QLZCClust: Quaternary Lempel-Ziv Complexity based Clustering of the RNA-seq Read Block Segments

Ashis Kumer Biswas, Baoju Zhang, Xiaoyong Wu and Jean X. Gao

Abstract— **The Next Generation Sequencing platform, RNA-seq provides quantitative expression data that exhibit distinctive sequence patterns in the segments of the short-reads level and are found useful in clustering of those segments. However, the result does not reflect the functional chemistry of the non-coding RNAs (ncRNAs). The functions of the ncRNAs are deeply related to their secondary structures. Thus by exploring the clustering in terms of structural profiles of the read block segments rather than their sequence patterns would be essential and useful. We proposed the QLZCClust (Quaternary Lempel-Ziv complexity based Clustering) method which is an extension to the popular Lempel-Ziv algorithm to compute pairwise secondary structure distance. We applied QLZCClust on the short-read segments obtained from the RNA-seq experient and found that it can separate most miRNAs and the tRNAs. Moreover, it can be used to detect structural similarities among different classes of ncRNAs. We compared our algorithm with the clustering of two other structural distance measures – SimTree edit distance and RNAz based distance, and found that our method performs superior.**

I. Introduction

RNA-seq is a revolutionary technology for profiling transcriptomes with which a very precise measurement of expression levels of transcripts and their isoforms can be accomplished [1]. The expression data reveal a vast number of possibilities to look into the transcriptomic details of an organism that may or may not have a well-studied genome. In most cases the reference genome is available and the analysis pipeline starts by mapping the RNA-seq short-read sequences to the genome [2]. In this study we limited our focus on the RNA-seq experiments that dealt with the microRNAs, the central topic for many therapeutic researches. The microRNA like small non-coding RNAs are produced from microRNA precursors, other structured RNAs and the dicers [3], [4].

The diversity of processing pathways urges the researchers exploring how the short read patterns in RNA-seq datasets relate to the processing of particular non-coding RNAs. For instance, the characteristic mutual positioning with a 3'-overhang of miR and miR* products that is a characteristic feature for dicer cleavage, the anomalous 5'-overhang observed for some microRNAs resulting from a distinct, dicer-dependent two-step mechanism, and the dicer-independent processing of mir-451 [5]. Therefore, we seek for profiles to represent distinct pathways.

Ashis Kumer Biswas and Jean X. Gao are with Department of Computer Science and Engineering, University of Texas at Arlington, Texas 76019, The United States of America ashis.biswas@mavs.uta.edu; gao@uta.edu

Baoju Zhang and Xiaoyong Wu are with School of Physics and Electronic Information, Tianjin Normal University, Tianjin, 300387, China

There are several studies on the identification of the profiles. The deepBlockAlign [5] introduced a two-step approach to align RNA-seq read patterns with the aim of quickly identifying RNAs that share similar processing footprints. Overlapping mapped reads are first merged to blocks and then closely spaced blocks are combined to block groups, each representing a locus of expression. In the second stage, block patterns are compared by means of a modified Sankoff algorithm that takes both block similarities and similarities of pattern of distances within the block groups into account. Hierarchical clustering of the block groups separated most miRNA and tRNA, and also identified about a dozen tRNAs clustering together with miRNA. However, the method only used RNA-seq short read sequence patterns in both the stages and did not consider any similarity measure in terms of the secondary structure. But, this structural similarity measurement is important because most of the RNAs in the dataset are structural. We, therefore, introduce QLZCClust, a hierarchical clustering of the same block groups, but instead of doing the sequence based distance calculation of the segment pairs, we propose a new pairwise structural distance measure – quaternary Lempel-Ziv complexity which is able to enhance the clustering results in the structural domain.

In the following section, we describe in detail the proposed QLZCClust scheme along with dataset preparation step. In section III we summarize our experiment results and comparing our performance with clustering results by different secondary structural distance metrics. Finally, in section IV we conclude our manuscript with a guide to future research direction.

II. Methods

A. Dataset Preparation

We worked with the same benchmark Illumina sequencing datasets used by "deepBlockAlign" [5] authors which is the Human_eb dataset [6]. At first, the short-read sequences from the RNA-seq experiment were mapped onto the reference genome using "segemehl" [7]. The mapped reads are then divided into blocks of consecutive reads using the "blockbuster" tool [8] (with parameters: `distance=30`, `minBlockHeight=1`, `minClusterHeight=50`, `scale=0.5`). The expression filter [5] was applied on the read block groups (segments) to keep only those segments having more than one block, at least 50 reads per group and the size range between 50 nt and 200 nt. At the end, 455 RNA read block segments remained.

