2015

**A Novel Structural Measure Separating Non-Coding RNAs from Genomic Backgrounds**

Yingfeng Wang  
*the School of Information Technology, Middle Georgia State University, Macon, GA 31206, USA.*

Russell L. Malmberg  
*Department of Plant Biology and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA.*

Liming Cai  
*Department of Computer Science and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA.*

---

Follow this and additional works at: https://tsinghuauniversitypress.researchcommons.org/tsinghua-science-and-technology

Part of the Computer Sciences Commons, and the Electrical and Computer Engineering Commons

---

**Recommended Citation**


This Research Article is brought to you for free and open access by Tsinghua University Press: Journals Publishing. It has been accepted for inclusion in Tsinghua Science and Technology by an authorized editor of Tsinghua University Press: Journals Publishing.
A Novel Structural Measure Separating Non-Coding RNAs from Genomic Backgrounds

Yingfeng Wang*, Russell L. Malmberg, and Liming Cai*

Abstract: RNA secondary structure has become the most exploitable feature for ab initio detection of non-coding RNA (ncRNA) genes from genome sequences. Previous work has used Minimum Free Energy (MFE) based methods developed to identify ncRNAs by measuring sequence fold stability and certainty. However, these methods yielded variable performances across different ncRNA species. Designing novel reliable structural measures will help to develop effective ncRNA gene finding tools. This paper introduces a new RNA structural measure based on a novel RNA secondary structure ensemble constrained by characteristics of native RNA tertiary structures. The new method makes it possible to achieve a performance leap from the previous structure-based methods. Test results on standard ncRNA datasets (benchmarks) demonstrate that this method can effectively separate most ncRNAs families from genome backgrounds.

Key words: RNA secondary structure; RNA tertiary structure; Inside algorithm

1 Introduction

RNA molecules that do not yield proteins are termed non-coding RNAs (ncRNAs). In the past three decades, ncRNAs have been discovered to be involved in many important biological processes such as gene regulation, catalysis, and RNA splicing[1–3]. The recognition of the importance of ncRNAs has motivated much research towards the computational identification of ncRNA genes in genomes[4–6]. However, unlike protein-coding genes, ncRNA genes have been elusive; they have yet to demonstrate any statistically strong, universal sequential signals for detection. ncRNA gene finding remains one of the most challenging tasks in bioinformatics research. It would be desirable to have effective computational methods that can help narrow down the number of ncRNA candidates to be validated by much more time consuming and costly experimental determination means. Such computational methods, especially ab initio ones that use only genome sequence information, would need the capability to effectively distinguish ncRNA sequences from genomic backgrounds. RNA secondary structure has been the most exploited feature in ncRNA gene finding[4, 7, 8]. In particular, an ncRNA sequence is expected to have a thermodynamically more stable secondary structure than one predicted from a non-structural sequence; this has energized some leading groups to develop structure-based ncRNA gene finding tools[9–12]. Most of these tools compute the minimum free energy, using the thermodynamic energy model[13–16], as the fold stability of a given sequence; however, they rely on multiple genome sequences to incorporate additional information into their models. For example, RNAz considers sequence stability and secondary structure conservation[10, 11]. Evofold uses not only a sequence alignment, but also phylogenetic information.
incorporated into a Stochastic Context-Free Grammar (SCFG)\cite{12}. However, the fold stability measure may underperform relative to expectations because of the small difference in free energy between native ncRNAs and random sequences folded by chance\cite{6, 17–20}. Fold certainty, the Shannon entropy over alternative secondary structures defined with the Boltzmann ensemble\cite{21}, has also been used to characterize ncRNAs. The fold certainty is often approximated with the entropy defined over alternative base pairs\cite{22, 23}. The certainty of base pairs is expected to be low for native ncRNA sequences. The fold certainty measure has shown a strong correlation with the fold stability measure\cite{3, 24}; both gave diverse performances across different ncRNA data sets\cite{19, 24}. For example, they perform very well on miRNA precursors but poorly on tRNAs tested against randomly shuffled sequences.

