Haplotyping Problem, A Clustering Approach

Changiz Eslahchi[1,3], Mehdi Sadeghi[1,2,4], Hamid Pezeshk[1,5], Hadi Poormohammadi[1,3], Mehdi Kargar[1,6]

[1]Bioinformatics Group, School of Computer Science, IPM, Tehran, Iran.

[2]National Institute for Genetic Engineering and Biotechnology, Tehran-Karaj Highway, Tehran, Iran.

[3]Faculty of Mathematics, Shahid-Beheshti University, Tehran, Iran.

[4]Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.

[5]Center of Excellence in Biomathematics, School of Mathematics, Statistics and Computer Sciences, University College of Science, University of Tehran, Tehran, Iran.

[6]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran.

The sequencing of human genome is one of the most important fields in biology. Human genomes are a sequence of three billion letters from the nucleotide alphabet {A, C, G, T}. Comparing genome of different population indicates that differences only occur in about 1% of nucleotide positions called Single Nucleotide Polymorphism (SNP). In human genome there are two copies of each chromosome. A SNP sequence from one of two chromosomes is called haplotype. Given a set of SNP fragments obtained by sequencing two copies of chromosomes, Minimum Error Correction (MEC) finds and corrects minimum number of error positions. MEC is an NP-hard problem.

In this paper we proposed a novel algorithm based on clustering analysis in data mining for haplotyping problem. In contrast to MEC model which is NP-hard, our heuristic algorithm has polynomial time complexity and could be applied to large datasets. Based on maximum normalized hamming distance, our iterative algorithm produces two clusters of fragments. Using normalized hamming distance, in each iteration, the algorithm approximates the best fragment to a given cluster. Compared to other methods, in this novel approach, both the clusters of fragments are used to reconstruct each haplotype. Our results show that the algorithm has less reconstruction error rate in comparison with other algorithms. Also this reconstruction error rate is very close to zero when this is applied to actual biological data. A discussion on several input parameters influencing reconstruction error rate is also presented.