Intelligent Data Recognition of DNA Sequences Using Statistical Models

Jitimon Keinduangjun[1], Punpiti Piamsa-nga[1], and Yong Poovorawan[2]

[1] Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok, 10900, Thailand
{jitimon.k, punpiti.p}@ku.ac.th
[2] Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok, 10400, Thailand
yong.p@chula.ac.th

Abstract. The intelligent data acquisition in biological sequences is a hard and challenge problem since most biological sequences contain unknowledgeable, diverse and huge data. However, the intelligent data acquisition reduces a demand on the use of high computation methods because the data are more compact and more precise. We propose a novel approach for discovering sequence signatures, which are sufficiently distinctive information in identifying the sequences. The signatures are derived from the best combination of the *n*-grams and the statistical scoring models. From our experiments in applying them to identify the Influenza virus, we found that the identifiers constructed by too short *n*-gram signatures and inappropriate scoring models get low efficiency since the inappropriate combinations of *n*-gram signatures and scoring models bring about unbalanced class and pattern score distribution. However, the other identifiers provide accuracy over 80% and up to 100%, when they apply an appropriate combination. In addition to accomplishing in the signature recognition, our proposed approach also requires low computation time for the biological sequence identification.

1 Introduction

The rapid growth of genomic and sequencing technologies during the past few decades has facilitated the incredibly large size of diverse genome data, such as DNA and protein sequences. However, most biological sequences contain very little known meaning. Therefore, techniques for knowledge acquisition from sequences become more important for transforming the sequences into useful, concise and compact information. These techniques generally consume long computation time; their accuracy usually depends on data size; and there is no best known solution. Many sequence processing projects still have some common stages for experiments, which are recognition of the most significant characteristics (Intelligent Data -- Signatures).

Signatures are short informative data that can identify types of the sequences. Recently, several biological research areas demand the informative signatures as one of important keys of success in the research areas since the signatures help reduce the computation time and also the data are more compact and more precise [4]. Many

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 630 – 635, 2005.

computational techniques are used in biological research areas such as *Sequence Alignment*, *Inductive Learning* and *Consensus Discovery*.

The well-known tool using the *Sequence Alignment*, such as BLAST [5], aligns uncharacterized sequences with the existing sequences in database and then assigns the uncharacterized sequences to the same class with one of the sequences in database that gets the best alignment score. This technique has to perform directly on all sequences in database, whose sizes are usually huge; therefore its processing time is much higher than other techniques. The *Inductive Learning* [9] and the *Consensus Discovery* [8] perform their tasks in a pre-process to derive rules that are used to perform the tasks during processing time. Although the use of the rules attains low processing time, the procedure for deriving rules still has too long computation time.

Our approach is to discover DNA signatures in the pre-process as the inductive learning and the consensus discovery. However, the proposed approach has much less pre-processing time. We apply an *n*-gram method and statistical scoring models for the signature discovery and evaluate the signatures over the influenza virus. Our system and methods are described in Section 2. Section 3 has a discussion on our experiments. Finally, we summarize the proposed approach in Section 4.

2 System and Methods

The signature recognition is a significant task in the Computational Biology research since the biological sequences are zero-knowledge based data. We do not know any "knowledge" of the biological data. Therefore, the recognition of the knowledge or informative signatures becomes more important. Our signature discovery framework of biological data is depicted in Fig. 1 as the four following steps.

Step I: *Pattern Generation* is to transform training data into *n*-gram patterns.

Step II: *Candidate Signature Selection* is to find the most significant *n*-gram patterns predicted as candidate signatures.

Step III: *Identifier Construction* is to create identifiers using the candidate signatures.

Step IV: *Performance Evaluation* is to estimate the goodness of the candidate signatures by comparing each identifier, generated by the different candidate signatures. If the accuracy of any identifiers is high, the candidate signatures used for constructing the identifiers are predicted as "Signatures" of the DNA sequences.