The mixed success of the traditional structure-based measures suggests that the RNA secondary structure space defined with the energy model may not have captured well some features of ncRNA sequences. Because ncRNAs functions may be determined by their tertiary structure\cite{25, 26}, incorporating tertiary structure characteristics of ncRNAs into structural measures may improve performance in ncRNA detection. Indeed, earlier work on RNA secondary structure prediction showed improved results when coaxial stacking of helices was incorporated\cite{27}. Other tertiary motifs, for instance tetra-loops, have been considered in some of secondary structure prediction programs\cite{16, 28, 29}. To apply the idea to ncRNA detection, TRIPLE, a program developed in our previous research\cite{30}, assumed stems in RNA secondary structure must contain at least three consecutive canonical base pairs, reflecting the energetic stability of helices in the tertiary structure. With the Shannon base pair entropy measure, TRIPLE significantly improved energy-based programs (e.g., NUPACK) in their ability to distinguish all 13 native ncRNAs\cite{24} from randomly shuffled sequences. Although TRIPLE was not able to distinguish native ncRNAs from genomic backgrounds, it demonstrates an improved potential to effectively detect ncRNAs with secondary structure models which are constrained with tertiary elements.

The current paper presents a novel structure-based measure for ncRNAs, which has achieved the following performances. (1) It can effectively distinguish native structural ncRNAs from genomic sequences; (2) it has nearly the same performance across all the 13 ncRNAs datasets of Freyhult et al.\cite{24} and the 51 ncRNA benchmarks from Rfam selected by Nawrocki et al.\cite{31}. The new structural measure is based on a novel constrained RNA secondary structure ensemble and to compute for any given RNA sequence the overall capability to fold into native-like secondary structures. In a number of aspects, our method differs from the previous ones\cite{12, 32, 33} that underly the thermodynamic energy model. First, building blocks of the new secondary structure model are $k$-way junctions ($k \geq 1$), reflecting the characteristics of RNA tertiary structure. Second, lengths of unpaired loops are junction structure-specific and constrained with information obtained from the native ncRNA tertiary database\cite{34}. Third, the new structure is modeled with a weighted Context-Free Grammar (CFG) with production rules to weight equally all alternative structures. These novelities make it possible for our method to effectively distinguish native ncRNAs from genomic backgrounds. In particular, for almost every tested ncRNA dataset (of the 13 families\cite{24} and the 51 families\cite{31}), our method can detect more than 75% of the ncRNA sequences when they are compared to the Pyrococcus furiosus genomic background with about 80% specificity.

## 2 Method

Our new RNA structure ensemble was established as a constrained space of secondary structures with RNA junctions as atomic, building blocks. RNA junctions are secondary structure elements formed when one or more helices come together and an RNA secondary structure can be defined to consist of interconnected junctions. A junction can be a $k$-way junction, for some $k \geq 1$, which is a loop enclosed by $k$ helices. For example, a one-way junction, two-way junction, and $k$-way junction ($k \geq 3$) correspond to a stem-loop, internal-loop, and multi-loop in a secondary structure. A $k$-way junction is composed of a leading helix, $k$ unpaired single strands, and $k–1$ entries for other junctions. Thus, our secondary structure space was defined to contain various loops of $k$-way junctions, with the intention to capture the characteristics of the junction architecture derived from experimentally determined RNA secondary structures\cite{35, 36}. Figure 1 shows the tertiary structure of P4-P6 domain of Group
I intron (a) and its corresponding secondary structure (b) which consists of a number of two-way junctions (internal loops) and one-way junctions (stem-loops) and one three-way junction, being connected in nested and parallel fashions.

