# Alignment-free Sequence Comparison based on NGS Short-reads Neighbor Search

Phanucheep Chotnithi
SOKENDAI
The Graduate University for Advanced Studies
Tokyo, Japan
Email: phanucheep@nii.ac.jp

Atsuhiro Takasu
National Institutes of Informatics
Tokyo, Japan
Email: takasu@nii.ac.jp

***Abstract*—Next-generation sequencing (NGS) is becoming the mainstream format for genome-sequence data and creates new challenges in genome-sequence comparison. The multiple-sequence alignment approach is not suited to NGS data because of short-read assembly and computational resource problems. Therefore, alignment-free methods are needed for comparisons involving NGS data. Most alignment-free methods rely on $k$-mer-based distance measures. However, the characteristics of NGS data mean that $k$-mer-based alignment-free methods might not be optimal. NGS data contain substantial amounts of overlap among the NGS reads, which will affect the distances between the NGS sets for each input species as calculated by these methods. We propose a novel alignment-free sequence-comparison method, based on the number of neighbors in the NGS data, which aims to reduce the effect of the NGS-read overlap. We performed experiments that compared the proposed method with two existing methods. The results show that our method can distinguish the differences between diverse species better than the compared methods. Moreover, our method performs NGS data comparisons while showing robustness with respect to the $k$ parameter, in contrast to the compared methods.**

***Keywords–NGS; Phylogeny; Sequence comparison; Alignment-free***

## I. Introduction

Next-generation sequencing (NGS) is a method used to transform data from genome samples to a digitized data sequence, achieving a rapid throughput compared with traditional sequencing processes. Instead of one long sequence of genome data, NGS produces large numbers of sequence fragments called *reads* per genome sample. NGS can be applied to many biological problems, including *de novo* whole-genome sequencing and RNA-seq [1].

For most genome-sequence analysis applications, *short-read* data is a new challenge [2], where sequence comparison and phylogeny analysis are issues that we are interested in. Normally, sequence-comparison algorithms use one long genome sequence, such as 16S rRNA in mitochondrial DNA (mtDNA), and the whole genome when measuring the distance between sequences [3], [4], [5]. Clustering and classification algorithms are applied via distance matrices to produce a phylogenetic tree from which evolutionary relationships among species can be inferred. The emergence of NGS short-read methods, with their new form for genome sequences, will challenge the approach to genome-sequence analysis. In fact, existing methods and algorithms are no longer efficient for this new type of genome data [6].

The traditional method for sequence comparison is the multiple-sequence alignment (MSA) method, which has trouble dealing with a large proportion of NGS short-read data. Its approach is to reconstruct the short reads into one long sequence. In a process called *assemble*, NGS reads are mapped onto the template sequence, which involves significant computational cost. To assemble the genome without template sequences is very challenging because the reads are mostly short and contain large numbers of repeated genome data. Recently, the alignment-free method for sequence comparison has attracted attention from researchers because of its processing efficiency compared with the alignment-based method [6]. This method does not require an assembly process, and is therefore scalable to large numbers of NGS short reads, which avoids the main problem with MSA. Most alignment-free methods rely on $k$-mer frequencies as the sequence profile used to measure the distance between profiles [5]. However, alignment-free methods remain less accurate than MSA.

Several studies have proposed techniques that focus specifically on NGS short-read data. $CVTree$ [7], [8] and $d_2^S$ [9] have shown good results for distance measurements and phylogeny reconstruction with both NGS data and long genome sequences. For a given $k$, $CVTree$ and $d_2^S$ calculate the distance between two NGS samples (or two DNA sequences) based on the normalized $k$-mer frequencies. Because these methods rely on $k$-mers, we need to consider the random overlaps between NGS short reads. These overlaps affect the frequency of occurrence of $k$-mers within NGS sets, which could lead to an inaccurate distance matrix. The random overlaps between NGS short reads can cause differences between the $k$-mer frequency profiles of any two NGS sets obtained from the same species sample.

