APPLICATION NOTE

Genomics & Informatics
Vol. 9(2) 89-91, June 2011

DOI: 10.5808/GI.2011.9.2.89
pISSN 1598-866X eISSN 2234-0742

# A Clustering Tool Using Particle Swarm Optimization for DNA Chip Data

**Xiaoyue Han and Minsoo Lee***

Department of Computer Science and Engineering, Ewha Womans University, Seoul 120-750, Korea

## Abstract

DNA chips are becoming increasingly popular as a convenient way to perform vast amounts of experiments related to genes on a single chip. And the importance of analyzing the data that is provided by such DNA chips is becoming significant. A very important analysis on DNA chip data would be clustering genes to identify gene groups which have similar properties such as cancer. Clustering data for DNA chips usually deal with a large search space and has a very fuzzy characteristic. The Particle Swarm Optimization algorithm which was recently proposed is a very good candidate to solve such problems. In this paper, we propose a clustering mechanism that is based on the Particle Swarm Optimization algorithm. Our experiments show that the PSO-based clustering algorithm developed is efficient in terms of execution time for clustering DNA chip data, and thus be used to extract valuable information such as cancer related genes from DNA chip data with high cluster accuracy and in a timely manner.

***Availability:*** The codes for the developed algorithm may be obtained from the author under consent to intellectual property agreements. The code is a demo version and thus some configuration methods are done through editing configuration files. The results are shown in a text window which will be improved in the future. Contact mlee@ewha.ac.kr if you are interested in the demo version of the codes for possible collaboration.

***Keywords:*** clustering, DNA chip data

## Introduction

Recent advances in biological science are creating a revolutionary opportunity for understanding genome sequences for many different organisms. Such organisms are complex and genomes can be immense,and thus new and powerful technologies are being developed to analyze large numbers of genes and proteins as a complement to traditional methodologies that study a small number at a time. The recently developed DNA chips, in other words DNA microarrays, have emerged as a prime candidate for such high performance analysis methods (Debouck *et al.*, 1999; DeRisi *et al.*, 1997).

The clustering algorithms for DNA chip data could benefit from using algorithms that imitate the ecosystem. Algorithms that imitate the ecosystem have the following benefits. First, they provide a solution based on a solid statistical model. In other words, they do not rely on a single solution but have more flexibility due to the fact that they are based on a statistical method that considers several solutions at one time. This allows the algorithm to find good solutions that may be missed by other algorithms. Second, algorithms that imitate the ecosystem make use of the interaction among the possible solutions. For example, the genetic algorithm allows solutions to pair together and creates new solutions. Third, these algorithms allow exceptions. Therefore, solutions that are not typical but are actually better solutions can be found. Because of these reasons, the algorithms are suitable to solve the complex problem of analyzing mass amounts of complex biological data.

## Related Research

Clustering is the process of grouping together similar entities. Clustering is appropriate when there is no a priori knowledge about the data. In such circumstances, the only possible approach is to study the similarity between different samples or experiments. There are many existing clustering algorithms: hierarchical clustering, self-organizing maps, K-means clustering and many others.

Because of the characteristics of the DNA chip data such as the high complexity, large in amount, and changeable property, we propose a clustering algorithm which uses an algorithm that imitates the ecosystem.

## Particle Swarm Optimization

The Particle Swarm Optimization (PSO) algorithm (Parsopoulos *et al.*, 2002) that our system uses is a stochastic, population-based computer problem-solving

*Corresponding author: E-mail mlee@ehwa.ac.kr
Tel +82-2-3277-3401, Fax +82-2-3277-2306
Accepted 29 May 2011

algorithm. It is a kind of swarm intelligence that is based on social-psychological principles and provides insights into social behavior, as well as contributing to engineering applications. The swarm is typically modeled by particles in multidimensional space that have a position and a velocity. These particles fly through multidimensional space and have two essential reasoning capabilities. They are their memory of their own best position and knowledge of their neighborhood's best position. Here, the term "best" simply means the position with the smallest objective value. Members of a swarm communicate good positions to each other and adjust their own position and velocity based on these good positions. There are two main ways this communication is done. One is a global best that is known to all and immediately updated when a new best position is found by any particle in the swarm. The other is the "Neighborhood" bests where each particle only immediately communicates with a subset of the swarm about best positions. Suppose that the search space is *D*-dimensional, then the *i*-th particle of the swarm can be represented by a *D*-dimensional vector, *Xi* = (*xi*1, *xi*2, . . . , *xiD*). The velocity (position change) of this particle, can be represented by another D-dimensional vector *Vi* = (*vi*1, *vi*2, . . . , *viD*). The best previously visited position of the i-th particle is denoted as *Pi* = (*pi*1, *pi*2, . . . , *piD*). Defining g as the index of the best particle in the swarm, the PSO equation is the following.

