# NcRNA Homology Search Using Hamming Distance Seeds

Osama Aljawad
Dept. of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
aljawado@msu.edu

Yanni Sun *
Dept. of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
yannisun@msu.edu

Alex Liu
Dept. of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
alexliu@cse.msu.edu

Jikai Lei
Dept. of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
leijikai@msu.edu

## ABSTRACT

NcRNAs play important roles in many biological processes. Existing genome-scale ncRNA homology search tools identify ncRNAs in local sequence alignments generated by conventional sequence comparison methods. However, some types of ncRNA lack strong sequence conservation and tend to be missed by conventional sequence comparison methods.

In this paper, we propose an ncRNA identification framework that is complementary to existing sequence comparison tools. By integrating a filtration step based on Hamming distance and a local structural alignment program such as FOLDALIGN, we can identify ncRNAs that lack strong sequence conservation. We introduce a coding method by which the Hamming-distance based filtration can easily distinguish transition from transversion, which show different frequency in functional ncRNAs. Our experiments demonstrate that the carefully designed *Hamming distance seed* can achieve better sensitivity in searching for poorly conserved ncRNAs than conventional sequence comparison tools.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

## General Terms

Application

## 1. INTRODUCTION

Identifying non-coding RNAs (ncRNAs), which are transcribed but not translated into protein, has drawn tremendous attention recently for two main reasons. First, besides well-known functions in protein-synthesis, regulatory roles of small ncRNAs have been revealed in gene regulation [2] in a wide variety of species. Second, new members of annotated ncRNA families or novel ncRNAs have been identified due to advances of the next-generation sequencing technologies and RNA-seq. Understanding ncRNAs plays a key role in elucidating the complexity of regulatory network of both complicated and simple organisms.

The state-of-the-art methodology for ncRNA annotation is based on comparative analysis, which searches for evolutionarily conserved ncRNAs in related genomes or their transcriptomes. Existing genome-scale ncRNA identification methods [19, 25, 18] first employ conventional sequence comparison tools such as BLAST [1] to locate an initial set of alignments for further screening. Then, features such as secondary structure conservation, minimum free energy (MFE), sequence conservation, GC content, base or basepair substitution patterns etc. [25, 15] are employed to classify these local alignments as putative ncRNAs, protein-coding genes, and other genomic features. However, although BLAST-like sequence comparison tools have been successfully used to find protein-coding genes, segment duplications, and other genomic features, they are not well suited for ncRNA search. NcRNAs function through both their sequences and structures. Some types of ncRNA evolve faster in their sequences than in their secondary structures and thus have low sequence conservation. For example, RNase P is highly structured and cannot be found by conventional sequence similarity search tools [2]. Many lineage specific ncRNAs such as Xist or Air have very low sequence conservation [17] and pose hard cases for BLAST-like tools. Even some small ncRNAs such as tRNA have a wide range of sequence conservation. Figure 1 shows the histogram of sequence similarity between homologous tRNAs in the human and mouse genomes. More than half of the homologous tRNAs have similarity below 60%.

BLAST-like sequence comparison tools tend to miss these ncRNAs for two reasons. First, genome-scale sequence comparison tools use the seed-and-extend scheme, where effi-

*corresponding author


ACM-BCB '11, August 1-3, Chicago, IL, USA

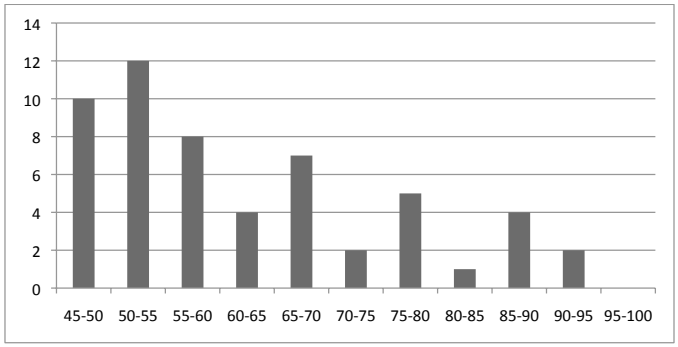


**Figure 1: The sequence similarity histogram of homologous tRNAs in the human and mouse genomes. The X-axis is the sequence similarity. The Y-axis is the number of homologous tRNA pairs. The sequences are obtained from the tRNA family in the Rfam [8] database. For each human tRNA, we report the highest sequence similarity.**

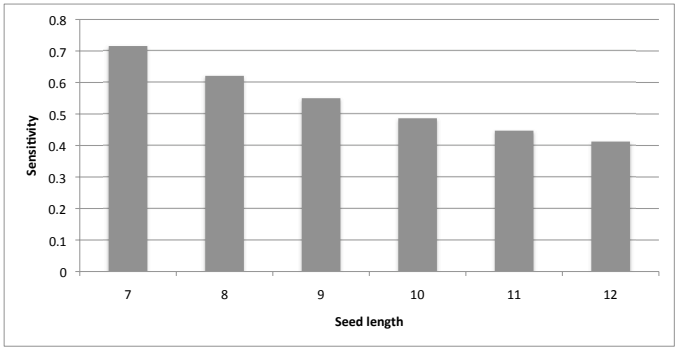


**Figure 2: The sensitivity of the BLAST seeds of different lengths on 3925 homologous ncRNAs between human and mouse genomes. X-axis is the length of seeds. Y-axis is the sensitivity.**