In this paper, we propose an assembly-free and alignment-free sequence-comparison method for NGS data called $d^{NS}$. The main aim of $d^{NS}$ is to reduce the effect of the overlap among NGS short reads in sequence comparisons. By grouping similar short reads together, we can assume that reads sharing the same overlap are likely to fall into the same group. Using a statistical assessment of the number of short reads included in the neighbor search with a set of queries, the method provides information about the similarity between NGS sets. We performed experiments with two simulated NGS datasets. According to the results using 29 mammalian mtDNA sequences [10], [11], $d^{NS}$ performed well when reconstructing the phylogenetic tree of a diverse-species dataset, which indicates that $d^{NS}$ can achieve sequence comparisons using NGS data. For a 29-member $Escherichia/Shigella$ whole-genome dataset [12], $d^{NS}$ outperformed $d_2^S$ and matched the performance of $CVTree$. In addition, the results showed that

$d^{NS}$ is more robust with respect to various values for $k$ than $d_2^S$ and $CVTree$, which indicates that $d^{NS}$ is robust against the effects of NGS short-read overlap on the $k$-mer frequency distribution. Because this neighbor-search-based alignment-free approach to sequence comparison is novel, there is plenty of scope for further development and possible improvements.

## II. Background and Related Work

Two $k$-mer frequency-based alignment-free methods are considered in this paper, namely $CVTree$ [7], [8] and $d_2^S$ [9]. Both $CVTree$ and $d_2^S$ focus on normalized $k$-mer frequencies. The difference is that $CVTree$ calculates the distance between two genome sequences or NGS short-read sets by using their normalized $k$-mer frequency vector, called the composite vector (CV), whereas $d_2^S$ is a statistical approach to modifying raw distance measures to produce measures better suited to NGS data.

### A. $CVTree$: CV alignment-free method

The $CVTree$ process is as follows. For a fixed length $k$, count separately the number of substrings of length $k$, $k-1$, $k-2$ on each input sequence. The initial CV is the number of $k$-mer items, which is $N = 4^k$ total dimensions for DNA sequences and $N = 20^k$ for protein sequences in lexicographic order. Calculate the $subtraction\ score$ for the $k$-mer:

$$a_i(\alpha_1\alpha_2...\alpha_k) \equiv \frac{f(\alpha_1\alpha_2...\alpha_k) - f^0(\alpha_1\alpha_2...\alpha_k)}{f^0(\alpha_1\alpha_2...\alpha_k)},$$

where $f(\alpha_1\alpha_2...\alpha_k)$ is the frequency of $k$-mer $\alpha_1\alpha_2...\alpha_k$ and $f^0(\alpha_1\alpha_2...\alpha_k)$ is the predicted frequency of the $k$-mer, calculated by using a $(k-2)$-th Markov assumption.

Let $CV_A = (a_1a_2...a_N)$ and $CV_B = (b_1b_2...b_N)$ be the CVs for the species $A$ and $B$, respectively. Finally, calculate the distance matrix for the modified CV:

$$D(A,B) = (1 - C(CV_A, CV_B))/2,$$

where

$$C(CV_A, CV_B) = \frac{\sum_{i=1}^{N} a_i \times b_i}{\sqrt{\sum_{i=1}^{N} a_i^2 \times \sum_{i=1}^{N} b_i^2}}.$$

### B. $d_2^S$ $k$-mer statistical alignment-free method

$d_2^S$ statistics is a modified version of $D_2$, $D_2^*$, and $D_2^S$ statistics [13], [14]. They are applicable to NGS data by considering the random processes of NGS data in terms of $D_2$, $D_2^*$, and $D_2^S$ to model the correct $k$-mer distribution of NGS data. NGS short reads are small fragments from the original long sequence, which means that the method of sampling those reads will affect the $k$-mer frequency distribution. Another characteristic of NGS data relevant to $d_2^S$ statistics is that an NGS short read can originate from the forward or reverse strand of the original genome, requiring consideration of not only the $k$-mer distributions of short-read data themselves but also their complementary sequences. $d_2^S$ can be calculated by:

$$d_S^2 = \frac{1}{2}\left(1 - \frac{D_S^2}{\sqrt{\sum_{w \in A^k} \tilde{X_w^2}/\tilde{Z_w}}\sqrt{\sum_{w \in A^k} \tilde{Y_w^2}/\tilde{Z_w}}}\right),$$

