Published online 1 April 2009

Nucleic Acids Research, 2009, Vol. 37, No. 9 ***e66***
doi:10.1093/nar/gkp206

Identification and classification of ncRNA molecules using graph properties

Liam Childs[1,*], Zoran Nikoloski[2], Patrick May[1] and Dirk Walther[1]

[1]Max-Planck Institute for Molecular Plant Physiology, Am Mühlenberg 1 and [2]University of Potsdam, Institute for Biology and Biochemistry, Karl-Liebknecht-Str. 24-25, Haus 26, 14476 Golm, Germany

Received December 22, 2008; Revised February 27, 2009; Accepted March 12, 2009

ABSTRACT

The study of non-coding RNA genes has received increased attention in recent years fuelled by accumulating evidence that larger portions of genomes than previously acknowledged are transcribed into RNA molecules of mostly unknown function, as well as the discovery of novel non-coding RNA types and functional RNA elements. Here, we demonstrate that specific properties of graphs that represent the predicted RNA secondary structure reflect functional information. We introduce a computational algorithm and an associated web-based tool (GraPPLE) for classifying non-coding RNA molecules as functional and, furthermore, into Rfam families based on their graph properties. Unlike sequence-similarity-based methods and covariance models, GraPPLE is demonstrated to be more robust with regard to increasing sequence divergence, and when combined with existing methods, leads to a significant improvement of prediction accuracy. Furthermore, graph properties identified as most informative are shown to provide an understanding as to what particular structural features render RNA molecules functional. Thus, GraPPLE may offer a valuable computational filtering tool to identify potentially interesting RNA molecules among large candidate datasets.

INTRODUCTION

Non-coding RNA genes (ncRNA) are integral components of many biological processes including translation (tRNA, rRNA), RNA splicing (ribozymes), gene regulation through mRNA hybridisation (miRNA, piRNA), gene regulation through metabolite binding (riboswitches) and RNA methylation and pseudouridylation (snoRNA) (1). Functions such as translation and RNA splicing have long been considered to be the sole role of ncRNA. However, new and unexpected functions have been discovered recently, revealing that RNA molecules assume highly diverse functions and are more actively involved in biological processes than previously thought (2). The intensified study of ncRNA and search for new functional roles of RNA is further propelled by the realisation that a larger portion of intergenic space than previously acknowledged is actually transcribed. For instance, 85% of the fruit fly genome (3), 62% of the mouse genome (4) and a staggering 93% of the human genome (5,6) have been reported as transcribed. Understanding the functional role of this otherwise seemingly wasteful transcription requires the analysis of large amounts of genomic sequence data. Thus, computational methods have a great potential to contribute significantly toward this goal by predicting potentially functional non-coding regions and their respective function.

The structure of ncRNA is thought to provide insight into the biological function (7). In the folding process, characteristic nucleotide base-pairing and stacking interactions play significant roles and are governed by molecular forces acting on and within any molecule in aqueous solutions (e.g. electrostatic interactions) (8). The adopted shapes or folds can be highly complex and are capable of carrying out a variety of molecular functions, such as binding metabolites and proteins with high specificity (9–14). RNA is particularly suited for hybridizing with nucleotide sequences allowing for highly specific targeting of genes and genomic regions (15–17). Furthermore, it is conceivable that two ncRNA molecules with completely different nucleotide compositions would still fold to form the same structure and have the same function. For example, the secondary structure of tRNA has a characteristic cloverleaf shape; however, the nucleotide composition of tRNA can vary to the degree that two tRNAs can have completely different sequences. Thus, methods that incorporate ncRNA structural, and not just sequence, information are required for an accurate prediction of function.

Due to the importance of RNA structure, several computational RNA folding tools have been developed, such as: mfold (18), RNAfold (19), vsfold (20), evofold (21) and sfold (22). The majority of these algorithms work on an input sequence to determine the folded secondary

*To whom correspondence should be addressed. Tel: +49 0 30 5678 624; Fax: +49 0 30 5678 136; Email: childs@mpimp-golm.mpg.de

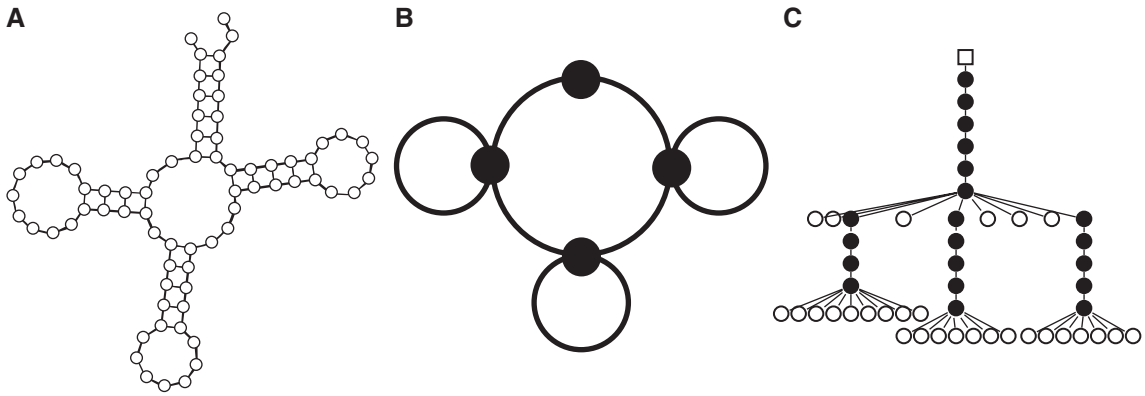

Figure 1. Representations of RNA structure using graphs. (**A**) A typical tRNA structure represented using the bracketed graph representation. Nucleotides are represented as nodes (open circles) and bonds (both base-base hydrogen bonds and backbone ester bonds) as edges. The secondary structure of the tRNA is reminiscent in the shape of the graph. This is the chosen graph representation in the current article. The values for the 20 chosen graph properties for this graph are shown in Table 1. (**B**) The dual-edge graph for the same tRNA is shown. Stems are converted to nodes and loops to edges. Information about dangling ends is lost in this representation. (**C**) Planar tree representation uses a special node for the root (5′-/3′-end of the structure) depicted here as an open square. Base pairs are converted to 'stem' nodes (closed circles) and loop nucleotides are converted to 'loop' nodes (open circles). The tree is built by following the strand from 5′ to 3′ and the order of children is important. Information about dangling ends is also lost in this representation.

structure that minimizes the free energy by optimizing the intramolecular base pairing. The input sequence may come from publicly available repositories, e.g. Rfam (23) which currently contains 636 138 sequences, grouped in 603 ncRNA families, that are largely computationally annotated (24).

The listed structure prediction tools are fast and accurate when operating on sequences of less than 200 bp; however, they are not suitable for longer sequences (25). The accuracy of the predicted structure has been improved by algorithms that use multiple sequence alignments to produce a consensus structure. Another, relatively recent class of algorithms are designed to fold pseudoknots—structures where each bonded base pair is not required to be bounded by another bonded pair of bases that are closer to the ends of the molecule. An example of a pseudoknot is the 'kissing hairpin' where the loops of two hairpins are bound to each other. The problem of folding pseudoknots was shown to be NP-complete and many tools do not attempt to fold them (26). There is also an algorithm that predicts the 3D structure (as opposed to secondary and pseudoknotted structures) of RNA molecules (27). This method involves using fragments of RNA whose 3D structure has been experimentally determined as building blocks to assemble the shape of an investigated molecule. Any methods developed to predict ncRNA function are fully reliant on the ability for these tools to predict the structure accurately.

