Australian Journal of Basic and Applied Sciences, 3(1): 277-282, 2009
ISSN 1991-8178

Solving Haplotype Reconstruction Problem in MEC Model with Hybrid Information Fusion

[1]Ehsan Asgarian, [2]M-Hossein Moeinzadeh, [3]Ammar Rasooli Valaghozi, [1]Shahrouz Moaven

[1]Department of Computer engineering, Sharif University of Technology, Tehran, Iran
[2]School of Mathematics, Statistics and Computer Science, University of Tehran, Iran
[3]Department of Applied Mathematics, Iran University of Science and Technology, Iran

Abstract: Single Nucleotide Polymorphisms (SNPs), a single DNA base varying from one individual to another, are believed to be the most frequent form responsible for genetic diferences. Genotype is the conflated information of a pair of haplotypes on homologous chromosomes. Although haplotypes have more information for disease associating than individual SNPs and genotype, it is substantially more dificult to determine haplotypes through experiments. Hence, computational methods which can reduce the cost of determining haplotypes become attractive alternatives. MEC, as a standard model for haplotype reconstruction, is fed by fragments as input to infer the best pair of haplotypes with minimum error to be corrected. It is proved that haplotype reconstruction in MEC model is a NP-Hard problem. Thus, reducing running time and obtaining acceptable result are desired by researchers. Heuristic algorithms and diferent clustering methods are employed to achieve these goals. In this paper, the idea of combining diferent methods is presented. A hybrid model, which is employed the eficiency of diferent serial and parallel models, is suggested. FCA, K-means and neural network are considered as its component. K-means clustering method is used to improve neural network eficiency. Then the results are compared in diferent datasets.

Keys words: Haplotype; SNP Fragments; MEC Model; SOM Neural Network; Fuzzy Clustering.

INTRODUCTION

The availability of complete genome sequence for human beings makes it possible to investigate genetic differences and to associate genetic variations with complex diseases (Venter *et al.,* 2001). It is generally accepted that all human beings share about 99% identity at the DNA level and only some regions of differences in DNA sequences are responsible for genetic diseases (Terwilliger and Weiss, 1998). Single Nucleotide Polymorphisms (SNPs), a single DNA base varying from one individual to another, are believed to be the most frequent form responsible for genetic differences (Bonizzoni *et al.,* 2003) and are found approximately every 1000 base pairs in the human genome and turn to be promising tools for doing disease association study. Every nucleotide in a SNP site is called an allele. Almost all SNPs have two different alleles, known here as 'A' and 'B'. The SNP sequence on each copy of a pair of chromosomes in a diploid genome is called a haplotype which is a string over {'A', 'B'}. A genotype is the conflated information of a pair of haplotypes on homologous chromosomes.

Although haplotypes have more information for disease associating than individual SNPs and genotype, it is substantially more difficult to determine haplotypes than genotypes or individual SNPs through experiments. Hence, computational methods which can reduce the cost of determining haplotypes become attractive alternatives. SNP fragments are composed of Gaps and errors. One question arising from this discussion is that how the distribution of gaps and error in the input data affects computational complexity. Minimum Error Correction (MEC) (Panconesi and Sozio, 2004), Longest Haplotype Reconstruction (LHR) (Greenberg *et al.,* 2004), Minimum Error Correction with Genotype Information (MEC/GI) (Moeinzadeh *et al.,* 2007), Minimum Conflict Individual Haplotyping (MCIH) (Venter *et al.,* 2001) and some other models have been discussed for haplotype reconstruction. MEC and its alternative (MEC/GI) as two standard models are decided to be central problem of our research. MEC, is a standard model for haplotype reconstruction, which is fed by fragments as an input to infer the best pair of haplotypes with the minimum error to be corrected. Genotype information is also used to improve the efficiency of methods in MEC/GI model (Asgarian *et al.,*2007). For MEC model two different procedures can be employed to resolve the problem. First, Partitioning

Corresponding Author: Ehsan Asgarian, Department of Computer engineering, Sharif University of Technology, Tehran, Iran
E-mali: asgarian@alum.sharif.edu

and clustering methods can be designed to divide the SNP fragments into two classes. In this approach, each class corresponds to one haplotype. To infer the haplotypes from each partition, another function is designed (described later). The second approach is based on inferring haplotypes directly from SNP fragments and correcting simultaneously the errors.