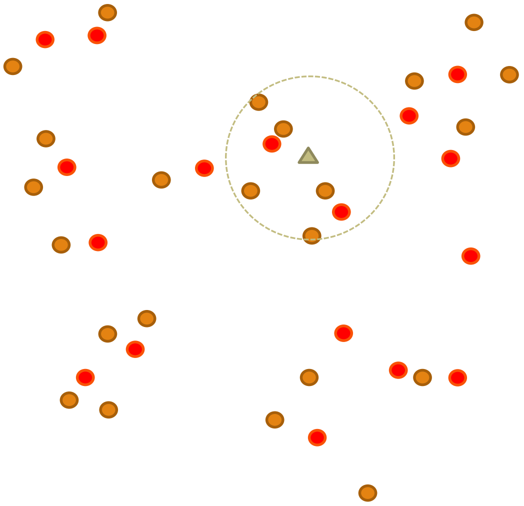

Figure 1. Neighbor search in NGS short reads

where

$$D_2^S = \frac{\tilde{X_w}\tilde{Y_w}}{\tilde{Z_w}}$$

and

$$\tilde{Z_w} = \sqrt{\tilde{X_w^2} + \tilde{Y_w^2}}.$$

Suppose that $M$ reads of length $\beta$ are sampled from a genome of length $n$. Let $X_w$ and $Y_w$ be the numbers of occurrences of word pattern $w$ in the $M$ pairs of reads from the first genome and the second genome, respectively. We define $\tilde{X_w^2} = X_w - M(b-k+1)(p_w + p_{\bar{w}})$ with $\tilde{Y_w^2}$ being defined analogously. Let $w = w_1w_2...w_k$ and $p_w = p_{w_1}p_{w_2}...p_{w_k}$, with $\bar{w}$ being the complement of word $w$. Consider two genome sequences taking $L$ letters (0, 1, ..., $L-1$) at each position. For the null model, we assume that the two genomes are independent and both are generated by models with $p_l$ being the probability of taking state $l$, $l$ = 0, 1,..., $L-1$.

## III. Proposed Method

The NGS data comprise a very large quantity of short reads that contain overlapping data. Particularly for whole-genome sequences, the number of overlaps and repeats can grow dramatically. Most existing research on alignment-free methods adopts $k$-mer frequencies to specify the profile of a sequence and when obtaining distances in NGS sets. However, the random overlap of short reads in NGS data will clearly affect the distribution of $k$-mer frequencies. This is the key problem we focus on in this research.

Because the problem is caused by the overlap and the repeating data, the key idea is to reduce their effect by grouping similar short reads. We can then use a statistical approach to calculate the evolutionary distance between NGS short reads. Fig. 1 shows a feature space spanned by the $k$-mers. (As mentioned above, the dimensionality of the space is $4^k$, but, for readability, we show a 2-D space.) Each dot represents an NGS short read. Dots of the same color indicate that the corresponding NGS short read comes from the same genome sequence. For a given short read $r$, its set of neighbors is defined as the set of short reads whose distance from $r$ is within a predefined threshold. The circle in Fig. 1 encloses the neighborhood of the short read represented by the triangle.

The assumption is that the short reads that are placed near each other in the feature space will have a high probability of sharing overlapping data. We define the difference between any two NGS sets by comparing the number of neighbor-search results that correspond to the same collection of search queries on their NGS short reads. Because this method does not consider $k$-mer frequencies in the similarity measures of NGS sets, any overlap effects on the final distance matrix are reduced.

### A. Notations and equations

Denote $d^{NS}(X, Y)$ as the pairwise distance between NGS sets $X$ and $Y$, where $X = \{x_1, x_2, ...x_n\}$ and $n$ is the number of NGS short reads of $X$. Similarly, $Y = \{y_1, y_2, ...y_m\}$ and $m$ is the number of NGS short reads of $Y$. For a query sequence $q$, let $R_X^q$ denote the number of neighbors of $q$ in $X$. $d^{NS}(X, Y)$ can then be calculated as follows:

$$d^{NS}(X, Y) = (D(X, Y) + D(Y, X))/2, \tag{1}$$

where

$$D(X, Y) = \sum_{i=1}^{n} \left( 1 - \frac{min\left(\frac{R_X^{x_i}}{n}, \frac{R_Y^{x_i}}{m}\right)}{max\left(\frac{R_X^{x_i}}{n}, \frac{R_Y^{x_i}}{m}\right)} \right) \times \left( \frac{R_X^{x_i}}{\sum_{i=1}^{n} R_X^{x_i}} \right). \tag{2}$$

