Journal of Engineering and Applied Sciences 13 (Special Issue 3): 3198-3203, 2018
ISSN: 1816-949X

A Relative Frequency-Based Signature Sequence Extraction Method for Two Contrasting Sequence Groups

Keon Myung Lee, [2]Chan Hee Lee and [3]Hyung Woo Youn
[1]Department of Computer Science, Chungbuk National University, Cheongju, 28644 Chungbuk, Korea
[2]Department of Microbilogy, Chungbuk National University, Cheongju, 28150 Chungbuk, Korea
[3]Department of Clinical Pathology, Chungbuk Health and Science University Cheongju,
28644 Chungbuk, Korea

Abstract: The advances in molecular technology enable to classify organisms based on their genetic sequence information. In classification, it is sometimes desirable to have a pattern for each class which characterizes the class and discriminates it from others. The objective of this research is to develop a method to extract signature sequences which is a sequential pattern with such characteristics from two contrasting sequence groups. It is assumed that there are two sequence groups, the self group and the other group and all the sequences from both groups are multiply aligned together, so that, they have the same length. To begin with for each group the relative base frequencies at each base location are computed and its consensus sequence is identified. From each base location of a group, the most frequent base is selected as a constituent of signature sequence only when its relative frequency is higher than that of the same base in the corresponding location of the contrasting group by at least the specified threshold called base frequency difference threshold, BDT. A candidate signature sequence is constructed by placing those selected bases in their location of a sequence. A desirable signature sequence for a sequence group is a sequential pattern which facilitates to discriminate its group from the other group and retains its unique group characteristics. In an experiment of virus sequences, the cross-validation study showed that the method produces consistent results and the generated signature sequences give the high sensitivity and high specificity for the sequence data set. The results indicate that the proposed signature sequence extraction method is useful in the characterization and classification of a group of sequences.

Key words: Sequence analysis, classification, feature extraction, signature sequence, data analysis, sequence extraction

INTRODUCTION

The classification is one of important tasks in sequence data analysis where classification accuracy are paid lots of attention and the patterns for classes acquired sometimes in the course of the task are regarded as their characterization. The classification approaches can be divided into two categories, based on how to use the data, sequence-based approaches and feature-based approaches. The sequence-based approaches use the sequences themselves as input to a classification system. The consensus sequence-based methods and the HMM (Hidden Markov Model)-based methods (Eddy, 1996) belong to this category. On the other hand in the feature-based approaches, some features expressed in numeric values are extracted from the sequences and then used as input to a classification system such as neural networks (Hornik *et al*., 1989), SVM (Support Vector Machine) (Furey *et al*., 2000). The sequence-based approaches usually produce sequential patterns as the representatives of the classes whereas the feature-based approaches either produce feature vectors as the class representatives or do not produce explicit representatives. A sequential pattern for a class contains some valuable information that can be used in motif identification, PCR primer design (Aluru, 2005), genetic therapy and so on. Lee *et al*. (2010, 2017) and Kim *et al*. (2014) such sequential pattern is called signature sequence (Baldi and Brunak, 2001).

A signature sequence for a group or class is desirable to facilitate to discriminate the group from the others and to retain the inherent characteristics of the group. This study is concerned with a method to identify such signature sequences for two groups of sequence data

Corresponding Author: Keon Myung Lee, Department of Computer Science, Chungbuk National University, Cheongju, 28644 Chungbuk, Korea

where one group is named the self group and the contrasting one is the other group. The proposed method makes use of the relative frequency information at each base location of sequence groups. It also measures how well a candidate signature sequence preserves the inherent characteristics of a group by inserting or deleting the signature information into or from the sequences.

MATERIALS AND METHODS

In order to get a signature sequence for a group of sequences, the proposed method makes use of the comparative information of a contrasting sequence group. The sequences from both groups are assumed to be multiply aligned together, so that, they have the same length. For the convenience of description, the following notations are used:

- Gs = {s_1, s_2, …, s_n}: the self group of size n and its sequences s_i
- s_i: (s^i_1, s^i_2, s^i_L) the ith sequence of length L in the self group Gs
  s^i_k : the base at base location k of sequence s_i
- Go = {o_1, o_2, …, o_m}: the other group of size m and its sequences o_j
- o_j = $(o^j_1, o^j_2, ..., o^j_3)$: the jth sequence of length L in the other group Go
- Cs = (cs_1, cs_2, …, cs_L): the consensus sequence for the self group Gs
- Co = (co_1, co_2, …, co_L): the consensus sequence for the other group Go
- Ss = (ss_1, ss_2, …, ss_L): a candidate signature sequence for the self group Gs
- So = (so_1, so_2, …, so_L): a candidate signature sequence for the other group Go
- Fs(x) = (fs_1(x), fs_2(x), …, fs_L(x)): relative frequencies of base x at each base location for Gs
- Fo (x) = (fo_1(x), fo_2(x), …, fo_L(x)): relative frequencies of base x at each base location for Go

For each group, the relative base frequencies are calculated over all base locations and then Fs(x) and Fo (x) are determined for each base, e.g., x∈{A, T, G, C} in case of DNA sequences. A base with a large relative frequency difference at the same base location for both groups could contribute a lot to the group discrimination. From each base location of a group, the most frequent base is selected as a constituent of signature sequence only when its relative frequency is higher than that of the same base in the corresponding location of the contrasting group by at least the specified threshold, called base frequency difference threshold, BDT. The following shows how to construct a candidate signature sequence Ss for the self group at BDT θ where $\text{argmax}_x fs_i(x)$ indicates the base x which maximizes $fs_i(x)$, i.e., the most frequent base at the location i in the self group and is the don't care base for the location in which no signature property is involved:

$$Ss = (ss_1, ss_2, ..., ss_L)$$

Where:

$$ss_i = \begin{cases} \text{argmax}_x fs_i\ (x) \text{ if } fs_i\ (\text{argmax}_x\ fs_i(x))- \\ Fo_i\ (\text{arg max}_x fs_i\ (x)) > \theta, \text{-, otherwise} \end{cases} \qquad (1)$$

The candidate signature sequences for the other group are constructed in the same way by changing the roles of the self group and the other group.

The value of BDT takes from the range 0-100%. Hence, various candidate signature sequences can be generated. This BDT-based candidate generation focuses only on the discrimination property which tells one group apart from the other. Optimal signature sequences are expected to have high discrimination capability to keep inherent characteristics of the group and to have small length. Here, the signature length is defined as the number of bases but the don't care base'-' in a signature sequence. In order to measure how well a candidate sequence retains the inherent characteristics of the group, the signature property is deleted from or inserted into sequences and it is observed how much they move away from or closer to the group. A way to delete signature property from a sequence is to replace the bases in the signature base locations with the corresponding bases of the consensus sequence for the contrasting group. The following shows how to generate a converted sequence s_i' by deleting the signature property from a sequence s_i of the self group:

$$s^{'}_I = (s^{i'}_1, s^{i'}_2, ..., s^{i'}_L)$$

Where:

$$s^{i'}_k = \begin{cases} co_k \text{ if } ss_k \neq \text{'-'} \\ o^i_k \text{ otherwise} \end{cases} \qquad (2)$$

In a similar way the signature property can be introduced into a sequence. The following formula shows how to produce a converted sequence $o^{'}_i$ from a sequence o_i of the other group by introducing the signature information:

$$o^{'}_i = (o^{i'}_i, o^{i'}_i, o^{i'}_i)$$

Where:

$$o_k^{i'} = \begin{cases} ss_k \text{ if } ss_k \neq \text{'-'} \\ o_k^i \text{ otherwise} \end{cases} \quad (3)$$

The closeness between two sequences or between a sequence and a signature sequence is measured by the distance measures which evaluate how much they have in common. The distance $d(a_1,a_2)$ between two sequences a_1 and a_2 is measured by the ratio of shared bases:

$$d(a_1, a_2) = 1\text{-}\frac{|a_1 \cap a_2|}{L} \quad (4)$$

The distance $d_c(a)$ of a sequence a to a signature sequence c is measured by how much the sequence contains the signature sequence:

$$d_c(a) = 1\text{-}\frac{|a \cap c|}{|c|} \quad (5)$$

Here, |c| indicates the number of bases but '-' in the sequence c and |a∩c| the number of locations whose bases are the same for both a and c.

Figure 1 shows the relationship between the relative signature length and BDT for the data set used in the experiment. The higher the value of BDT is the shorter the signature length is. The shortest signature sequence is not always best because the shortest one may not hold the inherent characteristics of its group even though it gives high accuracy. The proposed method regards as the best signature sequence the one whose removal from the self group sequences moves away from the self group as far as its insertion to the other group sequences moves closer to the self group.