To accurately account for $k$-way junctions in the new secondary structure ensemble, known native tertiary structures were investigated to retrieve the information of unpaired single strands (called loops) involved in such junctions. We were interested in prevalent constraints between lengths of involved loops in each type of junction. We obtained the frequencies of junctions with given loop length combinations by submitting queries to the available native ncRNA database RNA FRABASE 2.0. Based on the distributions of loop lengths, the constraints of loop lengths of $k$-way junctions ($1 \leq k \leq 5$) were set separately so as to cover at least 90% cases on each category. For example, our investigation shows that the loops of one-way junctions usually have 3 to 15 unpaired nucleotides, while most four-way and five-way junctions have at least two loops with up to 2 unpaired nucleotides, and other loops contain up to 7 nucleotides. The constraints we observed on loop lengths of two-way and three-way junctions are given in Tables 1 and 2, respectively. Additionally, some constraints of helices were also included based on an energy model to guarantee all detected helices are thermodynamically stable.

High-order junctions ($k \geq 6$) are not included in the current model definition because they occur rarely. In particular, more than 90% of the junctions available in public databases, i.e., PDB, are low-order $k$-way junctions ($1 \leq k \leq 5$). In addition, too little higher-order junction information is available in RNA FRABASE 2.0 to draw statistical conclusions. However, higher-order junctions may be approximated within the current framework (see the section of discussion).

### Table 1 Constraints of loop lengths (number of unpaired nucleotides) of two-way junctions.

<table>
<thead>
<tr>
<th>5' loop length</th>
<th>3' loop length</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1–4</td>
</tr>
<tr>
<td>1</td>
<td>0–3</td>
</tr>
<tr>
<td>2</td>
<td>0–4</td>
</tr>
<tr>
<td>3</td>
<td>0–7</td>
</tr>
<tr>
<td>4</td>
<td>1–7</td>
</tr>
<tr>
<td>5</td>
<td>1–7</td>
</tr>
<tr>
<td>6</td>
<td>3–5</td>
</tr>
<tr>
<td>7</td>
<td>2–7</td>
</tr>
</tbody>
</table>

### Table 2 Constraints of loop lengths (number of unpaired nucleotides) of three-way junctions.

<table>
<thead>
<tr>
<th>5' loop length</th>
<th>Middle loop length</th>
<th>3' loop length</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2–5</td>
<td>2, 3</td>
</tr>
<tr>
<td>1</td>
<td>3–9</td>
<td>2–8</td>
</tr>
<tr>
<td>2</td>
<td>1–5, 15</td>
<td>1–6</td>
</tr>
<tr>
<td>3</td>
<td>0–6</td>
<td>3–7</td>
</tr>
<tr>
<td>4</td>
<td>0–3</td>
<td>1–3</td>
</tr>
<tr>
<td>5</td>
<td>0–2</td>
<td>2, 3</td>
</tr>
<tr>
<td>6</td>
<td>3, 4</td>
<td>6, 7</td>
</tr>
<tr>
<td>7</td>
<td>1, 4, 10, 11</td>
<td>3, 4</td>
</tr>
<tr>
<td>7</td>
<td>3, 4</td>
<td>2, 3</td>
</tr>
</tbody>
</table>

3 Model and Algorithm

The new secondary structure ensemble was modeled with a weighted context-free grammar. Such a mathematical system has been widely used for RNA secondary structure modeling. Roughly speaking, a CFG is a rewriting system with rewriting (production) rules that defines a language of sentences. For an RNA “language”, the sentences are RNA sequences. Such production rules (of the format $X \Rightarrow Y_1Y_2\cdots Y_m$, for some $m \geq 0$) are used to replace an occurrence of variable $X$ with symbols $Y_1Y_2\cdots Y_m$. Where some
of the \( Y_i \) may be variables while others are terminal tokens (e. g., nucleotides for RNA sequences). A specific way to parse an RNA sequence with such a set of grammar rules yields a parsing tree interpreted as the corresponding secondary structure for the sequence. Different parses of the same sequences would give its alternative secondary structures. With a probability distribution assigned to the grammar rules, a CFG becomes stochastic and actually models an RNA secondary structure ensemble in which sequences and their structures are all associated with probabilities[40]. Stochastic CFGs have been widely used in profiling specific RNA secondary structures for ncRNA gene annotation[31, 41] and in RNA secondary structure prediction[42]. More recent work shows that a stochastic CFG is capable of modeling the thermodynamic energy based Boltzmann ensemble of RNA secondary structure[33].

