watson-Crick helix is also in cluster 1. The same observation is true for class II and cluster 2. As for class III, only 70 percent are in cluster 3. We have performed a similar cluster

analysis when removing the Watson-Crick double helix constraint (i.e., when accounting for all data points, even the non-Watson-Crick doublets) and have observed similar results.

Agreement of our clustering results with base content is very good. Exceptional cases, i.e., those for which the two classifications diverge, might represent the most interesting but nontrivial cases of base stacking. These will probably include stacking of uniaxial helices, junctions, and bulges [39]. These nontrivial motifs define the final three-dimensions structure of RNA. Correct classification of this kind of stacking would help in the understanding of the folding and the self-assembly of helical motifs into a functional well-defined three-dimensions structure.

Examination of the data space in the up-up group demonstrates the power of the Potts clustering method. Fig. 7 shows a three-dimensions projection of that data space without any filtering and reveals no clear underlying cluster structure at first sight. Using the Potts classifier as a filter to select the most populated clusters gives very satisfying results (see Fig. 9).

The next largest data group is the up-down group. Members of this group participate in the interstrand Watson-Crick double helical structure, as shown in Fig. 8. Following the usual procedure, we produced and analyzed our clustering results at temperature $T = 0.12$. We observe three major clusters that can be easily affiliated with three “content classes.” Quantitative results for these three clusters are shown in Table 7. The two tables show the same types of results as for the up-up case.

5.3.2 New Clusters—After validating the reliability of the algorithm (for the case of the COP parameterization), we have performed a full scan that was intended to find new nontrivial clusters in the RR0033 data base. The clustering was performed with both parameterization methods. The results of this clustering procedure are presented in Table 8.

The table describes the different clusters according to their affiliation to the face-face group. In this table, we provide the total number of base stacking doublets in the corresponding face-face group for both the COP and COM parameterizations. These numbers are a bit different because the criteria for base stacking are slightly different for the two representations. For each face-face group, the temperature that was used for the classification in both representations is also given. The next part of the table includes the specific clusters that have been identified. The first column of this part includes the cluster numerical annotation. We report the number of points (stacking doublets) that were identified in each cluster, as well as the number of points that are common to a given cluster in both representations.

It should be noted that only for the first cluster in the up-up group is the number of doublets in the COM parameterization much larger than the number of doublets in the COP representation. This discrepancy is due to the fact that the first cluster represents all of the intrastrand stacking conformations in a Watson Crick helix. As mentioned previously, in Section 5.2, the COM parameterization was designed to provide a good way to define the constraints of a Watson-Crick double helical arrangement of bases in the RNA.

The last three columns of the table give details on the characteristic features of each stack class. The first of these columns gives the base pair arrangement that each one of the residues is involved in (if such an interaction exists). The base pair interactions are arranged in the 3′-5′ order and the annotation that we are using is the same as in Section 5.2. The next column gives the majority content class of the doublets contained in the corresponding cluster. In the last column, we describe a typical geometrical feature of the cluster. In this

table, we also have included clusters that are parameterization specific. Cluster 10, the last cluster of the down-up group, has base doublets that do not qualify to be base pairs by the COM classification because the centers of mass of the two bases are too far apart to be defined as a base stack doublet by this technique. Cluster 2, found with the COM parameterization, contains many doublets that do not appear in the COP parameterization because of the sparsity of data points in this cluster. The same reasoning can be used to explain the failure of the Potts classification to find cluster 5 using the COP co-ordinates. Clusters 8 and 9 using COM are split into two clusters in the COP parameterization. Some differences in the content of the split clusters (8 versus 8a and 9 versus 9a) seems to indicate that the extracustering is necessary. At this stage, we have chosen to not consider the splitting, due to the small number of members in the split clusters. Finally, we would like to note that some of the new clusters are characterized by nontraditional stacking arrangements. While, we have not confirmed in this manuscript that all these clusters represent a preferred energy state, we nevertheless believe that these types of stacking are not artifacts and do not represent arbitrary arrangements. This subject will be the focus of future research. Fig. 10 shows a typical stacking doublet from each one of the nontrivial (non-Watson Crick) clusters. We have identified nine new clusters, among which seven appear in both parameterizations.

