The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009
Copyright © 2009 ORSC & APORC, pp. 229–236

Detection of Horizontal Gene Transfer in Bacterial Genomes*

Hui Ning[1] Bing Yang[1] Jun Cui[2] Ling Jing[1,†]

[1] College of Science, China Agricultural University, Beijing 100083, P.R. China.
[2] Xinjiang Radio & TV University, 830049, Urumuqi, P.R.China.

Abstract Most bacterial genes were acquired by horizontal gene transfer (HGT) from other prokaryotic organisms instead of being inherited by continuous vertical descent from an ancient ancestor. HGT is generally believed to be a major factor in microbiology evolution, allowing rapid diversification and adaptation. In this paper, we artificially simulate HGT by inserting phage genes into bacterial genomes, and then try to identify horizontally transferred genes from normal genes. Based on nucleotide composition analyses, Nonparallel Plane Proximal Classifier (NPPC) is applied to predict HGT for the first time. We also present a novel method, namely VQ-NPPC, which combines Vector Quantization (VQ) and NPPC. The results confirm that NPPC and VQ-NPPC show better performances than previous proposed methods i.e. C-Support Vector Classification(C-SVC) and One-class SVM for the detection of HGT.

Keywords Horizontal gene transfer; SVM; Vector Quantization; NPPC; bacterial genome

1 Introduction

Horizontal gene transfer (HGT), also called lateral gene transfer (LGT), is defined as movement of genetic material between different species, or across broad taxonomic categories. Although most thinking in genetics has focused on the more prevalent vertical transfer, there is a recent awareness that horizontal gene transfer is a significant phenomenon.

Recently, with a large amount of biological data have been generated, several approaches have been developed to identify HGT. The two most popular strategies for detecting HGT are phylogenetic methods and compositional methods. Both strategies have been successfully used to detect HGT events that have later been well-supported by many independent lines of evidence [1].

Traditionally, Phylogenetic methods typically compare the evolutionary history of each gene to the best estimate of the evolutionary history of the genome. They are particularly effective in identifying transfers and this has already been demonstrated by many researchers [2, 3].

* Supported by the Key Project of National Natural Science Foundation of China (No. 10631070).

† Corresponding author. E-mail address: jingling_student@163.com.

On the other hand, compositional methods hold that genes which appear atypical in their current genomic context are likely to have been introduced from a foreign source. Many of the previously published compositional methods were based on gene content, such as G+C content, nucleotide, oligonucleotide, amino acid usage or codon adaptation index, these features have been proposed as 'signatures' that would be characteristic for a genome, any gene deviating from the 'signature' can be marked as a horizontal gene transfer candidate. For an extensive discussion of various aspects of genome 'signature', the reader can refer to Aristotelis T. and Isidore R. [4], which provides an excellent summary to the topic.

Based on the previous studies, the HGT prediction can be formalized as a classification problem, both standard support vector machine (SVM) and one-class SVM have shown improvement performance for this issue [5]. However, although the phenomenon of HGT is common in microorganisms, we know that for a specific bacterial genome, the number of transferred genes is relatively less. When constructing a classifier for HGT prediction, extremely imbalanced training datasets which contain transferred and original genes dramatically decrease the performance of the SVM and One-class SVM. In this paper, nonparallel plane proximal classifier (NPPC) is firstly used for the detection of HGT, and then a new technique, namely VQ-NPPC which combines Vector Quantization (VQ) and NPPC is proposed for solving the unbalance problem in HGT prediction, which is also based on the compositional features of gene sequences. Experimental results on three bacterial genomes demonstrate the proposed method outperforms the others in recent research problem.

2 Data sources and processing

2.1 Manual simulation of HGT

For the problem of detecting horizontally transferred genes, however, the available information is only about some concrete examples especially in prokaryote, and it is generally unclear which genes are horizontally transferred in a whole genome of interested species. Consequently, in recent studies, artificial simulated gene transfer experiments are designed to the detection of HGT [4, 5, 6].