It was proved that haplotype reconstruction in MEC model is a NP-Hard problem (Venter *et al.,* 2001). Thus, reducing running time and obtaining acceptable result have been desired by researchers. Heuristic algorithm and different clustering methods are employed to achieve these goals. Combining different methods as information fusion techniques have also been attended by researches (Asgarian *et al.,* 2008; Wang *et al.,* 2005).

In this paper parallel and serial fusions were combined to make hybrid method. The idea of hybrid fusion leads us to obtain better results in simulation and real datasets. Therefore, we used reconstruction rate as the standard comparison method to demonstrate the efficiency of our proposed approach.

Biological definitions like SNP, SNP fragments, haplotype and genotype are formulated in next section (section 2). In section 3, FCA, GA, K-means, neural network as single methods and also serial and parallel methods as related works in MEC model are introduced. The FCA, K-means and neural network approaches are considered as supplemental methods for our solution. In section 4, the proposed approach is fully discussed in details. Experimental results on different datasets and conclusion are the next two sections.

Problem Definitions:

Suppose that there are m SNP fragments from a pair of haplotypes. Each SNP fragment corresponds to one of the two target haplotypes. For more convenience, fragment is used instead of SNP fragment in the rest of the paper. M=mij is defined as a matrix of fragments, which each entry mij has value ‘A’, ‘B’ or ‘-’ (‘-’ is missing or skipped SNP site which is called gap). The rows and column of the matrix Mn×m demonstrate fragments and SNP sites respectively. The length of fragments including their gaps is the same as the two haplotypes which is equal to n.

We use partition P(C1,C2) (C1 and C2 are two classes) to formulate the problem. P as an exact algorithm or clustering method divides fragments into C1 and C2 (Figure1). Each haplotype is reconstructed from the members of one of the classes with voting function. The function is performed on all fragment columns of each class in order to decide for the values on the corresponding SNP site of related haplotypes. The function is so defined: (NiA(M) (or NiB(M)) denotes the number of ‘A’s (or 'B's) in jth column of matrix M)

$$v_{ij} = \begin{cases} A & N_A^j(C_i) N_B^J(C_i) \\ B & Otherwise \end{cases}$$

$$i = 1,2$$

$$0 \le j \prec n$$

Reconstruction rate (shortly RR) is a very simple and popular mean to compare the results of designed algorithms on existing datasets. RR, which is based on Hamming Distance (HD), is the degree of similarity between the original haplotypes (h = (h1 ,h2)) and reconstructed ones (h'= (h'1 *,h'2)).d (x , y)* is defined as the difference of two alleles in one SNP site. Hamming distance of two fragments HD(fi, fj) and RR(h ,h') are formulated as:

$$d(h_{ij}, h_{kj}) = \begin{cases} +1 & (m_{ij} \neq m_{kj} \neq -) \\ 0 & Otherwise \end{cases}$$

$$HD(h_i, h_k) = \sum_{j=1}^{n} d(h_{ij}, h_{kj})$$

$$r_{ij} = HD(h_i, h'_j) \qquad i, j = \{1,2\}$$

$$RR(h, h') = 1 - \frac{\min(r_{11} - r_{22}, r_{12} + r_{21})}{2n}$$

HD1 and HD2 are considered as two distances obtained from comparison of fi and the two other fragments (f1 and f2). Another function is recommended to distinguish these two distances (HD1 and HD2). When HD1 and HD2, the following distance function was used.