There are very few tools that deal with the classification of functional versus non-functional RNA sequences. One attempt to develop such a tool investigated the idea that the minimum free energy (MFE) of functional RNA sequences should be lower than that of random, shuffled and non-functional genomic sequences (28). In the study, MFE was identified to be largely unhelpful except in a later study, which discovered that MFE can be used to identify miRNA (29). Other studies calculated the thermodynamic stability of multiply aligned structures as a means of identifying functional RNA (30) and have been applied to the genomes of *Saccharomyces cerevisiae* (31) and *Plasmodium falciparum* (32).

Assigning unannotated RNA sequences to an Rfam family is better investigated and there are a wide variety of ncRNA family specific predictors, most of which focus on miRNA (33–36). Covariance models, as general predictors, are used to identify nucleotide pairs which vary together across multiple alignments and are thus likely to be bonded in secondary structure (37). Such models require multiple alignments and are computationally time consuming, limiting the number and type of sequences that can be processed.

A large portion of recent RNA-related research applies concepts developed in graph theory to the analysis of RNA structure (38–40). A graph is an abstraction of the relationship among objects, which uses nodes to represent the objects and edges to represent the relationship between two objects. There are many ways to represent RNA structure with graphs (Figure 1), including the bracketed (where nucleotides a converted to nodes and bonds to edges), planar tree (where base pairs are converted to 'stem' nodes and loop nucleotides are converted to 'loop' nodes, while following the molecule from 5′ to 3′) and dual graph representations (where stems are converted to nodes, while loops to edges), each with different advantages and disadvantages including information loss and complexity of calculation (38). Graph topology derived from RNA structure has also been used to assign Rfam family (41). Although the ability to discriminate between functional and non-functional genes was not demonstrated, this approach appears quite successful in terms of classification.

The structure of a graph can be employed to define and analyse different properties that could reflect the characteristics of the process or entity modelled by the graph (Table 1; Supplementary Material, Section 1). A property can be defined on the level of graph constituents

Table 1. Graph properties calculated for a typical tRNA depicted in Figure 1

Number of articulation points	3	Average Burt's constraint	0.4234
Average path length	9.577	Variance of Burt's constraint	0.0161
Average vertex betweenness	313.1	Average degree	2.514
Variance of vertex betweenness	54817.6	Diameter	22
Average edge betweenness	278.2	Girth	4
Variance of edge betweenness	40784.5	Average coreness	1.959
Average cocitation coupling	0.0555	Variance of coreness	0.0394
Average bibliographic coupling	0.0555	Maximum coreness	2
Average closeness centrality index	0.1088	Graph density	0.0344
Variance of closeness centrality index	0.00048	Transitivity	0

We have chosen 20 graph properties to calculate and train the SVMs. These properties were chosen by considering the following criteria (i) polynomial-time computation, (ii) relevance to local and global levels of the graph and (iii) usage in complex network research. The values shown beside each property are the graph properties as calculated for the graph shown in Figure 1.

(i.e. nodes and edges) or on the level of the graph itself. Furthermore, computing a property may require limited or full knowledge of the graph. Based on these two criteria (level of detail and required knowledge of the graph), graph-theoretic properties may be classified into local (using limited knowledge of the graph and referring to a graph's constituent), local–global (using full knowledge of the graph and referring to a graph's constituent), and global (using full knowledge of the graph and referring to the graph itself). Thus, graph representations of RNA molecules offer a means to capture both local–global and global structural properties that can be used to deduce the large- and small-scale structural, and therefore functional, differences between molecules.

Here, we go another level of abstraction higher than previous methods and address the question of how a set of selected graph-theoretic properties derived from a graph representation for predicted RNA secondary structures can be used as characteristic features for the classification of RNA molecules. Among the immense number of existing graph-theoretic properties, we select several representatives based on the following three criteria: (i) polynomial-time computation, (ii) relevance to local and global levels of the graph and (iii) usage in complex network research. As a means of exploring the relationship between graph properties and Rfam families, we attempt to recall the Rfam families of ncRNA sequences using support vector machines (SVMs) trained on the selected graph properties. Furthermore, we show that graph properties can be employed to differentiate between functional and non-functional sequences as well as predict a likely function. In this study, a small number of graph properties are identified as most relevant for the correct classification of ncRNAs and their interpretation is demonstrated to shed light on structural properties that may render RNA molecules functional compared to their non-functional counterparts.

MATERIALS AND METHODS

The data set

Seed and full RNA sequence alignment datasets were obtained from Rfam release 9.0 (23) (http://www.sanger.ac.uk/Software/Rfam) and all redundant identical sequences were removed using CD-HIT (42) (http://bioinformatics.ljcrf.edu/cd-hi/), yielding 52 855 (full) and 18 974 (seed) unique sequences for analysis. These sequences were split into 210 Rfam and 8 compound families, which were formed out of several smaller related Rfam families (CD-box, HACA-box, internal ribosome entry sites, leader sequences, miRNA, riboswitches, ribozymes and scaRNA). All RNA sequences were folded into their predicted secondary structures using RNAfold (19).

Calculating graph properties

The bracketed graph representation was used to represent the predicted structure (Figure 1). It was calculated by converting all nucleotides to nodes and all bonds between nucleotides (both ester and hydrogen) to edges.

From the three different ways in which a property can be defined and calculated, here we used the summary statistics for the local–global properties, since they provide insight not only on the global level of the graph itself, but also on the level of its nodes and edges. The employed statistics (mean and variance) allow for a uniform way of summarizing the distribution of values an investigated local property may assume. For instance, the node-betweenness used in our analysis is given by the mean and variance of the distribution of node-betweenness values over all nodes of a graph. Similarly, we used bibliographic coupling as given by the mean and variance of bibliographic couplings over all pairs of nodes.

All properties were calculated using the *igraph* R package (43) (http://cneurocvs.rmki.kfki.hu/igraph) for complex networks with our own extensions to the presently implemented algorithms that facilitate the extraction of the graph representation and calculation of the necessary summary statistics. We focused on the following global properties: number of articulation points, diameter, girth, density and transitivity, together with the local–global properties (given by the mean and variance): Burt's constraint, path length, node betweenness, edge betweenness, degree, co-citation coupling, bibliographic coupling, coreness and closeness (a brief definition of all graph properties used in this study is provided in the Supplementary Material, Section 1).

SVM training and testing

We used the following procedure for training and testing all SVMs: First, we produced matched training/testing sets with randomly selected, but non-overlapping sequences and matching graph property sets. SVMs were then created from the training sets using libSVM software (44). All graph properties for the training sets were initially scaled between −1 and 1 to prevent graph properties with larger numerical ranges from dominating those with smaller ranges. A 10-fold cross validated grid search, based on the training set, was used to optimize

the initial parameters C (the cost parameter) and γ (the kernel width). In addition, the SVM was trained on the full training set using the optimised values. The radial basis function (RBF) kernel was employed as it is able to identify non-linear relationships between class-labels and features (graph properties), requires fewer hyper-parameters, and presents fewer numerical difficulties than other kernels. The testing sequences were in turn submitted to test the SVMs, and results are reported in the Results section. Each SVM was trained 100 times with different sets of random sequences.