Figure 2 shows how the deletion and insertion of signature information moves the converted self and other group sequences to the self group. In Fig. 2a self-self line indicates the average distance between self sequences computed by $d_{ss}(G_s)$:

$$d_{ss}(G_s) = \frac{1}{n(n\text{-}1)} \sum_{si \in Gs} \sum_{oj \in S,\ j \neq i} d(s_i, s_j) \quad (6)$$

Self-other line indicates the average distance between self sequences and other sequences computed by $d_{so}(G_s, G_o)$:

$$d_{so}(G_s, G_o) = \frac{1}{n \times m} \sum_{si \in Gs} \sum_{oj \in Go,\ j \neq i} d(s_i, o_j) \quad (7)$$

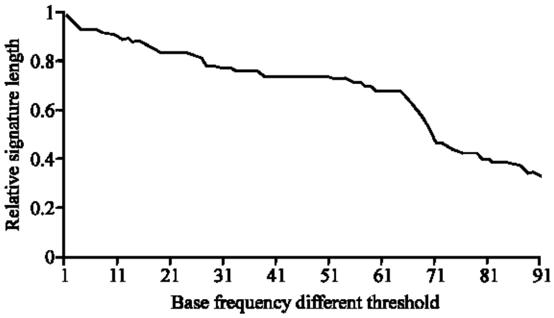

Fig. 1: The signature length trend over changing base difference rate

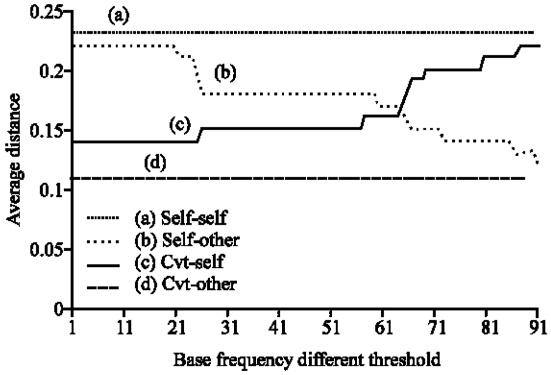

Fig. 2: Distance changes over the base difference rate

Cvt-self curve shows the average distance of the signature-deleted self sequences to the self group computed by $d_{cs}(G_s, S_{s\theta})$ where s_i' is the converted self sequence for s_i obtained with respect to the signature sequence Ss_θ at BDT^θ:

$$d_{cs}(G_s, S_{s\theta}) = \frac{1}{n(n\text{-}1)} \sum_{si \in Gs} \sum_{sJ \in Go} d(s_i, o_j') \quad (8)$$

Cvt-other curve shows the average distance of the signature-inserted other sequences to the self group computed by $d_{co}(G_s, G_o, Ss_\theta)$ where o_j is the converted self sequence for o_j obtained with respect to the signature sequence Ss_θ at BDT^θ:

$$d_{co}(G_s, G_o, Ss_\theta) = \frac{1}{n \times m} \sum_{si \in Gs} \sum_{oj \in Go} d(s_i, o_j') \quad (9)$$

Because the signature length gets shorter as the BDT gets larger (c) cvt-self curve tends to decrease while (d) cvt-other curve tends to increase, as shown in Fig. 2. If a signature sequence contains the essential characteristics of the self group, its deletion from the self sequences moves them away from the self group and its introduction to the other sequences moves them closer to the self group. Once the deletion curve (c) is the insertion curve (d), the corresponding signature sequences are not enough to move away the self sequences from the self group and to move closer the other sequences to the self group. That is, the signature sequences corresponding to the BDTs whose deletion curve is below the insertion curve, seem to be over-simplified not to hold the essential characteristics of the self group. In the same token, the signature sequences corresponding to the BDTs whose deletion curve is above the insertion curve, might contain unnecessary information as well as the essential characteristics of the self group. From these observations, the proposed method takes as a best one the signature sequence at the BDT with which the deletion curve (c) approaches most closely to the insertion curve (d). The best BDT θ_{best} is determined by the following equation where $\text{argmin}_{(\theta)D(\theta)}$ gives θ to minimize $D(\theta)$:

$$\theta_{best} = \text{argmin}_{\theta}\ D(\theta)$$

Where:

$$D(\theta) = d_{cs}(G_s, Ss_{\theta})\text{-}d_{co}(G_s, G_o, Ss_{\theta}) \text{ and } D(\theta)>0 \tag{10}$$