We have compared our results to the classification that was suggested by Sykes' work [20], which as mentioned earlier, proposed to use a mixture of k-means and agglomerative clustering methods. We have found that all of the representative doublets from the different stacking clusters discovered in Sykes' work belonged to the up-up and up-down trivial (Watson Crick) stacking. This finding is not surprising, since methods based on k-means are known to perform overclassification of the densely populated regions of the data space [9]. In future work, we intend to check the clustering on a larger data base to form an annotation scheme for the different stacking conformations.

6 Conclusions

In this paper, we proposed a clustering algorithm based on a Potts model. We used it to validate two documented structures, namely, the single nucleotide conformation and the base pair geometries, and one undocumented structure, namely, the base stacking. For the documented cases, we obtained reasonably good classifications by using the clustering procedure in a fully automated manner without employing any prior knowledge. However, some discrepancies between our clustering results and the previously published classifications were observed. Some of these turn out to appear in the fringes of the classes and do not seem to pose real issues. In several other cases, the discrepancies lead to the merging or splitting of clusters. By comparing these results to well-known structural motifs, we can conclude that those obtained with the Potts model seem to be finer than the previously developed classifications. For the case of base stacking we have established and validated a new classification scheme. This classification can be used as a base for a new structural annotation of RNA structures, which will enable a complete description of the backbone and base pair annotation schemes.

Finally, we have demonstrated the ease of use of our proposed method. We have shown through our examples that only one or two "temperatures" are typically needed for the analysis. The only cases that may require prior knowledge remain those for which the susceptibility diagram exhibits more than one peak. There, the more challenging step consists in determining what temperature defines the best clustering configuration. Despite this degree of uncertainty, we have demonstrated in the present work that this process constitutes a much simpler task than the choice of an optimal cut in a hierarchical clustering tree.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by grants from NSF, AFOSR, ARO, MURI, as well as by a grant from NIH (NAC P41 RR-13218) through Brigham and Women's Hospital. This work is part of the National Alliance for Medical Image Computing (NAMIC), funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54 EB005149. Information on the National Centers for Biomedical Computing can be obtained from <http://nihroadmap.nih.gov/bioinformatics>. This research was also supported by the National Science Foundation, Grant Nr. ECS-0535382, and by the National Institute of Health, through the Centers of Cancer Nanotechnology Excellence: Emory/GT Nanotechnology Center for Personalized and Predictive Oncology, Award Nr. 5-40255-G1: CORE 1. The authors would also like to thank Dr. Reuven Zeitak of Alcatel for some helpful conversations and to Prof. Neocles Leontis and Prof. Craig Zirbel, from Bowling Green State University, for supplying us with results from their FR3D code.