In our research, we simulated the HGT between bacterial and phage genomes which indeed happens in real nature [6]. For example, Lysogenic phages are able to integrate into the host genome and become part of the genetic material which make up of the host bacterium. Furthermore, transduction is regarded as one of three main mechanisms of gene transfer in prokaryotes, which is a bacteriophage-facilitated transfer of genetic material from one bacterial host to another [7]. Therefore, we used three complete bacterial genomes including Escherichia coli str. K-12, Bacillus cereus E33L and Borrelia burgdorferi B31 as host genomes, meanwhile we took the overall 1574 gene sequences of 27 phages as donor pool, and randomly sub-selected an appropriate fraction of these genes that were then incorporated into the bacterial host genome [6]. The intention of our experiment was to recover as many as possible of the inserted donor genes. All of these sequence data were downloaded from NCBI/GenBank.

2.2 Genome compositional features

Nucleotide composition is one of the specific properties of organism genomes, so we extract frequency of codon usage (FCU) of bacterial and phage genomes as the genome 'signature'， FCU is a simple measure of genomic nucleotide composition which can be calculated from the following equation:

$$FCU_i = \frac{obs_i}{total} \tag{1}$$

where， obs_i is the number of codon $\boldsymbol{i}$ appearing in the whole gene ($i = 1, \cdots, 61$), $total$ represents the total number of the codon in the whole gene. In other words, in terms of every given gene sample， we calculate the FCU of the all codons except for three stop codons, subsequently, we get a 61 dimensional vector as the compositional feature vector for each gene.

3 Methods

In this section, SVM and nonparallel plane proximal classifier (NPPC) are described and the main differences between them are shown briefly, In addition, our proposed algorithm (VQ-NPPC) is also explained.

3.1 C-Support Vector Classification(C-SVC)

The essential idea of C-SVC is to search a linear separating hyperplane which maximizes the distance between two classes of data to create a classifier [8]. For a binary classification problem, the training data set consists of l samples $x_i \in \Re^n$, $i = 1, \cdots, l$ and with corresponding class values $y_i \in \{1, -1\}$, $i = 1, \cdots, l$. The standard formulation is a linear softmargin algorithm which is to solve the following optimization problem:

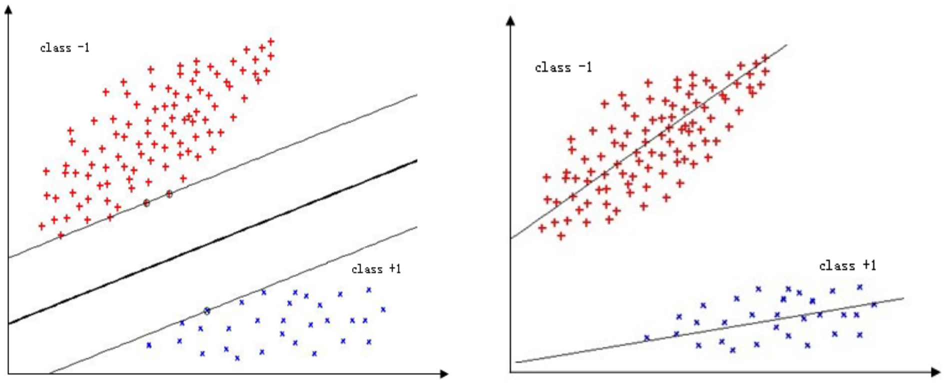
$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ s.t. \quad & y_i\left(\left(w \cdot x_i\right) + b\right) \ge 1 - \xi_i \, , i = 1, \cdots, l \\ & \xi_i \ge 0, i = 1, \cdots, l \end{aligned} \tag{2}$$

where the predefined parameter C is a trade-off between training accuracy and generalization, ξ_i is the slack variable, $w \in \Re^n$ is a weight vector, while b is a bias which moves the hyperplane parallel to itself. The decision function is presented as:

$$f(x) = \operatorname{sgn}\left(\left(w \cdot x\right) + b\right) \tag{3}$$

The solution of this optimization problem is given by solving the corresponding dual problem with introduced lagrange multiplier α .Generally, the solution α of the dual problem is sparse, then the corresponding decision hyperplane depends only on few "support vectors" (Fig.1a).

3.2 Nonparallel plane proximal classifier（NPPC）



a. Geometric interpretation of C-SVC b. Geometric interpretation of NPPC

Figure 1: For a binary classification problem, (a): the principle of C-SVC, the decision plane (thick line) is closely related to a few support vectors (circled points) and it classifies points by assigning them to one of two disjoint half-plane. (b): NPPC，It can find two nonparallel planes that are pushed apart as far as possible, and points are classified depending on which of the two planes they lies closest to.