$$v_{id}^{n+1} = v_{id}^{n} + cr_1^{n}(p_{id}^{n} - x_{id}^{n}) + cr_2^{n}(p_{gd}^{n} - x_{id}^{n})$$

$$x_{id}^{n+1} = x_{id}^{n} + v_{id}^{n+1}$$

Where *d* = 1, 2, . . ., *D*, *i* = 1, 2, . . . , *N*, and *N* is the size of the swarm,c is a positive constant, *r*1, *r*2 are random numbers, uniformly distributed in [0, 1], and *n* = 1, 2, . . ., determines the iteration number.

## DNA Chip Analysis System

DNA chips, in other words microarrays, are a tool for gene expression analysis. The DNA chip consists of the probe which is a single strand DNA printed on the solid substrate. The types of the chip or the name of the chip depends on the method of the chip fabrication. The idea behind the DNA chip is that the DNA in the solution that contains sequences complementary to the sequences of the DNA deposited on the surface of the array will be hybridized to those complementary sequences. Usually, this interrogation is done by washing the array with a solution containing DNA, called a target. To analyze the DNA chip, overviews of the steps that must be gone through are as follows. First, the researchers perform experiments using DNA chips. Afterwards the DNA analysis process is carried out by first scanning the results of the experiments. Next, quality control and normalization processes are carried out to revise errors. During this step, a lot of data is filtered and the quantity of the test data decreases. And then feature selection that selects specific parts is performed. The next step is the data mining work such as classification and clustering. Finally, the results are stored in the biological data warehouse and the warehouse system provides analysis results of integrated biological information to users.

Our tool focuses on the clustering process. Clustering is similar to classification in that data is being grouped. However, unlike classification, the groups are not predefined. Instead, the grouping is accomplished by finding similarities between data according to characteristics found in the actual data. The groups are called clusters. Clustering is appropriate when there is no a priori knowledge about the data. In such circumstances, the only possible approach is to study the similarity between different samples or experiments. Similarity is measured in many different ways, and the final result of the clustering depends on the formula used.

## PSO based clustering

Our clustering algorithm is based on the Particle swarm optimization (PSO), which is a population-based stochastic search process, modeled after the social behavior of a bird flock.

To perform clustering using the PSO-based Clustering System, the user needs to access the DNA chip data. In our system, the DNA chip data is stored in the database after going through the quality control and normalization. During the input stage, the system connects to the database and brings portions of the normalized biological data into the system memory. The next step is running the PSO clustering algorithm. In this step, the system initializes particles for each gene and then maps the particle and velocity. In order to decide the cluster number for each gene, the system calculates the fitness function. This work is done for all particles, and the system is updated with the global best value and local best values. And then the centroids of the clusters are updated. A cluster number for each gene is assigned by considering the fitness function value which is designed by the user. The last step is displaying the results. The system displays information such as the number of genes, the number of samples, iteration count, and the cluster number for each gene and so on. The overall system flow is shown in Fig. 1.

