Alignment-free genome sequence comparison method based on pair transition difference of k-words

Gyu-Bum Han[1] and Dong-Ho Cho*

***Abstract*— This paper proposes a novel alignment-free genome sequence comparison method using pair transition property of k-words. In the proposed method, difference between pair transition of k-words is extracted from genome sequence as feature vector. Using the statistical distances metric and the phylogenetic tree, the classification performance of proposed method is identified. Also, to reduce the computational complexity, the selection method of pair word is introduced. Finally, the performance of proposed method is compared with conventional alignment-free methods for 54 MT genomes with respect to the RF distance and the computational complexity.**

I. INTRODUCTION

Genome is whole set of genetic material which manages biological phenomena for an organism. Genome sequence consists of adenine (A), cytosine (C), guanine (G), thymine (T) nucleotides which are coded to guide phenotype of individual organisms [1]-[2]. In recent years, analysis of genome sequence grew substantially to study the relationship between genome and biological phenomena. With the development of genome sequencing technologies, comparison methods based on genome sequence played an important part in the classification of organism. [3] Traditionally, popular genome sequence comparison methods were pairwise or multiple sequence alignment schemes [4]. These methods generally gave high resolution results of genetic relationships when there exist reliably aligned sequences. A major limitation of these methods is that they have high amount of computation and complexity because it takes time proportional to product of input sequence lengths for alignment process. To overcome the disadvantages of alignment-based approaches, alignment-free methods have been proposed in recent years [5]. Because alignment-free methods have relative low amount of computation and complexity which are proportional to the sum of input sequence lengths, they can provide a linear running time. Most of alignment-free methods are based on the property of word count results or matches length [6]. We focus on word-count based comparison methods. Representative examples of methods using the property of sequence words are feature frequency profiles (FFP) and k-words relative measure methods. FFP approach is the frequencies profile comparison of feature word in genome [7]. Most of k-words relative measure methods compare genome sequences based on k-words position or frequency information [8]-[10]. Previous works related to alignment-free methods have a relatively low resolution because they compare genome sequences using restricted genome information. Also, in most of studies based on the profile of k-words, they considered frequency and position profile only. In this paper, therefore, concept of transition probability is employed to identify genetic relationship among organisms. An alignment-free sequence comparison method is proposed based on difference between pair transitions of k-word. Additionally, a selection method for essential pair words is introduced to reduce computational complexity. Then, the performance and complexity of proposed scheme is compared with that of conventional alignment-free methods using a set of mitochondrial genomes.

II. MATERIALS AND METHODS

A. Benchmark sequences

To identify the performance of the proposed sequence comparison method, we use a set of 54 mitochondrial (MT) genomes. The species and groups of all MT genomes are described in Table I. They were obtained from NCBI (http://www.ncbi.nlm.nih.gov/genomes/). The MT genome includes 3 groups which are mammals (M), fishes (F) and birds (B). The size of sequences ranged from 16Kb to 18Kb.

B. Distance metric

To obtain the distance matrix for input genome sequences $S_1, ..., S_N$, we employ statistical distances metric which is the Bhattachayya distance d_B. It is defined as

$$d_B(\mathbf{X}_{S_i}, \mathbf{X}_{S_j}) = -\ln\left(\sum_h \sqrt{X_{S_i}[h] X_{S_j}[h]}\right) \qquad (1)$$

the Bhattachayya distance measures amount of overlap between two statistical samples or populations.

C. Performance evaluation metric

To evaluate the performance of proposed alignment-free method based on transition of k-words, we test phylogeny reconstruction analysis using sequence set of 54 MT genomes. Also, for these sequences, the reliable reference tree is determined as phytogenetic tree is based on multiple sequence alignment whose result was achieved by Clustal W [11]. The performance metric is considered as the difference between

*This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.R-20160301-002998, Development of bio-marker, cancer diagnostics chip and precise diagnostics device based on normal and cancer genome)

[1] Gyu-Bum Han is with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro Yuseong-gu Daejeon, Korea gbhan@comis.kaist.ac.kr

*Dong-Ho Cho with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro Yuseong-gu Daejeon, Korea dhcho@kaist.ac.kr