The importance of the graph properties was calculated using the *F*-score (45). The *F*-score is a simple measure that discriminates between two sets of real numbers. Given m training vectors, x, and n_+ positive and n_- negative instances, then the *F*-score of the i-th feature is defined as:

$$F(i) \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\left(\begin{array}{c} 1/(n_+ - 1) \sum_{k=1}^{n_+} \left(x_{k,i}^{(+)} - \bar{x}_i^{(+)} \right)^2 \\ +1/(n_- - 1) \sum_{k=1}^{n_-} \left(x_{k,i}^{(-)} - \bar{x}_i^{(-)} \right)^2 \end{array} \right)}, \quad \mathbf{1}$$

where x_i, $x_i^{(+)}$, $x_i^{(-)}$are the average of the i-th feature of the whole, positive and negative data sets, respectively; $x_{k,i}^{(+)}$ is the i-th feature of the k-th positive instance, and $x_{k,i}^{(-)}$ is the i-th feature of the k-th negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the *F*-score, the more likely it is that this feature is more discriminative. This algorithm is available using *fselect* which is available on the libSVM internet site.

Functional versus non-functional RNA sequence prediction

SVMs were trained to differentiate functional from non-functional RNA using graph properties. Sequences available from Rfam are considered functional and comprise the set of all functional sequences (here, mRNA is considered non-functional). A non-functional set was created by shuffling each Rfam sequence once while preserving dinucleotide content using uShuffle (46). A 200 functional and 200 non-functional sequences were randomly chosen for the training and testing sets with each family having an equal chance of being chosen, yielding 400 sequences for each set. After classification, the important graph properties were determined by calculating the *F*-score.

Predictive power of graph properties

We attempted to determine whether graph properties alone can be used to recall Rfam families. To remove the influence of sequence similarity and length, we filtered the training and testing sequences in two ways.

First, to account for sequence similarity, we created diverging testing and training sets. A distance matrix for each family was created by an all-against-all comparison of sequences within a family using the similarity score provided by CLUSTALW pairwise alignments (47) (http://www.clustal.org/). Each family was then divided into diverging training and testing sets, where the greatest similarity between a member picked from a training set and a member chosen from the paired testing set would be less than or equal to a given threshold. We set the initial threshold to give a maximum similarity between the two sets of 90 percent identity (%id) to allow the training and testing sets to become highly similar but not identical. We then decreased this threshold in steps of 10%id. As any two completely random RNA sequences are expected to have 25%id due to random chance, we set the lower bound to 20%id, thus creating a total of eight sets (20, 30, 40, 50, 60, 70, 80, 90%id).

Second, graph properties were calibrated for the potential bias introduced by length and GC content (%G + C). A set of random sequences was generated, in which all combinations of the lengths 50–1000 nt (in steps of 50) and the %G + C 10–100% (in steps of 10) were represented 100 times, producing a matrix for each graph property with 10 000 entries. The graph properties of each sequence were then calibrated by dividing by the entry with the closest length and %G + C in the corresponding calibration matrix. The *F*-score was also used here to calculate the predictive power of each graph property.

As the maximum similarity between the training and testing set decreases, the number of available sequences also decreased and many families became too small to be used leaving, finally, 18 families for analysis (Supplementary Material, Section 2). Training sets were restricted to 50 random members, while testing sets were restricted to 20 from each family.

The sensitivity (Q^D) and specificity (Q^M) of SVM-based predictions for each individual family were calculated using the following equations:

$$Q_i^D = \frac{z_{ii}}{\sum_j z_{ij}} \quad \mathbf{2}$$

and

$$Q_j^M = \frac{z_{jj}}{\sum_j z_{ij}}, \quad \mathbf{3}$$

where Z_{ij} is an entry in a confusion matrix, i is an index for the actual family and j is an index for the predicted family.

We also investigated the possibility of combining the results of the SVM with the sequence-based assignment of Rfam family using BLAST in order to improve accuracy. For each sequence, the SVM produces probabilities ($p\text{score}_{\text{SVM}}$) that the sequence belongs to each ncRNA family. The sum of the $p\text{scores}_{\text{SVM}}$ totals to 1. Similarly, for each sequence, we produced an *E*-value for each family using BLAST. This *E*-value was adjusted to the same scale as the SVM pscore by calculating the inverse *E*-value as a fraction of the total inverse *E*-values:

$$p\text{score}_{\text{BLAST}} = \frac{1/e_i}{\sum_{k=0}^{n} \frac{1}{e_k}}, \quad \mathbf{4}$$

where e_i the *E*-value obtained for an individual family and e_k is the sum of *E*-values over all families. The two values were then combined linearly using a weighting factor, α, as follows:

$$p\text{score}_{\text{MERGE}} = (1-\alpha) \times p\text{score}_{\text{SVM}} + \alpha \times p\text{score}_{\text{BLAST}} \quad \mathbf{5}$$

As a result, we obtained for each sequence a merged *p*-score for each family that, although not considered a probability, indicates how likely the sequence belongs to that family. The family with the highest $p\text{score}_{\text{MERGE}}$ was assigned to the sequence.

Standalone WUBLAST (48) (http://blast.wustl.edu), the INFERNAL package (37) (http://infernal.janelia.org) and HMMER (49) package (http://hmmer.janelia.org/) were used to provide references to methods which are expected to perform either poorly and well on the diverging training sets. As the sets diverge, the performance of sequence comparison based methods, such as WUBLAST, should degrade, whereas structure based methods would ideally remain stable. The comparison was performed using the same training and testing sets. A description of how these methods were applied can be found in the next section.

COMPARISON TO OTHER METHODS

To compare our method to existing tools, we a chose representative method from each class of classifier presented in a previous study used to benchmark a number of other tools (50). We chose WUBLAST from the homology-based methods, HMMER from the Hidden Markov Model-based methods and INFERNAL from the covariance model-based methods. Training sets of 50, 100 and 200 seed sequences per Rfam family were generated, which resulted in 25, 8 and 3 Rfam families of sufficient size for each training set, respectively. All tools were used with the default settings following the same procedure described in the previous section.

For comparison with WUBLAST, each training set was split into the constituent Rfam families and converted into blastable databases. The testing set was then blasted against each database, using an *E*-value threshold of 100, resulting in a set of *E*-values for each sequence that measures how well it matched each Rfam family. Sequences were then classified according to the family with the lowest *E*-value.

A similar procedure was followed using INFERNAL and HMMER. Training sets were split into constituent Rfam families and aligned using MUSCLE (51) (http://www.drive5.com/muscle). From each family alignment, covariance and Hidden Markov models were built. The testing set was then searched using each of the models and each sequence was scored on how well it matched a given family. Sequences were classified according to the best identified matching family.

Performance measures

Prediction performances of classifiers was assessed using the Matthew's correlation coefficient (MCC) (Equation 6), and Receiver Operating Characteristic (ROC) and associated area under the ROC (AUC) reported in the Supplementary Material.

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}, \quad \mathbf{6}$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives.

RESULTS

In this work, we developed three approaches to investigate graph properties and their ability to reflect the functional information of RNA molecules. In the first approach, we tested the ability of graph properties to discriminate between functional and non-functional RNA molecules. In the second, we removed any bias that may be introduced through sequence similarity, length and GC content (%G + C) by using calibrated and diverging training and testing sets to test the predictive power of the graph properties alone when predicting the Rfam family of an ncRNA sequence. In the third, we removed the limitations imposed in the second approach and compared the ability for the developed method to predict Rfam family to other established tools.