$$D'_{mm}(m_i, m_k) = \sum_{j=1}^{n} d'(m_{ij}, m_{kj}),$$

$$d'(m_{ij}, m_{kj}) = \begin{cases} -1 & (m_{ij} = m_{kj} \neq -) \\ +1 & (m_{ij} = m_{kj} \neq -) \\ 0 & Otherwise \end{cases}$$

Related Work:

GA: To resolve haplotype assembly Wang *et al.* (2008) have proposed a genetic algorithm to cluster the fragments. The chromosomes are defined as a binary string *Chi* ~ {0, 1} of length m (number of fragments). When Chi is equal to 0 (or 1), it means that the ith fragment is considered to be one of the first (or the second) class members. Goodness and badness of individuals must be assigned based on number of error corrections required. But there is no cluster centers obtained yet. There is a fitness function recommended for evaluating the individuals by Wang *et al.* (2008), which computes the distance of all fragments with their class centers (the class centers in this problem are computed by a voting function method).

Table 1: The resuts of defferent method in different datasets

Gap	R. Error	ACE Database-MEC					SIM _0 Database-MEC					SIM_50 Database-MEC				
		K-Means	SOM	FCM	IFPC (K-Means+ AGC-HCC)	Hybrid Model	K-Means	SOM	FCM	IFPC (K-Means+ AGC-HCC)	Hybrid Model	K-Means	SOM	FCM	IFPC (K-Means+ AGC-HCC)	Hybrid Model
0.25	0.1	0.996	0.969	0.999	0.999	0.990	0.999	0.996	0.996	1.000	1.000	0.996	0.999	0.999	1.000	1.000
	0.2	0.952	0.915	0.963	0.971	0.963	0.952	0.965	0.965	0.979	1.000	0.965	0.952	0.954	1.000	0.990
	0.3	0.814	0.814	0.846	0.851	0.832	0.817	0.851	0.865	0.805	0.925	0.849	0.811	0.845	0.921	0.818
	0.4	0.650	0.667	0.678	0.739	0.746	0.651	0.644	0.639	0.684	0.672	0.618	0.655	0.621	0.698	
0.5	0.1	0.977	0.971	0.977	0.974	0.974	0.977	0.988	0.989	0.993	1.000	0.988	0.979	0.978	1.000	0.995
	0.2	0.898	0.886	0.889	0.969	0.940	0.910	0.919	0.920	0.966	0.992	0.919	0.905	0.919	0.979	0.967
	0.3	0.755	0.762	0.752	0.775	0.913	0.768	0.753	0.786	0.788	0.884	0.752	0.766	0.780	0.809	0.846
	0.4	0.637	0.632	0.649	0.694	0.751	0.640	0.556	0.554	0.684	0.794	0.556	0.627	0.647	0.690	0.808
0.75	0.1	0.887	0.877	0.885	0.972	0.974	0.908	0.916	0.915	0.957	0.998	0.912	0.915	0.915	0.990	0.987
	0.2	0.738	0.769	0.755	0.858	0.944	0.744	0.731	0.741	0.902	0.936	0.693	0.757	0.798	0.896	0.942
	0.3	0.680	0.677	0.667	0.786	0.922	0.675	0.608	0.616	0.837	0.907	0.628	0.675	0.665	0.812	0.923
	0.4	0.624	0.611	0.632	0.770	0.823	0.606	0.556	0.545	0.763	0.818	0.543	0.603	0.610	0.734	0.917

FCA: FCA (Fuzzy clustering approach) was proposed in (Moeinzadeh *et al.,* 2007). In this method, the degree of membership between SNP fragments and the clusters was defined. The SNP-fragments are clustered according to degree of membership and two haplotypes are inferred from the clusters. Therefore, two centers are made iteratively until the stop condition satisfied.

Neural Network:

A two layer competitive unsupervised neural network (UWNN) has been designed to solve mentioned models Moeinzadeh *et al.* (2007). Fragments are fed to the neural network consecutively. The first layer is made up of 'n' units (SNPs layer), each node related to one SNP site, while the second one is composed of two units. Two strings, called semi-haplotypes (haplotype which is demonstrated with decimal number) are reconstructed after appropriate epochs. One of secondary layer nodes, which are competed based on the similarity of each fragment and semihaplotypes, is marked as the winner. The weights of the winner node are updated in each epoch.