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

B. Pairwise Structural Distance of Read Block Segments

We extended the Lempel-Ziv (LZ) sequence comparing algorithm [9] to compute pairwise secondary structural distance between two RNA segments. In the following section we discuss LZ-complexity based structural distance calculation of two structures if the structures are represented using only two symbols (i.e., the simpler binary case). Following this section we introduce our Quaternary LZ-complexity based structural distance measurements.

1) Binary LZ-Complexity: Before aligning two RNA structures, we first converted the secondary structure of each transcript from bracket notation to dot plot representation. A dot plot is a two-dimensional graph in which there is a dot (or symbol "1") at position (i,j) if base at position i pairs with the base at position j, otherwise there is no dot (or symbol "0"). Fig. 1 shows both the predicted secondary structure and corresponding dot plot representation of the read block segment 405. In the dot plot, if scanned downward diagonally from left to right fashion and stopping at the symmetric border line and re-scan from the next column or row, we will get a binary sequence of 0s and 1s. In the binary sequence a block of consecutive 1s represent a stem of the secondary structure and block of consecutive 0s between two stems represent loop. We further replaced each block of 0s by a single "0" for simplicity. Thus, the characteristic binary sequence for the structure of segment 405 is "0111101011111110".

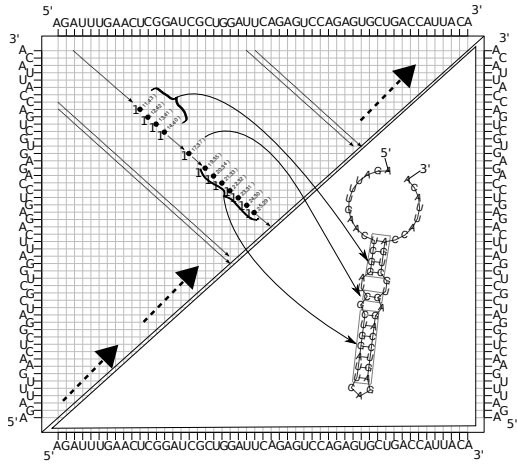

Fig. 1. Dot plot representation of the secondary structure of the read block segment (ID: 405). The lower right triangle contains the secondary structure and the upper left triangle is its dot plot representation. The mapping of the stems (consecutive base-pairs in the structure) are shown using the arrows from the dot plot to the secondary structure plot. The scanning direction starts from the lower left part of the upper triangle to its upper right part (shown as the dotted triangular arrow heads).

LZ algorithm was introduced to analyze the complexity of linear sequences by Lempel and Ziv in 1976 [9]. The LZ complexity of a finite sequence is related to the number of steps required by a production process that builds the original sequence. Let S, Q and R be sequences defined over an alphabet Σ, where $|S|$ denotes number of symbols in sequence S, $S[i]$ denotes the i^{th} symbol of sequence S and $S[i:j]$ denotes the substring of S composed of the elements of S between index i and j (inclusive). An extension $R = SQ$ of S is reproducible from S (denoted $S \rightarrow R$) if there exists an integer $p \leq |S|$ such that $Q[k] = R[p+k-1]$ for $k = 1,\ldots,|Q|$. That is, R can be obtained from S by copying elements from the p^{th} location in S to the end of S. As each copy extends the length of the new sequence beyond $|S|$, the number of symbols copied can be greater than $|S|-p+1$. Thus, this is a simple copy process of S starting from position p, which can carry over to the added part, Q. For example, 011100 $\rightarrow$ 011100110 with $p = 3$, 011100 $\rightarrow$ 01110011 with $p = 2$ and 011100 $\rightarrow$ 0111000111 with $p = 1$.

A sequence S is producible from its prefix $S[1:j]$, which is denoted by $S[1:j] \Rightarrow S$, if $S[1:j] \rightarrow S[1:|S|-1]$. For instance, 01001 $\Rightarrow$ 010010011 and 01001 $\Rightarrow$ 0100100100 with $p = 3$. Thus, the production allows an extra different symbol at the end of copy process which is not permitted in reproduction.

Any sequence S can be built using a production process where at its i^{th} step $S[1:h_{i-1}] \Rightarrow S[1:h_i]$, assuming an empty symbol produces the first symbol of S. An m-step production process of S results in a parsing of S in which $H(S) = S[1:h_1]\cdot S[h_1+1:h_2]\cdot\ldots\cdot S[h_{m-1}+1:h_m]$ is called the history of S and $H_i(S) = S[h_{i-1}+1:h_i]$ is called the i^{th} component of $H(S)$. For instance, $0\cdot1\cdot1\cdot0\cdot1\cdot0\cdot1\cdot1\cdot0\cdot0\cdot0\cdot1$ and $0\cdot1\cdot10\cdot1011\cdot00\cdot01$ are two different histories of the sequence 011010110001. If $S[1:h_i]$ is not reproducible from $S[1:h_{i-1}]$, then $H_i(S)$ is called exhaustive history. A history is called exhaustive if each of its components (except possibly the last one) is exhaustive. For example, the second history of the sequence 011010110001 is an exhaustive history. Moreover, every sequence S has a unique exhaustive history [9].