In the first approach, the classifier based on SVM and using graph properties as features was able to classify RNA sequences into functional and non-functional classes with Matthew's Correlation Coefficients (MCC) ranging between 0.61 and 0.98 with an average MCC of 0.87, and sensitivity and specificity of 0.73, respectively (Supplementary Material, Section 4). This indicates that graph properties can be used to identify functional RNA sequences and performs significantly better than random assignment (MCC = 0). The discriminatory power of each graph property was then calculated using a measure called the *F*-score (see Materials and Methods section) (Figure 2). This score revealed that the 'number of articulation points' possessed the most discriminatory power with an average *F*-score of 0.094 followed by the 'variance of coreness' (0.080), 'average coreness' (0.062), 'average Burt's constraint' (0.062) and 'average degree' (0.056). The *F*-score decreased significantly for the remainder of the graph properties along with the minimum free energy (MFE). 'Girth', 'maximum coreness' and 'transitivity' had little or no discriminatory power and were included to provide baseline support for high-scoring graph properties.

The second approach explored the idea that graph properties are able to reflect RNA structure and function in greater detail by attempting to recall the correct Rfam family without the influence of sequence similarity, length and %G + C. To control for sequence similarity we created diverging testing and training sets. To control for sequence length and %G + C, we performed a calibration using generated random sequences of various lengths and %G + C (see Materials and Methods section). SVMs trained on calibrated graph properties classify RNA sequences with an average MCC of 0.32 (Figure 3); i.e. substantially above the expected rate when guessing.

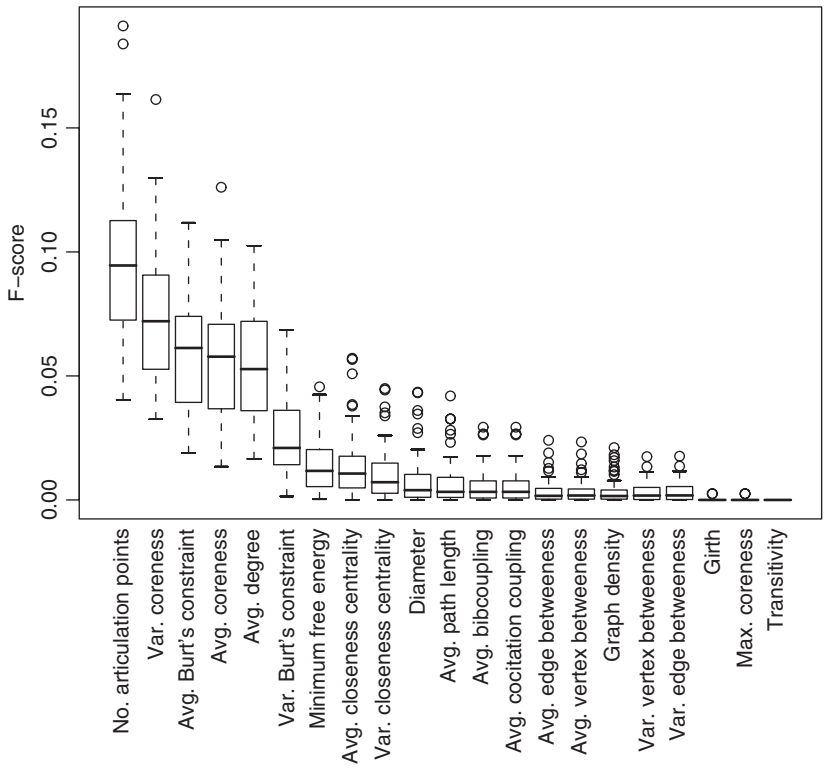

Figure 2. Graph property discriminatory power for functional vs. non-functional classification. The discriminatory power of each graph property was determined by calculating the *F*-score (Equation 1) with larger *F*-scores indicating more relevant properties. The distribution of *F*-scores is shown for each graph property as a box plot where the middle bar is the median, the outer edges are the 10 and 90 percentiles and the edges of the box are the 25 and 75 percentiles. Outliers are shown as circles. When classifying functional versus non-function RNA, we find that the 'number of articulation points', 'variance of coreness', 'average coreness', 'average Burt's constraint' and 'average degree' consistently have significantly higher *F*-scores than the other graph properties.

This value is relatively stable at all eight selected thresholds of sequence divergence as it varies between 0.29 and 0.37, and shows that the method is robust at all levels of sequence divergence.

To test whether purely sequence-homology-based methods can correctly identify family members under conditions of increasing sequence divergence, we ran WUBLAST on datasets of increasing sequence divergence as well. When using WUBLAST to classify the sequences based on sequence similarity, there is a significant drop in the average MCC from 0.48, obtained for sets with 90% maximum similarity between training and testing sets, to only 0.08 (20% maximum similarity) (Figure 3). Thus, sequence-based methods alone appear insufficient to correctly detect family members that have diverged at the sequence level. Such remote family members are likely to be identified by methods relying on structural aspects such as INFERNAL. Unexpectedly, and similar to WUBLAST, we observed a drop in the average MCC from 0.49 to 0.12 as the sets diverge (Figure 3) implying that within Rfam families, the structures predicted by RNAfold may not have sufficient similarity to each other be detected by INFERNAL. HMMER, a method based on Hidden Markov Models, was also used on the diverging training and testing sets and performs significantly worse at all levels of divergence than the other methods (Figure 3). While WUBLAST and INFERNAL

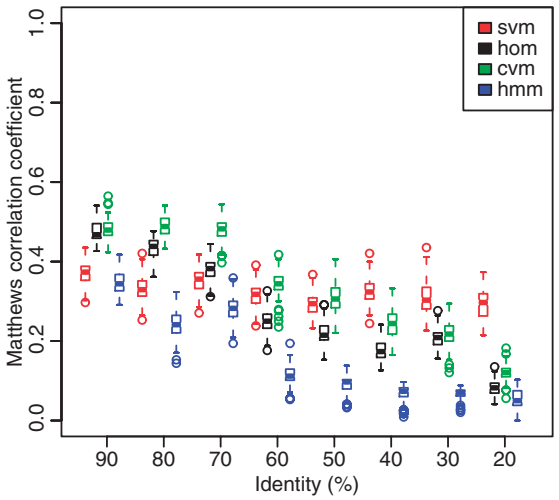

Figure 3. Rfam family classification results for calibrated, diverging RNA sequence sets. At high levels of sequence identity, the performance—as judged by the MCC—of the graph-property-based SVM method (svm) is worse than the other tested methods when classifying sequences calibrated for length and %G+C indicating that RNA families have distinct lengths and %G+C that affect the graph properties. However, the performance remains stable as the sequences diverge, whereas the performance of homology methods (hom), covariance models (cvm) and Hidden Markov Models (hmm) degrades sufficiently that SVMs still outperforms them at maximum similarities of 50% and below. In this performance comparison, 18 sufficiently large Rfam families were included in the training and testing protocol.