978-1-5090-4179-4/17/$31.00 ©2017 IEEE

TABLE I

54 MT GENOMES : GROUPS AND SPECIES NAMES

Group	Species names
Mammals	Wallaroo, Opossum, Platypus, Rabbit Squirrel, tarsius bancanus, Fat dormouse, lemur catta, Fruit bat Sheep, Cow, Pig, Fin whale Blue whale, Indian rhinoceros White rhinoceros, Donkey, Hippopotamus, Harbor seal, Gray seal Dog, Armadillo, Mouse, Rat Guinea pig, Pigmy chimpanzee, Common chimpanzee, Human, Gorilla Orangutan, Gibbon, Baboon Cebus albifrons, Hedgehog
Birds	Redhead duck, domestic chicken Ostrich, Rhea, Village indigobird Broadbill, Peregrine falcon
Fishes	Bichir, Coelocanth, gummy shark Dogfish, Spiny dogfish, Rainbow trout Atlantic salmon, Atlantic Cod, Goldfish Carp, Plecoglossus altivelis, Lungfish Sea lamprey

trees from proposed and reference method. The difference between trees is captured by the Robinson-Foulds metric [12]. Conventional alignment-free methods which are FFP [7] and k-word relative measure method [8] are applied for performance comparison with proposed method. For practical implementation, the range of k-words length is considered from 1 to 5.

III. Results and discussions

A. Transition probability of k-word in genome sequence

As usual, we assume that S represents a sequence, where $S[h] \in \{A, C, G, T\}$ denotes h-th component of S and L is length of sequence S. Let $w_i^k = w[1]w[2]...w[k]$ be a k-word, where $w[h] \in \{A, C, G, T\}$ is h-th element of w for $1 \leq h \leq k$. Frequencies of w_i^k are captured based on a sliding window of length k from a genome sequence S. To examine frequency profile of w_i^k exhaustively, sliding window is run through a sequence S from start position 1 to $L - k + 1$. From sliding window results, we can achieve frequencies of k-word w_i^k for $1 \leq i \leq 4^k$ for a sequence S. In this condition, transition frequency between w_i^k and w_j^k is defined as $f_{w_i^k - w_j^k}$. Then, transition probability between w_i^k and w_j^k is represented by

$$P_{w_i^k - w_j^k} = \frac{f_{w_i^k - w_j^k}}{\sum_{h=1}^{4^k} f_{w_i^k - w_h^k}} \tag{2}$$

B. Difference between pair transition of k-words

To extract property vector of genome sequence S_n, we use difference value between pair transitions of k-words, which is

$$t_{i,j} = |P_{w_i^k} - P_{w_j^k}| \tag{3}$$

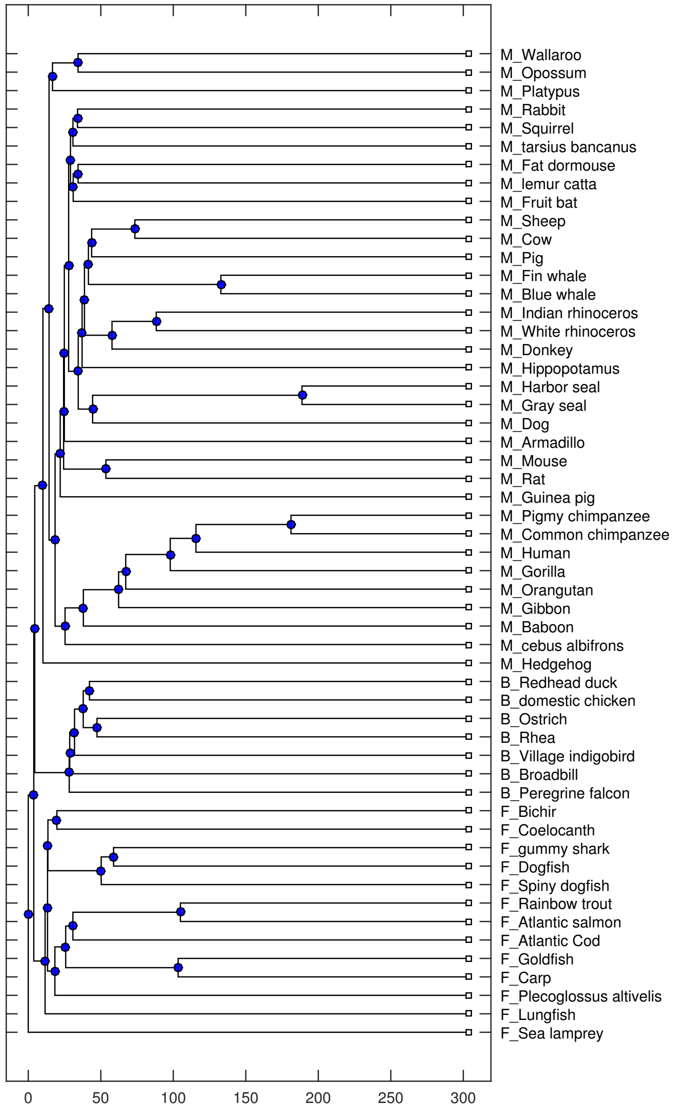

Fig. 1. Phytogenetic tree result of 54 MT genomes for k = 4

For S_n, difference values between transitions of all pairs can be expressed as matrix form T_{S_n}. (i, j) component of T_{S_n} is equals to $t_{i,j}$.