Instead of finding a single hyperplane in C-SVC, Nonparallel Plane Proximal Classifier (NPPC) [10] which is an improvement formulation of Twin Support Vector Machine (TWSVM) [9] finds two nonparallel hyperplanes such that each plane is clustered around one particular class data (Fig.1b). The formulation of NPPC for binary data classification is based on two identical mean square error (MSE) optimization problems which lead to solving two small systems of linear equations in input space. Thus it eliminates the need of any specialized software for solving the quadratic programming problems [10].

$$\text{(NPPC1)} \quad \begin{array}{ll} \min\limits_{(w_1,b_1,\xi_2)} & \frac{1}{2}\left\|Aw_1+e_1b_1\right\|^2 + C_1 {e_2}^T \xi_2 + \frac{C_2}{2}\xi_2^T\xi_2 \\ s.t. & -\left(Bw_1+e_2b_1\right)+\xi_2 = e_2 \end{array} \quad (4)$$

$$\text{(NPPC2)} \quad \begin{array}{ll} \min\limits_{(w_2,b_2,\xi_1)} & \frac{1}{2}\left\|Bw_2+e_2b_2\right\|^2 + C_3 {e_1}^T \xi_1 + \frac{C_4}{2}\xi_1^T\xi_1 \\ s.t. & \left(Aw_2+e_1b_2\right)+\xi_1 = e_1 \end{array} \quad (5)$$

where matrix $A \in \Re^{m_1 \times n}$ represent the data points of class +1 and matrix $B \in \Re^{m_2 \times n}$ represent the data points of class -1 and they contain m_1 and m_2 training patterns respectively in n dimensional space and $m_1 + m_2 = l$, $w_1, w_2 \in \Re^n$ are weight vectors and $b_1, b_2 \in \Re$ are bias terms of respective planes. $C_1, C_2, C_3, C_4 > 0$ are regularization parameters, $e_1 \in \Re^{m_1}$ and $e_2 \in \Re^{m_2}$ are vectors of ones, $\xi_1 \in \Re^{m_1}$ and $\xi_2 \in \Re^{m_2}$ are error variable vectors due to classes +1 and -1 data, respectively.

Then two non-parallel hyperplanes $w_1^T x + b_1 = 0$ and $w_2^T x + b_2 = 0$ can be obtained from the solution of NPPC1 and NPPC2. A new data sample $x \in \Re^n$ is assigned to class +1or -1 depending on which of the two hyperplanes lies closest to the point in terms of perpendicular distance. Finally, the decision function can be written as:

$$Class\ k = \underset{k=1,2}{Min} \left| w_k^T x + b_k \right| \tag{6}$$

In this paper, the training data sets are large and imbalanced for detection of HGT. In artificial simulated experiments, for example, the genome of Bacillus cereus E33L contains 103 positive data samples and 5134 negative data samples in which there are noises. In other words, the training dataset is highly imbalanced and there are some positive data in negative class and we don't know which one is false, also we know the false negative ones may distribute closed to true positive data samples. Accordingly, the decision plane of C-SVC (Fig.1a) learned by this training dataset would be not effective for new test data samples so its generalization capability would decrease. In order to reduce the influence of false negative samples, we chose linear NPPC algorithm to predict the horizontally transferred genes for 3 bacterial genomes and the results showed that NPPC algorithm was more suitable for this study.

3.3 VQ-NPPC

VQ-NPPC combines Vector Quantization (VQ) and nonparallel plane proximal classifier (NPPC) in order to rebalance datasets without significant information loss.