Let $c_H(S)$ be the number of components in the history of S. Then the LZ complexity of S is $c(S) = \min\{c_H(S)\}$ over all histories of S. It can be shown that $c(S) = c_E(S)$, where $c_E(S)$ is the number of components in the exhaustive history of S.

Now given two sequences Q and S, and SQ is the concatenation of S and Q. By definition, the number of components needed to build Q when appended to S is $c(SQ)-c(S)$. This number will be less than or equal to $c(Q)$ because at any given step of the production process of Q (in building the sequence SQ), we will be using a larger search space due to the existence of S. Therefore the copying process can only be longer which in turn would reduce the number of exhaustive components. This can also be seen from the additivity of the LZ complexity: $c(SQ) \leq c(S)+c(Q)$. Thus, the quantity that how much $c(SQ)-c(S)$ is less than $c(Q)$ denotes the degree of similarity between S and Q. For instance, we are given three sequences $S = 0111101110111110$, $R = 0110110111111110$ and $Q = 0111011110111110$. The exhaustive histories of the three sequences are:

$$H_E(S) = 0\cdot1\cdot1110\cdot11101111\cdot10$$
$$H_E(R) = 0\cdot1\cdot10\cdot110111\cdot111110$$
$$H_E(Q) = 0\cdot1\cdot110\cdot1111\cdot0111110$$

Thus we can see that $c(S) = c(R) = c(Q) = 5$. The exhaustive histories of the sequences SQ and RQ would be:

$$H_E(SQ) = 0 \cdot 1 \cdot 1110 \cdot 11101111 \cdot 100 \cdot 111011110 \cdot 111110$$
$$H_E(RQ) = 0 \cdot 1 \cdot 10 \cdot 110111 \cdot 111110 \cdot 01110 \cdot 111101 \cdot 11110$$

Here, $c(SQ) = 7$ and $c(RQ) = 8$. It took one more step in the production process of RQ than SQ. The reason behind is because Q is close to S than R. In this example, we can see that S and Q share patterns 1111, 111 and 11111. We can forumate the number of steps it takes to generate a sequence Q from a sequence S by $c(SQ) - c(S)$. Thus if S is closer to Q than R then we would expect $c(SQ) - c(S)$ to be smaller than $c(RQ) - c(R)$.

There are several distance measures between two linear sequences S and Q defined by Otu el.[10]. In our study we used the following normalized distance function $d(S,Q)$:

$$d(S,Q) = \begin{cases} \dfrac{c(SQ) - c(S) + c(QS) - c(Q)}{\frac{1}{2}\left[c(SQ) + c(QS)\right]} & \text{if } Q \neq S \\ 0 & \text{otherwise.} \end{cases}$$

2) Quaternary LZ-Complexity: The Binary LZ complexity based distance measure does not consider base compositions into account in the stem sites, that is, it treats characteristic sequences of AU or UA, GC or CG, GU or UG pairs without their order of occurrences. In Quaternary LZ complexity, we took the order of the base-pair compositions into consideration. We prepared the dot plot from the secondary structure in the same way as was prepared in the binary case, except that we assign in the $(i,j)^{th}$ cell a 1 if (i,j) base pair is a AU or UA, a 2 if it is a GC or a CG base pair, a 3 if it is a GU or a UG base pair, and otherwise 0 to represent a no base pair. Then we extracted the characteristic sequence out of the dot plot, and applied LZ-complexity algorithm to deduce the pairwise normalized distance score between two structures.

C. Clustering the Read-Block Segments

The structural distance measure, QLZC (Quaternary Lempel-Ziv Complexity distance) defined in the previous section was applied on the 455 read block segments to perform a hierarchical clustering to find structural similarities among different classes of non-coding RNAs, and explore whether the method is capable of separating major classes of structured RNAs. In the hierarchical clustering, we employed the complete linkage as the agglomeration method.