$D(X, Y)$ is a divergence measurement calculated by summation of the rational difference between the number of neighbors in NGS sets $X$ and $Y$ for all NGS short reads $x_1, x_2, ...x_n \in X$. The $min$ to $max$ ratio of two normalized values $\frac{R_X^{x_i}}{n}$ and $\frac{R_Y^{x_i}}{m}$ in Eq. (2) indicates the rational similarity between those two values. If the normalized numbers of neighbors for $X$ and $Y$ are the same, this term will be equal to 1. Subtracting the term from 1 makes it a divergence measurement. For each short read in $X$ and $Y$, the distance is weighted by the normalized number of the neighbors for that query. Because $D(X, Y)$ is an asymmetric function, we define the distance $d^{NS}(x, y)$ as the average value of $D(X, Y)$ and $D(Y, X)$.

In the current implementation, we use locality-sensitive hashing (LSH) [15] for the neighbor search because of its lightweight nature. Minhash [16] was originally used to compare the similarity between documents. This algorithm provides a fast approximation of the Jaccard similarity between two sets by using their Minhash signatures and simply counts the number of components of the signatures that are equal. Let $h$ be the hash function for mapping an integer to another different integer, with no collisions. Apply $n$ hash functions in $H = h_1, h_2, ..., h_n$ to the set of integers. For each $h_i$ from $i = 1$ to $n$, the minimum hash value produced by $h_i$ will be assigned to the $i$th component of the Minhash signature. We use this process to obtain the Minhash signature of an NGS short read. The set of $k$-mers that appear in an NGS short read are transformed into a set of integers to enable the hash functions to be applied. These hash functions are randomly generated with various values for the parameters that produce different hash functions. LSH is a process for finding a group of items whose Minhash signature is similar to a query's signature. It separates the Minhash signature into a series of bands, each comprising a set of rows. For example, 200 Minhash signatures might be separated into 20 bands of 4 rows each. Each band is then hashed to a "bucket. If two sets have the same Minhash signature in a band, they will be hashed to the same bucket, and will therefore be considered candidate pairs. In our approach, utilizing LSH with Minhash enables us to search for similar NGS short reads easily. However, $d^{NS}$ could adopt alternative neighbor-search algorithms because the distance measurements in $d^{NS}$ are based on the results of neighbor search, rather than its method.

## IV. Evaluations and Results

### A. Experiment setup

Two datasets, comprising 29 mammalian mtDNA sequences [10], [11] and 29 $Escherichia/Shigella$ [12] genomes were used to evaluate $d^{NS}$ by comparing it with two existing $k$-mer-based alignment-free methods, namely $CVTree$ and $d_2^S$. Because both datasets were originally made up of long sequences, we used a tool called $MetaSim$ [17] to simulate NGS short reads from long genome sequences. We used three error models, namely 454, Empirical(Illumina), and Sanger, which enabled us to simulate the NGS high-throughput sequencing results from three different NGS platforms. These sequenced the actual samples into NGS data. In the following discussion, the term "Exact" refers to the non-error case in simulating NGS short reads from long genomic sequences. We used sampling depths of 1, 5, 10, and 30, where the sampling depth means the average number of occurrences of the character at each position in the original sequences appearing in the NGS set. The length of NGS short reads was set to 100, with a default parameter for the error distribution for each model. For the parameter $k$, we considered using $k$ values in the range 6 to 10. Although a larger $k$ should give a better result, the processing time to map each NGS short read to the feature space would increase significantly. We planned our experiments to use this range of $k$ values for several reasons. One reason was that $CVTree$ and $d_2^S$ proponents have suggested it as a suitable range. Second, for $d^{NS}$, $k$ values out of this range would affect the efficiency of the neighbor-search process.

MSA was used as the benchmark method for comparison with the alignment-free methods to evaluate their performance on phylogeny reconstruction. We used the $ClusterOmega$ tool [18], followed by the $dnadist$ tool in the PHYLIP package [19], on aligned sequences from MSA to calculate distance matrices.

For a distance matrix, either from MSA or from an alignment-free method, we used the $neighbor$ tool in the PHYLIP package to construct a phylogenetic tree using the neighbor-joining method [20]. We used the popular Robinson–Fould distance (RF) [21] for evaluation, as described in [22]. The RF value can be calculated by counting the internal nodes that appear in one tree but not in the other. A small RF value means that the shape of the trees is close to the benchmark tree. The values for RF range from 0, meaning two tree are exactly the same, to $2(n - 3)$ where $n$ is the number of leaf nodes.