References

1. Blatt M, Wiseman S, Domany E. Data Clustering Using a Model Granular Magnet. *Neural Computation*. 1997; vol. 9:1805–1842.
2. Cech T. Ribozyme, the First 20 Years. *Biochemical Soc. Trans.* 2001; vol. 30:1162–1166.
3. Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh S-H, Srinivasan AR, Schneider B. The Nucleic Acid Database: A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. *Biophysical J.* 1992; vol. 63:751–759.
4. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P. The Protein Data Bank. *Nucleic Acids Research*. 2000; vol. 28:235–242. [PubMed: 10592235]
5. Noller H. RNA Structure: Reading the Ribosome. *RNA*. 2005; vol. 309:1508–1514.
6. Sponer, J.; Lankas, F. *Computational Studies of RNA and DNA*. Springer; 2006.
7. Moore P. Structural Motifs in RNA. *Ann. Rev. of Biochemistry*. 1999; vol. 68:287–300.
8. Hershkovtis E, Tannenbaum E, Howerton S, Sheth A, Tannenbaum A, Williams L. Automated Identification of RNA Conformational Motifs: Theory and Application to the HM LSU 23S rRNA. *Nucleic Acids Research*. 2003; vol. 1:6249–6257.
9. Hershkovtis E, Sapiro G, Tannenbaum A, Williams L. Statistical Analysis of RNA Backbone. *IEEE/ACM Trans. Computational Biology and Bioinformatics*. Jan-Mar;2006 vol. 3(no. 1):33–46.
10. Murray L, Arendall W, Richardson D, Richardson J. RNA Backbone Is Rotameric. *Proc. Nat'l Academy of Sciences USA*. 2003; vol. 100:13904–13909.
11. Richardson J, Schneider B, Murray L, Kapral G, Immormino R, Headd J, Richardson D, Ham D, Hershkovtis E, Williams L, Keating K, Pyle A, Micallef D, Westbrook J, Helen M, Berman H. RNA Backbone: Consensus All-Angle Conformers and Modular String Nomenclature. *RNA*. 2008; vol. 14:465–481. [PubMed: 18192612]
12. Duarte C, Pyle A. Stepping through an RNA Structure: A Novel Approach to Conformational Analysis. *J. Molecular Biology*. 1998; vol. 284:1465–1478.
13. Duarte C, Wadley L, Pyle A. RNA Structure Comparison, Motif Search and Discovery Using a Reduced Representation of RNA Conformational Space. *Nucleic Acids Research*. 2003; vol. 31:4755–4761. [PubMed: 12907716]
14. Schneider B, Moravek Z, Berman H. RNA Conformational Classes. *Nucleic Acids Research*. 2004; vol. 32:1666–1677. [PubMed: 15016910]
15. Gautheret D, Major F, Cedergren R. Modeling the Three-Dimensional Structure of RNA Using Discrete Nucleotide Conformational Sets. *J. Molecular Biology*. 1993; vol. 229:1049–1064.
16. Leontis N, Westof E. Analysis of RNA Motifs. *Current Opinion in Structural Biology*. 2003; vol. 13:300–308. [PubMed: 12831880]
17. Leontis N, Westof E. Geometric Nomenclature and Classification of RNA Base Pairs. *RNA*. 2001; vol. 7:499–512. [PubMed: 11345429]
18. Fukunaga, K. *Introduction to Statistical Pattern Recognition*. second ed.. Academic Press; 1990.

