*International Journal of Innovative Computing, Information and Control*
Volume **4**, Number **5**, May **2008**

ICIC International ©2008 ISSN 1349-4198
pp. 1203–1212

# A PROXIMITY APPROACH TO DNA BASED CLUSTERING ANALYSIS

Rohani Binti Abu Bakar, Junzo Watada
Graduate School of Information, Production and Systems
Waseda University
2-7 Hibikino, Wakamatsu, Kitakyushu, Fukuoka 808-0135, Japan
rohani@ump.edu.my; junzow@osb.att.ne.jp

Witold Pedrycz
Department of Electrical And Computer Engineering
University of Alberta
Edmonton AB T6R 2G7, Canada

Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland
pedrycz@ece.ualberta.ca

Received February 2007; revised August 2007

Abstract. *Clustering deals with huge amounts of data and aims at the discovery of their structure which becomes expressed in terms of a collection of clusters – information granules capturing the underlying topology of the data. The objective of this paper is to propose an algorithm to support clustering realized in the form of bio-soft or DNA computing. This approach is of particular interest when dealing with large and heterogeneous data sets and when being faced with an unknown number of clusters. We present the details of the algorithm of proximity clustering and show how the overall computing is supported by the individual mechanisms of DNA processing. We offer a numerical example to illustrate essential aspects of the DNA-based clustering.*

**Keywords:** DNA computing, Optimization, Clustering, Cluster validity, Proximity

## 1. Introduction.

The primordial objective of clustering is to discover a structure in data by forming a finite number of clusters (groups). The underlying principle is evident: we expect that similar objects (patterns) will be placed in the same cluster while different objects are assigned to different clusters. Clustering is widely used in various areas such as machine learning, image analysis, data mining, and bioinformatics, in particular when dealing with a very large database.

Currently, a great deal of interest was focused on proximity algorithms to cluster (structure) different sources of information and involve specific domain knowledge pertinent to the problem at hand. For instance, Oehler and Gray have developed a clustering technique to solve a problem in signal processing and vector quantization [10]. Shopbell et al. proposed a clustering technique to cluster objects in the sky for astronomy [13]. Jiang and Tuzhillin focused on clustering customers interests studying relations with a certain marketing problem [7]. Jimmy et al. addressed several issues of clustering medical data [8].

The ultimate challenge of clustering associates with a combinatorial explosion of the search space. Another challenge comes with the fact that almost all clustering techniques require that a number of clusters is provided in advance. While a number of enhancements