Published online 7 January 2009

Nucleic Acids Research, 2009, Vol. 37, No. 3 ***999–1009***
doi:10.1093/nar/gkn1054

Genome-wide searching with base-pairing kernel functions for noncoding RNAs: computational and expression analysis of snoRNA families in *Caenorhabditis elegans*

Kensuke Morita[1], Yutaka Saito[1], Kengo Sato[1,2,3], Kotaro Oka[1], Kohji Hotta[1] and Yasubumi Sakakibara[1,*]

[1]Department of Biosciences and Informatics, Keio University, 3–14–1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223–8522, [2]Japan Biological Informatics Consortium (JBIC), 2–45 Aomi, Koto-ku, Tokyo 135–8073 and [3]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2–42 Aomi, Koto-ku, Tokyo 135–0064, Japan

Received July 03, 2008; Revised November 25, 2008; Accepted December 17, 2008

ABSTRACT

Despite the accumulating research on noncoding RNAs (ncRNAs), it is likely that we are seeing only the tip of the iceberg regarding our understanding of the functions and the regulatory roles served by ncRNAs in cellular metabolism, pathogenesis and host-pathogen interactions. Therefore, more powerful computational and experimental tools for analyzing ncRNAs need to be developed. To this end, we propose novel kernel functions, called *base-pairing profile local alignment (BPLA) kernels*, for analyzing functional ncRNA sequences using support vector machines (SVMs). We extend the local alignment kernels for amino acid sequences in order to handle RNA sequences by using STRAL's scoring function, which takes into account sequence similarities as well as upstream and downstream base-pairing probabilities, thus enabling us to model secondary structures of RNA sequences. As a test of the performance of BPLA kernels, we applied our kernels to the problem of discriminating members of an RNA family from nonmembers using SVMs. The results indicated that the discrimination ability of our kernels is stronger than that of other existing methods. Furthermore, we demonstrated the applicability of our kernels to the problem of genome-wide search of snoRNA families in the *Caenorhabditis elegans* genome, and confirmed that the expression is valid in 14 out of 48 of our predicted candidates by using qRT-PCR. Finally, highly expressed six candidates were identified as the original target regions by DNA sequencing.

INTRODUCTION

Postgenomic transcriptome analysis has revealed the existence of a large number of transcripts which lack protein-coding potential, called noncoding RNAs (ncRNAs), and has shown that only 2 % of the human genome encodes protein-coding RNAs, while 60–70 % of the remainder is transcribed into ncRNAs (1). Hence, despite the accumulating research on ncRNAs, it is likely that we are seeing only the tip of the iceberg regarding our understanding of the functions and the regulatory roles served by ncRNAs in cellular metabolism, pathogenesis and host-pathogen interactions.

Several computational methods based on stochastic context-free grammars have been developed for modeling and analyzing functional RNA sequences (2–7). These grammatical methods have succeeded in modeling typical secondary structures of RNAs, and are commonly used for structural alignment of RNA sequences. However, such stochastic models are not capable of discriminating the member sequences of RNA families from nonmembers with sufficiently high accuracy to detect nRNA regions in genome sequences.

*To whom correspondence should be addressed. Tel: +81 45 566 1791; Fax: +81 45 566 1791; Email: yasu@bio.keio.ac.jp
Present address:
Kensuke Morita, IBM Japan, Ltd. 19–21 Nihonbashi Hakozaki-cho, Chuo-ku, Tokyo 103–8510, Japan

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Recently, following the genome sequencing of various species, several computational methods based on the comparative approach have been developed for finding ncRNA sequences (5,7–9). Rivas and Eddy (5) have developed QRNA, which can classify a given pairwise alignment as one of three models using SCFGs: the coding model (COD), in which substitutions between synonymous codons occur frequently as a means of conserving amino acid sequences, the non-coding model (RNA), in which covariances of base pairs occur frequently in order to conserve secondary structures, and others (OTH). In addition, Pedersen *et al*. (7) have developed EvoFold on the basis of phylogenetic SCFGs (4), which assumes that any mutations on each column of a given multiple alignment occur under a given phylogenetic tree of sequences, and mutations in unpaired bases occur more frequently than those in base pairs in conserved secondary structures. These assumptions improve the accuracy of predicting secondary structures and provide the capability of predicting structurally conserved regions from multiple alignments on the basis of phylogenetic information. Furthermore, Washietl *et al*. (8,9) have developed RNAz, which detects structurally conserved regions from multiple alignments using support vector machines (SVMs). RNAz employs the averaged *z*-score of the minimum free energy (MFE) for each sequence and the structure conservation index (SCI). Assuming that the MFE for the common secondary structure is close to that for each sequence if a given multiple alignment is structurally conserved, SCI is defined as the ratio of the MFE for the common secondary structure to the averaged MFE for each sequence. The MFE for each sequence and the common secondary structure are calculated by RNAfold and RNAalifold in the Vienna RNA package (10). These comparative methods have succeeded in detecting many structurally conserved noncoding regions. However, the structural conservation criterion failed to detect some families, such as H/ACA and C/D snoRNAs.

Contrary to the aforementioned methods, several works have been developed with an emphasis on certain families, such as snoRNAs (11–15) and miRNAs (16–19). SnoRNA finders based on pattern recognition algorithms, such as Snoscan (11) for C/D snoRNAs as well as SnoGPS (12) and the MEF-based method (13) for H/ACA snoRNAs, can identify snoRNAs using target site information about rRNA modifications including 2′-*O*-ribose methylation or pseudouridylation in the posttranscriptional process. Therefore, these methods can screen only guide snoRNAs which modify known targets. A successor package called SnoSeeker (14) has been designed for the detection of not only guide snoRNAs, but also of orphan snoRNAs whose targets are unknown, and has been successfully applied to the human genome. Furthermore, an SVM-based snoRNA finder called SnoReport (15) uses several features tailored specifically for snoRNAs, such as constraint MFEs under typical secondary structures of snoRNAs and the match scores of the box motifs of snoRNAs, and it does not require target site information about rRNA modifications.

On the other hand, SVMs and other kernel methods are being actively studied, and have been proposed for solving various problems in many research fields, including bioinformatics (20). These methods are more robust than other existing methods. For example, Saigo *et al*. (21) have proposed local alignment kernels for amino acid sequences, and their kernels with SVMs have outperformed other state-of-the-art methods in benchmarks for remote homology detection of amino acid sequences. Therefore, we considered using kernel methods, including SVMs, for the analysis of functional ncRNAs.

For the purpose of analyzing ncRNAs using kernel methods, including SVMs, we have already proposed *stem kernels*, which extend the concept of string kernels to allow measurement of the similarities between two RNA sequences from the viewpoint of secondary structures (22). The feature space of the stem kernels is defined by enumerating all possible common base pairs and stem structures of arbitrary lengths. However, since the computational time and the memory footprint of stem kernels are of the order of $O(n^4)$, where n is the length of the inputted RNA sequence, applying stem kernels directly to large data sets of ncRNAs is impractical.

Therefore, we propose novel kernel functions, called *base-pairing profile local alignment (BPLA) kernels*, for discrimination and detection of functional RNA sequences using SVMs. We extend the concept of local alignment kernels in such a way that it can handle RNA sequences using STRAL's scoring function (23). The local alignment kernels measure the similarity between two sequences by summing the scores over all possible local alignments with gaps. STRAL's scoring function takes into account sequence similarities as well as upstream and downstream base-pairing probabilities, which enables us to model secondary structures of RNA sequences. Note that unlike SnoReport, BPLA kernels do not depend on any family-specific features.