For our purpose, we introduced a CFG consisting of production rules to recursively describe all possible junction-based RNA secondary structures. In particular, rules for producing (i. e., parsing) a new junction have the signature form \( T \Rightarrow LJO \) that defines a \( k \)-way junction following the enclosing base pair of the leading helix, where nonterminals \( L \), \( J \), and \( O \) define the first 5' loop, the second helix (which is the leading helix of another junction), and the other component to the right of the helix. For example, the three-way junction in Fig. 1 would be produced by the CFG with the following rule applications: \( T \Rightarrow LJO \Rightarrow LJL'O \Rightarrow LJL'J'L'' \), where \( J \) and \( J' \) are the P5c and P5b helices (1-way junctions) with three unpaired loops \( L \), \( L' \), and \( L'' \) between them and the leading helix, while \( O \) and \( O' \), components to the right of the corresponding helices, are replaced by helices and loops in the derivation. A pseudo-knot can be recognized as two or more alternative substructures separately.

Our CFG is weighted, where weights of the rules in the CFG are such that the alternative secondary structures produced by the rules have an equal weight. Therefore, the weights do not necessarily abide by probability distributions. In addition, since we use a finite number of rules to produce infinite number of structures, the weight equality can only be approximate. In particular, all unpaired nucleotides are treated equally. Our model assigned weights 0.25, 0.25, 0.17, 0.17, 0.08, and 0.08 to canonical base pairs C-G, G-C, A-U, U-A, G-U, and U-G, respectively. All other non-canonical base pairs were assigned zero weight.

With the above setting, the weights of production rules were set following the principle that the weight gain of producing two unpaired nucleotides should be equal to that of generating one base pair, on average. This principle guaranteed our weight setting shows no preference between folding and non-folding substructures. In our current grammar, the weight gain of generating an unpaired nucleotide is 1. So the gain of two unpaired nucleotides is \( 1 \times 1 = 1 \). If there is a production rule producing a canonical base pair within C-G, G-C, A-U, U-A, G-U, and U-G, then the average weight gain of this base pair is \( w \times q_{bp} \), where \( q_{bp} \) is the geometric average of base pair weights, and can be calculated by the following expression.

\[
q_{bp} = \sqrt[6]{0.25 \times 0.25 \times 0.17 \times 0.17 \times 0.08 \times 0.08} \quad (1)
\]

According to our principle, we have

\[
w \times q_{bp} = 1 \quad (2)
\]

So the weight \( w \) of this rule is 6.65. Weights of other production rules generating base pairs were calculated similarly.

Under this structure model, we are able to compute the summation of weights of all alternative structures fitting our structure space by adopting the inside algorithm, which was originally developed for SCFG[40]. For a given nonterminal \( S \), the inside probability-like weight is defined as

\[
\alpha(S, i, j, x) = \text{Weight}(S \Rightarrow^* x_i x_{i+1} \cdots x_j) \quad (3)
\]

i. e., the summation of weights of alternative substructures specified by \( S \) for the sequence segment \( x_i x_{i+1} \cdots x_j \). So, \( \alpha(S_0, 1, n, x) \) was the summation of weights of all sequence \( x \)'s alternative structures under the model, where \( S_0 \) was the initial nonterminal. The inside algorithm can efficiently compute \( \alpha(S_0, 1, n, x) \) based upon dynamic programming with time complexity \( O(mn^2) \) for a grammar containing \( m \) nonterminals and rules and a sequence having \( n \) nucleotides.