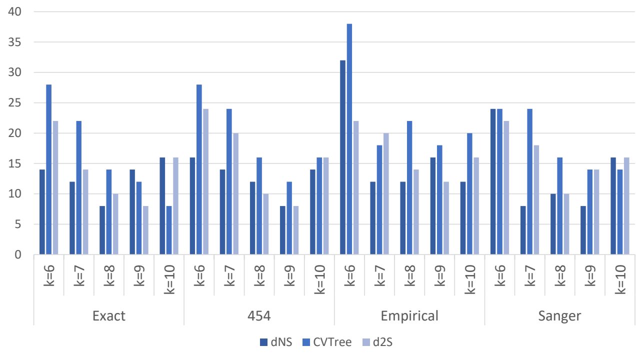

Figure 2. The RF of phylogenetic tree results for each method on NGS reads of 29 mammalian mtDNA sequences with a sampling depth of 5

TABLE I. BEST RF RESULT FOR ANY $K$ PARAMETER ON NGS READS OF 29 MAMMALIAN MTDNA SEQUENCES WITH A SAMPLING DEPTH OF 5

| | $d^{NS}$ | $CVTree$ | $d_2^S$ |
|---|---|---|---|
| Exact | **8** | **8** | **8** |
| 454 | **8** | 12 | **8** |
| Empirical | **12** | 18 | **12** |
| Sanger | **8** | 14 | 10 |

### B. Experimental results

*1) The 29 mammalian mtDNA sequences:* The 29 mammalian mtDNA sequences are a well-studied dataset, being widely used for the evaluation of existing sequence-comparison methods. The MSA tree for this dataset is therefore a reliable benchmark for our experiments. Because the evolutionary relation between each species in this dataset is diverse, a sequence-comparison method should be able to reconstruct a phylogenetic tree almost identical to that for MSA to offer confidence in the performance of the method.

We applied three alignment-free methods, namely $d^{NS}$, $CVTree$, and $d_2^S$, to simulated NGS short-read data. We compared the resultant phylogenetic trees with the benchmark tree obtained from MSA with mtDNA sequences. At a sampling depth of 1, the phylogenetic trees obtained from the three alignment-free methods were very different from the MSA benchmark tree because of the shallow sampling depth.

Fig. 2 shows the RF between the MSA benchmark tree and the phylogenetic tree obtained by $d^{NS}$, $CVTree$, and $d_2^S$ on four types of NGS reads, using a sampling depth of 5 and various $k$ parameter values. The figure shows that $d^{NS}$ produces a more accurate tree than either $CVTree$ or $d_2^S$ in most cases.

Table I summarizes the most accurate result for each alignment-free method shown in Fig. 2. Note that RF can be up to 52 for this dataset. The best RF result in the table among all three methods is 8, which means that the rational distance between the tree obtained via alignment-free methods and the benchmark tree is 8/52 = 0.154. We can therefore consider that $d^{NS}$ and the other two alignment-free methods all perform well using this dataset. $d^{NS}$ produced the best result among the alignment-free methods across all NGS error models. Regarding the sampling depth, we found no significant differences with 10 and 30 sampling, as shown in Fig. 3 for $d^{NS}$. The same result was found for $CVTree$ and $d_2^S$ [22].

We investigated how parameter values affect the performance of $d_{NS}$. Fig. 3 shows the result of $d^{NS}$ on NGS reads of 29 mammalian mtDNA sequences with a parameter setup that included four NGS error models, $k$ values from 6 to 10, and sampling depths of 1, 5, 10, and 30. With a sampling depth of 1 for any NGS error model, $d^{NS}$ could not produce an accurate phylogenetic tree for this dataset. The reason could be that the numbers of queries used in the neighbor search are too small to retrieve good distance measurements. According to this result for $d^{NS}$, we can infer that a more suitable value for the $k$ parameter would be 8 or 9.

*2) The 29 $Escherichia/Shigella$ whole-genome sequences:* We used this dataset to evaluate the performance of $d^{NS}$ on the whole genomes of species that are close to each other in evolutionary terms. The 29 whole-genome sequences come from two main genera, namely $Escherichia$ and $Shigella$, which are from the same $Enterobacteriaceae$ family in the $Bacteria$ kingdom. Because the dataset is large, MSA's lack of scalability prevents it from being applied. We obtained the benchmark tree for this dataset from [12]. This involved concatenating the alignments of the 2034 core genes of the $Escherichia/Shigella$ genomes, then using a maximum-likelihood method to construct the phylogenetic tree for this dataset.