In order to test the performance of our kernels, we applied them to the problem of discriminating members of an RNA family from nonmembers using SVMs. The results indicated that the discrimination ability of our kernel functions is stronger than that of existing methods. Furthermore, we performed several experiments regarding the prediction of functional RNA regions using SVMs together with our kernel functions. The experimental results showed that our kernel functions enable us to efficiently discern individual RNA families within genome sequences. Finally, in order to confirm their performance for practical use, we used the proposed kernels in implementing a search for snoRNA families in the *Caenorhabditis elegans* genome, which has been studied extensively by computational and expression analyses (24–27). Our prediction was tested by qRT-PCR, and the verified snoRNA candidates were further confirmed by DNA sequencing.

METHODS

Local alignment kernels for amino acid sequences

Before proposing our new kernels, we briefly describe the concept of local alignment kernels, which has been proposed by Saigo *et al*. (21). A local alignment kernel

is a kind of string kernel which can calculate the similarity between a pair of sequences.

Let Σ be a finite set of symbols ($|\Sigma| = 20$ for amino acid sequences, or $|\Sigma| = 4$ for nucleotide sequences). For two sequences $\mathbf{x}, \mathbf{y} \in \Sigma^*$, a concatenation of $\mathbf{x}$ and $\mathbf{y}$ is denoted as $\mathbf{xy}$. Let $|\mathbf{x}|$ be the length of $\mathbf{x}$.

A local alignment kernel between a pair of sequences $\mathbf{x}$ and $\mathbf{y}$ is defined by decomposing them into simple pieces and convoluting them. Given two string kernels K_1 and K_2, we can define a convolution kernel $K_1 * K_2$ as follows:

$$K_1 * K_2(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{x}=\mathbf{x}_1\mathbf{x}_2, \mathbf{y}=\mathbf{y}_1\mathbf{y}_2} K_1(\mathbf{x}_1, \mathbf{y}_1) K_2(\mathbf{x}_2, \mathbf{y}_2). \quad \mathbf{1}$$

If both K_1 and K_2 are valid kernels, $K_1 * K_2$ is also a valid kernel (28).

Three atomic string kernels are defined as follows. The first is the constant kernel which models a null contribution to a local alignment:

$$K_0(\mathbf{x}, \mathbf{y}) = 1 \quad \text{for } \forall \mathbf{x}, \mathbf{y} \in \Sigma^*. \quad \mathbf{2}$$

The second is the kernel between two residues:

$$K_a^{(\beta)}(\mathbf{x}, \mathbf{y}) = \begin{cases} 0 & \text{if } |\mathbf{x}| \neq 1 \text{ or } |\mathbf{y}| \neq 1 \\ \exp[\beta s(\mathbf{x}, \mathbf{y})] & \text{otherwise,} \end{cases} \quad \mathbf{3}$$

where $\beta \geq 0$ is a constant and $s : \Sigma^2 \to \mathbb{R}$ is a substitution scoring function between two residues. If the matrix $[s(a, b)]_{a,b\in\Sigma}$ is conditionally positive semi-definite, $K_a^{(\beta)}(\mathbf{x}, \mathbf{y})$ is a valid kernel. The last kernel is the one which models the affine gaps:

$$K_g^{(\beta)}(\mathbf{x}, \mathbf{y}) = \exp\{\beta[g(|\mathbf{x}|) + g(|\mathbf{y}|)]\}, \quad \mathbf{4}$$

where $\beta \geq 0$ is a constant and $g(n)$ is a gap cost function for sequences of length n given by

$$g(n) = \begin{cases} 0 & \text{if } n = 0 \\ d + e(n-1) & \text{if } n > 0, \end{cases} \quad \mathbf{5}$$

where d and e are the gap opening penalty and the gap extension penalty, respectively.

Let $\Pi(\mathbf{x}, \mathbf{y})$ be a set of all possible local alignments of $\mathbf{x}$ and $\mathbf{y}$. Given a local alignment $\pi \in \Pi(\mathbf{x}, \mathbf{y})$, we can define $K_\pi^{(\beta)}$ as a string kernel for the local alignment π with n matching columns as follows:

$$K_\pi^{(\beta)} = K_0 * \left(K_a^{(\beta)} * K_g^{(\beta)}\right)^{(n-1)} * K_a^{(\beta)} * K_0. \quad \mathbf{6}$$

This kernel decomposes the alignment π into an initial part (whose similarity is measured by K_0), n aligned residues (whose similarities are measured by $K_a^{(\beta)}$), gaps (whose similarities are measured by $K_g^{(\beta)}$) and a terminal part (whose similarity is measured by K_0).

Finally, in order to compare two sequences with respect to all possible local alignments, a local alignment kernel $K_{LA}^{(\beta)}$ sums over all $K_\pi^{(\beta)}$ for all local alignments:

$$K_{LA}^{(\beta)}(\mathbf{x}, \mathbf{y}) = \sum_{\pi\in\Pi(\mathbf{x},\mathbf{y})} K_\pi^{(\beta)}(\mathbf{x}, \mathbf{y}). \quad \mathbf{7}$$

According to the state transition diagram of pairwise local alignments shown in Figure 1, a local alignment kernel (equation 7) can be calculated from the following recursive equations:

$$M(i,j) = \begin{cases} 0, & i = 0 \text{ or } j = 0 \\ \exp[\beta s(x_i, y_j)][1 + I_x(i-1, j-1) + I_y(i-1, j-1) + M(i-1, j-1)], & \text{otherwise} \end{cases}$$

$$I_x(i,j) = \begin{cases} 0, & i = 0 \text{ or } j = 0 \\ \exp(\beta d) M(i-1, j) + \exp(\beta e) I_x(i-1, j), & \text{otherwise} \end{cases}$$

$$I_y(i,j) = \begin{cases} 0, & i = 0 \text{ or } j = 0 \\ \exp(\beta d)[M(i, j-1) + I_x(i, j-1)] + \exp(\beta e) I_y(i, j-1), & \text{otherwise} \end{cases}$$

$$R_x(i,j) = \begin{cases} 0, & i = 0 \text{ or } j = 0 \\ M(i-1, j) + R_x(i-1, j), & \text{otherwise} \end{cases}$$

$$R_y(i,j) = \begin{cases} 0, & i = 0 \text{ or } j = 0 \\ M(i, j-1) + R_x(i, j-1) + Ry(i, j-1), & \text{otherwise.} \end{cases}$$

Finally, we can obtain the kernel value as follows:

$$K_{LA}^{(\beta)} = 1 + R_x(|\mathbf{x}|, |\mathbf{y}|) + R_y(|\mathbf{x}|, |\mathbf{y}|) + M(|\mathbf{x}|, |\mathbf{y}|). \quad \mathbf{8}$$

The complexity of this calculation is of the order of $O(|\mathbf{x}||\mathbf{y}|)$ with respect to both time and memory.

The local alignment kernel can be regarded as a special case of the partition function of all possible alignments between two sequences. Miyazawa *et al.* (29) pioneered the use of the partition function for a reliable sequence alignment, which is known as a centroid estimator (30). Mückstein *et al.* (31) employed the partition function for stochastic pairwise alignments by which suboptimal alignments can be drawn from an ensemble of all possible alignments of two sequences. The partition function plays a crucial role in the reliability of these methods as well as the local alignment kernel.