K-means:

A heuristic clustering method (based on MEC model) has been published by Wang *et al.* First, two fragments are selected as the primitive centers. So the other fragments are clustered according to hamming distance of them and specified centers. Each iteration, the centers are updated according to newly constructed clusters and voting function. So in the next iteration, the distance between the new centers and all the fragments has to be computed for clustering. Numerical results approve the efficiency of this method.

Proposed Approaches:

Haplotypes Reconstruction is considered as a Multiple Clustering System (MCS) which consists of a set of individual clustering approaches (Zhang approach Venter *et al.,*(2001)). For this system we define a fusion method to combine simple clustering outputs and make the final decision.

$MCS= \{W(C\ 1\ ),W(C2),W(C3\ ),...,W(C_n\ )\}$

In this paper, we try to design special composition of clustering methods based on our problem models. Then K-means, FCA and neural network were selected as our model components. In the following section the hybrid design of mentioned single method is explained.

The best clustering methods are not necessarily the ideal choice due to noisy and incomplete inputs. So we decided to combine the result of the clustering approaches to infer better haplotypes. Serial and parallel form of clustering was discussed in (Asgarian *et al.,* 2008; Wang *et al.,* 2005). In mentioned papers GA and K-means were decided as serial classifier components. A greedy algorithm was combined with K-means in parallel form in (Wang *et al.,* 2005). In this paper implementation of Hybrid classifiers are focused.

Serial Form in Hybrid Model:

A Serial classifier has two main components, classifiers and information fusion function. In our problem model, information fusion functions are designed based on classifiers properties. Therefore K-means and NN were selected in serial form. The efficiency of NN and its dependency to its primitive initial values leads us to select it as the second classifier. Good initial centers can greatly affect the results of neural networks. K-means clustering method finds good solution in first step. Our approach is based on classification SNP fragments by K-means to generate two acceptable centers. Therefore, good initial weights of neural network can help the algorithm to converge to better results.

Parallel Form in Hybrid Model:

A parallel classifier is proposed which combines K-means-NN and FCA. Information fusion combiner function decides according to the majority of classifier decisions. The algorithm ignores noisy input data with the following formulation:

$$\underset{i=1..m}{Fragment} = \begin{cases} 1^{st}\,Class & 1^{st}\,Classfier = 2^{nd}\,Classfier = 1^{st}\,Class \\ 2^{nd}\,Class & 1^{st}\,Classfier = 2^{nd}\,Classfier = 2^{nd}\,Class \\ omitted & otherwise \end{cases}$$

As shown in previous formula, some haplotypes (which might be useful for reconstruction of final haplotypes) are eliminated.

The explanation behind this is to eliminate those noisy haplotypes, on which the two classifiers decisions do not match. So the following algorithm was designed.

Experimental Results:

We use Visual C++ 6.0. in order to implement our methods. The proposed approaches are tested on standard real and simulation datasets. Each dataset has different error and gap rates (Error Rate = 10%, 20%, 30% and 40% and Gap rate = 25%, 50%, 75%). ACE (Angiotensin Converting Enzyme) as real dataset, includes 24 different test cases for each error rate. SIM0 and SIM50 are simulation datasets which are made for special purpose. In SIM0, there is no similarity between the pairs of original haplotypes (30 test files for each error rate). In the same way, 50% of the haplotypes are used to generate the incorrect and gapped fragments (30 test files for each error rate). Our algorithms were totally tested on 336 test files. We implemented a clustering method which is proposed in (Venter *et al.,* 2001). (as Zhang approach). Neural network and FCA were also implemented in (Wang *et al.,* 2007) and (Mocinzadeh *et al.,* 2007).The reliability of our implementation is proved by comparison between our and the mentioned papers results.