D. Evaluating the clusters

The read block segments are compared to known annotation of non-coding RNAs for overlaps. The known annotation data sources we used are – (i) miRBase for microRNA loci, (ii) tRNA loci from gtRNAdb, (iii) snoRNA loci from UCSC annotation. Among the 455 read block groups, 437 had overlap with eight classes of non-coding RNA – (i) miRNA, (ii) tRNA, (iii) rRNA, (iv) scRNA, (v) snRNA, (vi) C/D box snoRNA, (vii) H/ACA box snoRNA and (viii) scaRNA. The remaining 18 segments were not annotated in any of the known data sources. The Rand Index[11] between the ground truth clustering T of the given RNA segments and a clustering result R can be employed for validation. Suppose, TP represents the number of pairs of segments that are in the same cluster in T and also in the same cluster in R, TN denotes the number of segment pairs that are in different clusters in T and also in different clusters in R, FP represents the number of segment pairs that are in different clusters in T set, but are in the same cluster in the R set. Finally, the FN represents the number of segment pairs that are in the same clusters in the T set, but are in the different cluster in the R set. Thus, the Rand Index is defined in Equation 1.

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

The Rand Index lies between 0 and 1. The index scores close to 1 indicate better performance of the clustering algorithm, because it indicates how close the clustering result is to the ground truth.

III. Results and Discussion

Fig. 2 illustrates the dendrogram after applying the hierarchical clustering algorithm of the 455 block segments from the Human_eb dataset based on the Quaternary LZ distance measure. It can be seen in the dendrogram that our proposed distance measure can clearly separate two broad classes of non-coding RNAs – tRNA and miRNA. We highlighted these two branches with blue and red boxes respectively.

There are some significant clusters having tRNAs, snoRNAs or unannotated segments clustering together with microRNAs. An earlier study discovered the fact that there is a set of individual and characteristic tRNA-derived fragments which are actively derived from mature tRNAs by specific endonucleotic cleavage or exonuclease digestion by a number of enzymes [5], [12]. In addition, it was shown that dicer-dependent small tRNA fragments, along with other small RNAs from a number of non-miRNA sources, can potentially bind to Argonaute complexes and thereby unfold trans-silencing capacities [4]. Therefore, we examined fifteen tRNAs that clustered within the microRNA cluster. By taking a closer look at these candidates, we identified three of these have been reported in literature[12].

We performed hierarchical clustering with complete linkage using several other structural distance measures. We applied the SimTree edit distance measure by Eden et al. [13], which takes into account secondary structure similarities in addition to sequence similarities. It first transforms the given two RNA secondary structures into labeled trees and then computes the distance between the two trees resulting in a similarity score. We also applied RNAz 2.0 [14] each of the pairs and obtained the mean z-score and structure conservation index (SCI). We first used the SCI as the distance metric, then combined the mean z-score and the the SCI score to deuce a new distance score which is equal to $\{(1 - \text{mean z-score}) + \text{SCI}\}$. The final score lies between 0 and 2, where a pairwise structure distance score closer to 2 indicates high similarity between the two structures.

In order to investigate the separability of the clustering based on these pairwise structural distance measures, we

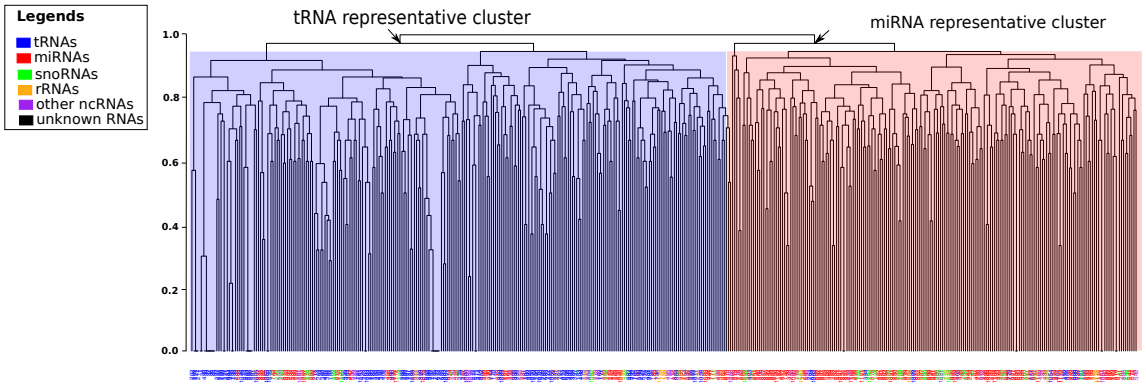

Fig. 2. Hierarchical clustering results on the 455 block segments based the QLZ distance measure.. The leaf labels are shown in different colors and we put the color legends on the top-left corner. It can be found that if we cut the dendrogram to get two clusters, we can clearly separate two broad classes of non-coding RNAs, namely miRNAs and tRNAs.