BPLA kernels for ncRNAs

Here, we propose new kernels for ncRNA sequences which take into account secondary structures by utilizing STRAL's scoring function (23).

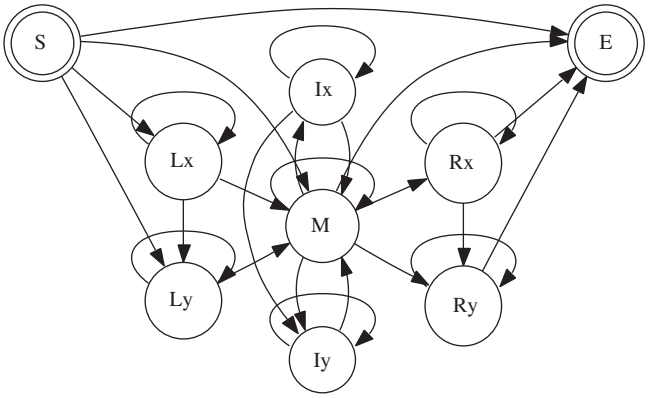

Figure 1. A state transition diagram of pairwise local alignments. S is the initial state, L_x and L_y are unaligned states before the alignments, M is the match state, I_x and I_y are the gap states, R_x and R_y are unaligned states after the alignments and E is the final state.

Local alignment kernels can calculate the similarity between a pair of sequences by taking into account only sequence homology. However, it is well known that ncRNAs form secondary structures with important functions. Therefore, it is necessary to consider the secondary structures of RNAs in order to compare two RNA sequences.

Let us introduce a certain kind of secondary structure information into the match scores of local alignments. For each sequence, we first calculate a base-pairing probability matrix using the McCaskill algorithm (32). The base-pairing probability matrix for a sequence $\mathbf{x}$ consists of the base-pairing probabilities P_{ij} that the i-th and the j-th nucleotides form a base pair, which is defined as:

$$P_{ij} = \mathbb{E}[I_{ij}|\mathbf{x}] = \sum_{y \in \mathcal{Y}(\mathbf{x})} p(y|\mathbf{x}) I_{ij}(y), \quad \mathbf{9}$$

where $\mathcal{Y}(\mathbf{x})$ is an ensemble of all possible secondary structures of $\mathbf{x}$, $p(y|\mathbf{x})$ is the posterior probability of a secondary structure y given $\mathbf{x}$, and $I_{ij}(y)$ is an indicator function, which equals 1 if the i-th and the j-th nucleotides form a base pair in y and 0 otherwise. In this study, we employ the Vienna RNA package (10) for computing the expected counts (equation 9) using the McCaskill algorithm.

Subsequently, for each position i, we categorize the base-pairing probabilities into three kinds of sums: the probability $P_i^{\text{left}} = \sum_{j>i} P_{ij}$ that a pair is formed with one of the downstream nucleotides, the probability $P_i^{\text{right}} = \sum_{j<i} P_{ji}$ that a pair is formed with one of the upstream nucleotides, and the probability $P_i^{\text{unpair}} = 1 - (P_i^{\text{left}} + P_i^{\text{right}})$ that the nucleotide is unpaired. A probability distribution consisting of these three probabilities is called a *base-pairing profile* (33).

In this case, in accordance with STRAL (23), the match score between two nucleotides x_i and x_j is defined using base-pairing profiles as follows:

$$\begin{aligned} s(x_i, y_j) &= \alpha(S_{\text{struct}}) + S_{\text{seq}} \\ &= \alpha\left(\sqrt{P_{x_i}^{\text{left}} P_{y_j}^{\text{left}}} + \sqrt{P_{x_i}^{\text{right}} P_{y_j}^{\text{right}}}\right) + \\ &\quad \sqrt{P_{x_i}^{\text{unpair}} P_{y_j}^{\text{unpair}}}\, d(x_i, y_j), \end{aligned} \quad \mathbf{10}$$

where α is a weight parameter for structural information, and $d : \Sigma^2 \to \mathbb{R}$ is a substitution scoring function between two nucleotides. The first term of (equation 10) measures the structural similarity between $\mathbf{x}$ and $\mathbf{y}$ by taking the inner product of the base-pairing profiles, and the second term captures the sequence-level homology as well as the standard local alignment kernels. If d is positive semi-definite, (Equation 10) is obviously also positive semi-definite (28). We use a modified RIBOSUM 85-60 (34) as the substitution matrix d. Since the original RIBOSUM 85-60 cannot satisfy the condition of positive semi-definiteness, we subtract the smallest eigenvalue of the matrix from each of its diagonal elements in order to transform the matrix into a positive semi-definite one. Therefore, we can employ s as the match scores for the local alignment kernels (3). We call this *the BPLA kernel*. The computational complexity of the base-pairing probability matrices is of the order of $O(|\mathbf{x}|^3 + |\mathbf{y}|^3)$ with respect to time and $O(|\mathbf{x}|^2 + |\mathbf{y}|^2)$ with respect to memory, and calculating the kernel value shows a complexity of the order of $O(|\mathbf{x}||\mathbf{y}|)$ for both time and memory.

Experimental verification by qRT-PCR

Expression analysis by qRT-PCR was performed in order to verify the validity of our snoRNA candidates predicted in the *C. elegans* genome. The reason for the choice of qRT-PCR was that high throughput of the method enabled to analyze the expression level of our candidate in the relatively large scale compared with the other methods such as northern blotting.

Total RNA was extracted from mixed developmental stages of *C. elegans* using RNAqueous-4PCR kit (Ambion). In order to remove genomic DNA contaminants, the total RNA was treated with 6 U of DNase I for an hour which is more thorough than 2 U for 30 min in the manufacturer's instruction. This substantially reduced the background noise of fluorescence intensities in RT-PCR especially at the later cycle phase and enabled us to determine the C_t values more accurately.

Templates for RT-PCR were produced by poly(A) tailing and oligo(dT) priming using Ncode miRNA First-Strand cDNA Synthesis kit (Invitrogen) following the manufacturer's protocol except that we did not use Universal qPCR Primer included in the kit. It might be confusing to readers that the name of this kit contains the word 'miRNA' despite our experiments for verication of the snoRNA candidates. However, the kit does not include any miRNA-specific features in its principle and thus can be applied to any kind of small RNA families. The positive cDNA template, denoted by RT (+), was synthesized with the reverse transcriptase SuperScript III RT/RNaseOUT Enzyme Mix, and the negative control template, denoted by RT (−), was prepared by adding DEPC-treated RNase-free water (Invitrogen) instead of the reverse transcriptase.

RT-PCR experiments were performed using StepOne Real-Time PCR System (ABI). The reaction was carried out in 20 µl with 1 µl of the template, 0.25 µM of the specific primer pair (see below) and Power SYBR Green PCR

Master Mix (ABI) at 95°C for 15 s and 60°C for 1 min for 28 cycles followed by a melting curve analysis. The chart of fluorescence intensities was analyzed using the instrument's system software v2.0 (ABI). We confirmed that amplification of RT (−) was not detected for all the primer pairs. In the experiments of RT (+), every amplified candidate showed the specific PCR product indicated by the melting curve.