cient exact matching for short seeds is used as the filtration step to locate regions that are likely to be true homologs. Full dynamic programming is only applied to regions around seed hits. However, as the sequence similarity decreases, the probability that homologous features contain a match to the seed is also decreased fast. As a result, these ncRNAs will be missed in the filtration step. In order to quantify how the seeding heuristic in BLAST affects ncRNA homology search, we extracted 3925 pairs of homologous ncRNAs from the human and mouse genomes from Rfam 10 [8]. For each pair of homologous ncRNAs, we test whether they match a seed of different length. The result is summarized in Figure 2. When we use the default seed size 11 in BLAST, there are only 1755 (i.e. 45%) pairs of ncRNAs passing the filtration step. We also tested the seed in BLASTZ on the same data set. The sensitivity is 0.517. BLASTZ adopts the optimal spaced seed (1110100110010101111) designed by PattermHunter [16], but allows a transition mutation in one of matching positions. Although spaced seeds [16, 4, 22] have been used to improve BLAST's sensitivity, ncRNAs lack sequence signatures or characteristics such as the triplet amino acid code for protein coding gene detection, posing great challenges for seed design. The second problem of using BLAST-like tools for ncRNA identification is that they do not incorporate structural similarity. Deriving secondary structure on pure sequence alignment has limited accuracy. Previous work [7] has shown that the final alignments generated by BLAST and structural alignment tools such as FOLDALIGN [9, 10] can be quite different.

In order to conduct ncRNA search efficiently and accurately, we propose a new approach that integrates a more sensitive filtration step with a local structural alignment step for identifying homologous ncRNAs. The filtration step locates substrings with Hamming distance smaller than a given threshold. By carefully choosing the length and distance threshold for Hamming distance, we can locate all regions within a range of sequence similarity. In the second step, the regions generated by filtration will be folded and aligned simultaneously to maximize both sequence and structural similarity. NcRNAs that may be missed by pure sequence comparison tools have higher probability to be identified using the structural alignment programs.

We applied this approach to ncRNA homology search between intergenic regions in human and mouse genomes [24], and between the *Burkholderia cenocepacia* J2315 genome and the *Ralstonia solanacearum* genome [6]. We compared our method with BLAST and QRNA [19]. The experimental results demonstrate that our approach is efficient and is more sensitive than conventional sequence alignment tools for ncRNA search.

## 2. RELATED WORK

There are a number of ncRNA alignment tools that incorporate both sequence and structural similarity. However, most of them are based on global alignment, requiring known starting and ending positions of ncRNAs. Identifying ncRNAs in genomes or transcriptome data sets requires local ncRNA alignment. FOLDALIGN [9, 10] is a highly sensitive local structural alignment tool that can identify ncRNAs with very low sequence similarity (<40%). Using heuristics such as dynamic programming matrix pruning, FOLDALIGN is faster than the accurate implementation of the Sankoff algorithm [20]. However, it is still CPU-intensive on large data sets. When it is applied for ncRNA search between the intergenic regions of the human and mouse genomes, FOLDALIGN took about 5 months on 70 2-GB-RAM nodes in a linux cluster [24]. Thus, it is not practical to directly apply FOLDALIGN to large sequence sets. Recently, a posterior-probability based ncRNA local alignment tool PLAST-ncRNA has been implemented [5]. However, it is designed to align a relative short query sequence with a long target sequence rather than between two genomes. Thus, it cannot be directly applied to genome-scale ncRNA search without manually dividing a long genome into numerous small segments.

Because of the cost of structural alignment, existing genome-scale ncRNA search tools [19, 25, 18] still rely on conventional sequence alignment programs such as BLAST [1]. As one of seeded alignment tools, BLAST relies on its seeding heuristics to achieve efficiency of local similarity search between long genomes. Both the theoretical analysis and empirical experiments [16, 23] have shown that choice of

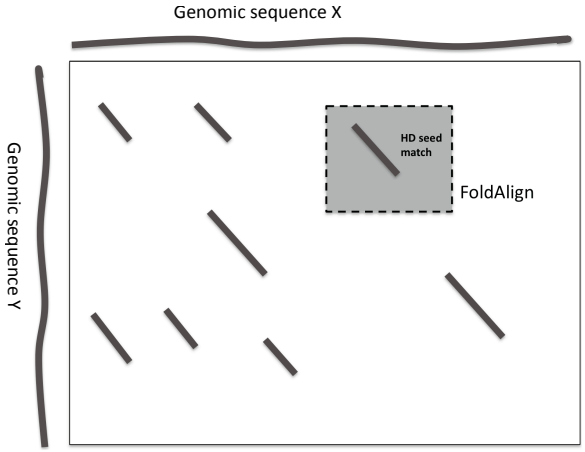


**Figure 3: The framework of genome-scale ncRNA search using HD seeds. In the first step, HD seed hits (represented by diagonal lines) are identified. Then more sensitive but slower local structural alignment tools such as FOLDALIGN are applied in the region surrounding a seed hit. Subsequent analyisis can be conducted on the output of FOLDALIGN.**

the seeding heuristics affects the sensitivity of local alignments. While BLAST requires consecutive matching, PatternHunter [16] allows spaced seeds, which can incorporate biological features of the underlying alignments. For example, spaced seeds designed for coding regions allow a mismatch following two exact matches, indicating the less strictly specified base in a codon. However, it is much more difficult to design useful spaced seeds for ncRNA search because 1) ncRNAs do not preserve strong sequence characteristics; 2) we lack enough training sequences for seed design. A more advanced seed type than spaced seed distinguishes transition and transversion as many functional genomic features including ncRNAs show a higher frequency of transition than transversion [21, 23, 11]. This type of seed is adopted by sequence comparison tool BLASTZ [21]. It uses the optimal spaced seed designed by PatternHunter but allows a transition mutation (A-G, G-A, C-T, or T-C) at any one of the inspected positions in the seed.

