Representative Pattern Extraction for Nucleotide Sequence Groups Using Base Frequency Differences

Kyoung Soon Hwang, Keon Myung Lee, Sung Soo Kim,
School of Electrical and Computer Engineering
Chungbuk National University, Korea

Chan-Hee Lee
Department of Microbiology,
Chungbuk National University, Korea

Hyung Woo Yoon
Department of Clinical Laboratory Science,
Juseong College, Korea

Sung Duk Lee
Department of Statistics,
Chungbuk National University, Korea

Abstract

Advances in high-throughput technology in molecular biology have been producing lots of sequence data on various organisms. Some organisms like virus have various variances in their nucleotide sequences and could be categorized into several subtypes. A sequential pattern which characterizes a subtype and discriminates it from other subtypes is called signature. This paper proposes a method which extracts signature from a collection of sequences data. Based on position specific relative base frequency deference between one subtype data set and the other subtype data set, the proposed method examines discrimination capabilities for the potential signatures. A tool has been developed which implements the proposed method and applied to an experiment to extract signatures for HIV-1 virus subtypes.[1]

1. Introduction

With the help of high throughput sequencing technologies and efficient computational tools, the genomes for dozens of organisms have been successfully sequenced and archived into biological databases. A genome is the complete genetic content of an organism, which consists of base pairs of DNA. The DNA bases are molecules which are called adenine(A), thymine(T), guanine(G), or cytosine(C). The genetic information of any life form is coded in nucleotide sequences and is transferred to its descents through the inheritance mechanism.[1,5] The changes in the nucleotide sequences have incurred differentiation and variations in species. Such changes are frequently observed in lower organisms such as virus. Especially, in virology, lots of efforts have been paid to sequence analysis for virus origin estimation, virus subtyping, phylogenetic analysis, vaccine development, and so on.[6] A sequential pattern which characterizes a subtype and discriminates it from other subtypes is called *signature*.[5] Signature extraction is one of important issues in the subtype analysis for likely varying organisms such as virus.

Once a collection of nucleotide sequences are given, the molecular biologists categorize them into groups according to their subtype label, align each group with a multiple sequence alignment tool, and then extract manually a signature based on their expertise. Some HMM(hidden Markov model)-based probabilistic models have been developed to represent a subtype, yet such models focus on a specific collection of sequence data set without comparing to the other subtypes.[2,4] Therefore, such probabilistic models have some weakness in description of unique features for a specific subtype and in addition they are not easily understandable because they are expressed as uninterruptible states and probability values. The consensus sequence for a collection of sequence data set could be used to extract features of a subtype, yet it is not enough to characterize a subtype in a way to discriminate the subtype from others. In the lights of these observations, we have proposed a new signature extraction method that generates a signature in a sequential pattern which characterizes and discriminate a subtype from others. In the proposed method, signature characters which makes up a signature are determined based on position-specific base frequency difference, specificity and sensitivity are taken into account in order to measure the discrimination capability.

[1] This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD) (The Regional Research Universities Program/Chungbuk BIT Research-Oriented University Consortium).

The remainder of this paper is organized as follows: Section 2 presents some related worked to signature extraction in the nucleotide sequence analysis. Section 3 describes the proposed signature extraction method and Section 4 presents how to classify sequences using signature pattern. Section 5 discusses how to choose the best signature and its control parameters. Section 6 presents the developed tool which implements the proposed method and some experiment. In final, Section 7 draws the conclusions.

2. Related Works

For the purpose of characterizing the features of a collection of biological sequence data set, various feature extraction methods have been used. A simple approach is manual extraction which is carried out as follows: For the given collections of a subtype and other subtypes, a biological analyst applies to each subgroup a multiple sequence alignment tool like ClustalW which produces biologically meaningful multiple sequence alignments of divergent sequence by calculating the best match for the selected sequences and lining them up so that similarities could be seen. Comparing the multiple aligned sequences for each subtype, she manually extracts the signature characters based on her expertise. This approach is widely used in practice, yet it is rather hard to justify the obtained signature due to its subjective determination.

HMM-based model[4] could be used to describe a group of sequence data set, in which an HMM is trained for a sequence data set of a subtype and such models is used to tell to which subtype a given sequence is closest. An HMM is a stochastic model that consists of a set of nodes of which state is hidden, transitions between nodes take places in a probabilistic way, and the outputs at nodes are generated probabilistically. Due to its probabilistic nature, it is relatively difficult for the analysts to figure out the signature embedded in an HMM.