For each PCR target, a sequence-specific primer pair (forward and reverse) was designed using Primer3 (35) and BLAST search against the *C. elegans* genome. Primer3 was executed with the parameters recommended by Takara Bio which is more stringent than the default setting about base composition biases and self/pair complementarity that cause cross-hybridization and amplification of primer dimers. Amplification of nonspecific products was further tested by several times of preliminary experiments, and the primers were redesigned if needed. Since some of the targets were apparently difficult for designing good primers, we discarded such regions. We first selected the 67 regions (top 50 candidates of our prediction, one known snoRNA `CeN45` discovered by Deng *et al.* (26) as a positive control, and randomly selected 6 intronic and 10 intergenic regions as negative controls), and after this primer design procedure retained the 59 targets for the subsequent experiments (48 candidates, one positive control and 4 intronic and 6 intergenic negative controls). The list of the primers designed for these 59 targets is available as a Supplementary Material.

Amplimers of the six candidates detected under the C_t value of 25 were separated by electrophoresis on a 6 % gel of NuSieve 3:1 Agarose (Lonza), and verified in terms of the product length (Figure 8). These candidates were further confirmed by DNA sequencing with the following protocol. The DNA fragments were excised and purified using QIAquick Gel Extraction Kit (Qiagen) and ligated into T vectors with pGEM-T Easy Vector System I (Promega) according to the manufacturer's protocol. DH5 α competent cell (Takara Bio) was used in transformation, and the RT-PCR products were sequenced after O/N culture and plasmid purification. The sequencing reaction was performed using BigDye Terminator v3.1 Cycle Sequencing Kit (ABI) in 20 μl with 2 μl of the template and 0.8 μM of M13 primer at 95°C for 10 s, 50°C for 5 s and 60°C for 4 min for 25 cycles. The reaction was then cleaned up on PERFORMA Gel Filtration Cartridge (Edge Biosystems).

RESULTS

Availability

Our implementation of the BPLA kernels is freely available at http://bpla-kernel.dna.bio.keio.ac.jp/ under the GNU public license. It takes a set of RNA sequences and calculates a kernel matrix, which can be used as an input for a popular SVM tool called LIBSVM (36). Furthermore, our software is capable of parallel processing using the Message Passing Interface (MPI) (37).

Computational predictions using BPLA Kernels

In order to confirm the accuracy of our new kernels, we carried out several experiments in which SVMs with our kernels attempted to detect known ncRNA families. The accuracy was assessed in terms of precision (prec) and sensitivity (sens), which were defined as follows:

$$\mathrm{prec} = \frac{TP}{TP + FP}, \quad \mathrm{sens} = \frac{TP}{TP + FN},$$

where TP is the number of correctly predicted positives, FP is the number of incorrectly predicted positives and FN is the number of incorrectly predicted negatives. Furthermore, the area under the precision–sensitivity curve was also used for evaluation. The precision–sensitivity curve plots the sensitivity as a function of the precision for varying decision thresholds of a classifier.

In our first experiment, the discrimination ability and the execution time of our kernels were tested on our previous data set used in (22), which includes five RNA families: tRNAs, miRNAs (precursor), 5S rRNAs, H/ACA snoRNAs and C/D snoRNAs. We chose 100 sequences from each of the above RNA families from the Rfam database (38) as positive samples, such that the pairwise identity was not above 80% for any pair of sequences, and 100 randomly shuffled sequences with the same dinucleotide composition as the positives were generated as negative samples for each family. The parameters for the BPLA kernels were $d = -27$, $e = -0.1$, $\alpha = 1$ and $\beta = 0.1$. Furthermore, the discrimination performance was evaluated using the 10-fold cross-validation method, and the experimental results shown in Table 1 indicate that the match scores based on base-pairing profiles employed by the BPLA kernels improve the discrimination accuracy as compared with the local alignment kernels.

Next, we compared our kernels with previous methods including SnoReport (15) and miPred (18). SnoReport utilizes SVMs with several features tailored specifically for snoRNAs. This differs from our general approach, which does not depend on any family-specific features. We evaluated the results using the same measures as in (15), namely specificity (spec) and sensitivity (sens), which are defined as follows:

$$\mathrm{spec} = \frac{TN}{TN + FP}, \quad \mathrm{sens} = \frac{TP}{TP + FN},$$

where TN is the number of correctly predicted negatives, and calculated the Matthews correlation coefficient (MCC):

$$\mathrm{MCC} = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}.$$

We experimented with the same data set used in (15), which contains 135 positives and 1770 negatives of C/D snoRNAs, and 81 positives and 89 negatives of H/ACA snoRNAs. Table 2 shows that our kernels significantly outperformed SnoReport. MiPred also utilizes SVMs with 29 global and intrinsic folding attributes as features for pre-miRNAs. We experimented on the same

Table 1. Comparison between the BPLA kernels and the other existing kernels in terms of their discrimination ability with respect to five RNA families: tRNAs, miRNAs, 5S rRNAs, H/ACA snoRNAs and C/D snoRNAs

Family	BPLA kernel				Local alignment kernel				Stem kernel			
	AUC	prec	sens	time (s)	AUC	prec	sens	time (s)	AUC	prec	sens	time (s)
tRNA	**1.00**	**0.97**	**0.94**	7.9×10^{-4}	0.97	0.96	0.85	6.2×10^{-4}	0.96	0.95	0.89	0.9
miRNA	**0.99**	**0.95**	**0.92**	1.1×10^{-3}	0.87	0.90	0.65	8.2×10^{-4}	0.92	0.69	0.90	1.6
5S rRNA	**1.00**	**1.00**	0.92	1.8×10^{-3}	**1.00**	**1.00**	**0.95**	1.4×10^{-3}	0.98	0.99	0.72	5.1
H/ACA snoRNA	**0.89**	0.88	0.68	3.0×10^{-3}	0.88	**0.92**	**0.70**	2.5×10^{-3}	0.83	0.85	0.39	12.8
C/D snoRNA	**0.92**	**0.93**	0.75	1.5×10^{-3}	0.88	**0.93**	0.66	1.2×10^{-3}	0.79	0.61	**0.80**	4.7
Total	**0.96**	**0.95**	**0.84**	1.6×10^{-3}	0.92	0.94	0.76	1.3×10^{-3}	0.90	0.82	0.74	5.0

Family, name of the target ncRNA family; AUC, area under the precision–sensitivity curve; prec, precision of discriminating the target ncRNA family; sens, sensitivity of discriminating the target ncRNA family; time, average computation time for each kernel on a 2.0 GHz AMD Opteron processor. For each evaluation measure, the most accurate kernel value is indicated by a bold-faced number.

Table 2. Comparison between the BPLA kernels and SnoReport in terms of their discrimination ability with respect to the data set used in (15)

Family	BPLA kernel			SnoReport		
	spec	sens	MCC	spec	sens	MCC
C/D snoRNA	**1.00**	0.71	**0.83**	0.91	**0.96**	0.62
H/ACA snoRNA	**1.00**	**0.83**	**0.84**	0.89	0.78	0.68

Family, name of the target ncRNA family; spec, specificity of discriminating the target ncRNA family; sens, sensitivity of discriminating the target ncRNA family; MCC, Matthew's correlation coefficient of discriminating the target ncRNA family. For each evaluation measure, the most accurate value is indicated by a bold-faced number.

data set used in (18), which contains 323 human pre-miRNAs and 8494 pseudo hairpins from human RefSeq genes. After 200 positives and 400 negatives were randomly selected from pre-miRNAs and pseudo hairpins, respectively, we evaluated the accuracy using the 5-fold cross-validation as well as in the experiment in (18). Our method yielded the accuracy of 0.90, 0.92 and 0.97 for sensitivity, specificity and AUC, respectively, whereas miPred achieved that of 0.88, 0.98 and 0.98. These results indicate that our method is highly accurate, and is competitive with miPred.