In our work, we design a filtration strategy based on Hamming distance. There are a number of existing implementations that search for substrings satisfying a pre-defined Hamming distance threshold. For example, in the ungapped short read mapping problem, short reads generated from next-generation sequencing platforms are aligned to the reference genome by allowing a couple of mismatches. Techniques such as neighborhood generation and the pigeon hole theory have been applied to transform inexact match to exact match in order to improve the search speed. Although a number of efficient read mapping programs [14, 13] exist, they cannot be used as the filtration step in ncRNA search because read mapping usually only allows a very small number of mismatches. In addition, they are specifically designed to align a set of short reads with a long reference genome.

# 3. METHODS

Hamming distance is the number of mismatches in two strings of equal length. Based on Hamming distance, we define *HD seeds* (Hamming distance seeds) as a 2-tuple <L, T>, where L is the length of the seed and T is the threshold. A Hamming seed <L, T> *matches* a pair of strings of equal length L if the Hamming distance between two inputs is equal to or less than $T$. According to the definition of Hamming distance, any pair of input strings of length L with sequence similarity at least $\frac{L-T}{L}$ can be matched by the HD seed <L,T>. Thus, by choosing appropriate L and T, we can use HD seed matching as the filtration step to locate possible ncRNAs with low sequence conservation. Then we extend the seed hit to both directions and apply a local structural alignment method in the vicinity of the seed hit for more sensitive ncRNA screening. The pipeline of this method is illustrated in Figure 3.

In the remaining part of this section, we first describe the coding system that can distinguish transition from transversion in Hamming distance seeds. Then we present optimal HD seed generation.

## 3.1 Design a coding system to distinguish transition from transversion

Transition mutations are less likely to result in amino acid changes. Thus, it is expected that transitions are observed at higher frequency than transversions in homologous protein-coding genes. This fact has been adopted by sequence alignment tools such as BLASTZ to improve the performance of homology search. Similar observations have been made in homologous ncRNAs as well. In the score table RIBOSUM designed by Klein and Eddy [12], transitions in both single stranded regions and between base pairs have higher scores than transversions. Higgs [11] reported that the substitution rate between a base pair (such as AU) and its double transition base pair (such as GC) is significantly higher than other mutations. Thus, it is desirable to distinguish transition from transversion in our HD seeds. However, the Hamming distance defined on DNA or RNA bases treat each mismatch equally. In order to favor transition over transversion in HD seeds, we formulate the following coding problem.

First, all bases are encoded by binary strings of equal length. Let the length be *s*. For each base x, let x.code denote the encoded binary string. Let the function *D(x,y)* be the hamming distance of x.code and y.code, where x and y are two bases. For bases A, C, G, T, we need to determine their codes such that the following equations are satisfied:

$$\begin{aligned} D(A,G) &== D(C,T); \\ D(A,C) &> D(A,G); \\ D(A,C) &== D(A,T) \\ &== D(C,G) \\ &== D(G,T); \end{aligned} \quad (1)$$

Multiple codes exist. The shortest codes for the above problem are presented in Table 1. In the coded binary strings, the distance of exact match is zero; the distance for transition is 2; the distance for transversion is 3. As a result, the Hamming distance not only depends on the number of sub-

**Table 1: Converting bases into bits**

| Base | Binary codes |
|---|---|
| A | 1111 |
| C | 0001 |
| G | 1100 |
| T(U) | 0010 |

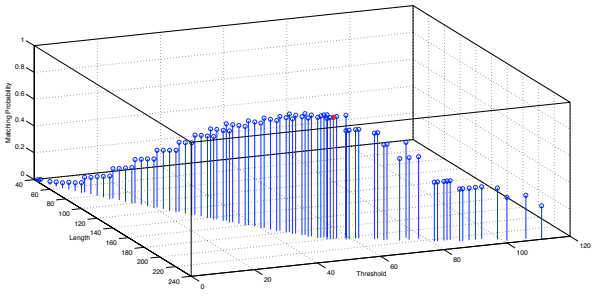


**Figure 4: Matching probabilities of HD seeds of different length L and threshold T in true ncRNA homologs. To make points distinguishable, a large number of seeds with matching probability close to zero (i.e. low sensitivity) are not shown.**

stitutions in a pair of input, but also the ratio of transition to transversion. For example, string "CCCCC" has Hamming distance 3 with both "CUCUU" and "CGCGG". After encoding, the corresponding bit strings have Hamming distance 6 and 9, respectively. Generally speaking, for two genomic sequences with equal length, if there are $x_1$ matches, $x_2$ transitions, and $x_3$ transversions, the HD distance is $2x_2+3x_3$ on two binary strings with length $4\times(x_1+x_2+x_3)$.