With the close evolutionary relationship between the $Escherichia$ and $Shigella$ species, all alignment-free methods tested in this experiment failed to obtain an accurate RF result when comparing their resultant trees with the benchmark tree. As shown in Table II, the best RF value was 16, with the rational distance between the result tree and the benchmark tree being 16/52 = 0.3. The performances of all three methods were below a satisfactory level. There was no significant difference among the $d^{NS}$, $CVTree$, and $d_2^S$ methods. In fact, $d_2^S$ performed better for the Exact error model, whereas $d^{NS}$ and $CVTree$ performed better for the other error models.

A point to note is that $d^{NS}$ appears more robust with respect to variations in the $k$ parameter than $CVTree$ or $d_2^S$, as shown in Fig. 4. For most $k$, and for each error model, $d^{NS}$'s phylogenetic tree is more accurate than those of the other methods, with the RF value being at the same level. For example, although $d_2^S$ performs best on the Exact model with RF of 18 when $k$ = 9 and 10, the RF values are much bigger for other $k$ values. Robustness against the parameter $k$ is beneficial because it makes parameter tuning easier and we can optimize the processing efficiency by choosing a smaller value for $k$. The reason for this effect is that the $k$ parameter does not directly affect how $d^{NS}$ calculates the distance between each species. It uses the $k$ value only for constructing the feature space. Because of limited computing resources, we examined only the case of the 1 sampling depth.

The main aim of this research is to introduce a novel approach to performing NGS data comparisons. It is to be expected that the computational efficiencies of $CVTree$ and $d_2^S$ would exceed that of $d^{NS}$ in its current implementation. Table III confirms that $d^{NS}$'s runtime is slower than the others. However, the $k$ parameter value does not affect the runtime of $d^{NS}$, unlike those for $CVTree$ and $d_2^S$. In particular, $d_2^S$'s runtime grows dramatically between $k$ = 6 and $k$ = 10. It is an

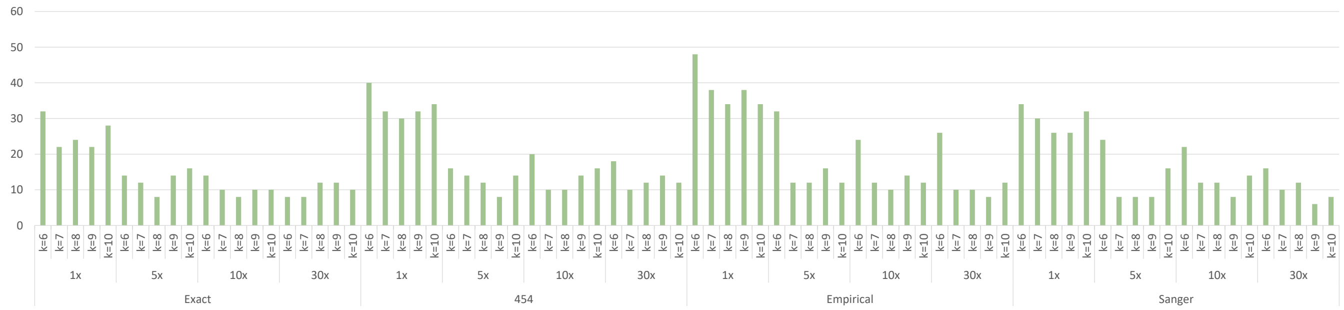

Figure 3. The RF of phylogenetic tree results for $d^{NS}$ on NGS reads of 29 mammalian mtDNA sequences using four NGS error models, $k$ = 6–10, and sampling depths of 1, 5, 10, and 30

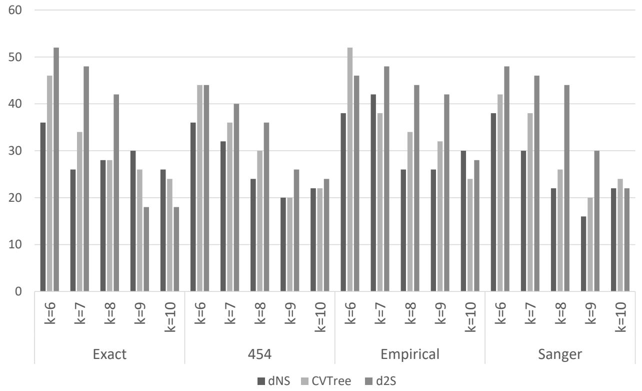

Figure 4. The RF results for NGS reads of 29 $Escherichia/Shigella$ whole-genome sequences with a sampling depth of 1

