Journal of Bioinformatics and Sequence Analysis Vol. 3(4), pp. 63-69, May 2011
Available online at http://www.academicjournals.org/JBSA
DOI: 10.5897/JBSA09.017
ISSN 2141-2464 ©2011 Academic Journals

Full Length Research Paper

Information fusion and multiple classifiers for haplotype assembly problem from SNP fragments and related genotype

M. Hossein Moeinzadeh[1] and Ehsan Asgarian[2]*

[1]Department of Computer Science, University of Tehran, Iran.
[2]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran.

Accepted 28 April, 2011

Most positions of the human genome are typically invariant (99%) and only some positions (1%) are commonly invariant which are associated with complex genetic diseases. Haplotype information has become increasingly important in analyzing fine-scale molecular genetics data, due to the mutated form in human genome. Haplotype assembly is to divide aligned single nucleotide polymorphism (SNP) fragments, which is the most frequent form of difference to address genetic diseases, into two classes, and thus inferring a pair of haplotypes from them. Minimum error correction (MEC) is an important model for this problem but only effective when the error rate of the fragments is low. MEC/GI as an extension to MEC, employs the related genotype information besides the SNP fragments and so results in a more accurate inference. The haplotyping problem, due to its NP-hardness, may have no efficient algorithm for exact solution. In this paper, we focus to design serial and parallel classifiers with two classifiers. Genetic algorithm and K-means were two components of our approaches. This combination helps us to cover the single classifier's weaknesses.

Key words: Multiple classifier systems, parallel classifiers, serial classifiers, haplotype, SNP fragments, genotype information, classification, reconstruction rate.

INTRODUCTION

The availability of complete genome sequence for human beings (Venter, 2001) makes it possible to investigate genetic differences and to associate genetic variations with complex diseases (Zhang, 2006). It is generally accepted that all human share about 99% identity at the deoxyribonucleic acid (DNA) level and only some regions of differences in DNA sequences are responsible for genetic diseases (Terwilliger et al., 1998; Chakravarti, 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 (Wang, 2005) 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 an single nucleotide polymorphisms site is called an allele. Most SNPs have two different alleles, known here as 'A' and 'B'. The SNP sequence information 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 association than individual SNPs and also more than genotype information, but it is substantially more difficult to determine haplotypes than to determine genotypes or individual single nucleotide polymorphisms through experiments. Hence, computational methods that can reduce the cost of determining haplotypes become attractive alternatives. Hole and error

*Corresponding author. E-mail: asgarian@alum.sharif.edu Tel: + (98) 9155084478. Fax: + (98) 511 8939527.

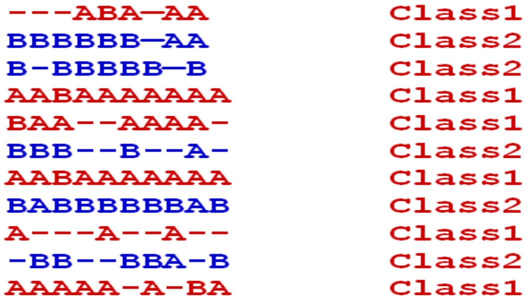
Abbreviations: MEC, Minimum error correction; **MEC/GI,** minimum error correction with genotype information; **MCS,** multiple classifier system; **DNA,** deoxyribonucleic acid; **GA,** genetic algorithm; **RR,** reconstruction rate; **SNP,** single nucleotide polymorphism.

consist in SNP fragments.

One question arising from this discussion is how the distribution of holes and error in the input data affects computational complexity. Minimum error correction (MEC), longest haplotype reconstruction (LHR), minimum error correction with genotype information (MEC/GI) and some other models have been discussed for haplotype assembly (Terwilliger, 1998; Chakravarti, 1998; Wang, 2005, 2007; Zhang, 2006, 2007). MEC and also MEC/GI are two standard models for haplotype reconstruction based on SNP fragments and genotype information as an input data to infer the best pair of haplotypes with the minimum error to be corrected. It is proved that MEC is a NP-hard problem (Bonizzoni, 2003; Zhang, 2006), so heuristic methods are used to reduce running time of this problem (Moeinzadeh et al., 2007). This problem was solved by some classification and heuristic methods. Zhang (2007) introduces a classification algorithm based on two distances (Hamming and a proposed distance) to compare SNP fragments together. An algorithm was implemented to solve MEC model (Zhang, 2007). Real and simulation data sets are available as two standard databases. The input in these databases contains an error rate between 10 and 40%. The method by Zhang (2007) is based on K-means algorithm. Although the result and algorithm's running time were acceptable, it is widely believed that K-means does not work well for noisy inputs. Solving MEC and MEC/GI models for haplotype assembly with genetic algorithm (GA) was published by Wang (2005) and Zhang (2006). The results in haplotyping were not only better than K-means but also it takes more execution time. On the other hand, GA has an adaptive behavior in terms of error rate and it approximately guarantees not to get stuck local minima. A variety of approaches that each of them might have its own strengths and weaknesses made us to combine the classifier's results.