Vector quantization, also called "block quantization" or "pattern matching quantization", is often used in lossy data compression. Given a training dateset consisting of ***L*** vectors: $T = \{x_1, x_2, \cdots, x_L\}$, the vector quantizer maps ***T*** into a finite set of vectors $C = \{c_1, c_2, \cdots, c_M\}$, each vector $c_i \ (i = 1, \cdots, M)$ is called a codevector, and the set of all the codevectors is called a codebook [11]. Then the whole region is partitioned by the codevectors into a set of sub-regions, so-called "Voronoi Region": $V = \{V_1, V_2, \cdots, V_M\}$, and it is defined by:

$$V_i = \left\{ x \in T : \left\| x - c_i \right\| \le \left\| x - c_j \right\|, \forall j \ne i \right\}, i = 1, \cdots, M \tag{7}$$

$$\bigcup_{i=1}^{M} V_i = T, \quad \bigcap_{i=1}^{M} V_i = \phi \tag{8}$$

Vectors within a region $\boldsymbol{V_i}$ are represented by their codevector $\Phi(x_l) = c_m$ if $x_l \in V_m$, to find ***C*** and ***V*** is to minimize the average distortion which can be given by:

$$D = \frac{1}{Mn} \sum_{n=1}^{L} \left\| x_l - \Phi(x_l) \right\|^2 \tag{9}$$

where n is the dimension of training vectors [12].

Our proposed approach VQ-NPPC can be implemented by two stages: At first, we select representative negative training samples from the majority class using VQ, in this case, the imbalance ratio must be under consideration. Then combine these selected negative data and original positive data to compose a new training dataset. Secondly, NPPC is applied on this new obtained dataset. Our experimental results show that VQ-NPPC outperforms the previous methods for detection of HGT, while the training time is significantly decreased. The concrete procedure is shown in Table1.

Table1: the procedure of VQ-NPPC

```
    Input:
        Positive training data P (minority group)
        Negative training data N (majority group)
        Set parameter sets {CodebookSize, $C_1, C_2, C_3, C_4$ }
        /* CodebookSize is the number of codevectors in VQ, $C_1, C_2, C_3, C_4$ are
  regularization parameters of NPPC. */
  Stage 1: Compress the majority group in order to select a new negative set
         $N_1 \leftarrow VQ\left(N, CodebookSize\right)$;
         New training dataset $T \leftarrow Combine\left(N_1, P\right)$.
    Stage 2: Classification based on the new training dataset
         $Hyperplane\ 1 \leftarrow NPPC\ 1\left(New\ training\ dataset, C_1, C_2\right)$;
         $Hyperplane\ 2 \leftarrow NPPC\ 2\left(New\ training\ dataset, C_3, C_4\right)$.
```

4 Experiments and Results

For each of the 3 host bacterial genome in turn, we carried out 100 experiments of simulated transfers from a gene pool composed of 1574 gene sequences of 27 phages [4, 5, 6]. In each experiment, the number of added genes was chosen to be a fixed percentage of the number of genes in the host genome, as to the genome of Escherichia coli str. K-12 and Bacillus cereus E33L, the transfer percentages were both 2%, but we chose 8% for the genome of Borrelia burgdorferi B31 because its genome size was smaller. The transferred genes were selected at random from the donor pool, the objective is to recover as many of the artificially transferred genes as possible, without using any a priori knowledge about the host genome or the donor genes.

For each training procedure, discrimination functions whose coefficients could be calculated were used to discriminate the test data consisting of only phage genes. Consequently, we also used ‘hit ratio (HT)’ [6] to denote the proportion between the number of artificially inserted genes and the number of genes recovered by a particular procedure. HT can be calculated by

$$HT = \frac{1}{100}\sum_{i=1}^{100}\frac{PT_i}{NT_i} \tag{10}$$

where NT_i and PT_i represent the number of genes and the number of predicted horizontally transferred genes in test dataset respectively. So HT was used to evaluate the reasonability of a method for the detection of HGT.

In this study, NPPC algorithm and VQ-NPPC were used to detect the horizontally transferred genes, furthermore, we compared our methods against other methods including C-SVC and OC-SVM which were used by Aristotelis T., Isidore R. in 2005 [4, 5] and Jiansheng Wu et al. in 2007 [6]. In experiments we used NPPC and VQ-NPPC implemented by writing procedure in Matlab. Table 2 shows the comparison of the HT by our two methods and previous proposed methods on each bacterial genome. It can be seen from the results that our proposed methods outperform all the other methods in this case.