$$T_{S_n} = \begin{bmatrix} t_{1,1} & t_{1,2} & \dots & t_{1,4^k} \\ t_{2,1} & t_{2,2} & \dots & t_{2,4^k} \\ \dots & \dots & \dots & \dots \\ t_{4^k,1} & t_{4^k,2} & \dots & t_{4^k,4^k} \end{bmatrix} \tag{4}$$

Here, T_{S_n} is $4^k \times 4^k$ matrix and diagonal components are equal to zero. Also, it is symmetric matrix because $t_{i,j}$ is the same as $t_{j,i}$ according to equation (3). Therefore, when genome sequences are compared, we use triangular matrix of T_{S_n} to reduce computational complexity.

C. Phylogenetic tree result of proposed sequence comparison

In Fig. 1, a phylogenetic tree of the proposed method are represented for $k = 4$. We use unweighted pair group method with arithmetic mean (UPGMA) method to achieve phylogenetic tree. When k is lower than 4, several classification errors exist. For $k = 4$, the division of 54 MT genome into mammals (M), fishes (F) and birds (B) groups has a clear except for only one sample. In our experiments, conventional methods [7]-[8] don't have clear 3 groups even though $k = 5$. Each group is separated by several groups. Tree of proposed method has only one exceptional sample which is named as Sea lamprey. It included in Cephalaspidomorphi class only among samples of fish. Also, classifying Sea Lamprey is in argument[14].

D. Selection method of pair words for low computational complexity

Given word length k, the number of word transitions is 4^{2k}. Also, proposed difference between pair transition has $\frac{4^{2k}}{2}$ components for each sequence. This result leads to high computational complexity when genome sequences are compared. Therefore, we propose selection method of essential pair words to reduce computational complexity. To select essential pair words, we use variance of transition frequency in set of input genome sequences. Variance of $w_i^k - w_j$ transition is expressed as $v_{w_i^k - w_j^k}$, which is defined by

$$v_{w_i^k - w_j^k} = \frac{1}{N}\sum_{n=1}^{N}(f^n_{w_i^k - w_j^k} - E[f_{w_i^k - w_j^k}])^2 \tag{5}$$

where N is the number of input genome sequences and $E[f_{w_i^k - w_j^k}]$ is mean value of $f_{w_i^k - w_j^k}$ in input genome sequences. Because $f_{w_i^k - w_j^k}$ isn't equal to $f_{w_i^k - w_j^k}$, $v_{w_i^k - w_j^k}$ isn't same as $v_{w_j^k - w_i^k}$. Therefore, to clarify criteria of essential pair, we consider summation value of variation for pair words, which is defined as

$$V_{(i,j)} = v_{w_i^k - w_j^k} + v_{w_j^k - w_i^k} \tag{6}$$

Using $V_{(i,j)}$ values, we select essential pair words which have higher $V_{(i,j)}$ value relatively among all k-words. In other words, set of pair words is sorted by $V_{(i,j)}$, then we select specific number of essential words pairs according to sorting order. We select 25%, 50% and 75% ratios to identify effect of selection. For example, when selection ratio equals to 50%, half of word pairs are selected to compare sequences. Distance measure and principle of phylogenetic tree are same as comparison based on all of word pairs. In Figure 2 and 3, we can show classification and complexity performance of selection method with varying word length. In Figure 2, the performance of classification result is captured by Robinson-Foulds (RF) distance metric to observe effect of pair word selection. Because RF distance means difference between reliable reference tree and proposed tree, the performance of classification is higher when RF distance is lower value. When k is smaller than 4, all pair based algorithm (100%) has best performance. However, when k is higher than 4, 75% selection based algorithms has better performance than that of all pair based algorithm. This is because all of transitions doesn't have statistical meaning when k is high value. Also, we can show performance degradation of 25% and 50% selection based algorithm when $k = 5$. From these results, we can know that genome property isn't extracted if essential words aren't selected enough. Additionally, complexity of proposed algorithms is depicted in Fig 3. Complexity of all pair based algorithm is $k\sum_{n=1}^{N} L_n + \frac{4^k}{2}N^2$ where L_n is length of sequence S_n. Similarly, complexity of selection algorithm is $k\sum_{n=1}^{N} L_n + p\frac{4^k}{2}N^2$ where p is word selection ratio such as $25\%, 50\%$ and 75%. We can identify the complexity of selection algorithms decreases linearly with low selection ratio.