TABLE I

RAND INDICES OF DIFFERENT HIERARCHICAL CLUSTERINGS

Distance Metric Used	**Rand Index**	***n(miRNAs)* (of 193)**	***n(tRNAs)* (of 157)**
Binary LZ distance	0.19	41	113
SimTree edit distance[13]	0.22	37	129
RNAz SCI distance[14]	0.25	100	76
RNAz mean-z score & SCI based distance[14]	0.28	99	88
Quaternary LZ distance	**0.57**	**127**	**138**

computed the Rand Indices for each of the clustering results focusing on how well each can separate the two broad classes of transcript segments – miRNAs and tRNAs. Table I shows the Rand Indices for each of the clusterings along with how many miRNAs and tRNAs could these clusterings group correctly in the "*n(miRNA)*" and "*n(tRNA)*" columns respectively. Here we can see that QLZC based distance measure that we proposed here performs better than the other distance measures (127 out of 193 miRNAs, 138 out of 157 tRNAs were grouped correctly, thereby the accuracy of our QLZCClust is about 76%, with a rand index 0.57).

IV. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

In this manuscript we presented an alignment-free pairwise secondary structure distance metric – Quaternary Lempel-Ziv Complexity distance which can be used in the hierarchical clustering of the RNA-seq read block segments to partition them in structural dimension. It opens up an opportunity to identify structural similarities among different classes of structural RNAs and essentially will help the researchers to retrieve structural motifs from the given set.

B. Future Works

There are several future research directions worth pursuing. The proposed distance measure could be employed in clustering to effectively partition different classes of non-coding RNAs transcripts and also to identify novel non-coding RNAs.

REFERENCES

[1] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[2] C. Trapnell, L. Pachter, and S. L. Salzberg, "Tophat: discovering splice junctions with RNA-seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[3] R. J. Taft, E. A. Glazov, T. Lassmann, Y. Hayashizaki, P. Carninci, and J. S. Mattick, "Small RNAs derived from snoRNAs," *RNA*, vol. 15, no. 7, pp. 1233–1240, 2009.

[4] A. M. Burroughs, Y. Ando, M. L. de Hoon, Y. Tomaru, H. Suzuki, Y. Hayashizaki, and C. O. Daub, "Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin," *RNA Biology*, vol. 8, no. 1, pp. 158–177, 2011.

[5] D. Langenberger, S. Pundhir, C. T. Ekstrøm, P. F. Stadler, S. Hoffmann, and J. Gorodkin, "deepblockalign: a tool for aligning RNA-seq profiles of read block patterns," *Bioinformatics*, vol. 28, no. 1, pp. 17–24, 2012.

[6] R. D. Morin, M. D. OConnor, M. Griffith, F. Kuchenbauer, A. Delaney, A.-L. Prabhu, Y. Zhao, H. McDonald, T. Zeng, M. Hirst *et al.*, "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells," *Genome Research*, vol. 18, no. 4, pp. 610–621, 2008.

[7] S. Hoffmann, C. Otto, S. Kurtz, C. M. Sharma, P. Khaitovich, J. Vogel, P. F. Stadler, and J. Hackermüller, "Fast mapping of short sequences with mismatches, insertions and deletions using index structures," *PLoS Computational Biology*, vol. 5, no. 9, p. e1000502, 2009.

[8] D. Langenberger, C. Bermudez-Santana, J. Hertel, S. Hoffmann, P. Khaitovich, and P. F. Stadler, "Evidence for human microRNA-offset RNAs in small RNA sequencing data," *Bioinformatics*, vol. 25, no. 18, pp. 2298–2301, 2009.

[9] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 75–81, 1976.

[10] H. H. Otu and K. Sayood, "A new sequence distance measure for phylogenetic tree construction," *Bioinformatics*, vol. 19, no. 16, pp. 2122–2130, 2003.

[11] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[12] Y. S. Lee, Y. Shibata, A. Malhotra, and A. Dutta, "A novel class of small RNAs: tRNA-derived RNA fragments (tRFs)," *Genes & Development*, vol. 23, no. 22, pp. 2639–2649, 2009.

[13] E. Eden, I. Wallach, and Z. Yakhini. SimTree: A Tool for Computing Similarity Between RNA Secondary Structures. [Online]. Available: http://bioinfo.cs.technion.ac.il/SimTree/

[14] A. R. Gruber, S. Findeiß, S. Washietl, I. L. Hofacker, and P. F. Stadler, "RNAz 2.0: Improved noncoding RNA detection," in *Pacific Symposium on Biocomputing*. World Scientific Publishing, 2010, pp. 69–79.