Once a candidate sequence is selected as the signature sequence, it can be used to decide whether a new sequence belongs to the group. In order to get the criteria for the group membership, the matching scores of the sequences to the signature sequence of the group are computed by the function $m_s(a) = 1\text{-}d_s(a)$. The sample mean $\bar{m}$ and the sample standard deviation $\bar{s}$ are then computed for the matching scores. The following shows how they are computed in case of the self group:

$$\bar{m} = \frac{1}{n}\sum_{sj\in G_s} m_s(s_j),\ \bar{s} = \sqrt{\frac{1}{n\text{-}1}\sum_{Sj\in G_s}(m_s(s_j)\text{-}\bar{m})^2} \tag{11}$$

When a new sequence a is given to a group with the signature sequence s, it is classified to the group only if $m_s(a) \geq m\text{-}k\times s$ where k is a constant to be specified by the analysts.

RESULTS AND DISCUSSION

To evaluate the proposed signature sequence extraction method, we applied it to the Korean HIV-1 (Human Immunodeficiency Virus type 1), isolates (Park *et al*., 2008) which are retrieved from Gene Bank. Among them, the HIV-1 nef sequences was extracted and grouped into the Korean clade of size 264 and the non Korean clade of size 71. For the sake of consistent description, the Korean clade is designated as the self group and the non-Korean clade as the other group. The 335 sequences from both group was multiply aligned together and came to have the length of 621.

The 10 fold cross validations (Refaeilzadeh *et al*., 2009) were carried out four times with the different random partitioning of the data set. In each cross-validation, the signature sequences were constructed upon the training set (i.e., 9/10 of the total sequences) and evaluated over the remaining set (i.e., 1/10 of the total sequences). At each run, a best signature sequence is determined according to the proposed method. For the validation sets of the self group and the other group, their matching scores to the signatures were evaluated and the score threshold for the group membership was set to m-2. Then, the sensitivity and the specificity were computed for both groups. For the self signature sequences, the sensitivity was 98.8% and the specificity was 100% where sensitivity is the percentage with which the self sequences are classified into the self group and specificity is the percentage with which the other sequences are classified into the other group. For the other signatures, the sensitivity was 98.2% and the specificity was 100%.

Figure 3 shows the histogram of the matching scores of self and other sequences to the consensus sequence of the self group where matching scores were computed using the function $m(a_1, a_2) = 1\text{-}d(a_1, a_2)$.

Figure 4 shows the histogram of their matching scores to a signature sequence of the self group. From these histograms, we observe that compared to the

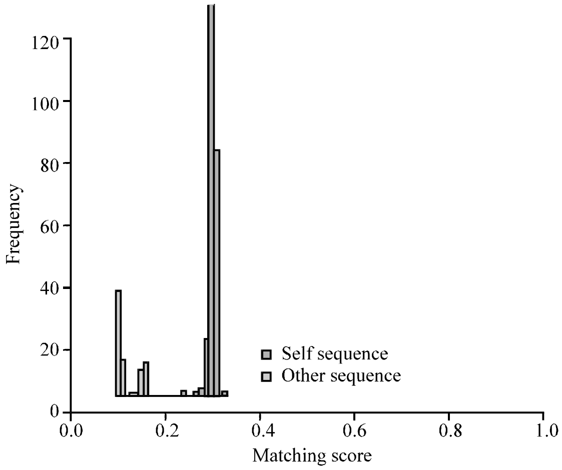

Fig. 3: Histogram of the matching scores of self sequences (red ones) and other sequences (gray ones) to the consensus sequence of the self group

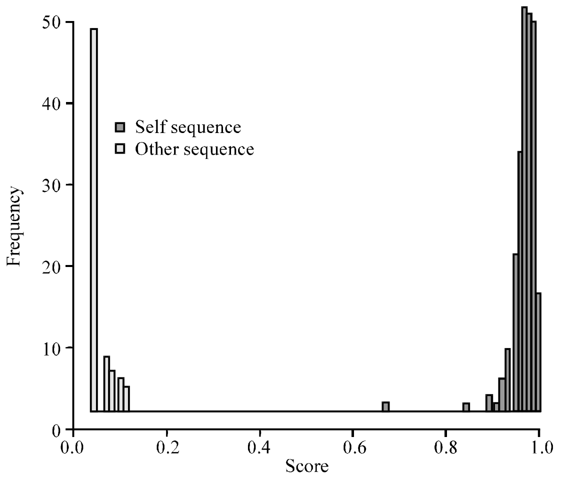

Fig. 4: Histogram of the matching scores of self sequences (red ones) and other sequences (gray ones) to a signature of the self group