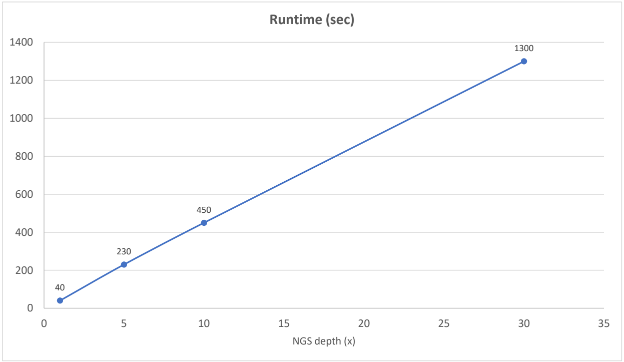

Figure 5. Computational runtime (seconds) for $d^{NS}$ on the 29 mammalian mtDNA dataset with NGS sampling depths of 1, 5,10, and 30

TABLE II. BEST RF RESULTS FOR ANY $K$ VALUE ON NGS READS OF 29 $ESCHERICHIA/SHIGELLA$ WHOLE-GENOME SEQUENCES WITH A SAMPLING DEPTH OF 1

| | $d^{NS}$ | $CVTree$ | $d_2^S$ |
|---|---|---|---|
| Exact | 26 | 26 | **18** |
| 454 | **20** | **20** | **20** |
| Empirical | 26 | **24** | 28 |
| Sanger | **16** | 20 | 22 |

advantage that the $k$ value has little effect on the runtime of our proposed method. In addition, the runtime of $d^{NS}$ shows linear growth with varying sampling depth, as shown in Fig. 5

## V. Conclusion and Future Work

In conclusion, we propose a novel approach for an alignment-free method $d^{NS}$ that is focused on NGS short-read data and based on neighbor searching. Its main advantage is that it is an accurate alignment-free sequence-comparison method for reconstructing a phylogenetic tree more consistently than other $k$-mer-based alignment-free methods. Although it might lose significant information in the NGS data when ignoring the $k$-mer frequencies, the method is able to specify the distance between NGS sets with good accuracy when a sufficient number of queries is used.

According to our experimental results on mammalian mtDNA and $Escherichia/Shigella$ whole-genome sequences with simulated NGS short reads, the $d^{NS}$ method can construct a phylogenetic tree that is almost as accurate as a benchmark tree. In addition, $d^{NS}$ is able to deal with NGS short-read faults in the $k$-mer distribution, as shown in our results. However, the main drawback of $d^{NS}$ is the computational inefficiency of the neighbor-search process, which consists of many NGS short-read comparisons.

Because this method is a novel approach to NGS short-read comparison, there are many aspects of it that we can develop to make the method more accurate and more computationally efficient. First, we should consider modifying $d^{NS}$ toward a parameter-free approach. To improve the $d^{NS}$ accuracy, we should consider applying different processes for mapping the NGS reads to a high-dimensional space during neighbor search, rather than using a $k$-mer frequency vector. This modification would seek to obtain more reliable grouping of overlapping NGS reads. Second, we should modify the equation for distance measurement used in this approach. We have noticed that calculating distances in the NGS data using only the number of neighbor query results might be insufficient to achieve better results. This information is quite coarse in comparison with normalized $k$-mer frequencies. We should consider combining our approach with other alignment-free methods to achieve higher accuracy. Finally, we need to optimize this method to be more scalable with respect to computational efficiency by considering alternative neighbor-search algorithms. In fact, we might consider a completely different approach, other than neighbor search, to the grouping of overlapping NGS reads.

TABLE III. COMPUTATIONAL RUNTIME FOR EACH ALIGNMENT-FREE METHOD (SECONDS) WITH 5 COVERAGE FOR MAMMALIAN mtDNA AND 1 COVERAGE FOR THE $ESCHERICHIA/SHIGELLA$ WHOLE-GENOME DATASET

| | $d^{NS}$ | $d_2^S(k=6)$ | $d_2^S(k=10)$ | $CVTree(k=6)$ | $CVTree(k=10)$ |
|---|---|---|---|---|---|
| 29 mammalian mtDNA | 230 | 4 | 780 | 2 | 5 |
| 29 $Escherichia/Shigella$ whole genome | 8600 | 30 | 1050 | 25 | 180 |

## References

[1] M. L. Metzker, “Sequencing technologies–the next generation,” Nature reviews. Genetics, vol. 11, no. 1, 2010, p. 31.