We hypothesized that the inside probability-like weight of real ncRNAs would be significantly higher than that of genomic backgrounds, since real ncRNAs are assumed to have more alternative structures fitting our structure space than does the background of genomic sequence.

4 Results

A program named JUNCTION was implemented for the inside probability-like weight calculation. This program was tested on ncRNA datasets, and its
performance was compared with that of the Minimum Free Energy (MFE) calculated by RNAfold 1.8.4\[16, 28\], which is a state-of-the-art model.

Our basic test was to obtain known ncRNAs, compare their scores with genomic sequences of the same length, then determine the detection accuracy.

4.1 Data preparation

We chose standard ncRNA sequence families which have been used by others for similar tests. All 13 ncRNAs datasets of Freyhult et al.\[24\] and 51 ncRNA benchmarks from Rfam selected by Nawrocki et al.\[31\] were downloaded. About 2.61% sequences having nucleotides other than A, C, G, U, and T were removed. For each of the Freyhult et al.’s 13 ncRNA datasets, we randomly picked 50 sequences, if there were more than 50 sequences. For each of Nawrocki et al.’s 51 ncRNA datasets, we randomly picked 100 sequences, if there were more than 100 sequences.

We prepared genomic background sequences using a Pyrococcus furiosus sequence (GenBank Accession: AE009950.1, GI: 18980902) from which we removed the annotated genes. For each native ncRNA sequence, 100 genomic background segments with the same length were randomly obtained from the Pyrococcus furiosus genome. All genomic background segments were further queried against the sequences in Rfam using the default settings\[43, 44\] to ensure they had no match.

4.2 Interquartile range based score

The tests generated 101 values for each ncRNA sequence being examined, in which one value was from the ncRNA sequence itself, while the other 100 values were based on the corresponding genomic backgrounds. We needed a measurement to evaluate the performances of the two methods in distinguishing real ncRNAs from its genomic backgrounds. Therefore, we sorted all MFEs of all 100 genomic backgrounds in ascending order, while the MFEs of real ncRNAs are expected to have low ranks. Let $Q_1$ be the value of the first quartile (above 25% values) and $Q_3$ be the value of the third quartile (above 75% values). The IQR is then $Q_3 - Q_1$. We can compute the IQR-based score for JUNCTION by

$$\text{Score} = \frac{\text{real} - Q_3}{Q_3 - Q_1}$$

and the IQR-based score for MFE by

$$\text{Score} = \frac{Q_1 - \text{real}}{Q_3 - Q_1}$$

where “real” is either the inside probability-like weight or the MFE. If IQR is zero, the “Score” is set to be 10 000, when real $> Q_3$ for the inside probability-like weight, or real $< Q_1$ for the MFE. If IQR is zero, the “Score” is set to be –10 000, when real $\leq Q_3$ for the inside probability-like weight, or real $\geq Q_1$ for the MFE. High IQR-based scores on real ncRNAs mean high sensitivities of distinguishing ncRNAs, while low IQR-based scores on genomic backgrounds indicates high specificities.

4.3 Comparison

Both the Freyhult et al.’s 13 ncRNA families and Nawrocki et al.’s Rfam benchmarks (51 ncRNA datasets)\[24, 31\] were tested with JUNCTION and RNAfold against Pyrococcus furiosus genome sequences. To thoroughly compare performance of these methods, we first examine IQR score distributions for tested sequences. Since the IQR score may be across a large range of values, we plotted histograms of $\ln(\text{Score} + c)$ of real ncRNAs and background sequences, where $c$ is a constant to ensure that logarithm was taken on a value no less than 1. Figures 2 and 3 respectively show comparisons between JUNCTION and RNAfold in their IQR scores for miRNAs (RF00104) and for tRNAs (RF00005). In early studies of Rivas and Eddy\[18\] and Clote et al.\[46\], the distributions of Z-scores of MFE were close to normal distribution. IQR score distributions appear different characters. The distributions show a significant improvement of performance on tRNAs by JUNCTION over RNAfold while both programs have similar abilities to distinguish miRNAs from genomic backgrounds. IQR score distributions of all tested ncRNAs are given in the supplementary material.