Consensus sequence might be regarded as a kind of signature which is a single sequence delineated from an alignment of multiple constituent sequences that represents a best for all those sequence.[6] In consensus sequence construction, a voting or other selection technique is used to determine which base is selected at a given position. A consensus sequence reflects the features for the subtype under consideration without reference to other subtype, so that it does not contain enough information to discriminate a subtype from other subtypes.

3. Nucleotide Sequence Signature Extraction based on Position-specific Base Frequency Differences

The proposed signature extraction method makes use of relative base frequency information for the groups of aligned sequences. To measure the quality of potential signatures, some measures are used to characterize the discrimination capacity.

For the convenience of description, the following notations are used:

$SG = \{SG_1, SG_2\}$: a collection of sequence groups

$SG_i = \{N_1^i, N_2^i, ..., N_{E_i}^i\}$: the set of aligned sequences in group i, where E_i denotes the number of sequences in SG_i

$N_j^i = (n_1, n_2, ..., n_L)$: the j-th sequence in SG_i

$m(k, j)$: the nucleotide with the maximum frequency at position j in group k

$s(k, j)$: the relative frequency of $m(k, j)$

$f(k, j, n)$: the relative frequency of nucleotide n at position j in group k

Figure 1 is an example to show the relationships for $f(k, j, n)$, $m(k, j)$, and $s(k, j)$.

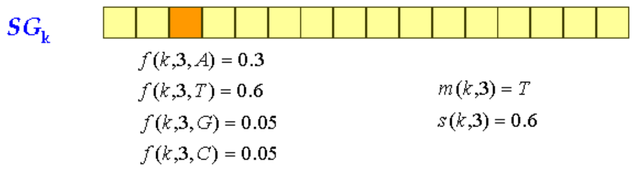


Figure 1. Relationships for $f(k, j, n)$, $m(k, j)$, and $s(k, j)$

In the proposed method, it is assumed that there are two groups of sequences which are aligned with a multiple sequence alignment tool. The biological observations show that genetic mutations for a specific group of virus sequences tend to take place in some localized regions.[7] In order to select discriminating features for a group, we have paid attention to the relative base frequencies in each position of aligned sequences. For each position, the bases with the maximum frequency are compared and the corresponding positions are regarded as being a candidate to constitute a signature once either the compared bases are different or their relative frequency difference is greater than or equal to a pre-specified threshold called *discrimination threshold* δ. Eq. (1) shows the condition that the position i for group SG_1 becomes a candidate position used to define a signature.

$$f(1, i, m(1, i)) \geq f(2, i, m(1, i)) + \delta \quad (0 < \delta < 1) \quad (1)$$

The maximum frequency base character $m(1, i)$ holding the above condition becomes a signature character and its relative frequency $s(1, i)$ is used as its weight in the signature description.

The signature S_k for group k is defined as follows: The signature as shown in Eq.(2) is composed of a sequence of pairs of a signature character and its weight.

$$S_k = (s_1^k, s_2^k, \ldots, s_L^k) \quad (2)$$

where

$$s_i^k = \begin{cases} <m(k,i), s(k,i)> & \text{if } m(k,i) \text{ is a signature character} \\ <'-', 0> & \text{otherwise} \end{cases}$$

The following procedure *BuildSignaturePatterns* describes the steps for building signature patterns. Here, N denotes the length of multiple assigned sequences used in the signature extraction. All sequences in the sets SG are assumed to have the same length since they are expected to be multiple aligned in the preprocessing phase.

```
procedure BuildSignaturePatterns($SG$, $\delta$, $S$)
  input:
    $SG = (SG_1, SG_2)$: a collection of sequence groups
    $\delta$ : discrimination threshold
  output:
    $S = \{S_1, S_2\}$: the set of signature patterns
begin
  for $k = 1$ to 2 do
    for $j = 1$ to $L$ do
      for each $n \in \{A, G, T, C\}$
        compute $f(k, j, n)$
      determine $m(k, j)$
      set $s(k, j)$
  for $k = 1$ to 2 do
    $c = -k + 3$
    for $j = 1$ to $L$ do
      if $f(k, j, m(k, j)) > f(c, j, m(k, j)) + \delta$
        $b_j^k = 1$
      else
        $b_j^k = 0$
  for $k = 1$ to 2 do
    for $j = 1$ to $L$ do
      if $b_j^k = 1$
        $s_i^k =< m(k, j), s(k, j) >$
      else
        $s_i^k =<' -', 0 >$
  return $S = \{S_1, S_2\}$, where $S_k = (s_1^k, s_2^k, ..., s_L^k)$.
end.
```