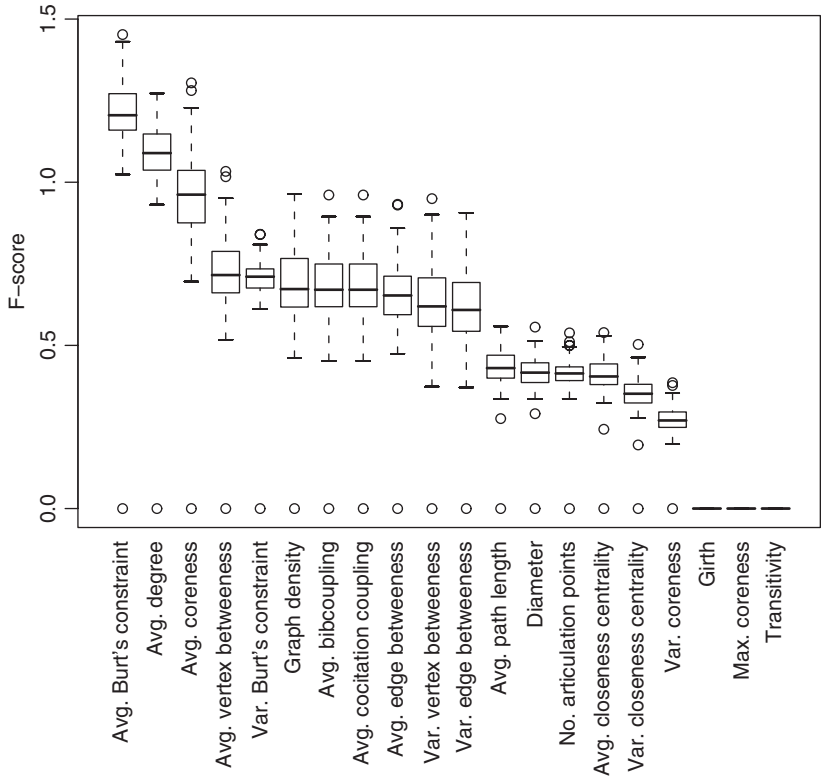

Figure 4. Graph property discriminatory power for Rfam family assignment. The discriminatory power as measured by the *F*-score (see Materials and Methods section) of each graph property was calculated to identify the important graph properties. When discriminating among the 18 Rfam families used in this analysis, the most important properties are the 'average Burt's constraint', 'average degree' and 'average coreness'.

performed better when applied to very sequence-similar sequences than our graph-property-based SVM method, they performed worse at greater sequence divergence with the SVM method displaying greater robustness with regard to increasing sequence separation.

In the cases, where the training and testing sets have at most 20% similarity, the *F*-scores of the graph properties signifying their predictive power, fall into four broad categories (Figure 4). The 'average Burt's constraint', 'average degree' and 'average coreness' have the highest *F*-scores; i.e. have greatest predictive power, while 'girth', 'maximum coreness' and 'transitivity' do not contribute at all to the SVM. The remaining graph properties fall into two, roughly equal groups that have average *F*-scores around 0.6 and 0.4. Thus, the important graph properties which determine functional versus non-functional RNA sequences (Figure 2) and those that determine the Rfam family (Figure 4) differ slightly. While the 'average Burt's constraint', 'average degree' and 'average coreness' remain among the most important, the 'number of articulation points' and the 'variance of coreness', which were important for functional versus non-functional classification, are ranked among the least important for assigning Rfam families.

The SVM method does not work evenly across all Rfam families (Table 2). When the sets are maximally divergent, the families SECIS (0.96 sensitivity, 0.58 specificity), Intron gp II (0.72, 0.73), 5S rRNA (0.63, 0.70), tRNA (0.42, 0.83) and MIRNA (0.79, 0.46) all perform well with high specificity, high sensitivity or both. IRES, LEADER and SRP are associated with the worst sensitivity and perform only slightly better than random assignment of Rfam families; 0.07 versus 0.05 for random predictions.

As the graph-property-based prediction approach may capture relevant aspects of RNA molecules that are not properly reflected by sequence similarity searches alone (as demonstrated by the more robust behaviour of our graph-based method when tested on diverging sequence sets; Figure 3), combining both methods may result in increased prediction performance compared to each individual approach. By combining a *P*-value calculated from the WUBLAST E-value to capture a sequences based score, and the SVM *P*-value to reflect graph-properties in a linear fashion with a properly chosen weighting factor, α, with $\alpha = 0.5$, MCC values higher than those produced by each method individually (Figure 5) were obtained. The average MCC for the combined methods is 0.446 and ranges from 0.313 in sets that are 20% similar to 0.567 in sets that are 90% similar.

Although we applied rigorous calibration to the sequences to identify whether the graph properties themselves were responsible for prediction or the influences from sequence similarities and length, it would be imprudent not to use this information when constructing an SVM intended for actual classification of RNA sequences outside the testing protocol. Thus, for comparison with other methods, the third approach used non-calibrated

Table 2. Confusion matrix for most divergent calibrated sets

True family \ Predicted family	CD-BOX	HACA-BOX	IRES	LEADER	MIRNA	5S rRNA	5.8S rRNA	tRNA	6S RNA	SRP	tmRNA	Intron gp I	Intron gp II	SECIS	SSU rRNA 5	T-box	RIBOSWITCH	RIBOZYME	Q^D
CD-BOX	242	149	80	416	100	49	52	5	21	161	74	40	14	0	23	209	298	47	0.12
HACA-BOX	114	886	47	247	16	65	125	4	73	55	14	80	17	13	3	88	115	18	0.45
IRES	76	41	145	86	0	29	130	1	83	57	173	185	0	0	159	336	150	329	0.07
LEADER	264	222	10	132	0	219	911	0	0	0	0	32	15	22	0	133	20	0	0.07
MIRNA	117	16	51	0	1568	24	35	38	2	8	2	2	67	12	4	4	28	2	0.79
5S rRNA	481	25	1	10	8	1247	78	13	1	0	0	0	1	8	0	26	73	8	0.63
5.8S rRNA	11	6	142	14	4	8	247	0	16	11	40	539	1	0	82	234	115	510	0.12
tRNA	28	0	0	0	1	1	11	831	0	0	0	0	354	754	0	0	0	0	0.42
6S RNA	106	31	490	1	6	2	5	0	768	57	20	52	1	0	50	336	12	43	0.39
SRP	21	1	1	51	1734	0	8	0	4	141	0	0	6	11	0	0	2	0	0.07
tmRNA	10	9	128	15	0	2	23	0	163	41	890	251	0	0	83	53	9	303	0.45
Intron gp I	42	41	174	46	0	12	54	2	18	196	307	384	0	0	125	188	105	286	0.19
Intron gp II	0	0	0	0	0	0	0	52	0	0	0	0	1431	497	0	0	0	0	0.72
SECIS	2	0	0	0	9	0	0	30	0	0	0	0	43	1896	0	0	0	0	0.96
SSU rRNA 5	28	64	94	32	0	2	204	1	50	53	456	308	0	0	227	60	39	362	0.11
T-box	29	97	247	36	0	36	70	0	49	94	205	437	4	0	143	290	43	200	0.15
RIBOZYME	104	60	135	17	0	83	168	22	73	15	124	319	0	33	124	186	206	311	0.10
RIBOSWITCH	48	47	102	19	0	0	86	4	151	92	439	234	3	0	120	176	96	363	0.18
Q^M	0.14	0.52	0.08	0.12	0.46	0.70	0.11	0.83	0.52	0.14	0.32	0.13	0.73	0.58	0.20	0.13	0.16	0.13	0.33

For each Rfam family, the classification of each RNA sequence from their actual class (y-axis) into their predicted classes (x-axis) is shown. Q^M is the specificity of the method and Q^D the sensitivity (see Materials and Methods section for definition). This confusion matrix was calculated from training and testing sets that were at most 20% similar and used calibrated graph properties. We observed that the chosen graph properties calculated for the chosen graph representation are best able to reflect the Rfam families SECIS, Intron gp II, 5S rRNA, tRNA, MIRNA and HACA-BOX. Entries are coloured by number classified starting from 0 (white) to 2000 (green).