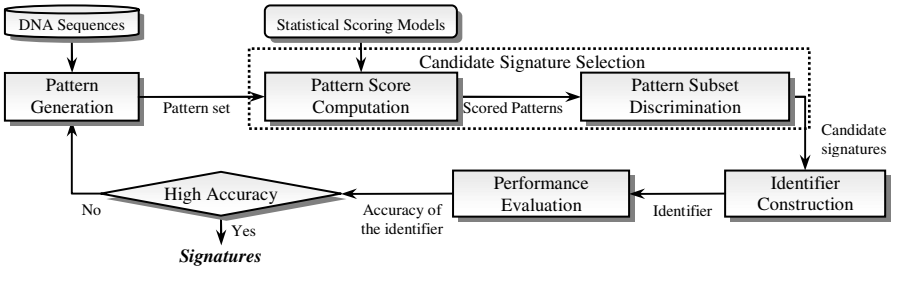

Fig. 1. Signature discovery framework

2.1 Pattern Generation

The Pattern Generation process is to find a data representation of DNA sequences. The data representation, called a pattern, is a substring of the DNA sequences. DNA sequences include four symbols {A, C, G, T}; therefore, the members of a pattern are also restricted to the four symbols. Let Y be a sequence $y_1y_2...y_M$ of length M over the four symbols of DNA. A substring $t_1t_2...t_n$ of length n is called an n-gram pattern [3]. The n-gram method generates the n-gram patterns representing the DNA sequences with different n-values. There are M-n+1 patterns in a sequence of length M for generating the n-gram patterns, but there are only 4^n possible patterns for any n-values. Notice that 1-gram patterns have 4 (4^1) possible patterns; while 24-gram patterns have 2.8×10^{14} (4^{24}) possible patterns. The numbers of possible patterns obviously vary from 4 to 2.8×10^{14} patterns. High n-values are not necessary since each pattern does not occur repeatedly and scoring models cannot discover signatures from these patterns. However, in our experiments, the patterns are solely generated from 1- to 24-grams since the higher n-values do not improve any performance.

2.2 Candidate Signature Selection

This process is to measure the significance of each n-gram pattern, called a score, using statistical models and then sort all n-gram patterns as their scores. Finally, the process selects the ten highest-score patterns as "Candidate Signatures". Let Pattern P represent a set of m members of n-gram patterns p_1, p_2 ... p_m generated by training data. The score of each pattern p_i, $score(p_i)$, is to measure the significance of each pattern using a statistical model. Let S be the candidate signatures, selected from the k highest-score patterns. The Pattern P and the candidate signatures S are defined by

$$P = \{p_1,...,p_{m-1},p_m \mid score(p_{i-1}) \le score(p_i) \wedge i = 2,3,...,m\}. \quad (1)$$

$$S = \{\forall_j s_j \mid s_j = p_{m-k+j} \wedge 1 \le j \le k\}. \quad (2)$$

For the statistical scoring models, we propose nine models to evaluate the goodness of biological data. Nine proposed models are *Term Frequency* (*TF*), *Rocchio* (*TF-IDF*), *DIA association factor* (*Z*), *Cross Entropy* (*CE*), *Mutual Information* (*MI*), *Information Gain* (*IG*), *NGL coefficient* (*NGL*), *Chi-Square* (X^2) and *Weighted Odds Ratio* (*WOR*) [1,7]. We compare formulas of the models across four several criteria [6] that concentrate on pattern and class value as follows

Criteria I: *Common Patterns* computes pattern scores using the number of times pattern P occurs, $Freq(p)$, or the probability pattern P occurs, $P(p)$.

Criteria II: *Pattern Absence* computes pattern scores using the number of times or the probability that pattern P does not occur, $\overline{P}$.

Criteria III: *Class Value* computes pattern scores using the probability of the i^{th} class value, $P(C_i)$, or the conditional probability given that pattern P occurs, $P(C_i \mid p)$.

Criteria IV: *Target Class Value* is a measure which emphasizes on difference between the target or positive class value, pos, and the non-target or negative class value, neg.

The criteria comparison of nine models is illustrated in Table 1. Where $Freq(p)$ is the number of times pattern P occurs; $DF(p)$ is the number of sequences pattern P

occurs at least once; $|D|$ is the total number of sequences; $P(p)$ is the probability pattern P occurs; $\overline{P}$ means that pattern P does not occur; $P(C_i)$ is the probability of the i^{th} class value; $P(C_i \mid p)$ is the conditional probability of the i^{th} class value given that pattern P occurs; $P(p \mid C_i)$ is the conditional probability of pattern occurrence given the i^{th} class value; $P(pos \mid p)$ is the conditional probability of the 'positive' class value given that pattern P occurs; $P(neg \mid p)$ is the conditional probability of the 'negative' class value given that pattern P occurs; $P(p \mid pos)$ is the conditional probability of pattern occurrence given the positive class value; and $P(p \mid neg)$ is the conditional probability of pattern occurrence given the negative class value.