## 3.2 Hamming distance seed design

To design an HD seed, we need to determine L and T to maximize its matching probability in ncRNA homologs while keeping the matching probability to random sequences as low as possible. Given a pair of true ncRNA homologs, the probability that the input pair contains a match to the given HD seed is proportional to the seed's sensitivity. Given a pair of random sequences, the probability that the input pair contains a match to the given seed is proportional to the seed's false positive (FP) rate. Thus, computing the matching probability allows us to compare performance of different seeds. As there are a large number of valid combinations of L and T, an efficient method is needed for the matching probability computation. In this work, we use a simple i.i.d. model to describe distributions of exact matches, transitions, and transversions in a pair of sequences. The theoretical HD seed matching probability can be efficiently computed based on the i.i.d. model.

The i.i.d. model $\mathcal{M}$ is defined as a 3-tuple $<p_1, p_2, p_3>$, where $p_1$, $p_2$, and $p_3$ are the probabilities of exact match, transition, and transversion, respectively. Thus, $p_1 + p_2 + p_3 = 1.0$. In order to compute the matching probability of an HD seed $<L, T>$, we start with the probability that a pair of sequences of length $l$ contain $x_1$ exact matches, $x_2$

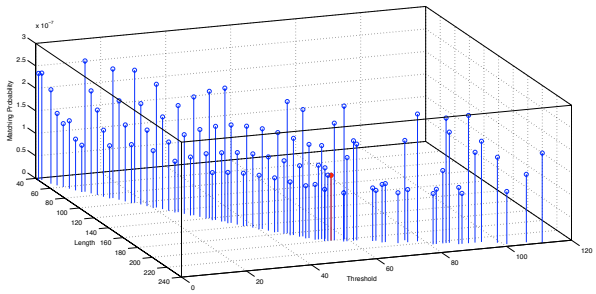


**Figure 5: Matching probabilities of HD seeds of different length L and threshold T in random sequences. To make points distinguishable, a large number of seeds with matching probability close to 1 (i.e. high FP rate) are not shown.**

transitions, and $x_3$ transversions as follows:

$$\begin{aligned} Pr^M(x_1,x_2,x_3) &= \binom{l}{x_1} p_1^{x_1} \binom{l-x_1}{x_2} p_2^{x_2} p_3^{x_3} \\ &= \frac{l!}{x_1 x_2 x_3} p_1^{x_1} p_2^{x_2} p_3^{x_3} \end{aligned} \quad (2)$$

where $l = x_1 + x_2 + x_3$. As we convert bases into binary codes according to rules in Table 1 before applying HD seed matching, the matching probability of an HD seed <L, T> can be represented using $Pr^{\mathcal{M}}(x_1, x_2, x_3)$ as below:

$$Pr^M(L,T) = \sum_{x_1+x_2+x_3=L/4;\ 2*x_2+3*x_3\le T} Pr^M(x_1,x_2,x_3) \quad (3)$$

For an HD seed <L,T>, there are multiple combinations of $x_1$, $x_2$, and $x_3$ satisfying the above equation. The matching probability must sum over all combinations. In the above equations, $l$ is the number of bases in genomic sequences and $L$ is the number of bits after coding.

The choice of L and T heavily depends on probabilities of matching and transition in $\mathcal{M}$. To compute matching probabilities in true ncRNA homologs, we train $\mathcal{M}$ on pairwise ncRNA alignments from seed families in Rfam version 10. $M$ = <0.68, 0.15, 0.16>. In order to compute HD seed matching probability in random sequences, which indicates the false positive rate, we assume that the four bases occur with the same probability. Thus, in the i.i.d. model $\mathcal{M}'$, $p_1 = 0.25$, $p_2 = 0.25$, and $p_3 = 0.5$. By applying $\mathcal{M}$ and $\mathcal{M}'$ to Eqn. 3, we can use values of $Pr^{\mathcal{M}}(L,T)$ and $Pr^{\mathcal{M}'}(L,T)$ to quantify the performance of HD seeds with different length and threshold. There are total 5551 different HD seeds with length smaller than 60 bases (i.e. 240 bits). After removing seeds which can incur FP rate near 1 or sensitivity near 0, we plot $Pr^{\mathcal{M}}(L,T)$ and $Pr^{\mathcal{M}'}(L,T)$ for the remaining seeds in Figures 4 and 5. These two figures illustrate how the seed length and threshold affect the seed's matching probabilities.

Based on the two figures, we determine L and T with the best tradeoff between $Pr^{\mathcal{M}}(L,T)$ and $Pr^{\mathcal{M}'}(L,T)$. The chosen seed is <200,55>, which is highlighted in Figures 4 and

5. Its matching probability in true ncRNA homologs is 0.906 and its matching probability in random sequences is 1.45E-07. The seed <200,55> represents a similarity $\frac{200-55}{200}$=72.5% on coded bit strings. According to the coding table 1, for genomic sequence of length 50 = 200/4, the seed <200,55> allows 26 transition and 1 transversion mutation. This combination gives the lowest DNA-level similarity 46% = (50 − 26 − 1)/50. Thus, this chosen seed is able to detect highly structured ncRNAs which have very low sequence conservation.

### 3.3 Softwares for HD seed matching and local structural alignment

There are a number of tools that can implement HD seed matching. We chose a randomized algorithm LSH-ALL-PAIRS [3], which is based on locality sensitivity hashing. Although it is an approximation algorithm, it has achieved high sensitivity in detecting DNA homologs with similarity as low as 63%. More importantly, it is fast enough to apply to whole genomes even when the allowed substitutions T increases.

For a pair of substrings that contain a match to the HD seed, we apply FOLDALIGN [9] to conduct local structural alignment. Thus, homologous ncRNAs with low sequence similarity have a higher probability to be identified.