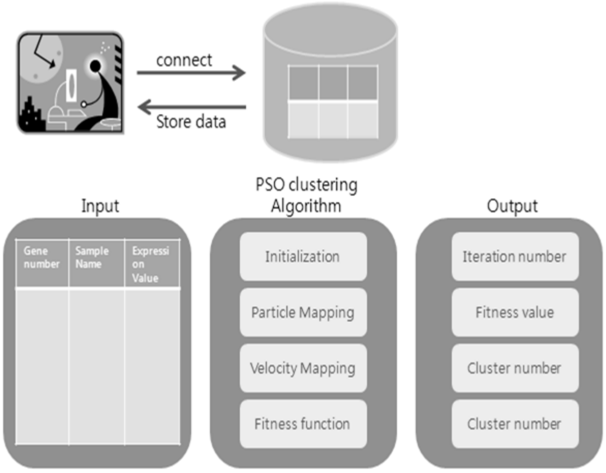
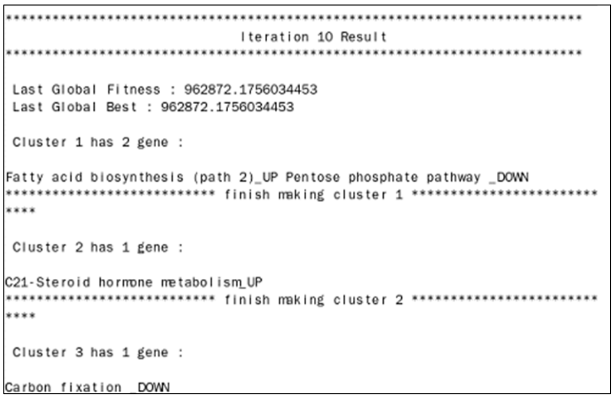


**Fig. 1.** Flow of PSO-based clustering system.



```
**************************************************************************
                              Iteration 10 Result
**************************************************************************

 Last Global Fitness : 962872.1756034453
 Last Global Best : 962872.1756034453

 Cluster 1 has 2 gene :

Fatty acid biosynthesis (path 2)_UP Pentose phosphate pathway _DOWN
************************** finish making cluster 1 ***********************
****

 Cluster 2 has 1 gene :

C21-Steroid hormone metabolism_UP
************************** finish making cluster 2 ***********************
****

 Cluster 3 has 1 gene :

Carbon fixation _DOWN
```

**Fig. 2.** Cluster formation while running PSO-based clustering system.

## Implementation & Experimental Results

The displayed results for the PSO-based Clustering System show the cluster and corresponding genes. The algorithm running is shown in Fig. 2.

We used the AB 1700 mouse chip (1- dye) which was provided from Macrogen Inc. We experimented with a test data set of 10 to 100 genes and 24 samples. We performed a comparison of the execution time with the K-means algorithm (Yuqing *et al*., 2003) and show that it has superior performance. The performance results are shown in Fig. 3.

## Conclusion

A DNA chip clustering system based on the Particle Swarm Optimization algorithm has been designed and implemented. The PSO algorithm uses a form of swarm intelligence. Clustering is accomplished by encoding particles to represent clustering results of the genes. Fitness functions are designed to calculate cluster distances via centroids. The implemented DNA chip system effectively clusters the data as verified by the experiments. In the future, we will optimize the algorithm to improve on the accuracy and execution speed.

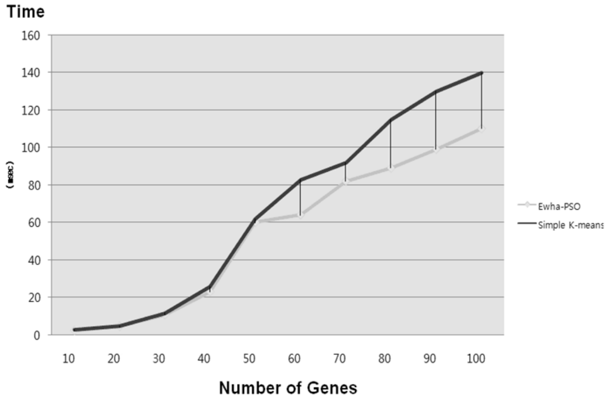


**Fig. 3.** PSO vs. K-means clustering execution time comparison.

### Acknowledgements

This work was supported by Mid-career Researcher Program through NRF grant funded by the MEST (No. 2009-0083992).

## References

Debouck, C., and Goodfellow, P.N. (1999). DNA microarrays in drug discovery and development. *Nat. Genet.* 21, 48-50.

DeRisi, J.L., Iver, V.R., and Brown, P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680-686.

Parsopoulos, K.E. and Vrahatis, M.N. (2002). Recent approaches to global optimization problems through Particle Swarm Optimization. *Nat Comput.* 1, 235-306.

Yuqing, P., Xiangdan, H., and Shang, L. (2003). The K-means Clustering Algorithm based on Density and Ant colony. In Proceedings of the IEEE Int. Conf. Neural Networks & Signal Processing, December 14-17, Nanjing, China.