19. Jain, A.; Dubes, R. Algorithms for Clustering Data. Prentice-Hall; 1988.
20. Sykes M, Levitt M. Describing RNA Structure by Libraries of Clustered Nucleotide Doublets. *J. Molecular Biology*. 2005; vol. 351:26–38.
21. Lemieux F, Major F. Automated Extraction and Classification of RNA Tertiary Structure Cyclic Motifs. *Nucleic Acids Research*. 2006; vol. 34:2340–2346. [PubMed: 16679452]
22. Agrawal H, Domany E. Potts Ferromagnets on Coexpressed Gene Networks: Identifying Maximally Stable Partitions. *Physical Rev. Letters*. 2003; vol. 90:158102.1–158102.4.
23. Wang J, Swendsen R. Cluster Monte Carlo Method. *Physica A*. 1990; vol. 167:565–579.
24. Florian P, Domanyiegler G. Kissing Numbers, Sphere Packing and Some Unexpected Proofs. *Notices of the Am. Math. Soc*. 2004; vol. 51:873–883.
25. Calinski T, Harabasz J. A Dendrite Method for Cluster Analysis. *Comm. in Statistics, Simulation and Computation*. 1974; vol. 3:1–27.
26. Ban N, Nissen P, Hansen J, Moore P, Steitz T. The Complete Atomic Structure of the Large Ribosomal Subunit at 24aa Resolution. *Science*. 2000; vol. 289:905–919. [PubMed: 10937989]
27. Sarver M, Zirbel C, Stombaugh J, Mokdad A, Leontis N. FR3dD. *J. Math. Biology*. 2008; vol. 56:215–252.
28. Cremer D, Pople J. A General Definition of Ring Puckering Coordinates. *J. the Am. Chemical Soc*. 1975; vol. 97:1354–1358.
29. Feffrey, A. An Introduction to Hydrogen Bonding. Oxford Univ. Press; 1997.
30. Lemieux S, Major F. RNA Canonical and Non-Canonical Base Pairing Types: A Recognition Method and Complete Repertoire. *Nucleic Acids Research*. 2002; vol. 30:4250–4263. [PubMed: 12364604]
31. Klein D, Moore P, Steitz T. The Roles of Ribosomal Proteins in the Structure Assembly, and Evolution of the Large Ribosomal Subunit. *J. Math. Biology*. 2004; vol. 340:141–177.
32. Saenger, W. Principles of Nucleic Acid Structure. Springer-Verlag; 1984.
33. Waller M, Robertazzi A, Platts J, Hibbs D, Williams P. Hybrid Density Functional Theory for π -Stacking Interactions: Application to Benzenes, Pyridines, and DNA Bases. *J. Computational Chemistry*. 2006; vol. 27:491–504.
34. Sponer J, Leszczynski J, Hobza P. On the Nature of Nucleic Acid Base Stacking. Nonempirical ab Initio and Empirical Potential Characterization of 10 Stacked Base Pairs. Comparison of Stacked and H-Bonded Base Pairs. *J. Physical Chemistry*. 1996; vol. 100:5590–5596.
35. Gupta G, Sasisekharan V. Theoretical Calculations of Base-Base Interactions in Nucleic Acids: Stacking Interactions in Free Bases. *Nucleic Acid Research*. 1978; vol. 5:1639–1653.
36. Bugg C, Thomas J, Sundaralingam M. Stereochemistry of Nucleic Acids and Their Constituents. X. Solid-State Base-Stacking Patterns in Nucleic Acid Constituents and Polynucleotides. *Biopolymers*. 1971; vol. 10:175–219. [PubMed: 5544940]
37. Luo R, Gilson H, Potter M, Gilson M. The Physical Basis of Nucleic Acid Base Stacking in Water. *Biophysical J*. 2001; vol. 80:140–148.
38. Friedman R, Honig B. A Free Energy Analysis of Nucleic Acid Base Stacking in Aqueous Solution. *Biophysical J*. 1995; vol. 69:1528–1535.
39. Klein D, Schmeing T, Moore P, Steitz T. The Kink-Turn: A New RNA Secondary Structure Motif. *European Molecular Biology Organization J*. 2001; vol. 20:4214–4221.

Biographies

Xavier Le Faucheur received the PhD degree in electrical engineering, in 2010 and the MSc degree in industrial engineering, in 2009 from the Georgia Institute of Technology,

Atlanta. In 2005, he also graduated from the Ecole Supérieure d'Electricité, Gif-sur-Yvette, France. He is currently working toward PhD degree in electrical and computer engineering at the Georgia Institute of Technology under the supervision of Allen Tannenbaum. His research interests include the study of RNA conformations, 3D surface analysis, and wavelet signal decomposition.

Eli HersHKovits received the BA degree in mathematics and physics, in 1988 from the Hebrew University, the MSc degree in physics from the Weizmann institute of Science, Rehovot, in 1991, and the PhD degree in physics from the Weizmann Institute of Science, in 1997. He is presently a research scientist in the School of Electrical and Computer Engineering at the Georgia Institute of Technology, Atlanta. His research interests include bioinformatics, the study of RNA conformations, and stochastic perturbation methods in physical chemistry.

Rina Tannenbaum is a faculty member in the School of Materials Science and Engineering at Georgia Tech, Atlanta. She works in nano and bio-materials, colloidal systems, polymers, and interfaces.

Allen Tannenbaum is a faculty member with the Schools of Electrical and Computer and Biomedical Engineering, Georgia Tech, Atlanta. He works in computer vision, control, and bioinformatics.

Fig. 1.