LSH-ALL-PAIRS and FOLDALIGN were downloaded from the authors' websites [1].

## 4. EXPERIMENTS AND RESULTS

For ncRNAs with high sequence similarity, BLAST and other seeded alignment tools suffice to identify them between related genomes. The goal of our tool is to provide complementary ncRNA identification method to conventional sequence comparison tools. In this section, we focus on testing HD seeds' ncRNA search performance in data sets with low sequence conservation.

The focus of the first experiment is to search for putative structural ncRNAs in genomic regions in human that could not be aligned with mouse. Torarinsson et al. [24] directly applied FOLDALIGN for ncRNA search in a set of intergenic regions in the two genomes. Structural ncRNAs with high confidence are revealed. From the paper's website, we downloaded 1297 alignments, which have high probabilities to be functional ncRNAs. These ncRNA pairs have low sequence similarity (48% on average) and a majority of them cannot be aligned by BLAST. We apply BLAST, BlastZ, and Hamming seeds to this data set and quantify their *sensitivity* and *false positive rate* (FP rate). Sensitivity evaluates the percentage of true homologs (i.e., 1297 alignments) that can be aligned by these programs. FP rate evaluates how many pairs of random sequences can be aligned by these programs. In order to compute the FP rate, we generated 10,000 pairs of random sequences assuming each base has the same probability. The sensitivity and FP rate are summarized in Table 2. According to Table 2, HD seed has the best sensitivity and also low FP rate. BlastZ has the higher sensitivity than BLAST. This experiment shows that using HD seeds to locate possible ncRNA homologs is more sensitive than using conventional sequence comparison programs.

[1]LSH-ALL-PAIRS: http://www1.cse.wustl.edu/∼jbuhler/pgt/; FOLDALIGN: http://foldalign.ku.dk/index.html

**Table 2: Comparison of Hamming seeds, BLAST, and blastZ**

| | HD seed | BLAST | BlastZ |
|---|---|---|---|
| sensitivity | **0.6** | 0.07 | 0.17 |
| FP rate | **0.0009** | 0.0011 | 0.0054 |

**Table 3: Comparison of the HD seed hits with putative ncRNAs reported by Coenye et al.**

| | Putative ncRNAs | HD seed hits | Overlapped |
|---|---|---|---|
| Chr1 | 78 | 162311 | 78 |
| Chr2 | 116 | 14336 | 106 |
| Chr3 | 19 | 2740 | 19 |

### 4.1 NcRNA search in the *Burkholderia cenocepacia* J2315 genome

In the second experiment we focus on ncRNA identification in the *Burkholderia cenocepacia* J2315 genome by comparing it with the *Ralstonia solanacearum* genome. *Burkholderia cenocepacia* is clinically important because they can cause lung infections in cystic fibrosis (CF) patients [6]. There are multiple members in *Burkholderia cenocepacia*. Coenye et al. conducted ncRNA search by applying BLAST and QRNA between *B. cenocepacia* strain J2315 and related genomes including the *Ralstonia solanacearum* genome. As BLAST can miss highly structured ncRNAs, we conducted a complementary analysis using HD seeds and FOLDALIGN. First, we downloaded the three chromosomes (accession IDs: NC_011000, NC_011001, NC_011002) of the *Burkholderia cenocepacia* J2315 genome from NCBI. Their sizes are 3,870,082 nt, 3,217,062 nt, and 875,977 nt, respectively. Similarly we downloaded the *Ralstonia solanacearum* GMI1000 genome (NC_003295) from NCBI. The single chromosome has length 3,716,413 nt. Using BLAST and QRNA, Coenye et al. [6] reported 78, 116, and 19 putative ncRNAs on the three chromosomes of J2315.

We first masked all low-complexity repeats and annotated protein-coding genes in input sequences. Then we applied our HD seed <200,55> and FOLDALIGN between the three chromosomes and the genome of *Ralstonia solanacearum*. Between every pair of input chromosomes, the total number of possible match positions is bounded by the product of the input sequences' sizes. For example, for a seed of size 50 bases, there could be at most $(3,870,082-49)\times(3,716,413-49)$ distinct seed matching places. Thus, in general, when the sizes of input sequences increase, more seed hits are expected. The total number of seed hits and the ones that overlap with reported putative ncRNAs by Coenye et al. are summarized in Table 3. Our HD seed detected all putative ncRNAs on chromosome 1 and 3. The HD seed missed 10 putative ncRNAs on chromosome 2 because they are either masked as low-complexity repeats or heavily overlap with existing coding regions. Thus the corresponding regions are masked and will not be scanned by the HD seed. Previous literature [24] on ncRNA search suggests that most ncRNAs are in intergenic regions in bacterial genomes. It needs ex-