Statistically, an observation with an IQR-based score no less than 1.5 is distant from other data\[47\]. If
we set the IQR-based score threshold to be 1.5, sequences with an IQR-based score no less than 1.5 will be predicted as ncRNAs, while sequences of IQR-based scores below this threshold will be identified backgrounds. Sensitivities of RNAfold and JUNCTION on Nawrocki et al.’s datasets with IQR-based score threshold $\geq 1.5$ are given in Fig. 4, giving the evidence that the overall sensitivity of JUNCTION is significantly higher than that of RNAfold. Figure 5 shows Receiver Operating Characteristic (ROC) curves of average performances of JUNCTION and RNAfold on datasets of Frehult et al. and Nawrocki et al. with IQR-based score no less than 1.5. The complete test results (sensitivities and specificities on various IQR score thresholds) on both data collections are given in the supplementary material. We observe that program JUNCTION achieves consistently high sensitivities across almost all ncRNAs families, while RNAfold seems to maintain a higher average specificity.

5 Discussion

This study introduced a new ensemble of RNA secondary structures upon which a novel structural measure was defined. With this measure, native structural ncRNA sequences can be effectively separated from genomic backgrounds. A good performance has been shown consistently across two large, standard ncRNA datasets. The comparison between the implemented program JUNCTION of ours
and RNAfold shows evidence of a performance leap by our method over traditional energy-based methods.

We note that there are a few exceptional ncRNA cases on which JUNCTION does not achieve high sensitivities. One such case is Hammerhead ribozyme type I (Hh1) in the Freyhult et al.’s benchmarks[24]. Some sequences in this Hh1 dataset are incomplete, typically lacking one hairpin loop, which might have confused our model. Likewise, our method is not effective in small RNAs without significant secondary structures. For instance Iron response elements (RF00037 in the Rfam benchmarks) are unlikely to have significant structures or junctions. Another example is C/D box snoRNAs that contain two short conserved sequence motifs, whose secondary structure consists of a “very weak” stem at the base that does not seem energetically stable. We believe small RNAs with significant conserved sequence motifs can be better recognized by sequence profile based search tools (e.g., Refs. [48, 49]) used in conjunction with our ab initio method.

We also note that the false positive rate of JUNCTION on many ncRNA datasets is higher than that of RNAfold. There may be two explanations for this. One is that our measure computes overall capability for the query sequence to fold into native-like secondary structures. A genome background sequence has a higher tendency to get a higher score under this measure, which may potentially be remedied with more accurate constraining conditions for junction structures. On the other hand, the higher false positive rate could have also been caused by the contribution from the genomic background of previously non-annotated genes. One Pyrococcus furiosus genomic background region with high inside probability-like weight was subsequently identified as a real ncRNA that had not been annotated before (Drs. Michael Terns and Rebecca Terns, personal communication).

To consider potential applications of JUNCTION in genome scanning to identify ncRNAs, one may question how GC contents of a genome and the scanning window size may have impacts on the inside probability like weight and IQR thresholds. Much like MFE being substantially affected by the length of

---

**Fig. 4** Sensitivity comparison between JUNCTION and RNAfold on Nawrocki et al.’s Rfam benchmarks[31] with IQR-based scores threshold $\geq 1.5$.

**Fig. 5** ROC curves of average performances of RNAfold and JUNCTION on Freyhult et al.’s datasets[24] (a) and Nawrocki et al.’s datasets[31] (b) with IQR-based scores no less than 1.5.
sequences, inside probability like weight can be window size-dependent. Interestingly, however, given a fixed window size, an uniform IQR score is achievable to guarantee a specificity value (e.g., above 90%) across different GC contents. The supplementary material shows statistics of inside probability like weights with four sets of random sequences of lengths 100, 150, 200, 250, respectively.