4 Sequence Classification based on Signatures

Signatures for a group of sequences are patterns discriminating the group from others. Therefore, they can be used to determine whether some sequence belongs to a specific group. A signature can be considered as a template to represent a group and we need a score function to measure the similarity of a given sequence to a template for the group.

In the proposed method, the following score function $G(N, S_k)$ is used to measure the matching degree of N to S_k. Here $N = (n_1, n_2, ..., n_L)$ is a sequence of which group is not known yet, and S_k is a signature for group k:

$$G(N, S_k) = \frac{\sum_{j=1}^{L} s(k,j) \cdot 1(n_j = m(k,j))}{\sum_{j=1}^{L} s(k,j)} \quad (3)$$

In the above equation, $1(n_j = m(k,j))$ is 1 when $n_j = m(k,j)$, 0 otherwise. The input sequence N is assumed to be aligned with the sequences SG_k of group k and thus it has the same length L. The score function gives the matching degree on the range $[0, 1]$.

When a new sequence N is given to determine its group, its matching score $G(N, S_k)$ is computed with respect to the signature S_k of each group. If it is greater than the threshold called *stringency threshold* θ_k, N is regarded as belonging to the group k.

$$\text{if } S(N, S_k) > \theta_k, N \text{ is classified to group } k \quad (4)$$

It is possible for a sequence to belong to multiple group at the same time depending on the specified stringency thresholds. In that situation, the judgement is left to the analysts. The choice of the proper stringency thresholds θ_k is one of important issues in the sequence classification based on signatures.

5 Selection of the Best Signatures

As mentioned in previous sections, the proposed signature extraction and classification method has two control parameters which strongly affects the qualities of extracted signatures and classification. One is the discrimination threshold δ_k used in Eq.(1), which controls which bases to be selected as signature characters. The other is the stringency threshold θ_k used in Eq.(4), which specifies the minimum matching degree with which a sequence has so as to be a member of group k.

The best signature would be the simplest ones which could discriminate the corresponding group from other groups. For a group k, the proposed signature extraction method generates a signature $S_k(\delta_i)$ with respect to the given discrimination threshold δ_i. The signature size $|S_k|$ for S_k is defined as the number of positions with signature information. That is, $|S_k| = |\{s_i^k \mid s_i^k =< m(k,i), s(k,i) > \text{ and } m(k,i) \neq' -'\}|$.

It is necessary to have some measures for signature quality in the perspective of discrimination capability. To choose an appropriate signature, the proposed method uses the sensitivity and the specificity of signatures. The sensitivity $\sigma_k(\theta_i)$ of signature S_k at the stringency threshold

θ_i estimates the percentage that the signature S_k correctly classifies the members of the sequence group k into the very group.

$$\sigma_k(\theta_i) = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Negative}} \quad (5)$$

The specificity $\psi_k(\theta_i)$ of signature S_k at the stringency threshold θ_i estimates the percentage that the signature S_k correctly classifies the members of the other group into the other group.

$$\psi_k(\theta_i) = \frac{\textit{True Negative}}{\textit{True Negative} + \textit{False Positive}} \quad (6)$$

The most desirable signature is one of which sensitivity and specificity are equal to 1. Due to inherent trade-off relationship between sensitivity and specificity, we choose the signature $S_k(\delta_k^{best})$ which maximizes the sum of sensitivity and specificity. The stringency threshold for a signature is determined at the point to maximize the sum of sensitivity and specificity. The stringency threshold θ_k^{best} for a signature S_k at the discrimination threshold δ_i is defined as follows:

$$\theta_k^{best} = argmax_{\theta_i}\ \{\sigma_k(\theta_i) + \psi_k(\theta_i)\} \quad (7)$$

When there are multiple signatures to give the same quality, the simplest one with the smallest size is preferred. In the situation that there are multiple signatures with the same size and discrimination quality, the ROC(receiver of characteristic) curve is considered. ROC curve is a 2D graph of which x-axis corresponds to $1-\psi_k(\theta_i)$ and y-axis to $\sigma_k(\theta_i)$. The area under an ROC curve ROC_{area} is a useful measure to reflect the quality of classifier, which indicates the accumulated area under the curve in the ROC curve graph. The signatures with larger ROC_{area} is preferred.