[2] A. Phillips, D. Janies, and W. Wheeler, “Multiple sequence alignment in phylogenetic analysis,” Molecular phylogenetics and evolution, vol. 16, no. 3, 2000, pp. 317–330.

[3] M. S. Waterman, Introduction to computational biology: maps, sequences and genomes. CRC Press, 1995.

[4] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press, 1998.

[5] S. Vinga and J. Almeida, “Alignment-free sequence comparisona review,” Bioinformatics, vol. 19, no. 4, 2003, pp. 513–523.

[6] C. X. Chan and M. A. Ragan, “Next-generation phylogenomics,” Biology direct, vol. 8, no. 1, 2013, p. 3.

[7] Z. Xu and B. Hao, “Cvtree update: a newly designed phylogenetic study platform using composition vectors and whole genomes,” Nucleic acids research, vol. 37, no. suppl_2, 2009, pp. W174–W178.

[8] J. Qi, H. Luo, and B. Hao, “Cvtree: a phylogenetic tree reconstruction tool based on whole genomes,” Nucleic acids research, vol. 32, no. suppl_2, 2004, pp. W45–W47.

[9] K. Song, J. Ren, Z. Zhai, X. Liu, M. Deng, and F. Sun, “Alignment-free sequence comparison based on next-generation sequencing reads,” Journal of computational biology, vol. 20, no. 2, 2013, pp. 64–79.

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

[11] Y. Cao, A. Janke, P. J. Waddell, M. Westerman, O. Takenaka, S. Murata, N. Okada, S. Pääbo, and M. Hasegawa, “Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders,” Journal of Molecular Evolution, vol. 47, no. 3, 1998, pp. 307–322.

[12] Z. Zhou, X. Li, B. Liu, L. Beutin, J. Xu, Y. Ren, L. Feng, R. Lan, P. R. Reeves, and L. Wang, “Derivation of escherichia coli o157: H7 from its o55: H7 precursor,” PloS one, vol. 5, no. 1, 2010, p. e8700.

[13] G. Reinert, D. Chew, F. Sun, and M. S. Waterman, “Alignment-free sequence comparison (i): statistics and power,” Journal of Computational Biology, vol. 16, no. 12, 2009, pp. 1615–1634.

[14] L. Wan, G. Reinert, F. Sun, and M. S. Waterman, “Alignment-free sequence comparison (ii): theoretical power of comparison statistics,” Journal of Computational Biology, vol. 17, no. 11, 2010, pp. 1467–1490.

[15] A. Gionis, P. Indyk, R. Motwani et al., “Similarity search in high dimensions via hashing,” in VLDB, vol. 99, no. 6, 1999, pp. 518–529.

[16] A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher, “Min-wise independent permutations,” Journal of Computer and System Sciences, vol. 60, no. 3, 2000, pp. 630–659.

[17] D. C. Richter, F. Ott, A. F. Auch, R. Schmid, and D. H. Huson, “Metasim: A sequencing simulator for genomics and metagenomics,” Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches, 2011, pp. 417–421.

[18] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding et al., “Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega,” Molecular systems biology, vol. 7, no. 1, 2011, p. 539.

[19] J. Felsenstein, “Phylip–phylogeny inference package (version 3.2) cladistics. 1989; 5: 164–166,” DOI: citeulike-article-id, vol. 2344765.

[20] N. Saitou and M. Nei, “The neighbor-joining method: a new method for reconstructing phylogenetic trees.” Molecular biology and evolution, vol. 4, no. 4, 1987, pp. 406–425.

[21] D. F. Robinson and L. R. Foulds, “Comparison of phylogenetic trees,” Mathematical biosciences, vol. 53, no. 1-2, 1981, pp. 131–147.

[22] N. H. Tran and X. Chen, “Comparison of next-generation sequencing samples using compression-based distances and its application to phylogenetic reconstruction,” BMC research notes, vol. 7, no. 1, 2014, p. 320.