Table 1. Comparing nine statistical scoring models across four several criteria

Statistical Scoring Models	Common Patterns	Pattern Absence	Class Value	Target Class Value
$TF(p) = Freq(p)$.	Yes	-	-	-
$TF - IDF(p) = TF(p) \cdot \log(\frac{\|D\|}{DF(p)})$.	Yes	-	-	-
$Z(p) = P(C_i \mid p)$.	-	-	Yes	-
$CE(p) = P(p)\sum_i P(C_i \mid p)\log\frac{P(C_i \mid p)}{P(C_i)}$.	Yes	-	Yes	-
$MI(p) = \sum_i P(C_i)\log\frac{P(p \mid C_i)}{P(p)}$.	Yes	-	Yes	-
$IG(p) = P(p)\sum_i P(C_i \mid p)\log\frac{P(C_i \mid p)}{P(C_i)} + P(\overline{p})\sum_i P(C_i \mid \overline{p})\log\frac{P(C_i \mid \overline{p})}{P(C_i)}$.	Yes	Yes	Yes	-
$NGL(p) = \frac{\sqrt{\|D\|} \cdot [P(pos \mid p) \cdot P(neg \mid \overline{p}) - P(neg \mid p) \cdot P(pos \mid \overline{p})]}{\sqrt{P(p) \cdot P(\overline{p}) \cdot P(pos) \cdot P(neg)}}$.	Yes	Yes	Yes	Yes
$X^2(p) = \frac{\|D\| \cdot [P(pos \mid p) \cdot P(neg \mid \overline{p}) - P(neg \mid p) \cdot P(pos \mid \overline{p})]^2}{P(p) \cdot P(\overline{p}) \cdot P(pos) \cdot P(neg)}$.	Yes	Yes	Yes	Yes
$WOR(p) = P(p) \cdot \log\frac{P(p \mid pos) \cdot (1 - P(p, neg))}{(1 - P(p, pos)) \cdot P(p \mid neg)}$.	Yes	-	Yes	Yes

2.3 Identifier Construction

This process uses the candidate signatures to formulate a similarity scoring function. When there is a query sequence, the function is used to estimate the significance of candidate signatures of each class of DNA types (*SimScore*). If the *SimScore* of any class is maximal, the query is identified as a member of the class. On the other hand, if *SimScore* of every class is zero; the query is not assigned to any classes. Let X be a query sequence with cardinality m. Let $p_{x1}, p_{x2} \ldots p_{xd}$ be a set of d $(m\text{-}n\text{+}1)$ n-gram patterns generated by the query sequence. Let $s_{y1}, s_{y2} \ldots s_{ye}$ be a set of e candidate signatures in a class Y. Then, $sim(p_x, s_y)$ is a similarity score of a pattern p_x and a candidate signature s_y, where $sim(p_x, s_y)$ is 1, if p_x is similar to s_y, and $sim(p_x, s_y)$ is 0, if not similar. The similarity score of a class Y, $SimScore(X, Y)$, is a summation of similarity scores of every pattern p_x and every candidate signature s_y as follows

$$SimScore(X,Y) = \sum_{1 \le i \le d, 1 \le j \le e} sim(p_{x_i}, s_{y_j}). \tag{3}$$

3 Experiments

In our experiments, we compare the accuracies of identifiers by distinguishing the eight inner genes of Influenza in GenBank Database [2]. The accuracy is the ratio of the number of sequences correctly identified to the total number of test sequences. Each identifier is produced from the different sets of candidate signatures which each set is derived from one *n*-gram and one statistical model. Following the use of different criteria of models, we divide the experimental results into 3 ranges of *n*-grams, 1- to 6-grams, 6- to 11-grams and 11- to 24-grams, for more precise analysis.

The identifiers, constructed by 1- to 5-gram signatures in every scoring model, get poor results as shown in Fig. 2a; whereas the identifiers, constructed by 6- to 11-gram signatures in most models, achieve good results, except the models based on *Pattern Absence*, such as *IG*, *NGL* and X^2, as illustrated in Fig. 2b. In Fig. 2c, the use of signatures which are longer than 11-gram provides 10% less accuracy than the 6- to 11-gram, nevertheless the results are quite stable and good in several models, except the models based on *Target Class Value*, such as *NGL*, X^2 and *WOR*.