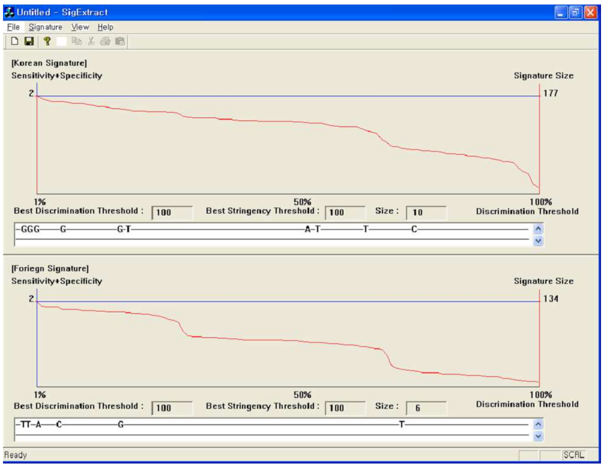
The following procedure *FindSignatureThresholds* describes how to determine the best signature and its corresponding discrimination and stringency thresholds.

```
procedure FindSignatureThresholds($SG$, $S_1^{best}$, $\sigma_1^{best}$, $\theta_1^{best}$)
  input:
    $SG = (SG_1, SG_2)$: a collection of sequence groups
  output:
    $S_1^{best}$ : the best signature
    $\sigma_1^{best}$ : the best discrimination threshold
    $\theta_1^{best}$ : the best stringency threshold
begin
  for $\delta = 0.01$ to 1 step 0.01 do
    Determine the signature $S_1(\delta)$ using the procedure
      BuildSignaturePatterns.
    for $\theta = 0.01$ to 1 step 0.01 do
      Classify the sequences $SG$ using the rule given
        in Eq.(4).
      Compute the sensitivity $\sigma_1(\theta)$ and specificity
        $\psi_1(\theta)$ based on the classification results.
    end for
    Determine the best stringency threshold $\theta_1^{best}(\delta)$ at
      the discrimination threshold $\delta$.
  end for
  Compute $\Delta_1^{best} = \{\delta' | \delta' = argmax_\delta \theta_1^{best}(\delta)\}$.
  Choose the set $\Delta_1^{best'}$ of discrimination thresholds whose
    size is the smallest among $\Delta_1^{best}$.
  if $|\Delta_1^{best'}| > 2$
    Choose the one with the largest $ROC_{area}(\delta)$
      as the best discrimination threshold $\delta_1^{best}$.
  else
    Choose the value in $\Delta_1^{best'}$ as the best
      discrimination threshold $\delta_1^{best}$.
  end if
  Set the signature $S_k(\delta_1^{best})$ to the best signature $\delta_1^{best}$.
  Set the stringency threshold $\psi_1(\delta_1^{best})$ to $\psi_1^{best}$.
  return $S_k(\delta_1^{best})$, $\delta_1^{best}$, and $\psi_1^{best}$.
end.
```

6. Implementation and An Application

The proposed signature extraction method has been implemented as a tool with graphical user interface in Windows environment. The developed tool allows the analysts to manipulate parameters and to analyze the results visually. Figure 2 show the main interface of the developed tool which displays the trends of the quality and size of found signatures, the found best signature and its corresponding discrimination and stringency thresholds, and some graphs for analysis. Figure 3 shows the interface to display the trends of sensitivity and specificity over the range of stringency threshold values at a specific discrimination threshold. Figure 4 shows another visualization interface to display the matching scores of sequences to the found best signature and an ROC curve. With the help of these visualization functionalities, the analysts can get easily the pictures on the quality and performance of the extracted signatures.

In order to evaluate the applicability of the proposed method and developed tool, we have done an experiment to apply it to the signature extraction task for 264 HIV-1 Korean subtype viruses and 71 HIV-1 foreign subtype viruses. Half of the sequences have been used to construct signatures for the Korean subtype and the foreign subtype, and the remaining ones have been used for performance evaluation. The developed tool has produced a signature of size 10 for the HIV-1 Korean subtype viruses and a signature of size 6 for the foreign subtype viruses as the best ones. The extracted signatures have been applied to classify the evaluation sequence sets and it could classify all sequences by very high accuracy, around 95 percent.



