BioSystems 105 (2011) 73–82

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

A CLIQUE algorithm using DNA computing techniques based on closed-circle DNA sequences☆

Hongyan Zhang, Xiyu Liu*

School of Management and Economics, Shandong Normal University, Jinan 250014, China

ARTICLE INFO

Article history:
Received 20 August 2009
Received in revised form 11 March 2011
Accepted 14 March 2011

Keywords:
DNA computing
Data mining
Grid-based clustering
CLIQUE algorithm
Simulation

ABSTRACT

DNA computing has been applied in broad fields such as graph theory, finite state problems, and combinatorial problem. DNA computing approaches are more suitable used to solve many combinatorial problems because of the vast parallelism and high-density storage. The CLIQUE algorithm is one of the gird-based clustering techniques for spatial data. It is the combinatorial problem of the density cells. Therefore we utilize DNA computing using the closed-circle DNA sequences to execute the CLIQUE algorithm for the two-dimensional data. In our study, the process of clustering becomes a parallel bio-chemical reaction and the DNA sequences representing the marked cells can be combined to form a closed-circle DNA sequences. This strategy is a new application of DNA computing. Although the strategy is only for the two-dimensional data, it provides a new idea to consider the grids to be vertexes in a graph and transform the search problem into a combinatorial problem.

1. Introduction

In 1994, Adleman (1994) solved a 7-vertex Hamilton path problem (HPP) and it was a breakthrough in DNA computing. DNA computing shows a great potential to solve combinatorial problems in various areas of applications because of its great storage ability and parallel reactions.

Compared with silicon computers, DNA computing methods were more suitable to be used in complex computational problems (Lipton, 1995) such as the Hamilton path problem, maximal clique problem (Ouyang et al., 1997), satisfiability problem (Liu et al., 2000), and chess problems (Faulhammer et al., 2000). These biological techniques are also used to solve some real problems (Barreto et al., 2006; Yamamoto et al., 2000; Zhou et al., 2007, 2008). DNA computing makes use of DNA sequences generated on certain rules to combine with each other in some biological reactions such as hybridization and ligation in the test tube. The solution will be generated in the test tube. The advantage of these approaches is the huge inherent parallelism, which has the potential to yield vast speedups over conventional silicon computers for such search problems.

In this paper we present another research on clustering based on the idea of CLIQUE (Clustering in QUEst (Agrawal et al., 1998)) using DNA computing. The parallel ability and potential of solving combinatorial problem of DNA computing are employed in this study. We propose the basic idea of using DNA computing techniques to realize the CLIQUE algorithm based on the closed-circle DNA sequences and meanwhile provide the coding methods as well as bio-chemical operations design. We provide a new algorithm to simulate our idea and discuss the time complexities between the general CLIQUE algorithm and the new algorithm, by using the parallel strategy. In the experiments, we give two experiments to prove the feasibility of the idea in simple graph and complex graphs.

2. Motivation

Most clustering algorithms exhibit polynomial or exponential complexity. The problem becomes even far more challenging when the number of clusters is unknown and the data set become huge (Jain and Law, 2005). The appearance of DNA computing provides an interesting and viable alternative.

During clustering, we need to calculate and process all combinations of data points which contain the right clustering solution. Thus the clustering is the combinational problem of the patterns. While the database size increases, the number of patterns also increase. The silicon computer only uses some algorithms to solve the problem and the solution may be the optimal of the part. The sil-

☆ The research is supported by the Natural Science Foundation of China (No. 60743010) and the Science Research Innovation Foundation for Ph.D. Student (No. BCX1005).

* Corresponding author.
E-mail addresses: zswanz@yahoo.com.cn (H. Zhang), xyliu@sdnu.edu.cn (X. Liu).