Finally, we evaluated the ability of our kernels to detect fragments which contain known ncRNAs. In practical situations, and in contrast to the above experiments, we usually extract fixed-length fragments from the target genomic sequences since the boundaries of RNA genes in genomic sequences are unknown. Therefore, it is important to confirm that the proposed kernels can also predict ncRNAs from such fragments in addition to entire sequences. We chose 24 ncRNA families from the Rfam database (38). For each sequence, we produced a longer sequence by concatenating randomly generated subsequences with the same dinucleotide composition into its upstream and downstream, as illustrated in Figure 2. Subsequently, we extracted fixed-length fragments from those sequences, where the window size of the fragments was 120 nt and sliding was 40 nt. These fragments include partial or complete ncRNA sequences. Negative examples were randomly generated by dinucleotide shuffling such that they do not include any ncRNA sequences. The parameters for the BPLA kernels were $d = -14$, $e = -0.07$, $\alpha = 7$ and $\beta = 0.07$. Figure 3 shows that our kernels have sufficient ability to detect fragments containing ncRNAs as well as entire ncRNA sequences with known boundaries.

Genome-wide search of snoRNA families in the *C. elegans* genome

In addition, we attempted to detect novel snoRNAs in the *C. elegans* genome. First, we trained a support vector classifier for snoRNAs using known snoRNA sequences in *C. elegans*, which are annotated in WormBase (Release WS182) (24). We extracted positive fragments of known snoRNAs from the *C. elegans* genome, where the window size of the fragments was 240 nt and the sliding was 40 nt. Negative fragments were randomly generated by dinucleotide shuffling. As a result, we trained the support vector classifier of snoRNAs with 128 positive fragments and 512 negative fragments. Subsequently, we extracted all fragments of the same window size from both strands of the entire genome of *C. elegans*, which resulted in a total number of fragments of 5 014 018. For each fragment, the trained support vector classifier calculated an SVM class probability, which indicates the confidence regarding the affiliation of the fragment to a given snoRNA family. The time needed to scan all of the fragments was about a week on our Linux cluster comprising twenty 2.8 GHz dual-core AMD Opteron processors. Figure 4 shows the distribution of the SVM class probabilities for the known snoRNAs and the other fragments. It is clear that the distribution of known snoRNAs is biased towards high probability, which indicates that the trained classifier has strong ability to distinguish snoRNAs from other sequences. The list of snoRNA candidates predicted by our method with high probability is available as a Supplementary Material.

In order to show that genome-wide search is a much more difficult task than the discrimination task on a well-established training set and the above result is only achieved by our BPLA kernels, we compared the ability of the genome search for snoRNA families with SnoReport. We performed the aforementioned procedure with SnoReport, instead of our kernels, only in the

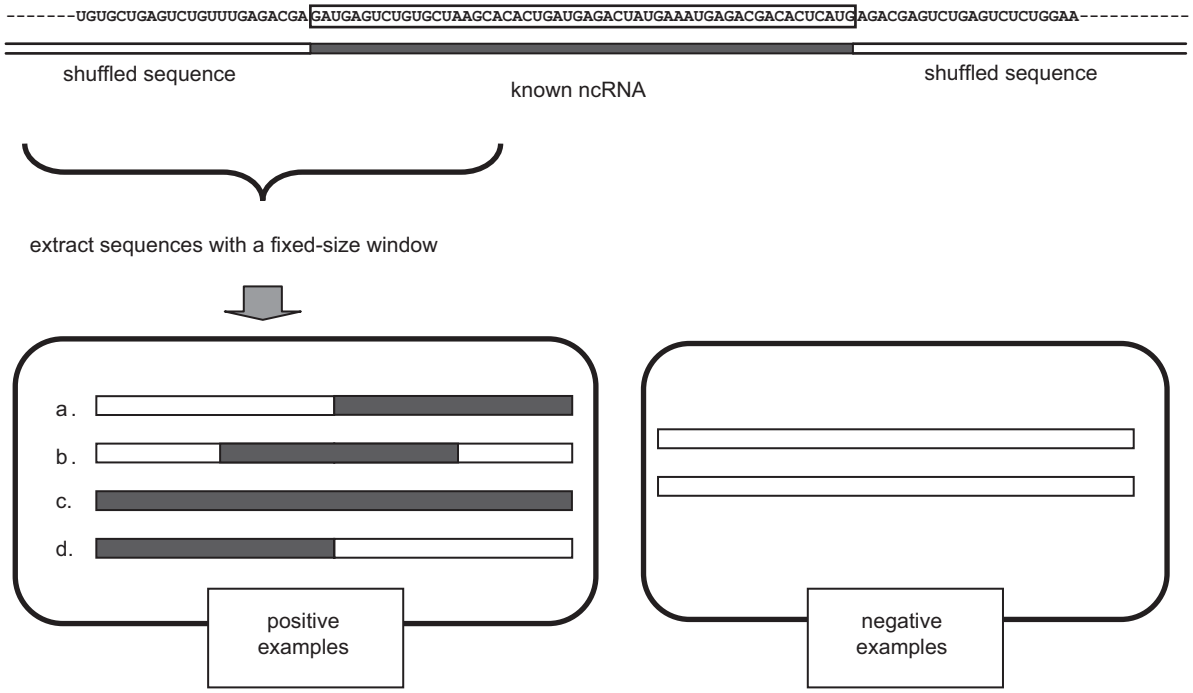

Figure 2. Generating a fragment data set. Given the ncRNA sequences, longer sequences are randomly generated by concatenating shuffled sequences, after which fixed-length fragments are extracted.

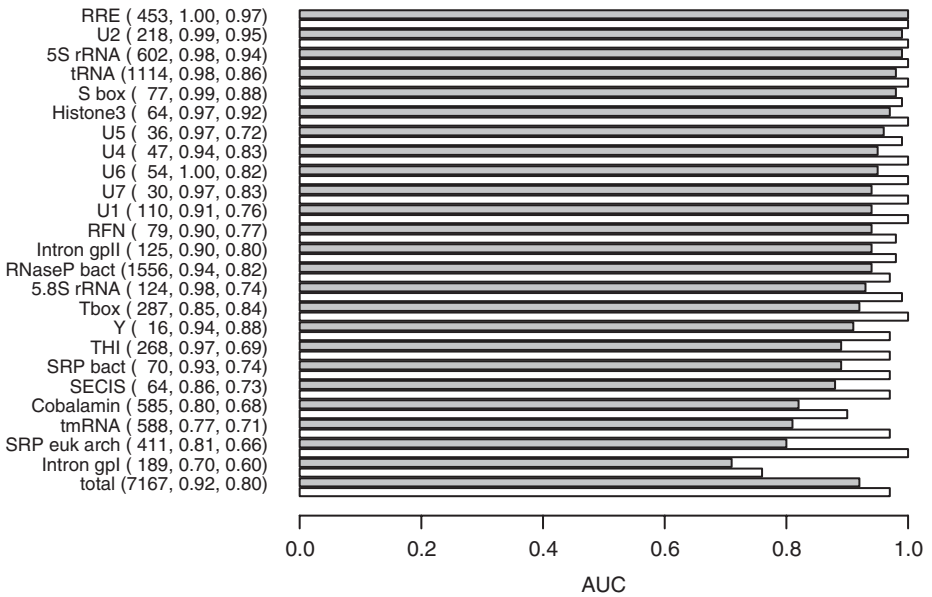

Figure 3. The accuracy of detecting 24 ncRNA families. Each bar indicates the area under the precision–sensitivity curve (AUC) for each family. The grey bars indicate the AUC of detecting fragmented ncRNA sequences with a fixed-size window, and the white bars indicate the AUC of detecting entire ncRNA sequences with known boundaries. The tuple of values following the name of each family represent the number of sequences, the precision and the sensitivity of detecting fragmented ncRNA sequences, respectively.