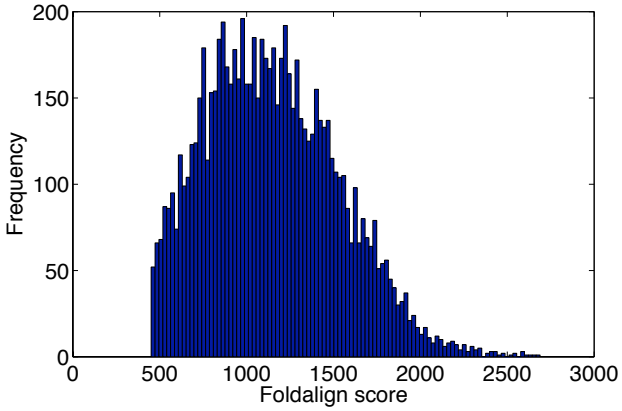


**Figure 6: The score distribution of FOLDALIGN alignments on chromosome 1.**

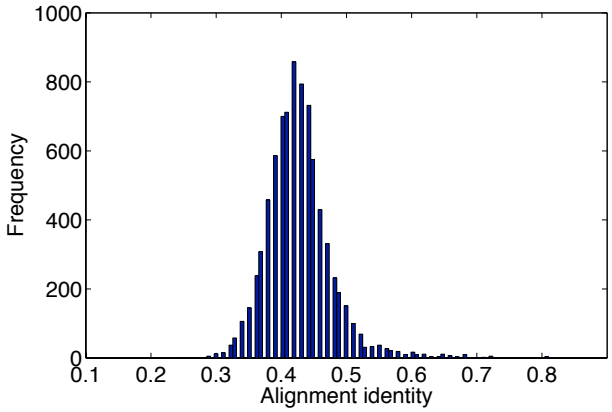


**Figure 7: The sequence identity distribution of FOLDALIGN alignments on chromosome 1.**

tensive investigation whether ncRNA genes overlap protein coding genes in bacterial genomes.

For each intergenic seed hit with identity no more than 60%, we extended it to left and right for 100 bases in each input. Then local structural alignment was conducted between extended substrings using FOLDALIGN. As chromosome 2 and chromosome 3 are much larger than chromosome 3 and may have more putative ncRNAs, we only present results of FOLDALIGN on chromosome 1 and chromosome 2. All programs run on a 128-node cluster, where each node contains 2 dual-core AMD Opterons running at 2.2GHz with 8GB of memory. The running time of HD seed matching using LSH-ALL-PAIRS is 8250 and 6850 seconds for chromosome 1 and chromosome 2, respectively. The running time of FOLDALIGN is 15 hours and 14 hours for chromosome 1 and chromosome 2, respectively.

For all FOLDALIGN output, we remove an alignment if it satisfies one of the following conditions: 1) the alignment overlaps with adjacent protein-coding genes; 2) the alignment score is smaller than 450; and 3) the alignment length is smaller than 55. We apply the above constraints

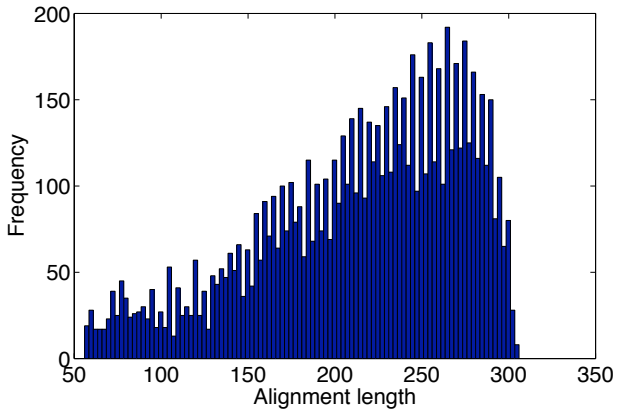


**Figure 8: The length distribution of FOLDALIGN alignments on chromosome 1.**

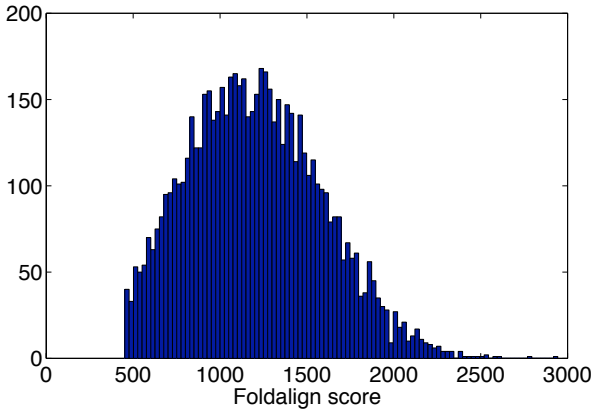


**Figure 9: The score distribution of FOLDALIGN alignments on chromosome 2.**

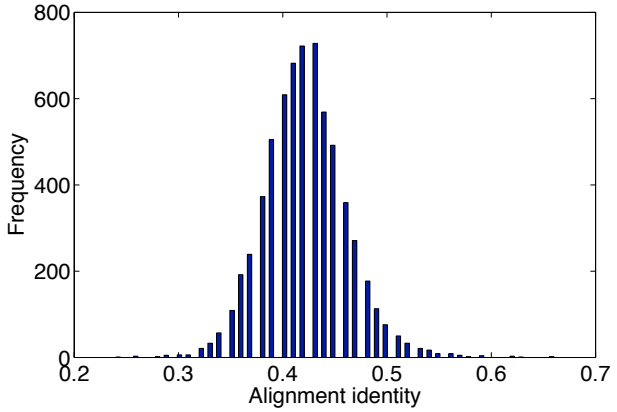


**Figure 10: The identity distribution of FOLDALIGN alignments on chromosome 2.**

**Table 4: Properties of three putative ncRNAs on chromosome 1 of J2315. All of them are conserved in R. solanacearum.**

| ID | FOLDALIGN score | Start | End | identity | p-value | 5' gene | 3' gene | 5' distance | 3' distance |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2591 | 3278580 | 3278849 | 0.52 | 0.022 | BCAL2989 | BCAL2990 | 50 | 53 |
| 2 | 2538 | 548365 | 548654 | 0.42 | 0 | BCAL0496 | BCAL0497 | 55 | 267 |
| 3 | 1792 | 3834502 | 3834660 | 0.46 | 0.3 | BCAL3494 | BCAL3495 | 125 | 85 |