[Korean Signature]

Best Discrimination Threshold : 100 Best Stringency Threshold : 100 Size : 10

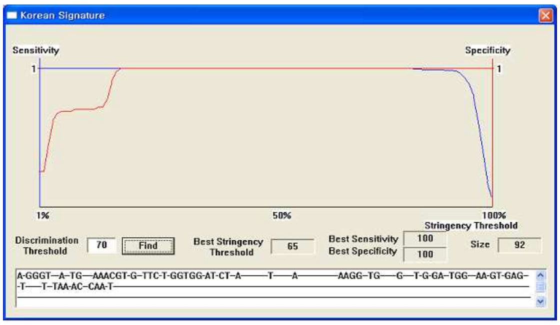
```
-GGG——G————G-T————————————A–T————T————C————
```

[Foriegn Signature]

Best Discrimination Threshold : 100 Best Stringency Threshold : 100 Size : 6

```
-TT–A—C————G——————————————————T————————
```

Figure 2. The developed tool for signature extraction



Discrimination Threshold 70

Best Stringency Threshold 65

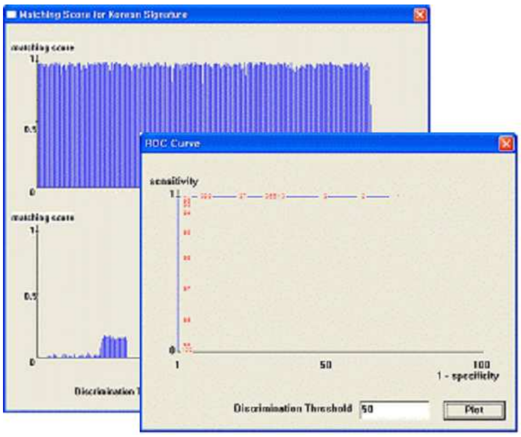
Best Sensitivity 100

Best Specificity 100

Size 92

```
A-GGGT—A–TG—AAACGT-G–TTC-T-GGTGG-AT-CT–A———T——A————AAGG–TG——G—T-G-GA–TGG—AA-GT-GAG–
-T——T–TAA-AC–CAA-T——————————————————————
```

Figure 3. The trends of sensitivity and specificity over the range of stringency threshold values.



Discrimination Threshold 50

Figure 4. A ROC curve and the matching score graph

7. Conclusions

A signature for a group of nucleotide sequences is a characteristic pattern for the group. We has proposed a signature extraction method which takes into account the relative frequencies at each base and discrimination capability. The behavior of the proposed method is affected by two control parameters, discrimination and stringency thresholds. The proposed method is facilitated to automatically determine those parameters. The proposed method comes up with a simplest signature with high discrimination capability. The tool embodied the proposed method has been developed and successfully applied to a real problem to construct the signatures for two HIV-1 types.

This study has focuses on finding signatures which best discriminates a group from other groups. Some biologists in the field express their expectation to extract signatures with not only high discrimination capability but also inherent characteristics of group itself. As a further study there remains to develop some method to find such signatures.

REFERENCES

[1] S. Aluru, Handbook of Computational Molecular Biology (eds.), Chapman & Hall/CRC, 2006.

[2] M. Kanehisa, Post-Gemome Informatics, Oxford, 1999.

[3] J. Barrera, R.M. Cear, C. Humes, D.C. Martins, D.F. Patrao, P.J. Silva, H. Brentani, A feature selection approach for identification of signature genes from SAGE data, BMC Bioinformatics, No.8, 169, 2007, May.

[4] A.K. Jain, R.P.W. Duin RPW, J. Mao, Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol22, pp.4–37, 2000.

[5] P. G. Higgs, T. K. Attwood, Bioinformatics and Molecular Evolution, Blackwell Publishing, 2005.

[6] G. B. Fogel, D. W. Corne, Evolutionary Computation in Bioinformatics (eds.), Morgan Kaufmann Publihers, 2003.

[7] C. H. Cannon, C. S. Kua, E. K. Lobenhofer, and P. Hurban Capturing genomic signatures of DNA sequence variation using a standard anonymous microarray platform Nucleic Acids Res.,Vol34, Oct. 2006.