Conclusion:

In this paper, we try to combine different methods. MCS (Multiple Clustering System) is employed to solve MEC model. A hybrid clustering approach is designed in order to use information fusion. So first we tried to improve the result of NN by employing K-means to obtain initial weight of neural network. The idea of combination in parallel form was used to cover the weaknesses of single clustering methods. By this method, noisy fragments are also established and then eliminated. All the aforementioned methods are implemented and tested on MEC model problem which are intended to infer haplotypes with high accuracy. We compare the results of all methods in terms of Reconstruction Rate. Consequently, it is proved by the experience that using appropriate clustering approaches as the components and information fusion techniques can improvethe results.

Algorithm: Parallel clustering

Input: SNP fragments
Output: two haplotypes
Step0: Initialize parameters

Parallel step

Step1.1: Perform K-Means algorithm to produce two centers
Step1.2: Initialize neural network weights with K-means results.
Step1.3: Run Neural network to obtain haplotypes
Step1.4: Cluster fragments based on New obtained haplotypes

Parallel step

Step1.1: Perform FCA algorithm to produce two clusters.
Step2:Eliminate noisy fragments which are not confirmed with both clustering approaches
Step3: Obtain two haplotypes from new classes

Algorithm1:. Pseudo code parallel Classifier.

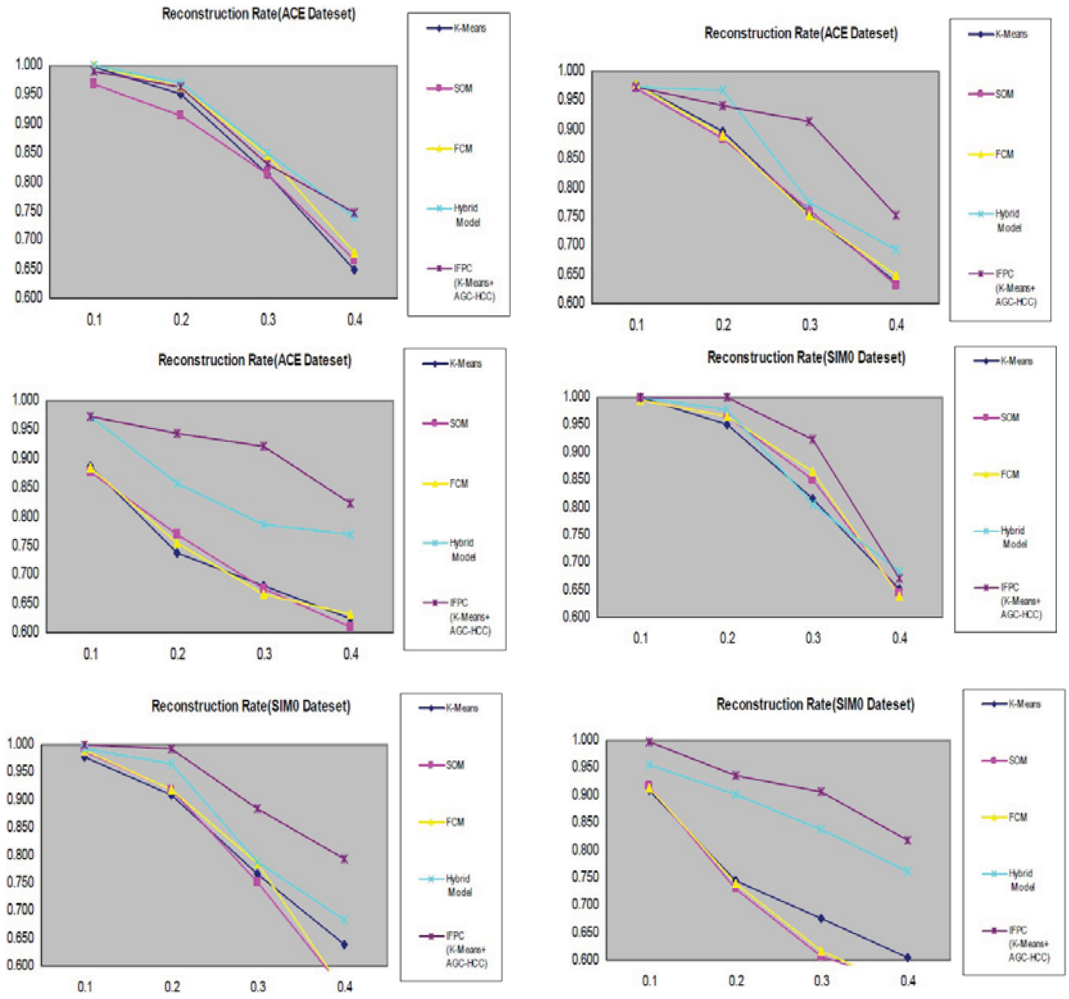

Fig. 1: Comparison of Reconstruction in ACE dataset

REFERENCES

Asgarian, E., M.H. Moeinzadeh, A.A. Najafi, R.S. Sharifian, J. Habibi, J. Mohammadzadeh, 2007. Solving MEC and MEC/GI Problem Models, Using Information Fusion and Multiple Classifiers, the 4th IEEE International Conference on Innovations in Information Technology, pp: 397-401.

Asgarian, E., M.H.A. Moeinzadeh, R.S. Sharifian, A. Najafi, A. Ramezani, J. Habibi, J. Mohammadzadeh, 2008. Solving MEC Model of Haplotype Reconstruction Using Information Fusion, Single Greedy and Parallel Clustering Approaches, The sixth ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-08), pp: 15-19.