We attribute the performance boost by our model to the constrained structure ensemble in which all secondary structures are junction-based. Signatures of building block junctions were characterized by the combinatorial preferences of the unpaired single strands (i.e., loops) involved in such junctions. These constraints are intended to set secondary structure conditions for RNA sequences to fold correctly in order to adopt such tertiary structures. Our method will become more effective as more known native structures are discovered.

There are a few technical ways that our method can be improved to gain a higher performance. First, in software JUNCTION, we only used the Weighted Context-Free Grammar (WCFG) to describe signatures of \( k \)-way junctions for \( k \) up to 5. However, we can approximate the signature of a higher-order junction with those of two lower-order junctions and do so with WCFG rules. Second, a more significant improvement may be achieved by including nucleotide correlation requirements for tertiary interactions within or across junctions, for example, the interaction between the tetra-loop of helix P5b and its receptor shown in Fig. 1. Such correlations, nevertheless, are context-sensitive and beyond the capability of CFG and we would need a model that can incorporate tertiary interactions. Third, our JUNCTION is still a prototype and its speed may not be suitable for genome-wide scanning. Because the inside probability like weight that JUNCTION computes is essentially the count of alternative structures, this computation measures the size of the ensemble of each given sequence. Efficient sampling algorithms may be developed to accurately estimate the ensemble size as the inside probability like weight.

6 Conclusions

Our study built a constrained secondary structure space by incorporating structure preferences of RNA junctions derived from experimentally determined RNA structures. Our software based on this structure space consistently distinguished native ncRNA sequences from genomic backgrounds on standard ncRNA datasets. These test results show the incorporation of tertiary structure favored structural elements into the structure model can help to design effective ncRNA gene finding tools.

7 Supplementary Material

The supplementary material is available at http://facultyweb.mga.edu/yingfeng.wang/research/ncrna/junction-supplement.pdf.

Acknowledgements

The authors thank Michael Terns and Rebecca Terns for helpful discussions. This research project was supported in part by NSF MRI 0821263, NIH BISTI R01GM072080-01A1 grant, NIH ARRA Administrative Supplement to NIH BISTI R01GM072080-01A1, and NSF IIS grant of award No 0916250.

References


Yingfeng Wang received his bachelor and master degrees in computer science from Hohai University (Nanjing, Jiangsu) in 2002 and 2005, respectively, and his PhD degree in computer science from the University of Georgia (Athens, Georgia) in 2012. He was a Postdoctoral Research Associate at Oak Ridge National Laboratory from 2012 to 2014. He also worked at Oracle as a software engineer in 2014. He currently is an assistant professor of Information Technology at Middle Georgia State University. His research interests include bioinformatics, graph grammar, and machine learning.

Russell L. Malmberg received an AB in biology from Oberlin College in 1972 and a PhD degree in genetics from the University of Wisconsin in 1976. He was previously employed at Michigan State University, Cold Spring Harbor Laboratory, and is now at the University of Georgia with the titles of Associate Dean and University Professor. He has received research grant funding from US NSF, DOE, USDA, NIH, and has published more than 110 papers. He is an elected Fellow of AAAS, and a member of ISCB and the Genetics Society of America. His current research interests are in RNA biology and bioinformatics, and in ecological and evolutionary genetics.

Liming Cai received his BS and MS degrees in computer science from Tsinghua University (Beijing) in 1984 and 1986, respectively, and his PhD in computer science from Texas A&M University (College Station, Texas) in 1994. He was in the Computer Science Faculty at East Carolina University (1994-1996) and Ohio University (1996-2001) before he joined University of Georgia (Athens, Georgia) in 2001, where he now is a full professor of computer science. His research focuses are algorithms, computational biology, and theory of computation with more than 80 publications in these areas. He has been the principle investigator for a number of research grants awarded by the National Science Foundation (NSF) and the National Institutes of Health (NIH) of the USA.