(a) 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 traditional nucleotide residue goes from phosphate to phosphate, whereas an RNA suite, which is more appropriate for local geometry analysis goes from sugar to sugar (or base to base). (b) A base interaction with a possible parameterization.

Fig. 2. Susceptibility graph 10 different random generated data sets. No real clustering phase is observed.

Fig. 3. Susceptibility graph for residue conformation (RR0033). We observe the three successive peaks that are responsible for the major transitions in the cluster configuration. Temperatures for clustering are taken at the local minima that immediately follow each one of these peaks.

Fig. 4.

(a) The COM parameterization for base pair interactions uses the rotation angle around the COM (red circle), that is, needed to align the two bases and the (x, y) -projection of the distance between the two glycosidic nitrogen (blue circle) on the plane of the lower base. (b) COP co-ordinate system in the plane of the base.

Fig. 5.
Possible edges for base pair interaction in a base.

Fig. 6.

Two-dimensions projection of the clustering for the base pair geometry case with up-down configuration. The coordinates of the projection are the normalized azimuthal and the rotational angles needed to align the two bases. (a) All the data points before clustering. (b) The major clusters obtained with the proposed method and that correspond to the LW classification ([16]). See Table 4 for more details.

Fig. 7. Data points for the base stacking, with up-up configuration and using the COP parameterization. Possible clusters are difficult to detect.

Fig. 8. Double helix structure exhibiting base stacking structure. We can observe the different classes: I) Pyrimidine-purine, II) pyrimidine-pyrimidine or purine-purine and III) purine-purine.

Fig. 9.
Three major clusters for the up-up case of base stacking in the COP parameterization.

Fig. 10.

In this figure, we show examples of the seven new stacking geometries that were discovered using the Potts classification. (a) The single new up-up cluster (**cluster 2**). (b) The first new cluster of the down-down group. (c) The second cluster of the same group (**cluster 4**).

Typical geometries from the first (**cluster 5**) and the second (**cluster 6**) clusters of the up-down group are, respectively, shown in figures (d) and (e). Typical geometries from the first (**cluster 8**) and the second (**cluster 9**) clusters of the down-up group are, respectively, shown in figures (f) and (g).

TABLE 1
Residue Conformation Classification: Correspondence between Binning and Potts Clustering

Bin letter	Bin number	Peak	Potts cluster	Cluster/Bin	Bin/Cluster	Remarks
<i>a</i>	3111	third	<i>a</i>	1545/1766	1545/1545	A-form RNA helix
<i>a</i>	3111	second	<i>a'</i>	11	11	base on the 3' side bulge out
<i>e</i>	3112	third	<i>e</i>	40/160	40/40	3' end of an A-form RNA ζ is in g+
<i>i</i>	2211	second	<i>i</i>	105/113	105/107	crank shaft in A-form RNA
<i>r</i>	3122	second	<i>r</i>	104/133	104/104	interstrand stacking
<i>d</i>	1322	first	<i>d</i>	11/18	11/15	take part in kink-turn
<i>d</i>	1322	second	<i>d'</i>	6	6	ζ in t
<i>c</i>	1121	first	<i>c</i>	32/32	32/38	take part in kink-turn
<i>n</i>	3121	second	<i>n</i>	33/40	33/40	non contigious stack both 3' 5' directions
<i>o</i>	2111	second	<i>o</i>	61/68	61/71	the turn in Tetra loop
<i>I</i>	1211	first	<i>I</i>	38/38	38/39	interstrand stack or (i,i-2) stack
<i>t</i>	1111	second	<i>t</i>	38/41	38/43	stacked between the 5' adjacent residue to 3' non adjacent
<i>u</i>	3211	second	<i>u</i>	24/32	24/26	hinge between two helical strands
<i>s</i>	2122	first	<i>s</i>	34/38	34/34	take part in e-loop and kink-turn
<i>h</i>	3222	first	<i>h</i>	9/15	8/9	take part in e-loop
<i>g,7</i>	2121,4121	first	<i>g</i>	8/9	8/17	bulged residue, base often non stacked
<i>v</i>	3311	first	<i>v</i>	7/11	7/7	
<i>m</i>	1122	first	<i>m</i>	10/18	10/14	
<i>f</i>	1112	first	<i>f</i>	8/15	8/8	
3	1221	first	3	8/8	8/8	
<i>y</i>	1311	first	<i>y</i>	7/15	7/15	crank shaft in A-form RNA
<i>y</i>	1311	first	<i>y'</i>	6/15	6/7	
1	3321	first	1	8/9	8/10	
0	1222	first	0	4/5	4/6	
	3312	first	3312	5/5	5/6	