Bonizzoni, P., G.D. Vedova, R. Dondi,, J. Li, 2003. The Haplotyping Problem: An overview of Computational Models and Solutions, Journal of Computer Science and Technology, 18(6): 675-688.

Chakravarti, 2001. It's raining, hallelujah? , Nature Genetics, 19: 216–217.

Greenberg, H.J., E.W. Hart, G. Lancia, 2004. Opportunities for Combinatorial Optimization in Computational Biology", INFORMS Journal on Computing, 16(3): 211-231.

Panconesi and M. Sozio, 2004. Fast Hare: A Fast Heuristic for Single Individual SNP Haplotype Reconstruction, Proceedings of 4th Workshop on Algorithms in Bioinformatics (WABI), LNCS Springer-Verlag, pp: 266-277.

Moeinzadeh, M.H., E. Asgarian, J. Mohammadzadeh, A. Ghazinezhad, A.A. Najafi, 2007. Three Heuristic Clustering Methods for Haplotype Reconstruction Problem with Genotype Information, the 4th IEEE International Conference on Innovations in Information Technology, pp: 402-406.

Moeinzadeh, M.H., E. Asgarian, M. Mohammad Noori, M. Sadeghi, A. Rasooli, J. Habibi, 2008. FCA: Designing a fuzzy Clustering Algorithm for Haplotype Assembly, 8th International Conference on BioInformatics and BioEngineering (BIBE).

Moeinzadeh, M.H., E. Asgarian, S. Sharifian, A. Najafi, J. Mohammadzadeh, 2007. Neural Network Based Approaches, Solving Haplotype Reconstruction in MEC and MEC/GI Models, Second Asia IEEE International Conference on Modeling & Simulation, pp: 934-939.

Terwilliger, J. and K. Weiss, 1998. Linkage disquilibrium mapping of complex disease: Fantasy and reality? Curr. Opin. Biotechnology, pp: 579–594.

Venter, J.C., M.D. Adams, *et al.*, 2001. The sequence of the human genome. Science, 291(5507): pp. 1304–1351.

Wang, R., L. Wu, Z. Li and X. Zhang, 2005. Haplotype reconstruction from SNP fragments by Minimum Error Correction. Bioinformatics, 21(10): 2456–2462.

Wang, Y., E. Feng,, R. Wang, 2007. A clustering algorithm based on two distance functions for MEC model, Computational biology and chemistry, 148: 150.