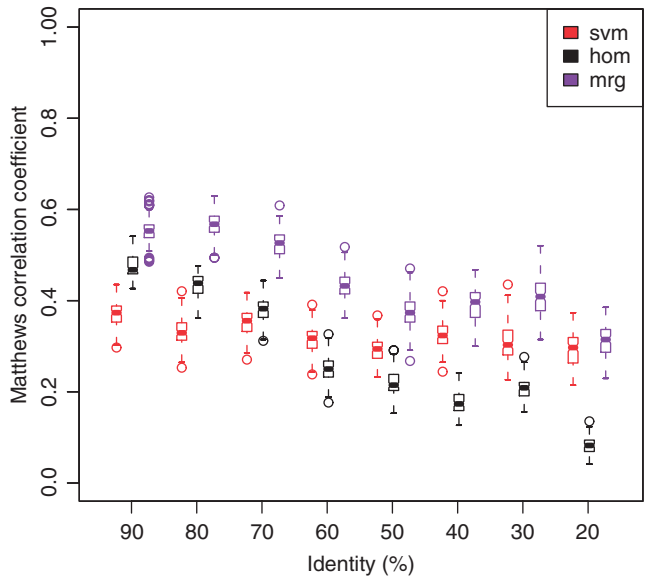

Figure 5. Linear combination of graph-property-based SVM and BLAST for calibrated, diverging RNA sequence sets. By merging the results of the WUBLAST (hom) and SVM (svm) methods, improved classification at all thresholds of divergence (mrg) were obtained. This indicates that both methods capture independent information that allows more accurate classification when combined. As in Figure 3, 18 Rfam families were considered.

graph properties and imposed no similarity restrictions between the training and testing sets.

Having established that graph properties have predictive value to correctly distinguish between Rfam families, we took a third approach that compared the performance of the graph-property-based SVM method to BLAST, INFERNAL and HMMER using training sets with 50, 100 and 200 sequences per Rfam family (Supplementary Material, Section 3). In the previous two approaches reported above, we used the full sets of Rfam family alignments. While this provided the largest possible as well as most diverse training sets, these sets can also be expected to include wrongly annotated RNA sequences. Therefore, for a fair and rigorous comparison to other methods, in the third approach we only included the seed-alignments associated with each Rfam family that can be assumed to constitute curated and more accurate datasets. The size of the training set and the number of families to be classified has a significant impact on the performance of the tested method (Figure 6). The SVM-based method shows performance increase from a median MCC of 0.88 for training sets of size 50 to 0.96 for training sets of size 100 and 0.98 for training sets of size 200. INFERNAL also show an increase (0.96, 0.99, 0.99), whereas WUBLAST remains stable (0.97, 0.95, 0.96) and HMMER shows a decrease (0.97, 0.94; 0.88). The results indicate that SVMs trained on graph properties are able to perform slightly better than homology-based methods and slightly worse than

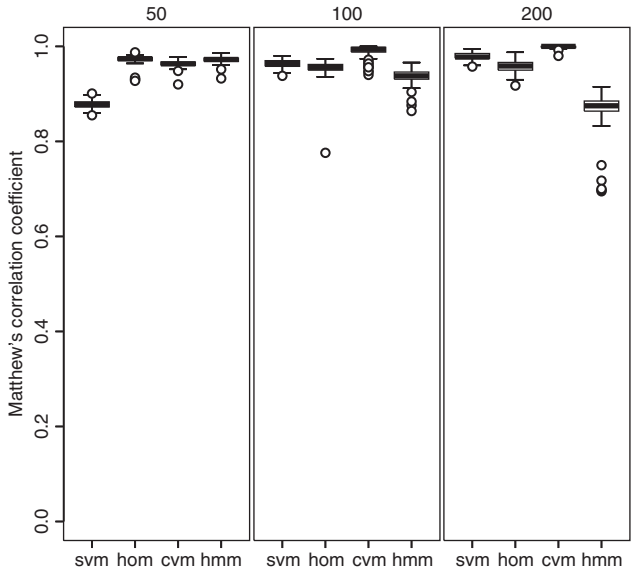

Figure 6. Comparison of graph-property-based SVMs to other methods. Four different classes of methods (50) were compared; SVM- (svm), homology- (hom, using WUBLAST), covariance model- (cvm, using INFERNAL), and Hidden Markov Model-based (hmm, HMMER). Performances vary depending upon the size of the training set. The SVM-based method performs with a median MCC of 0.88 when trained with sets of 50 sequences. The median MCC increases considerably for training sets with 100 and 200 sequences to 0.96 and 0.98, respectively. With larger training sets, SVMs compare favourably with the other methods such as homology methods, which had median MCC values of 0.97, 0.95 and 0.96, covariance models (0.96, 0.99, 0.99), and Hidden Markov Models (0.97, 0.94, 0.88) for the training sets of increasing sizes. Training sets of different sizes corresponded to sets of 50, 100 or 200 RNA sequences derived from 28, 8 and 3 Rfam families, respectively, with at least that many corresponding member sequences. Sequences were randomly sampled and the procedure was repeated 100 times.

covariance model-based methods when trained on sufficiently large datasets. The decreasing performance of Hidden Markov Models with increasing training set size may be explained by the impact that the increased variability may have on the transitional probabilities or decreasing quality of multiple sequence alignments used to derive the HMM. This illustrates the need for methods to detect different aspects of conservation than at the sequence level alone as attempted in this study.

DISCUSSION

Currently, there exist few predictors capable of assigning function to uncharacterised ncRNA molecules and even fewer that can predict whether or not an ncRNA molecule is functional. Available methods are based on sequence comparison (52), covariance models (53), graph topology (41) and structural alignments (30). However, the disadvantages of existing methods limit their application. Here, we present a novel method for *de novo* ncRNA and Rfam family prediction, which is based on a higher level of structural abstraction by using properties associated with RNA molecules when treating them as graphs, thereby addressing some of the problems of the existing methods, expanding the repertoire of available methods and, hopefully, contributing to the understanding of the RNA world.

By analysing the manner in which the graph properties are calculated, we may gain insight into how functional RNA is formed and what topological features render functional RNA molecules unique compared to their non-functional counterparts. Analysing the 'number of articulation points' may serve as an example as it is relatively easy to interpret. In graphs representing RNA secondary structure, there are only two situations, where removing a node in the graph can disconnect the graph: when a node is removed from the dangling 5′- or 3′-end, or when a node is removed from a bridge connecting potentially separable structures (Figure 7). From this study, we find that graphs with more articulation points are more likely to represent non-functional structures, indicating that functional structures minimize the length of the dangling ends and, in addition, either minimise the length of the bridges between separable structures or the number of separable structures. The remaining graph properties that were determined as important are calculated using more complex algorithms and a structural interpretation is more complex (54). Further biological interpretations of some of the other graph properties can be found in the Supplementary Material.