Table 2: HT of NPPC, VQ-NPPC and previous methods

species	Tsirigos' method	Wu's method	C-SVC	OC-SVM	NPPC	VQ-NPPC
Escherichia coli str. K-12	0.375	0.539	0.493	0.346	0.765	**0.879**
Bacillus cereus E33L	0.541	0.647	0.571	0.566	0.836	**0.949**
Borrelia burgdorferi B31	0.758	0.945	0.767	0.846	**0.962**	**0.962**

The results of first four columns in Table 2 are obtained from [6], the fifth column is the result of NPPC and the last one is the result of VQ-NPPC. Obviously, HT of NPPC is much higher than previous proposed methods, and the performance of VQ-NPPC is better than NPPC for this issue. But the result of NPPC and VQ-NPPC are not different for the third genome, the reason is that CodebookSize was set as the size of class -1 in the first stage of VQ-NPPC, in other words, VQ was not used on the third genome in our experiments because of its genome size was smaller, so it was unnecessary to compress the negative dataset by VQ before training.

It also can be observed that proposed methods get higher relative improvements on Escherichia coli str. K-12 and Bacillus cereus E33L genomes than on Borrelia burgdorferi B31 genome, the reason is that the codon usage in Borrelia burgdorferi B31 genome is strongly biased which is related to the translation and codon choice of a gene [13]. It proves that NPPC algorithm and VQ-NPPC procedure are both suitable for the detection of HGT.

5 Conclusion

In this paper, we proposed a composition-based framework for the detection of HGT. NPPC and VQ-NPPC were used to detect horizontally transferred genes for the first time, both of which reached higher accuracy than previously proposed schemes.

The main difference between our two methods and the previously reported methods is that NPPC algorithm can overcome the imbalanced problem due to its two nonparallel hyperplanes which mine the data information of classes +1 and -1respectively. Vector Quantization in VQ-NPPC is used to compress negative datasets at first, so VQ-NPPC has shown better performance than NPPC on the large imbalanced datasets, and for the detection of HGT, we can choose one of the two methods according to the size of a host genome.

However it requires further investigations about how to obtain more compositional features by other approaches and how to use comprehensive information of a genome for the detection of HGT. Except for three bacterial genome sequences in this study, more prokaryotic genome should be taken into account in our next research.

References

[1] Koonin EV, Horizontal gene transfer: the path to maturity, Mol Microbiol, 2003, 50(3): 725-727.

[2] Syvanen,M., Horizontal gene transfer: evidence and possible consequences, Annu. Rev. Genet. 1994, 28,237-261.

[3] In-Geol Choi, Sung-Hou Kim, Global extent of horizontal gene transfer, PNAS, 2007, 104(11): 4489-4494.

[4] Aristotelis T. and Isidore R., A new computational method for the detection of horizontal gene transfer events, Nucleic Acids Research, 2005, 33(3): 922-933.

[5] Aristotelis T. and Isidore R., A sensitive, support-vector-machine method forthe detection of horizontal gene transfers in viral, archaeal and bacterial genomes, Nucleic Acids Research, 2005, 33(12): 3699–3707.

[6] Jiansheng Wu, Jianming Xie, Tong Zhou, Jianhong Weng, Xiao Sun, Support Vector Machine for Prediction of Horizontal Gene Transfers in Bacteria Genomes, Progress in Biochemistry and Biophysics, 2007, 34(7):724–731.

[7] B.G. Kelly , A. Vespermann, D.J. Bolton, The role of horizontal gene transfer in the evolution of selected foodborne bacterial pathogens, Food and Chemical Toxicology, 2008, IN PRESS.

[8] V. Vapnik and C. Cortes,. Support vector networks, Machine Learning, 1995 20(3): 273-297.

[9] Jayadeva, Khemchandani, R., Chandra, S., Twin support vector machines for pattern classification, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(5), 905－910.

[10] Santanu Ghorai, Anirban Mukherjee, Pranab K. Dutta, Nonparallel plane proximal classifier, Signal Processing, 2009, 89,510–512.

[11] Gersho, A. and R.M. Gray, Vector Quantization And Signal Compression, Kluwer Academic Publishers,1992.

[12] Ting Yu, John Debenham, Tony Jan and Simeon Simoff, Combine Vector Quantization and Support Vector Machine for Imbalanced Datasets, TFTP International Federation for Information Processing, 2006, 217.

[13] Ikemura T, Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E.coli translational system, J Mol Biol, 1998, 151(3):389～409.