Following the results, the models based on frequencies of patterns solely, such as *TF* and *TF-IDF*, achieve good results; whereas the others depend on the length of *n*-gram signatures and the criteria used in the models. The models based on *Pattern Absence*, such as *IG*, *NGL* and X^2, between 6- to 11-grams produce poor results; whereas, longer than 11-gram, the model based on *Pattern Absence* but not based on *Target Class Value*, such as *IG*, gets good results. As discussed above, our discovery model selects the ten highest-score patterns to be signatures. The score comparison of each pattern, generated by the 6- to 11-grams, considers from frequencies of both presence and absence of pattern which may bring about unbalanced class and pattern score distribution. For longer than 11-gram, the score comparison mainly considers from frequencies of the pattern presence only, since the frequencies of pattern absence are very high in every pattern, because of the large number of possible patterns ($>4^{11}$ possible patterns) and the low probability of the same pattern occurrence. Hence, the performance of the models based on *Pattern Absence* in the long patterns is comparably equal with the models based on only *Common Patterns* owing to considering on the same only pattern presence.

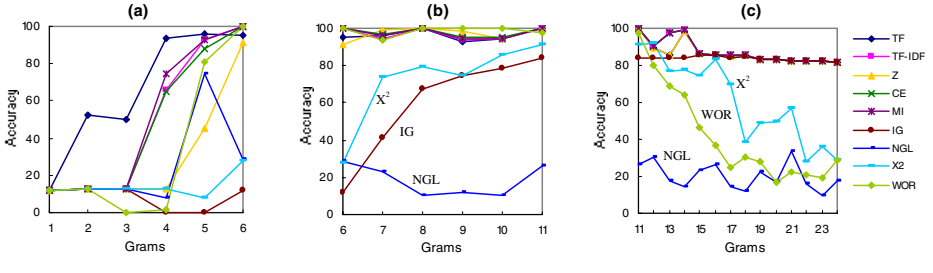

Fig. 2. Comparing the Accuracies of identifiers constructed using nine statistical scoring models and diverse *n*-grams: (a) 1- to 6-grams; (b) 6- to 11-grams; and (c) 11- to 24-grams

Next, we found that, longer than 11-gram, the accuracies of identifiers drop, when they use the models based on *Target Class Value*, such as *NGL*, X^2 and *WOR*, emphasizing on difference between the target and non-target classes. That the *Target Class Value* is not a good measure in the long *n*-gram patterns may be due to its unbalanced class and pattern score distribution. The use of long patterns generates the large number of possible patterns and the low probability of the same pattern occurrence; hence, the statistical analysis cannot handle with the too high pattern distribution which brings about the unbalanced score distribution.

However, from all results, several identifiers provide accuracy over 80% and up to 100% at 8- and 11-grams in several models. Hence, our proposed approach is highly possible to discover the actual "*DNA Signatures*" of the *Influenza virus*.

4 Conclusion

The signature discovery is an important task in the Computational Biology since the signature is more compact and more precise which helps reduce the high computation time. We propose a novel approach to discover "*Signatures*", which are sufficiently distinctive information for the sequence identification. We apply an *n*-gram method and statistical models in discovering the signatures. The signatures are derived from the best combination of the *n*-grams and the statistical models, which are evaluated by the identification system over the Influenza virus. The experimental results showed that the accuracies of several identifiers provide over 80% and maximal up to 100%, when the identifiers are constructed by the appropriate *n*-grams and statistical models. Hence, the candidate signatures used for the identifier construction are highly possible to be "Signatures" of the DNA sequences. Our approach succeeds in the signature discovery and also requires low computation time for the biological tasks.

References

1. Aalbersberg, I.: A document retrieval model based on term frequency ranks. Proc. of the 7th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1994) 163-172
2. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L.: GenBank. Nucleic Acids Research 28(1) (2000) 15-18
3. Brown, P.F., de Souza, P.V., Della Pietra, V.J., Mercer, R.L.: Class-based n-gram models of natural language. Computational Linguistics 18(4) (1992) 467-479
4. Chuzhanova, N.A., Jones, A.J., Margetts, S.: Feature selection for genetic sequence classification. Bioinformatics Journal 14(2) (1998) 139-143
5. Krauthammer, M., Rzhetsky, A., Morozov, P., Friedman, C.: Using BLAST for identifying gene and protein names in journal articles. Gene 259(1-2) (2000) 245-252
6. Mladenic, D., Grobelnik, M.: Feature selection for unbalanced class distribution and naïve bayes. Proc. of the 16th International Conference on Machine Learning (1999) 258-267
7. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1) (2002) 1-47
8. Wang, J.T.L., Rozen, S., Shapiro, B.A., Shasha, D., Wang, Z., Yin, M.: New techniques for DNA sequence classification. Journal of Computational Biology 6(2) (1999) 209-218
9. Xu, Y., Mural, R., Einstein, J., Shah, M., Uberbacher, E.: Grail: A multiagent neural network system for gene identification. Proc. of the IEEE 84(10) (1996) 1544-1552