chromosome I, and calculated the SVM class probability distribution. Then, we evaluated the ability to distinguish snoRNAs from other fragments using the receiver operating characteristic (ROC) analysis, in which positives are known snoRNAs, and negatives are the fragments from the genomic sequence. Only the fragments whose SVM class probability is above 0.5 were considered. As a result, the area under the ROC curve for our kernels was 0.873, whereas that for SnoReport was 0.686. This result indicates that even though the statistical method SVM is employed, the high performance cannot always be expected, and our well-designed BPLA kernels were

proved to be essential to the genome-wide search task. Note that we cannot exclude the possibility that some of fragments used for negatives here may be unknown snoRNAs. Therefore, the scores in the analysis represent lower bounds of them.

One of our predicted fragments which has an SVM class probability above 0.99 overlaps with an EST and the *C. briggsae* alignment, as shown in Figure 5. Note that this fragment has not been predicted by RNAz (25). We manually extracted a putative C/D snoRNA sequence from this fragment, and predicted its secondary structure using RNAfold (10) (Figure 6).

Expression analysis of predicted snoRNAs

We performed experimental verification of the predicted snoRNAs by qRT-PCR. We designed the primers for the 59 target regions containing the 48 candidates (see Methods section) which are not overlapped with either exons or ESTs.

Figure 7 shows the results of the cycle threshold (C_t) values in PCR amplification. Note that a smaller C_t value indicates that the corresponding candidate is expressed more strongly, and is more likely to be a region for a ncRNA gene. Since no negative control remained under the C_t value of 25, we determined this value as the threshold for significant expressions in the organism. By means of this threshold, we verified that 6 of the 48 candidates were in fact expressed as RNA genes. We emphasize that it is an extremely strict threshold which should disallow any expressions for negative controls (introns and intergenic regions). In contrast, Washietl *et al.* (39) have revealed that 43 of 175 predicted candidates in ENCODE regions were expressed by means of the threshold in which 4 of 38 negative controls were also expressed. If we allow that 1 of 10 negative controls was accepted as expressed, then 14 of

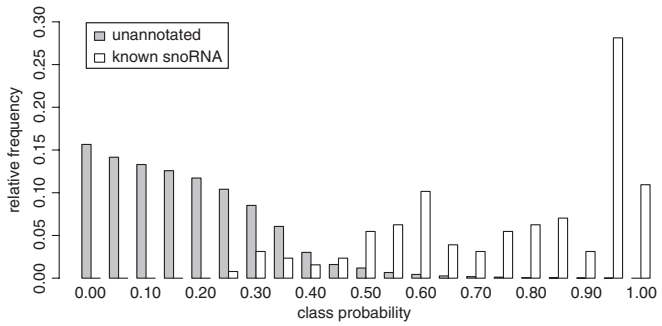

Figure 4. Distribution of the SVM class probabilities for known snoRNAs and the rest of the fragments in the *C. elegans* genome calculated by the trained SVM classifier.

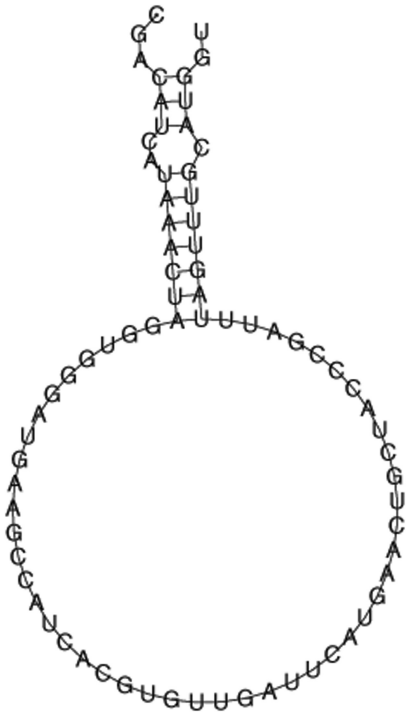

Figure 6. The predicted secondary structure of a putative C/D snoRNA in our predicted fragment which overlaps with an EST, as shown in Figure 5.

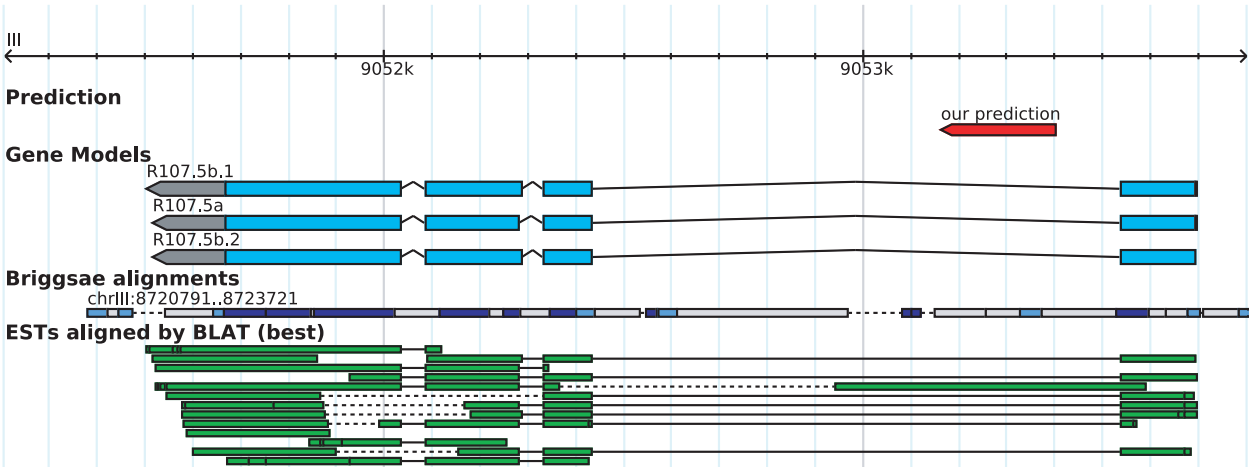

Figure 5. One of our predicted fragments overlaps with an EST and the *C. briggsae* alignment.

48 candidates were accepted as expressed below the C_t cycles of 26.5. This rate of expression is comparable with that of (39).

We separated the amplimers of the 6 most reliable candidates on a 6% agarose gel as shown in Figure 8. Then, the DNA fragments were sequenced, and we confirmed that all of these sequences were exactly identical to the original target regions by BLAST search on the WormBase. The list of the regions of the predicted candidates, the PCR targets and the mapped sequence reads is available as a Supplementary Material.