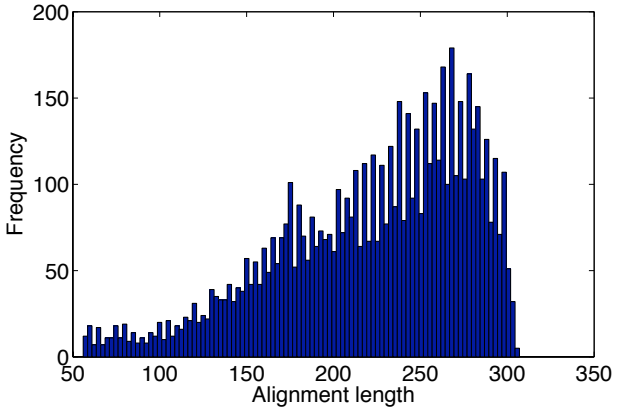


**Figure 11: The length distribution of FOLDALIGN alignments on chromosome 2.**

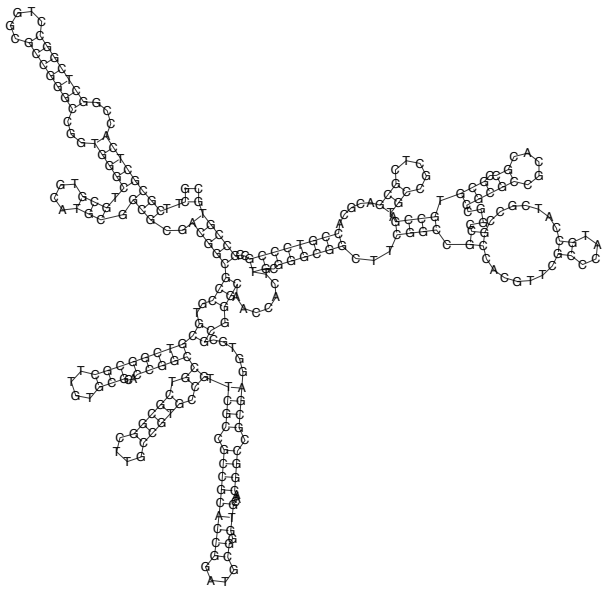

**Figure 12: The predicted secondary structure for putative ncRNA 1.**

based on the observed properties of annotated ncRNAs. After the filtration, we have 8112 and 6506 putative ncRNAs in chromsome 1 and 2, respectively. For the putative ncRNAs, we plot their FOLDALIGN scores, sequence identity, and alignment length for chromosome 1 and chromosome 2 from Figure 6 to Figure 11. The putative ncRNAs on chromosome 1 have average FOLDALIGN score around 1000, which is statistically significant according to the average FOLDALIGN scores of random sequences. They are usually highly structured and have low sequence identity. Note that although the lowest sequence identity allowed by our chosen HD seed <200,55> is 46%, FOLDALIGN is applied to bigger regions around each seed hit. As a local structural alignment, FOLDALIGN can report highly structured alignments with very low sequence conservation. This is shown in the identity distribution in Figures 7 and 10. Many of the putative ncRNAs on chromosome 1 are longer than annotated small ncRNAs. This is consistent to previous observation that small ncRNAs tend to have better sequence conservation than long ncRNAs [17].

According to the FOLDALIGN score distributions on random sequences [9], the alignments on the two chromosomes indicate strong sequence and structural conservation. Although extensive experiments are needed to evaluate whether they are functional ncRNAs, these alignments provide a promising set of putative ncRNAs. Figures 12, 13, and 14 show the secondary structures of three putative ncRNAs. Their properties including their positions, length, distance to adjacent protein-coding genes etc. are presented in Table 4.

## 5. CONCLUSIONS

Our experimental results show that using HD seed matching is an effective and efficient filtration step for genome-scale ncRNA search. Compared to conventional sequence comparison tools, HD seed matching is more sensitive in identifying ncRNAs with low sequence conservation. By designing a long HD seed, we can control the matching probability to random sequences. Thus, integrating HD seed matching and a sensitive local structural alignment tool provides a complementary ncRNA search method to existing sequence alignment-based implementations. Besides FOLDALIGN, other local ncRNA structural alignment tools or classification method that integrates more features can be applied to further examine HD seed hits.

We plan to apply this method to ncRNA identification in available transcriptome datasets. It has been reported that a large portion of transcript reads generated by RNA-seq cannot be mapped to annotated features such as protein-coding genes. It is unknown whether those reads are from functional ncRNAs. Our tool can be used to examine whether