0303-2647/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2011.03.004

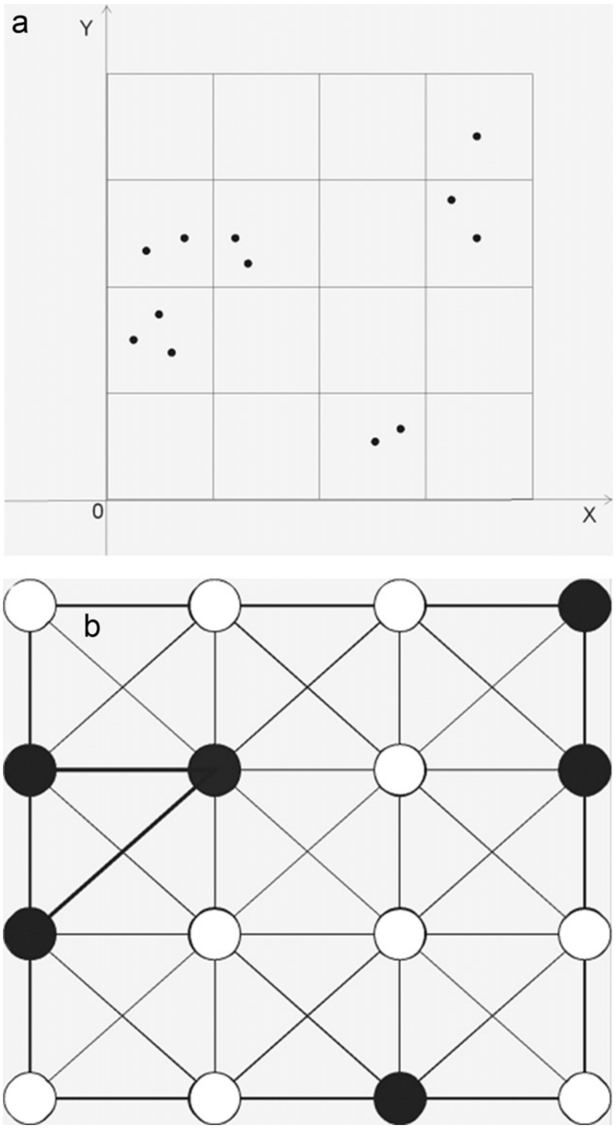

Fig. 1. The example graph of the 12 patterns.

icon computer might not be sufficient to face the enormous amount of calculations.

There are two advantages of DNA computing. One is the parallelism to process the calculations and another is the storage potentiality. An operation can take 10^{-24} s in a single test tube. Meanwhile, 1 bit information can be stored in 1 nm^3 in biological computing compared to a conventional storage media (Ezziane, 2005).

DNA computing is suitable to solve the combinatorial problems because it can obtain all the possible solutions. The first application of DNA computing is to solve the combinatorial problem (HPP). The researchers obtained many possible combinatorial solutions of the seven vertexes though some biological operations. The right solution was also in the test tube.

Considering the above analysis, DNA computing is also suitable to solve the clustering problem with its capability of massively parallel processing and greater storage. This is the new idea to use biological approach for the clustering problem. Once the biological techniques are mature, the advantages of DNA computing will appear.

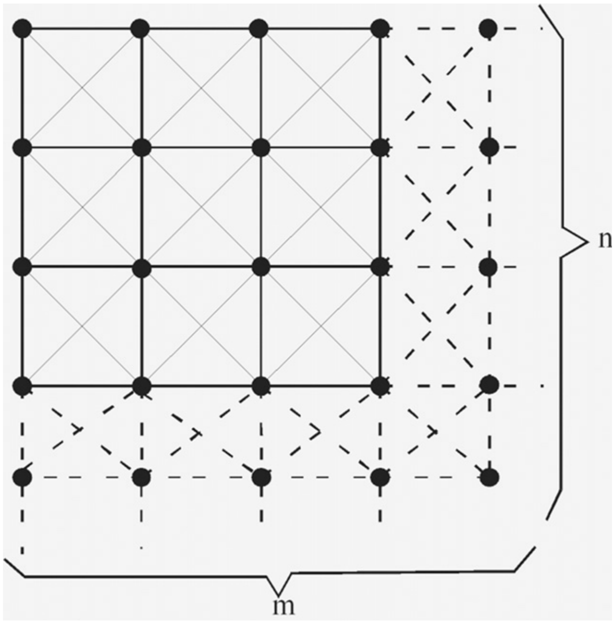

Fig. 2. The graph is the regular structure of the $n \times m$ vertexes forms a rectangle or square.

3. Background

3.1. CLIQUE algorithm

Grid-based clustering techniques are usually used for the more complex and high-dimension data. The main application is spatial data such as the geometric structure of objects in space, their relationships, properties and operations (Andritsos, 2002). The basic idea is to quantize the data set into a number of grids and then deal with objects belonging to these grids. This algorithm does not pay attention to the points but rather builds several hierarchical levels of groups of objects.

The CLIQUE algorithm is a grid-based method, which is a subspace clustering algorithm for high-dimension data. The CLIQUE algorithm first finds one-dimensional dense grids, then two-dimensional dense rectangles and so on, until all dense hyper rectangles of dimensionality *k* are found. Actually, the CLIQUE algorithm is an algorithm combined by density-based method and grid-based method (Han, 2000). The advantages of the CLIQUE algorithm are that it is unnecessary to require the order of the input data and it is very effective for the clustering of the spatial data in the high-dimension data. However, this algorithm has some disadvantages: (a) the grids can destroy the boundary of the group; (b) the threshold must be given; (c) the isolated point cannot be deleted automatically. Many researchers improved the CLIQUE algorithm in order to make it the best for clustering spatial data. Goil et al. (1999) presented the MAFIA algorithm, which improved the boundary shape and decreased the total run time. The ENCLUS (Entropy-based clustering) algorithm used entropy to compare each subspace (Cheng et al., 1999). Nagesh et al. (2000) proposed a scalable parallel subspace clustering algorithm for massive data sets, which was useful to cluster the pocket data. These improved CLIQUE algorithms have their characters whereas there is no CLIQUE algorithm which can be suitable for all spatial data. However, the combinatorial problems of the grids must be realized in any CLIQUE algorithm, which is the important step in the CLIQUE algorithm.