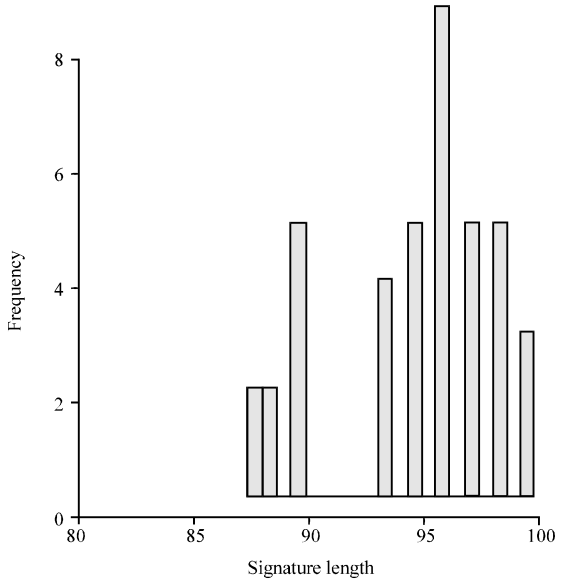

Fig. 5: Histogram of the signature lengths for the self group

consensus sequence, the signature sequences generated by the proposed method have more desirable property to cause the matching scores to be more apart between the contrasting groups.

Figure 5 and 6 show the histograms of signature lengths for the self group and the other group. While the length of the consensus sequence for the groups was 621, the average length of signature sequences for the self group was 91 and thus its length was 14.7% of the self group's consensus sequence and that for the other group the average length of signature sequences was 41.6 and thus its length was 6.6% of the other group's consensus sequence.

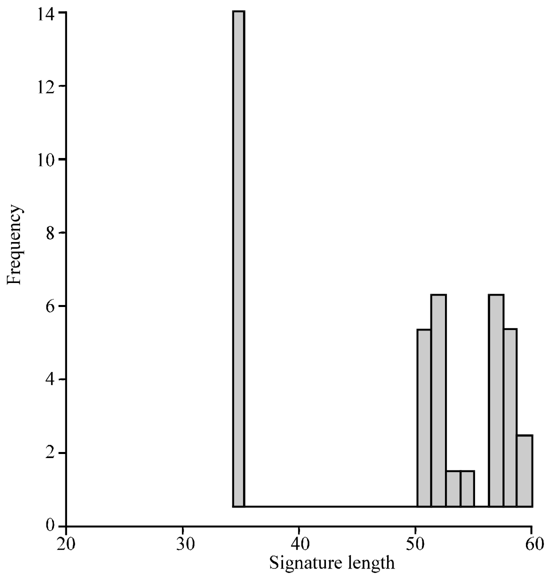

Fig. 6: Histogram of the signature lengths for the other group

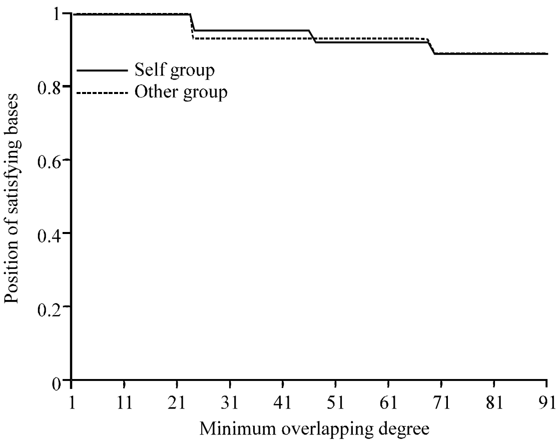

Fig. 7: The distribution of signature bases which meet the minimum overlapping criteria

Figure 7 shows the portions of the bases satisfying the minimum occurrences at each signature base locations over all the generated signature sequences. It indicates that the signature sequences are consistent in that signature bases appear over most of the signature sequences constructed in the course of cross-validation. Figure 8 and 9 show signature sequences extracted by the proposed method for HIV-1 nef Korean clade and non-Korean clade, respectively.