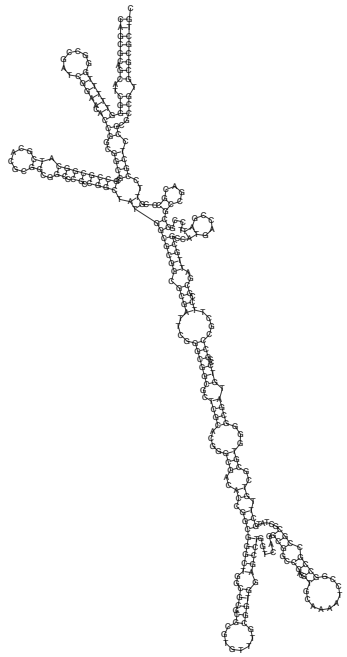

**Figure 13: The predicted secondary structure for putative ncRNA 2.**

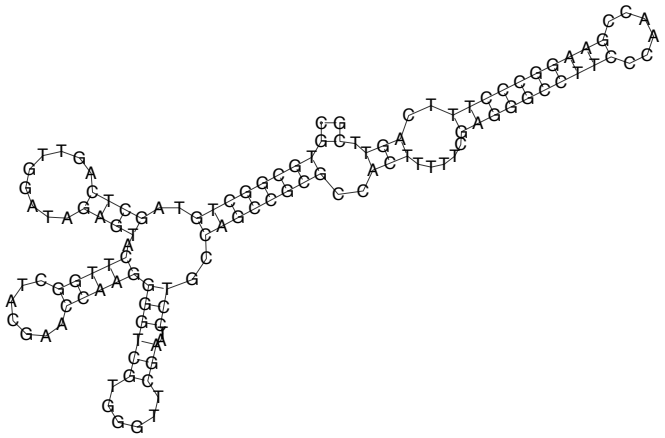

**Figure 14: The predicted secondary structure for putative ncRNA 3.**

the transcribed regions have structural conservation in related genomes when BLAST-like tools fail. We also plan to integrate more biological features to remove hits that are not likely to be ncRNAs.

## Acknowledgments

This work was supported, in part, by the NSF CAREER Grant DBI-0953738.

## 6. REFERENCES

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[2] A. F. Bompfunewerer, C. Flamm, C. Fried, G. Fritzsch, I. L. Hofacker, J. Lehmann, K. Missal, A. Mosig, B. Muller, S. J. Prohaska, B. M. Stadler, P. F. Stadler, A. Tanzer, S. Washietl, and C. Witwer. Evolutionary patterns of non-coding RNAs. *Theory in Biosciences*, 123(4):301 – 369, 2005.

[3] J. Buhler. Efficient large-scale sequence comparison by locality-sensitive hashing. *Bioinformatics*, 17(5):419–428, 2001.

[4] J. Buhler, U. Keich, and Y. Sun. Designing seeds for similarity search in genomic DNA. In *Proceedings of the seventh annual international conference on Computational molecular biology*, pages 67–75. ACM Press, 2003.

[5] S. Chikkagoudar, D. R. Livesay, and U. Roshan. PLAST-ncRNA: Partition function Local Alignment Search Tool for non-coding RNA sequences. *Nucleic Acids Research*, 38(suppl 2):W59–W63, 2010.

[6] T. Coenye, P. Drevinek, E. Mahenthiralingam, S. A. Shah, R. T. Gill, P. Vandamme, and D. W. Ussery. Identification of putative noncoding RNA genes in the Burkholderia cenocepacia J2315 genome. *FEMS Microbiology Letters*, 276(1):83–92, 2007.

[7] P. Gardner and R. Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5(1):140, 2004.

[8] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, 33(database issue):D121–D124, 2005.

[9] J. H. Havgaard, R. B. Lyngso, G. D. Stormo, and J. Gorodkin. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, 21(9):1815–1824, 2005.

[10] J. H. Havgaard, E. Torarinsson, and J. Gorodkin. Fast Pairwise Structural RNA Alignments by Pruning of the Dynamical Programming Matrix. *PLOS computational biology*, 3(e193), 2007.

[11] P. G. Higgs. RNA secondary structure: physical and computational aspects. *Quarterly Reviews of BioPhysics*, 33(3):199–253, 2000.

[12] R. Klein and S. Eddy. Rsearch: Finding homologs of single structured rna sequences. *BMC Bioinformatics*, 4(1):44, 2003.

[13] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[14] R. Li, Y. Li, K. Kristiansen, and J. Wang. Soap: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, 2008.

[15] Z. J. Lu, K. Y. Yip, G. Wang, C. Shou, L. W. Hillier, E. Khurana, A. Agarwal, R. Auerbach, J. Rozowsky, C. Cheng, M. Kato, D. M. Miller, F. Slack, M. Snyder, R. H. Waterston, V. Reinke, and M. B. Gerstein. Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res.*, 21:276–285, 2011.

[16] B. Ma, J. Tromp, and M. Li. PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18(3):440–445, March 2002.

[17] K. C. Pang, M. C. Fritha, and J. S. Mattick. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in Genetics*, 22(1):1–5, 2005.

[18] J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, W. Miller, and D. Haussler. Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Comput Biol*, 2(4):e33, 2006.

[19] E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2(1):8, 2001.

[20] D. Sankoff. Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825, 1985.

[21] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. Human-mouse alignments with BLASTZ. *Genome Research*, 13:103–107, 2003.

[22] Y. Sun and J. Buhler. Designing multiple simultaneous seeds for DNA similarity search. In *Proceedings of the eighth annual international conference on Computational molecular biology(RECOMB '04)*, pages 76–84. ACM Press, 2004.

[23] Y. Sun and J. Buhler. Choosing the best heuristic for seeded alignment of DNA sequences. *BMC Bioinformatics*, 7(1):133, 2006.

[24] E. Torarinsson, M. Sawera, Havgaard, M. Fredholm, and J. Gorodkin. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Research*, 16:885–889, 2006.

[25] S. Washietl, I. L. Hofacker, and P. F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A*, 102(7):2454–2459, 2005.