DISCUSSION

We investigated the overlap between the candidates of our kernels and those in previous studies (25,27). Missal *et al.* (25) predicted structurally conserved ncRNAs from the genomic alignment of *C. elegans* and *C. briggsae* using RNAz. He *et al.* (27) detected the transcribed fragments, or 'transfrags', focusing on polyadenylated or non-polyadenylated fractions of the transcriptome with *C. elegans* whole-genome tiling microarrays. Figure 9 illustrates the overlap between RNAz, our candidates whose SVM class probabilities are above 0.9, and all types of transfrags. There are only 10 candidates predicted by both RNAz and our kernels, and this low incidence is attributable to the radically different features of the two methods. The SCI used in RNAz directly assesses the structure conservation in multiple alignments in order to detect unknown structured ncRNAs, while BPLA kernels evaluate the structural similarity between ncRNA sequences using base-pairing profiles. Furthermore,

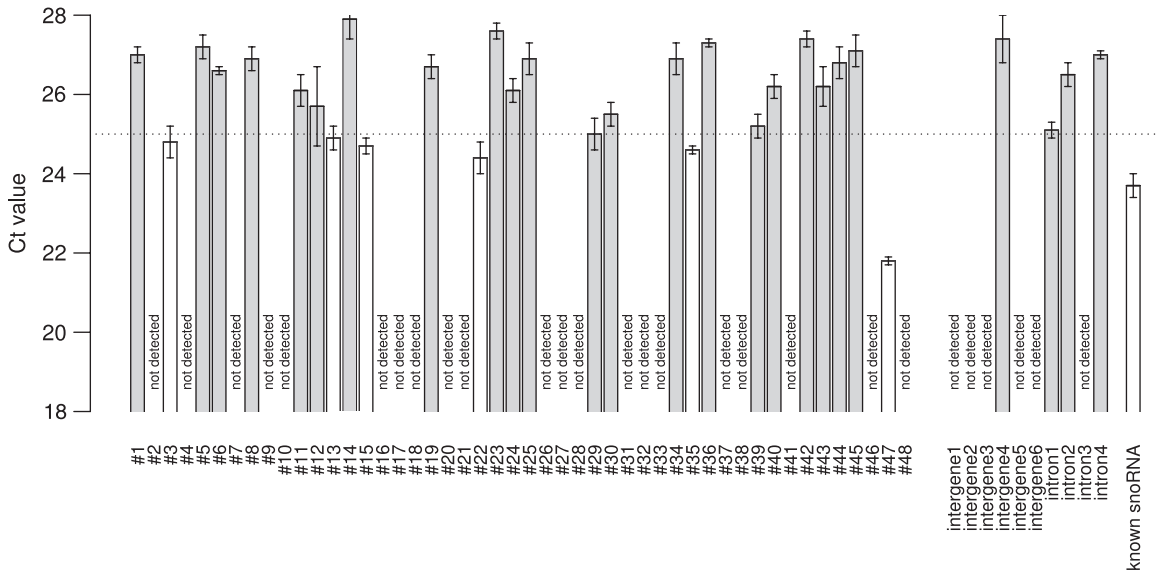

Figure 7. qRT-PCR verification of our predictions. The cycle threshold (C_t) values for the predicted candidates, the positive and the negative controls are shown. The horizontal dashed line indicates our threshold 25 of the C_t value to determine the significant level of expression. The white bars indicate the reliable candidates whose C_t values are under this threshold.

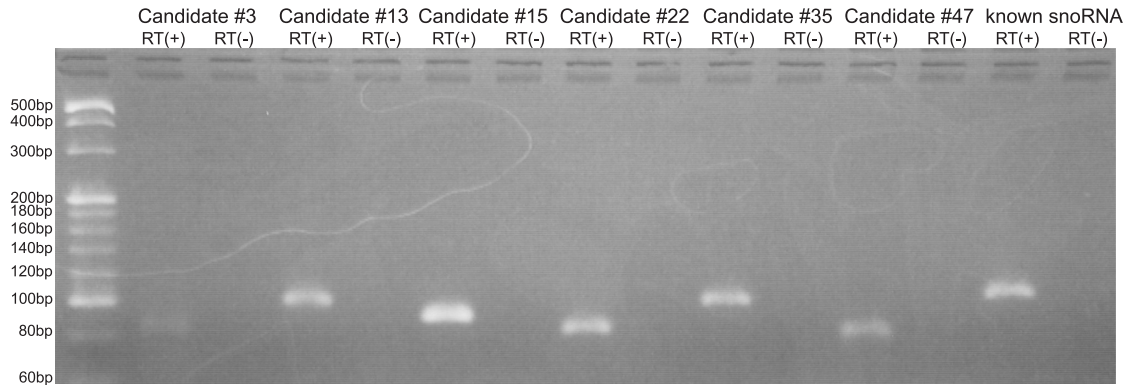

Figure 8. Agarose gel electrophoresis for the PCR products of the verified candidates. Amplimers of the reliable candidates under the threshold value were separated on an agarose gel in the lanes denoted by RT (+), and the corresponding PCR reactions with the negative template were applied into the lanes denoted by RT (−). No nonspecific PCR product was found in all of the lanes. The length of each product was shown to be identical to the original target region.

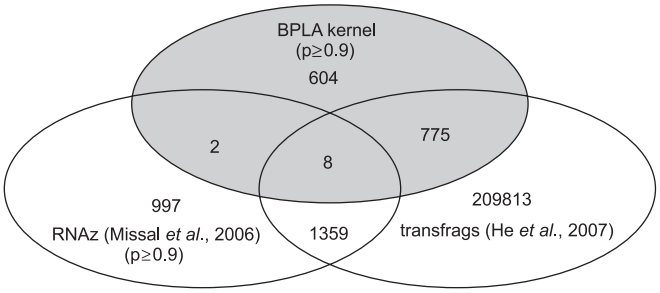

Figure 9. Comparison of our candidates with those in previous studies (25,27) in the *C. elegans* genome. For RNAz and the BPLA kernels, the respective number of fragments whose SVM class probabilities are above 0.9 are shown.

several works have reported that the ability of RNAz to detect snoRNA families is weaker than that for other ncRNA families (9,25). Therefore, it is very likely that RNAz predicted not only snoRNA-specific regions, but also structurally conserved regions which are not specific to snoRNAs, in the nematode alignment, whereas our kernels searched more precisely for snoRNA-like sequences in the *C. elegans* genome given the known snoRNAs.

As shown in Figure 9, ~60% of our predicted candidates whose SVM class probabilities are above 0.9 overlap with transfrags. At the same time, three of the six candidates confirmed by qRT-PCR also overlap with transfrags, although three confirmed candidates do not. This fact suggests that we cannot exclude the possibility that a part of the remaining >40% of the predicted candidates of snoRNAs located in the regions without any transfrags might be confirmed by qRT-PCR.

The relatively low precision values 14/48 ($C_t < 26.5$) and 6/48 ($C_t < 25.0$), which are the rates of the reliable candidates verified by qRT-PCR expression analysis, do not represent the true statistical significance. These values 14/48 and 6/48 can statistically be evaluated by using the *P*-value which follows the cumulative hypergeometric distribution:

$$p(k; N, m, n) = \sum_{l \geq k} \frac{\binom{m}{l}\binom{N-m}{n-l}}{\binom{N}{n}},$$

where k is the number of expressed candidates below the C_t value, m is the number of the selected candidates for the verification, n is the number of snoRNAs in the whole genome, and N is the number of nonexonic fragments in the whole genome. We can approximate that $n \approx 2700$ because Deng *et al.* (26) have revealed that the number of small noncoding transcripts in the *C. elegans* genome is at most 2 700, and $N = 5\,014\,018$ (the number of the fragments in the whole genome) $-1\,720\,000$ (the roughly estimated number of the exonic fragments) $\approx 3\,300\,000$. In accordance with this formulation, the *P*-value for 6/48, which corresponds to the probability to include six snoRNA coding regions among randomly selected 48 fragments from 3 300 000 fragments in the *C. elegans* whole genome, becomes 3.6e-12, and the *P*-value for 14/48 becomes 2.8e-32. These extremely small *P*-values indicate that six hits among 48 candidates selected from 3 300 000 fragments are not obtained just by chance and hence prove the effectiveness of our method.