TABLE 2

Delimitation of the Bins in the Binning Method

α	γ	δ	ζ
40–100	10–110	65–105	240–350
125–200	140–210	130–165	other
220–350	230–350	other	
other	other		

TABLE 3
Suite Conformation Classification for RNA05 (Seven-Dimensions): Comparison of Potts Results with Previous Nomenclature

Cluster	Points in cluster	Bin Ascl	Points in bin	Cluster 2008 (Consensus)	Example	δ_{-1}	ϵ_{-1}	ζ_{-1}	α	β	γ	δ
<i>a</i>	5544	<i>a</i>	5544	<i>1a,1m,1L,7a,9a,6g</i>	UR0020-11	82(3)	211(10)	289(8)	296(8)	174(9)	53(6)	81(3)
<i>a'</i>	17	<i>a</i>	15	<i>&a</i>	RR0082-1940	82(5)	193(7)	256(7)	301(8)	174(6)	51(9)	80(7)
<i>o</i>	111	<i>o</i>	111	<i>1g</i>	RR0082-1864	81(4)	220(9)	291(11)	166(9)	158(18)	53(5)	85(3)
<i>T</i>	54	<i>T,t</i>	48	<i>7d,3d</i>	RR0082-636	85(4)	240(20)	230(30)	68(18)	176(26)	58(11)	86(5)
<i>t</i>	13	<i>t</i>	11	<i>5d</i>	UR0020-49	81(6)	195(20)	59(13)	59(15)	136(7)	48(5)	84(8)
<i>u</i>	43	<i>u</i>	39	<i>1e</i>	UR0035-2665	81(3)	210(18)	286(19)	249(17)	85(12)	167(9)	85(3)
<i>u'</i>	16	<i>u</i>	13		RR0082-877	84(10)	230(23)	298(13)	291(11)	159(11)	182(12)	84(9)
<i>i</i>	522	<i>i</i>	517	<i>1c,1f</i>	UR0020-a28	81(5)	203(19)	281(24)	149(18)	203(24)	177(13)	84(5)
<i>i'</i>	8	<i>i</i>	8		PR0018-n69	87(8)	64(14)	293(19)	152(16)	184(30)	166(13)	89(4)
<i>L</i>	14	<i>L</i>	11	<i>5j</i>	AR0027-b17	96(18)	218(13)	73(15)	63(13)	104(17)	178(6)	83(3)
<i>n</i>	273	<i>n</i>	246	<i>1b,1l</i>	AR0023-b71	84(5)	217(12)	289(10)	302(11)	187(22)	57(9)	142(10)
<i>E</i>	12	<i>E</i>	12	<i>3b</i>	RR0082-904	84(2)	217(15)	169(15)	280(19)	163(18)	49(9)	144(6)
<i>E'</i>	10	<i>E</i>	9		PR0057-9	80(2)	227(14)	230(11)	290(14)	206(19)	49(10)	142(5)
<i>g</i>	19	<i>g</i>	g	<i>1z</i>	RR0082-1771	84(3)	207(9)	270(13)	197(18)	154(20)	53(8)	142(11)
<i>s</i>	43	<i>s</i>	42	<i>5z</i>	UR0026-3654	83(3)	207(6)	50(10)	164(7)	150(15)	51(5)	146(7)
<i>m</i>	11	<i>m</i>	9	<i>7p</i>	PR0033-b8	83(3)	229(20)	228(14)	67(9)	168(4)	55(6)	145(6)
11222	28	11 222	23	<i>1t</i>	PTE003-b907	83(5)	208(17)	290(15)	181(19)	184(23)	183(15)	143(10)
<i>d</i>	9	<i>d</i>	8	<i>7r</i>	RR0082-262	81(3)	218(8)	217(17)	55(11)	172(14)	296(3)	151(5)
<i>F</i>	108	<i>F</i>	97	<i>2a</i>	RR0082-1711	142(10)	261(13)	291(20)	286(15)	192(17)	53(8)	84(4)
<i>F'</i>	12	<i>F</i>	6		RR0082-1879	144(8)	192(9)	244(11)	287(21)	171(12)	51(7)	7 9(5)
<i>A</i>	112	<i>A</i>	101	<i>4a,0a,#a</i>	RR0082-2485	148(9)	228(23)	127(36)	280(18)	156(18)	43(10)	87(6)
<i>b</i>	25	<i>b/p</i>	21	<i>4g</i>	UR0012-a226	147(9)	256(19)	172(25)	205(14)	165(15)	49(7)	83(3)
<i>f</i>	26	<i>f</i>	14	<i>8d,4d</i>	RR0009-c1062	148(5)	265(15)	220(32)	75(17)	186(25)	56(8)	87(4)
<i>f'</i>	30	<i>f</i>	26	<i>6d</i>	RR0082-1879	140(13)	235(26)	87(21)	58(14)	159(23)	50(8)	84(4)
<i>I</i>	19	<i>I</i>	14	<i>4n</i>	RR0082-767	143(10)	226(13)	203(20)	74(10)	215(20)	193(11)	83(4)
<i>I'</i>	56	<i>I</i>	43	<i>0i,6n</i>	RR0082-940	143(11)	266(21)	84(18)	78(25)	201(31)	181(11)	89(14)