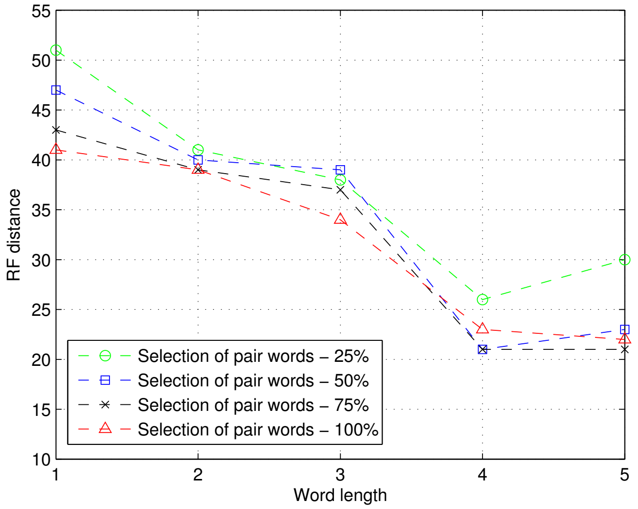

Fig. 2. RF distance vs word length for selection ratio of pair words

E. Comparison between proposed and conventional comparison methods

In Fig. 4, RF distance results of proposed and conventional methods can be shown. We can show RF distances of conventional methods such as FFP and k-word relative measure. FFP method is based on frequencies of k-words and k-word relative measure considers position of k-words. Representative difference of our proposed algorithm is to use transition property of word pairs . In overall range of word length, all pair based algorithm has better performance compared with conventional methods. In particular, these simulation results demonstrated that proposed all pair based method has up to 69.6% performance gain in terms of RF distance compared to conventional sequence comparison methods when $k = 4$. However, complexity of proposed algorithm is higher than conventional method. When $k = 4$, 48% complexity increment occurs in all pair based methods.

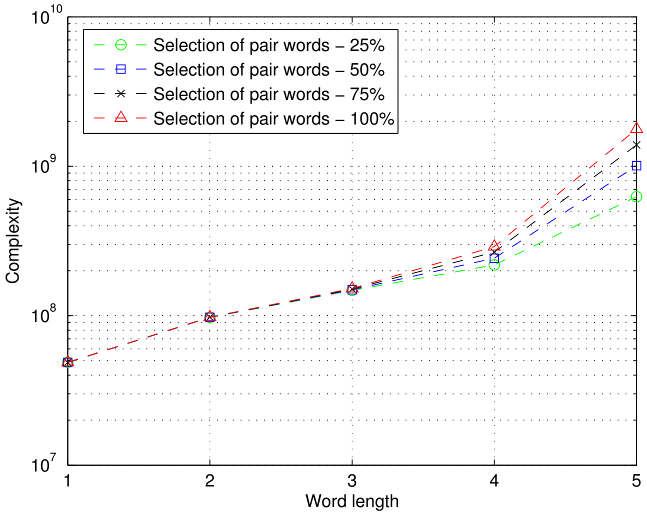

Fig. 3. Complexity vs word length for selection ratio of pair words

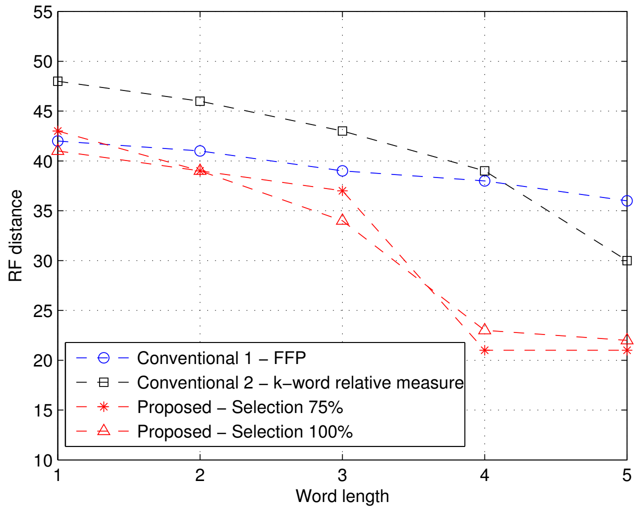

Fig. 4. RF distance vs word length for proposed and conventional comparison methods

In case of 75% selection based algorithm with, complexity increment equal to 36% as shown in Fig. 5.