The parameters d, e, α and β for the BPLA kernels are quite different from each of two cases: target sequences have flanking region or not. The optimal parameters for each case were calibrated by the grid search on all parameter combinations from selected search space for each parameter. However, this brute force approach should require huge computational time for large problems. Therefore, we are planning to implement an alternative approach, such as a gradient-based adaptation of parameters for given training data (40).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We wish to thank Chisato Ushida and Takashi Ideue for helpful comments and advice. We also thank our colleagues from the RNA Informatics Team at the Computational Biology Research Center (CBRC) for fruitful discussions.

FUNDING

New Energy and Industrial Technology Development Organization (NEDO) of Japan (Functional RNA Project); Ministry of Education, Culture, Sports, Science and Technology of Japan (Grant-in-Aid for Scientific Research on Priority Area "Comparative Genomics" No. 17018029) Funding for open access charge: Ministry of Education, Culture, Sports, Science and Technology of Japan (Grant-in-Aid for Scientific Research on Priority Area "Comparative Genomics" No. 17018029).

Conflict of interest statement. None declared.

REFERENCES

1. Mattick,J.S. (2004) The hidden genetic program of complex organisms. *Sci. Am.*, **291**, 60–67.
2. Sakakibara,Y., Brown,M., Hughey,R., Mian,I.S., Sjölander,K., Underwood,R.C. and Haussler,D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.*, **22**, 5112–5120.
3. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
4. Knudsen,B. and Hein,J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**, 446–454.
5. Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.
6. Dowell,R.D. and Eddy,S.R. (2004) Evaluation of several light-weight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 71.
7. 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.
8. 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.

9. Washietl,S., Hofacker,I.L., Lukasser,M., Hüttenhofer,A. and Stadler,P.F. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.*, **23**, 1383–1390.
10. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
11. Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.
12. Schattner,P., Decatur,W.A., Davis,C.A., Ares,M., Fournier,M.J. and Lowe,T.M. (2004) Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the Saccharomyces cerevisiae genome. *Nucleic Acids Res.*, **32**, 4281–4296.
13. Edvardsson,S., Gardner,P.P., Poole,A.M., Hendy,M.D., Penny,D. and Moulton,V. (2003) A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction. *Bioinformatics*, **19**, 865–873.
14. Yang,J.-H., Zhang,X.-C., Huang,Z.-P., Zhou,H., Huang,M.-B., Zhang,S., Chen,Y.-Q. and Qu,L.-H. (2006) snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Res.*, **34**, 5112–5123.
15. Hertel,J., Hofacker,I.L. and Stadler,P.F. (2008) SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, **24**, 158–164.
16. 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.
17. Hertel,J. and Stadler, P.F. (2006) Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, **22**, e197–e202.
18. Ng,K.L.S. and Mishra,S.K. (2007) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, **23**, 1321–1330.
19. Terai,G., Komori,T., Asai,K. and Kin,T. (2007) miRRim: A novel system to find conserved miRNAs with high sensitivity and specificity. *RNA*, **13**, 2081–2090.
20. Schölkopf,B., Tsuda,K. and Vert,J.P. (2004) *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA.
21. Saigo,H., Vert,J.-P., Ueda,N. and Akutsu,T. (2004) Protein homology detection using string alignment kernels. *Bioinformatics*, **20**, 1682–1689.
22. Sakakibara,Y., Popendorf,K., Ogawa,N., Asai,K. and Sato,K. (2007) Stem kernels for RNA sequence analyses. *J. Bioinform. Comput. Biol.*, **5**, 1103–1122.
23. Dalli,D., Wilm,A., Mainz,I. and Steger,G. (2006) STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics*, **22**, 1593–1599.
24. Rogers,A., Antoshechkin,I., Bieri,T., Blasiar,D., Bastiani,C., Canaran,P., Chan,J., Chen,W.J., Davis,P., Fernandes,J. *et al.* (2008) WormBase 2007. *Nucleic Acids Res*, **36**, D612–D617.
25. Missal,K., Zhu,X., Rose,D., Deng,W., Skogerbø,G., Chen,R. and Stadler, P. F. (2006) Prediction of structured non-coding RNAs in the genomes of the nematodes Caenorhabditis elegans and Caenorhabditis briggsae. *J. Exp. Zoolog. B. Mol. Dev. Evol.*, **306**, 379–392.
26. Deng,W., Zhu,X., Skogerbø,G., Zhao,Y., Fu,Z., Wang,Y., He,H., Cai,L., Sun,H., Liu,C. *et al.* (2006) Organization of the Caenorhabditis elegans small non-coding transcriptome: genomic features, biogenesis and expression. *Genome Res*, **16**, 20–29.
27. He,H., Wang,J., Liu,T., Liu,X.S., Li,T., Wang,Y., Qian,Z., Zheng,H., Zhu,X., Wu,T., *et al.* (2007) Mapping the C. elegans noncoding transcriptome with a whole-genome tiling microarray. *Genome Res.*, **17**, 1471–1477.
28. Haussler,D. (1999) Convolution kernels on discrete structures. *Technical Report UCSC-CRL-99-10*. Department of Computer Science, University of California at Santa Cruz.
29. Miyazawa,S. (1995) A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng.*, **8**, 999–1009.
30. Carvalho,L.E. and Lawrence,C.E. (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc. Natl Acad. Sci. USA*, **105**, 3209–3214.
31. Mückstein,U., Hofacker,I.L. and Stadler,P.F. (2002) Stochastic pairwise alignments. *Bioinformatics*, **18 (Suppl. 2)**, S153–S160.
32. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
33. Bonhoeffer,S., McCaskill,J.S., Stadler,P.F. and Schuster,P. (1993) RNA multi-structure landscapes. A study based on temperature dependent partition functions. *Eur. Biophys. J.*, **22**, 13–24.
34. Klein,R.J. and Eddy,S.R. (2003) RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4**, 44.
35. Rozen,S. and Skaletsky,H. (2000) Primer3 on the www for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
36. Fan,R.-E., Chen,P.-H. and Lin,C.-J. (2005) Working set selection using second order information for training support vector machines. *J. Mach. Learn. Res.*, **6**, 1889–1918.
37. Pacheco,P. (1996) *Parallel Programming with MPI*, Morgan Kaufmann, San Francisco, California.
38. 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.
39. Washietl,S., Pedersen,J.S., Korbel,J.O., Stocsits,C., Gruber,A.R., Hackermüller,J., Hertel,J., Lindemeyer,M., Reiche,K., Tanzer,A. *et al.* (2007) Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.*, **17**, 852–864.
40. Keerthi,S.S., Sindhwani,V. and Chapelle,O. (2007) An efficient method for gradient-based adaptation of hyperparameters in SVM models. In Schölkopf,B., Platt,J. and Hoffman,T. (eds.), *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, pp. 673–680.