When using graph properties, it is important to remove factors that may obscure the attempt to determine whether they reflect sufficient information about Rfam family to make a prediction; otherwise, simpler methods such as sequence alignment or predictions based on just the sequence length would suffice. After removing these confounding factors, the graph properties themselves are shown to maintain the predictive power. As a result, we gain insight into structurally important properties for functional RNA by interpreting the way graph properties are calculated in a biological context (demonstrated here with the number of articulation points). The graph properties also provide sufficient information for Rfam family prediction on highly divergent sequences. Combined with the aforementioned sequence properties, the accuracy of the method improves significantly.

ncRNA exhibits greater conservation on the secondary structure level than the primary structure (53), as is demonstrated by the large variety of tRNA molecules, and thus sequence similarity is a potentially suboptimal choice of a classification criterion. In many cases, the sequences are sufficiently dissimilar at the sequence- and even the secondary-structure (inferred from covariation) level that neither sequence alignment nor covariance models are able to recall the function, and yet a classifier built on the graph properties; i.e. based on a higher level of abstraction, still manages to perform accurately. This finding indicates that, to a certain degree, our method is sequence independent and that there are properties inherent in the structure of ncRNA indicative of function.

Of the many graph representations available, such as dual graph and planar tree representation, we chose bracketed graph representation. This representation was chosen over the others as it is more sensitive to small changes in the underlying RNA structure due to the

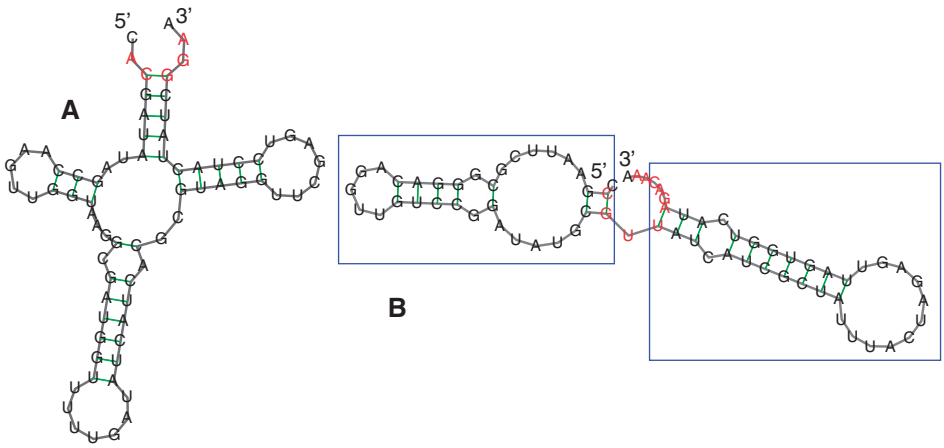

Figure 7. Structural relevance of articulation points. Articulation points are nodes that, if removed, will disconnect the graph. In a structural sense, these are either 'dangling' nucleotides at the 5′- and 3′-ends of a molecule or 'bridge' nucleotides that connect potentially separable structures. A tRNA molecule (**A**) and a shuffled counterpart (**B**) are shown. The red nucleotides are articulation points in the graph representation and the structures within each blue box are potentially separable secondary structures. Ester bonds are shown in grey and hydrogen bonds are coloured in green. If graphs with fewer articulation points tend to represent functional ncRNA, then functional RNA molecules are more likely to have fewer dangling/bridge nodes and potentially separable structures than non-function RNA molecules. The tRNA graph contains five articulation points and not separable structures, whereas the shuffled counterpart contains 10 articulation points and two separable structures.

greater number of nodes and edges. The graph space of the other representations is far smaller, potentially reducing the predictive power of the derived graph properties. This representation also minimizes information loss (e.g. dangling ends and, in dual graph representation, the length of the stems and loops) and is simpler to calculate. On the other hand, there are several potential disadvantages of our proposed method: (i) usage of sequences biased toward certain species, (ii) dependence on one folding algorithm of choice and (iii) usage of secondary structure prediction, thus neglecting pseudoknots and tertiary structures. Note that all of the identified issues are exogenous to our method, and one can account for them by conducting comparison tests—a task beyond the scope of this article.

The sequences available in Rfam are largely bacterial and viral, and thus the method we have developed will be biased towards the prediction of purely bacterial and viral Rfam families or, where the family occurs in several kingdoms, higher accuracy prediction in bacterial and viral sequences. When more ncRNA sequences become available, our method will benefit from being trained upon a more specific choice of sequences, e.g. purely plant or animal sequences.

Often the predicted structure of an ncRNA sequence is quite different from the experimentally determined structure. As we obtained all secondary structure assignments using RNAfold (19), our method is reliant on RNAfold producing consistent predictions. As the tools for RNA folding prediction improve, we plan to upgrade the folding algorithms, hopefully yielding higher accuracies and better understanding of the biological and structural relevance of graph properties. An immediate possible improvement is the use of multiple alignments which improves the accuracy of the predicted secondary structure (55).

We chose to calculate graph properties from secondary structure, rather than pseudoknots or predicted 3D RNA structure, which potentially limits the predictive power of the graph properties. By using secondary structure, the maximum degree (number of connected edges) for any node is limited to three; i.e. two ester bonds and a hydrogen bond. With pseudoknots and 3D structure, the possibility for much more complex graphs emerges with significant consequences on the graph properties. Such graphs are likely to better reflect functional information, which is a point for further study.

We expect further improvement through careful selection of the graph properties as potential discriminatory features. Of the current graph properties, we would ideally use only those that are most informative and perhaps more biologically relevant. There are also many more properties that can be calculated than the 20 chosen and experimentation with new graph properties may lead to improved accuracy and greater insight into ncRNA functionality. Including properties such as minimal free energy (which has been shown to be informative for miRNA), %G + C and perhaps dinucleotide frequencies in SVM training should provide a significant boost to the accuracy of the described method.

Many of the existing methods have shortcomings limiting their application. As many ncRNA families show little sequence homology, but high degree of structural conservation, homology-based methods would be unable to correctly identify all members of the family. Covariance models, although highly accurate, require long computation times and neither method are able to discriminate between functional and non-functional ncRNA sequences. The method developed in the current article, can cover more sequences than homology-based methods at quick speeds typical of SVM-based methods, which provides a good compromise between the two methods. Our method was shown to be robust with regard to increasing sequence divergence and performed at high accuracy levels when tested on both curated datasets (Rfam seed alignments)

and data sets based on electronic annotation (Rfam full alignments). It also exhibits the ability to identify functional RNA sequences. Finally, combining graph properties with other methods provides a significant boost to performance.

In conclusion, ncRNA is not simply a primitive form of molecule as it is active in a wide variety of roles not typical for proteins. We developed a computational method that represents a necessary first step for future ncRNA investigation tools. With a plethora of potential functions still undiscovered and many more molecules whose functional role is still unassigned, we believe that higher level structural abstraction and their respective properties will play a key role in discovering new ncRNAs and their plausible biological role.

Availability

The graph-property-based methods developed here has been made available as a web-based tool called the GRAph Property based Predictor and Likelihood Estimator (GraPPLE) at: http://grapple.mpimp-golm.mpg.de

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

BMBF-funded *Go*FORSYS project (to Z.N. and P.M.). Funding for open access charge: Max-Planck Society.

Conflict of interest statement. None declared.

