Available online at www.sciencedirect.com
ScienceDirect
Progress in Natural Science 19 (2009) 511–516
Progress in Natural Science
www.elsevier.com/locate/pnsc

# DNA splice site sequences clustering method for conservativeness analysis

Quanwei Zhang, Qinke Peng *, Tao Xu

*State Key Laboratory for Manufacturing Systems Engineering, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China*

Received 20 April 2008; received in revised form 28 May 2008; accepted 5 June 2008

**Abstract**

DNA sequences that are near to splice sites have remarkable conservativeness, and many researchers have contributed to the prediction of splice site. In order to mine the underlying biological knowledge, we analyze the conservativeness of DNA splice site adjacent sequences by clustering. Firstly, we propose a kind of DNA splice site sequences clustering method which is based on DBSCAN, and use four kinds of dissimilarity calculating methods. Then, we analyze the conservative feature of the clustering results and the experimental data set.


*Keywords:* Bioinformatics processing; Clustering; DBSCAN; Splice site

## 1. Introduction

Human Genome Project (HGP) aimed at elucidating the sequence of human being's three billion base pairs, discovering all the genes and their positions on chromosomes, and decrypting all the human being's genetic information. Along with the successful completion of HGP, bioinformatics comes into the post-genome era. Now, mining knowledge from the mass data has become much more important than obtaining biological data. Many studies are concentrated on the classification of sequences, discriminating coding section from non-coding section, structure prediction, function prediction, and so on. The main methods that researchers often use include neural network method [1], Bayes algorithm [2], HMM [3] (hidden Markov models), SVM [4] (support vector machine), and clustering methods [5].

The motive for clustering analysis is to divide the objects into several clusters; a cluster is a collection of objects that are similar to each other, and objects in different clusters are dissimilar from each other. Clustering is used in many applications such as text mining [6], web data analysis [7], and image processing [8]. Because similar DNA sequences usually have similar functions, researchers often predict function by sequence alignment. Therefore, in this study, we present a DBSCAN-based splice site sequences clustering method to find the feature of splice site sequences, which will be helpful to predict their functions. DBSCAN is chosen as the clustering method because it is better in the following aspects than K-means, which is one popular method in the field of bioinformation processing: (i) K-means must determine a cluster number before clustering, while DBSCAN does not need it, DBSCAN finds dense clusters automatically for a given density threshold; (ii) K-means is sensitive with noises, but DBSCAN handles noises well; (iii) K-means is only applicable to find circular or spherical clusters, but DBSCAN can find clusters with arbitrary shapes; (iv)

* Corresponding author. Tel.: +86 29 82663489; fax: +86 29 82667964.
*E-mail addresses:* zhang_quan_wei@126.com (Q. Zhang), qkpeng@xjtu.edu.cn (Q. Peng).

1002-0071/$ - see front matter © 2008 National Natural Science Foundation of China and Chinese Academy of Sciences. Published by Elsevier Limited and Science in China Press. All rights reserved.
doi:10.1016/j.pnsc.2008.06.021

K-means has the numeric-only limitation, but DBSCAN can work well with categorical data.

In this paper, we will exhibit some dissimilarity calculating methods between DNA sequences, and propose one new kind of dissimilarity calculating method. Then, after clustering, we analyze the conservative characterization of the clustering result and the experimental data set.

## 2. Cluster-based bioinformation processing

Clustering plays an important role in the field of bioinformation processing, and it is widely used in gene database knowledge discovery. It is used mainly in the following fields:

(1) Clustering can effectively partition different DNA sequence sets into different clusters by sequence alignment. Based on the clusters, further research can be developed. For example, Michael Eisen developed one microarray data analysis software called CLUSTER [9]. CLUSTER finds the most relevant gene pair by comparing genes with each other. Then, the average of the gene pair substitutes the gene pair, and the relevant matrix is calculated again. Repeat the process until the result is satisfactory.
(2) Clustering is applied in the field of function prediction. Several clustering methods are proposed to predict the secondary structures of RNA and protein [10]. Because the functions and the structures are closely related, these clustering methods are helpful to predict their functions. Clustering also does well in the identification and prediction of exons, introns and splice sites [11].
(3) Clustering also plays an important role in the research of molecule evolution theory [12]. Because evolution tree is the main method for analyzing molecule evolution, it provides an ideal opportunity for hierarchical clustering methods. Evolution trees are constructed by the dissimilarity between DNA sequences or protein structures, so it is obvious that clustering is one useful tool to construct evolution trees.