Cluster	Points in cluster	Bin Ascii	Points in bin	Cluster 2008 (Consensus)	Example	δ_{-1}	ϵ_{-1}	ζ_{-1}	α	β	γ	δ
r	58	r	44	2l	RR0082-264	143(11)	252(24)	292(23)	284(23)	209(20)	55(11)	145(8)
R	72	R	68	4b0b	RR0082-247	144(8)	237(24)	147(31)	297(24)	168(18)	45(13)	144(7)
c	47	c	37	6p	RR0082-96	145(8)	260(19)	93(24)	73(18)	176(22)	54(14)	147(6)
h	9	h	9	4s	UR0026-2655	150(2)	248(14)	170(10)	277(12)	84(6)	177(5)	148(9)
12231	14	12231	2		RR0082-2613	106(29)	230(44)	71(20)	198(38)	256(29)	278(34)	104(23)
y	50	y	38		RR0082-1206	83(5)	221(13)	266(11)	60(17)	196(23)	279(22)	96(9)

TABLE 4

Up-Down Base Pairs in RR0082 Conformations

Cluster name	Cluster size	Server's group	Server's group size	Typical content	Remarks
cWW(1)	671	cWW	668		
cWW(2)	10	cWW	0	GC basepairs (θ is larger than cluster 2)	
cWW(3)	19	cWW	19	GU base-pairs	
cWW(4)	19	cWW	19	GU base-pairs θ is larger than cluster 1	
cWW(5)	42	cWW	42	36*UG base-pairs + 6*UU base-pairs(cwW)	
tHW(1)	26	tHW	26	almost all base-pairs are AU	
tHW(2)	7	tHW	6	almost all base-pairs are AC	
tWH(1)	8	tWH	7	CA or AA base-pairs	
tWH(2)	7	tWH	7	UA base-pairs	
tHS(1)	35	tHS	34	AG base-pairs (AN6-G02')	
tHS(2)	11	tHS	11	AG base-pairs (AN6-G04')	
tHS(3)	7	tHS	0	AG N2-O2p interaction (i,i+3)	
tHS(4)	12	tHS	8	pyr-pur, pur-pyr	
tSH	43	tSH	42	AG base-pairs	
csS(1)	19	CsS, CSs	12,5	pur-pur(CsS) or UA (cSs)	
csS(2)	21	CsS, CSs	8,9	AC,CA,AU base-pairs	
CSs	8	cSs	6	AG,GG	
biff(1)	8	biff	8	CC base-pair	
biff(2)	8	biff	2	AC,CA,UA base-pairs	
cWS	-	-	13		This Sarver's group is not a Potts cluster
cSW	-	-	21		This Sarver's group is not a Potts cluster