REFERENCES

1. Meyers,B.C., Matzke,M. and Sundaresan,V. (2008) The RNA world is alive and well. *Trends Plant Sci*., **13**, 311–313.
2. Mattick,J.S. (2007) A new paradigm for developmental biology. *J. Exp. Biol*., **210**, 1526–1547.
3. Manak,J.R., Dike,S., Sementchenko,V., Kapranov,P., Biemar,F., Long,J., Cheng,J., Bell,I., Ghosh,S., Piccolboni,A. *et al*. (2006) Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nat. Genet*., **38**, 1151–1158.
4. Claverie,J.M. (2005) Fewer genes, more noncoding RNA. *Science*, **309**, 1529–1530.
5. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
6. Weinstock,G.M. (2007) ENCODE: more genomic empowerment. *Genome Res*., **17**, 667–668.
7. Mathews,D.H. and Turner,D.H. (2006) Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol*., **16**, 270–278.
8. Tinoco,I. Jr, Uhlenbeck,O.C. and Levine,M.D. (1971) Estimation of secondary structure in ribonucleic acids. *Nature*, **230**, 362–367.
9. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
10. Mironov,A.S., Gusarov,I., Rafikov,R., Lopez,L.E., Shatalin,K., Kreneva,R.A., Perumov,D.A. and Nudler,E. (2002) Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell*, **111**, 747–756.
11. Nahvi,A., Sudarsan,N., Ebert,M.S., Zou,X., Brown,K.L. and Breaker,R.R. (2002) Genetic control by a metabolite binding mRNA. *Chem. Biol*., **9**, 1043.
12. Schilling,O., Langbein,I., Muller,M., Schmalisch,M.H. and Stulke,J. (2004) A protein-dependent riboswitch controlling ptsGHI operon expression in Bacillus subtilis: RNA structure rather than sequence provides interaction specificity. *Nucleic Acids Res*., **32**, 2853–2864.
13. Winkler,W., Nahvi,A. and Breaker,R.R. (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, **419**, 952–956.
14. Winkler,W.C., Cohen-Chalamish,S. and Breaker,R.R. (2002) An mRNA structure that controls gene expression by binding FMN. *Proc. Natl Acad. Sci. USA*, **99**, 15908–15913.
15. Kurihara,Y., Matsui,A., Kawashima,M., Kaminuma,E., Ishida,J., Morosawa,T., Mochizuki,Y., Kobayashi,N., Toyoda,T., Shinozaki,K. *et al*. (2008) Identification of the candidate genes regulated by RNA-directed DNA methylation in Arabidopsis. *Biochem. Biophys. Res. Commun*., **376**, 553–557.
16. Brouns,S.J., Jore,M.M., Lundgren,M., Westra,E.R., Slijkhuis,R.J., Snijders,A.P., Dickman,M.J., Makarova,K.S., Koonin,E.V. and van der Oost,J. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, **321**, 960–964.
17. Nakashima,A., Takaku,H., Shibata,H.S., Negishi,Y., Takagi,M., Tamura,M. and Nashimoto,M. (2007) Gene silencing by the tRNA maturase tRNase ZL under the direction of small-guide RNA. *Gene Ther*., **14**, 78–85.
18. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*., **31**, 3406–3415.
19. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*., **31**, 3429–3431.
20. Dawson,W., Fujiwara,K., Kawai,G., Futamura,Y. and Yamamoto,K. (2006) A method for finding optimal rna secondary structures using a new entropy model (vsfold). *Nucleosides Nucleotides Nucleic Acids*, **25**, 171–189.
21. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol*., **2**, e33.
22. Ding,Y. and Lawrence,C.E. (1999) A bayesian statistical algorithm for RNA secondary structure prediction. *Comput. Chem*., **23**, 387–400.
23. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res*., **31**, 439–441.
24. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*., **33**, D121–D124.
25. Freyhult,E., Gardner,P.P. and Moulton,V. (2005) A comparison of RNA folding measures. *BMC Bioinformatics*, **6**, 241.
26. Lyngso,R.B. and Pedersen,C.N. (2000) RNA pseudoknot prediction in energy-based models. *J. Comput. Biol*., **7**, 409–427.
27. Das,R. and Baker,D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl Acad. Sci. USA*, **104**, 14664–14669.
28. Rivas,E. and Eddy,S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.
29. Bonnet,E., Wuyts,J., Rouze,P. and Van de Peer,Y. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.
30. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.
31. Steigele,S., Huber,W., Stocsits,C., Stadler,P.F. and Nieselt,K. (2007) Comparative analysis of structured RNAs in S. cerevisiae indicates a multitude of different functions. *BMC Biol*., **5**, 25.
32. Mourier,T., Carret,C., Kyes,S., Christodoulou,Z., Gardner,P.P., Jeffares,D.C., Pinches,R., Barrell,B., Berriman,M., Griffiths-Jones,S. *et al*. (2008) Genome-wide discovery and verification of novel

structured RNAs in Plasmodium falciparum. *Genome Res*., **18**, 281–292.
33. Cao,S. and Chen,S.J. (2006) Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res*., **34**, 2634–2652.
34. Lim,L.P., Lau,N.C., Weinstein,E.G., Abdelhakim,A., Yekta,S., Rhoades,M.W., Burge,C.B. and Bartel,D.P. (2003) The microRNAs of Caenorhabditis elegans. *Genes Dev*., **17**, 991–1008.
35. Myslyuk,I., Doniger,T., Horesh,Y., Hury,A., Hoffer,R., Ziporen,Y., Michaeli,S. and Unger,R. (2008) Psiscan: a computational approach to identify H/ACA-like and AGA-like non-coding RNA in trypanosomatid genomes. *BMC Bioinformatics*, **9**, 471.
36. Zhang,Y. (2005) miRU: an automated plant miRNA target prediction server. *Nucleic Acids Res*., **33**, W701–W704.
37. Eddy,S.R. (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**, 18.
38. Fera,D., Kim,N., Shiffeldrim,N., Zorn,J., Laserson,U., Gan,H.H. and Schlick,T. (2004) RAG: RNA-As-Graphs web resource. *BMC Bioinformatics*, **5**, 88.
39. Janssen,S., Reeder,J. and Giegerich,R. (2008) Shape based indexing for faster search of RNA family databases. *BMC Bioinformatics*, **9**, 131.
40. Kim,N., Shiffeldrim,N., Gan,H.H. and Schlick,T. (2004) Candidates for novel RNA topologies. *J. Mol. Biol*., **341**, 1129–1144.
41. Karklin,Y., Meraz,R.F. and Holbrook,S.R. (2005) Classification of non-coding RNA using graph representations of secondary structure. *Pac. Symp. Biocomput*., **10**, 4–15.
42. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
43. Csárdi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *Inter J. Complex Sys*., **1695**.
44. Chang, C.C. and Lin, C.J. (2001) LIBSVM: a library for support vector machines. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/
45. Chen,Y.W. and Lin,C.J. (2006) In Guyon,I., Gunn,S., Nikravesh,M. and Zadeh,L.A. (eds), *Feature Extraction: Foundations and Applications*. Vol. 1, Springer, New York.
46. Jiang,M., Anderson,J., Gillespie,J. and Mayne,M. (2008) uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, **9**, 192.
47. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al*. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
48. Gish, W. (1996–2003). Available at: http://blast.wustl.edu
49. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
50. Freyhult,E.K., Bollback,J.P. and Gardner,P.P. (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res*., **17**, 117–125.
51. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.
52. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
53. Yao,Z., Weinberg,Z. and Ruzzo,W.L. (2006) CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
54. Gross, J.L. and Yellen, J. (2004) *Handbook of Graph Theory*. CRC Press, Boca Raton FL, USA.
55. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol*., **319**, 1059–1066.