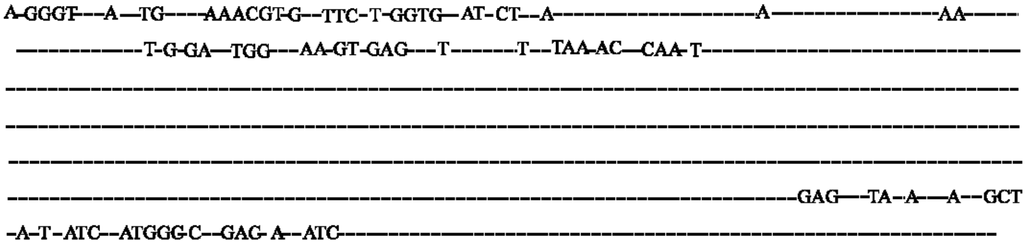
A-GGGT---A—TG----AAACGT-G--TTC-T-GGTG—AT-CT--A A AA
T-G-GA—TGG---AA-GT-GAG---T T--TAA-AC—CAA-T
GAG—-TA-A—A--GCT
-A-T-ATC--ATGGG-C--GAG-A--ATC

Fig. 8: A signature sequence of length 85 for HIV-1 nef Korean clade

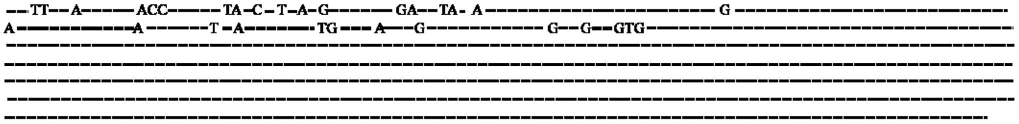

Fig. 9: A signature sequence of length 31 for HIV-1 nef non-Korean clade

CONCLUSION

This study proposed a method for signature extraction for two contrasting sequence groups. The proposed signature extraction method generates a signature sequence with the consideration of the classification accuracy and the preservation of the inherent characteristics of the groups. From the experiment on the Korean HIV-1 nef sequences data set, the method showed the capability to generate short signature sequences with high and consistent classification accuracy.

ACKNOWLEDGEMENT

This research was supported by Next-generation Information Computing Development Program through the National Research Foundation of Korea (Grant No.: NRF-2017M3C4A7069432).

REFERENCES

Aluru, S., 2005. Handbook of Computational Molecular Biology. CRC Press, Boca Raton, Florida, USA., ISBN:13-978-1-58488-406-4.

Baldi, P. and S. Brunak, 2001. Bioinformatics: The Machine Learning Approach. 2nd Edn., MIT Press, Cambridge, Massachusetts, USA., ISBN:978-0262-025065, Pages: 447.

Eddy, S.R., 1996. Hidden markov models. Curr. Opin. Struct. Biol., 6: 361-365.

Furey, T., N. Cristianini, N. Duffy, D. Bednarski, M. Schummer and D. Haussler, 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16: 906-914.

Hornik, K., M. Stinchcombe and H. White, 1989. Multilayer feedforward networks are universal approximators. Neural Networks, 2: 359-366.

Kim, Y.K., S.Y. Lee, S. Seo and K.M. Lee, 2014. Fuzzy logic-based outlier detection for bio-medical data. Proceedings of the 2014 International Conference on Fuzzy Theory and its Applications (iFUZZY), November 26-28, 2014, IEEE, Kaohsiung, Taiwan, ISBN:978-1-4799-4588-7, pp: 117-121.

Lee, K.M., K.S. Hwang and W.J. Kim, 2010. Contrasting cluster mining in microarray data. Proceedings of the 5th and 11th International Conference on Soft Computing and Intelligent Systems and Advanced Intelligent Systems, December 8-12, 2010, Okayama Convention Center, Okayama, Japan, pp: 938-941.

Lee, K.M., S.Y. Lee, K.M. Lee and S.H. Lee, 2017. Density and frequency-aware cluster identification for spatio-temporal sequence data. Wireless Pers. Commun., 93: 47-65.

Park, C.S., D.H. Lee, K.M. Lee and C.H. Lee, 2008. Characterization and signature pattern analysis of Korean clade HIV-1 using nef gene sequences. J. Microbiol., 46: 88-94.

Refaeilzadeh, P., T. Lei and H. Liu, 2009. Cross-Validation. In: Encyclopedia of Database Systems, Liu, L. and M.O. Tamer (Eds.). Springer, Berlin, Germany, ISBN:978-0-387-35544-3, pp: 532-538.