TABLE 5

Up-Up Base Pairs in RR0082 Conformations

Cluster name	Cluster size	Sarver's group	Sarver's group size	Typical content
tWW	20	tWW	20	mostly pur-pyr
cWh	6	cWh	5	mostly GU basepairs
cHW	37	cHW,cHS	12,11	T=0.019 two groups that match the ZL notation. cHS mostly pur-pyr Mostly,(i,i+1+),(i,i+2), (i,i+3)
cHS(1)	11	cHS	11	UG base-pairs
cHS(2)	8	cHS	6	mostly pur-pur base-pairs
cSH	6	cSH	1	other base-pairs do match the tSH geometry
twS	36	twS	20	mostly AG base-pairs most of the non defined ZL cases also have tWS geometry
tsw	13	tsw	6	mostly AG base-pairs most of the non defined ZL cases also have tSW geometry
thH	18	thH,tHH	16,2	almost all AA base-pairs
tsS	17	tsS	16	almost all are GA base-pairs
tSs	37	tsS,tSs	14,20	mostly pur-pyr,AG base-pairs

TABLE 6

Up-Up Group in Watson-Crick Double Helix Content Groups

Content class	majority cluster	# of bases in content class	# of bases in majority cluster
pyr-pyr	1	597	570
pur-pyr	2	210	178
pyr-pur	3	347	310

TABLE 7

Up-Down Group in Watson Crick Double Helix Content Groups

Content class	majority cluster	# of bases in content class	# of bases in majority cluster
pur-pur	1	130	111
pyr-pyr	2	73	67
pyr-pur	3	309	289

TABLE 8

Clusters of Base Stacking Doublets

Base orientation group	Points w/ C.o.P.	Points w/ C.o.M.	Temp w/ C.o.P.	Temp w/ C.o.M.	Cluster index	Points in C.o.P. cluster	Points in C.o.M. cluster	Common Points	Typical LW content	Base type	Geometry	Sequence order
up-up	1902	2128	0.001	0.006	1	1818	1991	1740	(cWW,cWW)			
					2		27		$9 \times (\text{cWW}, \text{tWH})$	pur-pur	im(i) above im(j)	
down-down	94	87	0.12	0.129	3	23	17	14	$10 \times (\text{cWW}, \text{cWW})$		Im(i) above sugar(j)	
					4	32	30	26	(tWS*), (tWH,*)		pyr(i) above pyr(j)	
up-down	858	715	0.111	0.098	5	-	19	-	$8 \times (\text{tWH}, *)$	10*UA	Im(i) above O2(i)	16 × (i+2,i)
					6	19	10	5	$7 \times (\text{tHS}, \text{cWW})$	pur-pur	pyr(i) above N6/O6(j)	
					7	527	465	398	(cWW,cWW)	pur-pur		
down-up	215	178	0.089	0.105	8	19	31	17	$10 \times \text{tHH}, \text{tHS}, \text{tWS}$	AC or AU	pyr(i) above N6/O6(j)	28 × (i+1, i)
					8a	10	same cluster	9		pyr-pyr		
					9	22	38	16	$7 \times (\text{tHS}, *)$	AA, pyr-pur, pur-pyr		
					9a	13	same cluster	10	$7 \times (\text{tHS}, \text{tWW})$	GG,AG,UG	pyr(i) above pyr(j)	
					10	55	-	-	(cWW,cWW)	pur-pur		far appart