In brief, comparison of dissimilarity between objects is the main method for mining knowledge in gene database, and so clustering is an indispensable method in the field of biological knowledge discovery.

## 3. Splice site adjacent sequences clustering

### 3.1. DBSCAN method description

DBSCAN (density-based spatial clustering of applications with noise) is a density-based algorithm. DBSCAN is based on the fact that clusters are of higher density than its surroundings. The key idea of DBSCAN is that for each object of a cluster, the neighborhood with a given radius $\varepsilon$ must contain at least a minimum number of *MinPts* objects, i.e. the cardinality of the neighborhood has to exceed a given threshold. In what follows, we will present the basic definitions of DBSCAN [13].

**Definition 1** (*neighborhood*)**.** The neighborhood of object $p$ is the set of objects in the circle, whose center is $p$ and radius is $\varepsilon$, i.e. $N_\varepsilon(p) = \{q \in D | dist(p,q) \leqslant \varepsilon\}$, where $D$ is the database of objects.

**Definition 2** (*directly density reachable*)**.** An object $p$ is directly density reachable from an object $q$ wrt $\varepsilon$ and *MinPts* if $p$ is within the neighborhood of $q$, i.e. $p \in N_\varepsilon(q)$; and $q$ is core object, i.e. $|N_\varepsilon(q)| \geqslant MinPts$ (core object condition).

**Definition 3** (*density reachable*)**.** An object $p$ is density reachable from an object $q$ wrt $\varepsilon$ and *MinPts* if there is a chain of objects $p_i$ $(i = 1, \ldots, n)$, and $p_i$ is directly density reachable from $p_{i+1}$; $p$ is $p_1$, and $q$ is $p_n$.

**Definition 4** (*density connected*)**.** An object $p$ is density connected to another object $q$ wrt $\varepsilon$ and *MinPts* if there is another object $p_t$ such that both $p$ and $q$ are density reachable from $p_t$ wrt $\varepsilon$ and *MinPts*.

**Definition 5** (*noise*)**.** Let $C_1, \ldots, C_k$ be the clusters wrt $\varepsilon$ and *MinPts*. If $p$ is an object of $D$, and it does not belong to any cluster $C_i$ $(i = 1, \ldots, k)$, then $p$ is a noise, i.e. noise set $= \{p \in D | \forall i : p \notin C_i\}, (i = 1, \ldots, k)$.

A cluster is one set of objects, which are density connected with each other. The algorithm is given as follows.

Algorithm **DBSCAN**

Input: objects $D=\{p_i\}$ $(i = 1, \ldots, N)$, $\varepsilon$, *MinPts*.

Output: clusters and noise set.

① Check the object $p$ that has not yet been processed (clustered or marked as noise). If $p$ is a core object, a new cluster $C$ is created, the neighbors of $p$ which are not yet contained in any cluster are added into cluster $C$.

② Check the object $q$ in $C$ which has not been checked, if $q$ is a core object, the neighbors of $q$ which are not yet contained in any cluster are added into cluster $C$.

③ Repeat ② until all the objects in $C$ are checked.

④ Repeat ①, ② and ③ until all the objects are classified as some cluster or noise.

### 3.2. Dissimilarity definitions

The definition of dissimilarity between objects is crucial for clustering, and now we will discuss some common dissimilarity definitions between DNA sequences and then introduce one dissimilarity definition that is proposed in this study.

#### 3.2.1. Direct alignment

Given two DNA sequences $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)$, at first, we compare each base pair at the corresponding position directly. If the base pair is the