IV. CONCLUSIONS

In this paper, we proposed the alignment-free genome sequence comparison method based on difference between pair transitions of k-words. We examined performance of classification using phylogenetic tree of 54 MT genome which are separated into 3 groups. Phylogenetic tree result from proposed method included mammals, birds and fishes groups clearly except for only one sample. Also, we introduced selection method of essential pair words to reduce computational complexity. Finally, we compared the performance of proposed method with that of conventional methods. Then, we identified the performance gain of classification in terms of RF distance metric.

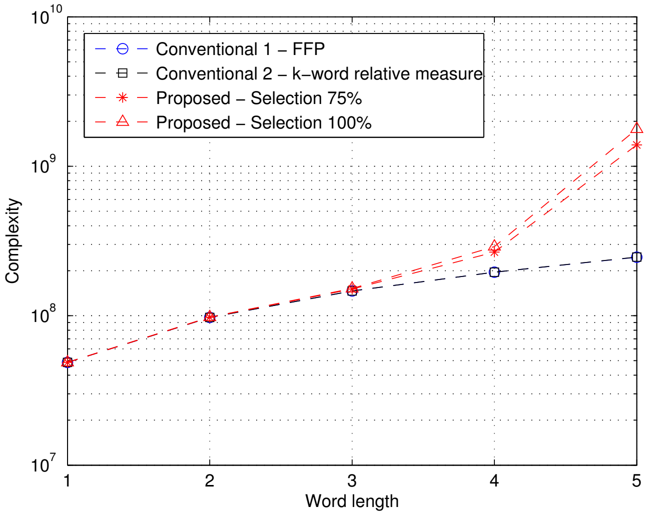

Fig. 5. Complexity vs word length for proposed and conventional comparison methods

REFERENCES

[1] Seberg Ole, “Genome analysis, phylogeny, and classification.”, *Plant systematics and evolution*, vol.166, no.3, pp.159-171, Sep. 1989.

[2] Crick, Francis HC, “The origin of the genetic code.”, *Journal of molecular biology*, vol.38, no.3, pp.367-379, Dec. 1968.

[3] Fuchs R, “From sequence to biology: the impact on bioinformatics.”, *Bioinformatics*, vol.18, no.4, pp.505-506, Apr. 2002.

[4] Waterman MS, “Introduction to computational biology: maps, sequences and genomes.”, *CRC Press*, 1995

[5] Vinga S, Almeida J, “Alignment-free sequence comparison - a review.”, *Bioinformatics*, vol.19, no.4, pp.513-523 Jul. 2002.

[6] Bernard, Guillaume, Cheong XC, and Mark AR, “Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer.”, *Scientific Reports*, vol.6, no.1, Jul. 2016.

[7] Sims GE, Kim SH, et al, “Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions.”, *Prod, National Academy of Sciences*, vol.106, no.8, pp.2677-2682, Feb. 2009.

[8] Tang J, et al, “A novel k-word relative measure for sequence comparison.”, *Computational biology and chemistry*, vol.53, part.B, pp.331-338, Dec. 2014.

[9] Leimeister, Chris A, et al, “Fast alignment-free sequence comparison using spaced-word frequencies.”, *Bioinformatics*, vol.15, no.30, pp.1991-1999, Jul. 2014.

[10] Gunasinghe, Upuli, Damminda A, and Susan B, “Extraction of high quality k-words for alignment-free sequence comparison.”, *Journal of theoretical biology*, vol.358, no.1 , pp.31-51, Oct. 2014.

[11] Thompson JD, et al, “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.”, *Nucleic acids research*, vol.22, no.22, pp.4873-4680, Sep. 1994.

[12] Robinson D and Foulds LR, “Comparison of phylogenetic trees.”, *Mathematical Biosciences*, vol.53, no.1, pp.131-147, Feb. 1981.

[13] Sokal R, and Michener C, “A statistical method for evaluating systematic relationships.”, *University of Kansas Science Bulletin*, vol.38, no.1, pp.1409-1438, Jan. 1958.

[14] King J, Everett L, “Classification of sea lamprey (Petromyzon marinus) attack marks on Great Lakes lake trout (Salvelinus namaycush).”, *Canadian Journal of Fisheries and Aquatic Sciences*, vol.37, no.11, pp.1989-2006, Jan. 1980.