We use information fusion techniques to improve our results. These techniques are information processes dealing with the association, correlation, and combination of data and information from single and multiple classifiers or sources to achieve refined estimation of parameters, characteristics, events, and behaviors. This approach is used to improve the result of the mentioned problem.

In this paper, we design serial and parallel classifiers. Utilizing multiple classifiers would help us to increase reconstruction rate (RR), which is described in the following sections, and also using error rate for the first time. In our research, we concentrate on K-means and GA properties and use them together. K.G (K-means.GA), G.K (GA.K-means) as two serial classifiers was implemented. In these two approaches, first classifier's answers and main input are fused to form the second classifier's inputs. We also implement K.G.K and G.K.G to study the results. In parallel classification, result



---ABA—AA	Class1
BBBBBB—AA	Class2
B-BBBBB—B	Class2
AABAAAAAAA	Class1
BAA--AAAA-	Class1
BBB--B--A-	Class2
AABAAAAAAA	Class1
BABBBBBBAB	Class2
A---A--A--	Class1
-BB--BBA-B	Class2
AAAAA-A-BA	Class1

Figure 1. Classification of the SNP fragments.

combiner, as an information fusion function, was the

Formulation and problem definition

Suppose that there are m SNP fragments from a pair of chromosomes, corresponding to two haplotypes with the length of n, defined $(M=m_{ij}`)$ as a matrix of SNP fragments, whose every entry m_{ij} has value 'A', 'B' or '-' ('-' is missing or skipped SNP site which is called gap). Each row of the matrix M is one SNP fragment and each column corresponds to one SNP site. The length of SNP fragments including their gaps is the same as its own haplotype. We use partition P (C1, C2), class C1 and class C2, to formulate the problem. P as an exact algorithm or classification method divides SNP fragments into two classes, C1 and C2. The SNP fragments in each class must combine with their own class members to reconstruct the haplotypes. We call this operation voting and define it completely in the following parts (Figure 1). Each genotype is the conflation of two haplotypes, depending on their sites. We define genotype as a string of 'A', 'B' and '-'. 'A' ('B') denotes that both haplotypes are the 'A' ('B') and '-'when they are heterozygous. Reconstruction rate (shortly RR) is a very simple and popular way for comparing the results of designed algorithms in existing databases with each other. RR which is based on Hamming distance is the degree of similarity between the original haplotypes $(h=(h_1,h_2))$ and the reconstructed ones $(h'=(h'_1,h'_2))$. It is defined as:

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

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

And also Hamming distance is based on the distance between two SNP fragments which we call it $d(x,y)$, and which is in turn, defined by the following formula:

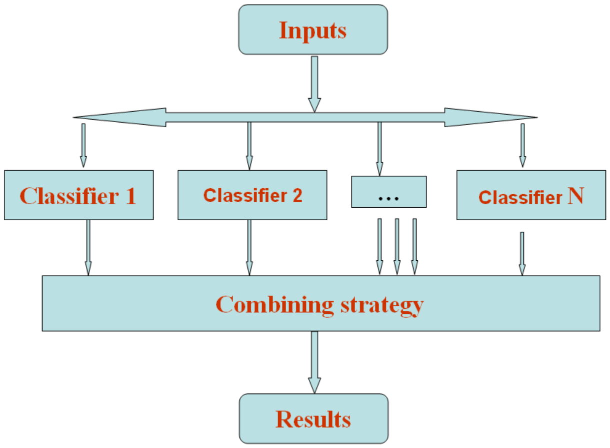


Figure 2. Parallel classifier.

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

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

And the usage of a second distance becomes necessary when the hamming distance between one and two other fragments are equal, which is defined as follows:

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

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

In this paper, we study MEC (Minimum Error Correction) model. In this model, a matrix of SNP fragments is available as an input. We try to decrease the number of haplotype errors in comparison with corresponding real haplotypes. For doing so, all the aforementioned algorithms in MEC/GI (Minimum error correction with genotype information) model were implemented.

METHODOLOGY

Haplotypes assembly is considered as a multiple classifier system (MCS) which consists of a set of individual classifiers like genetic algorithm (GA) and K-means. For this system, we define a fusion or selection method to combine simple classifiers outputs and make the final decision.

$$MCS = \{W(H), W(E), \ldots, W(C)\}$$

In this paper, we try to design special composition of classifiers based on our problem models. In the following section, some serial and parallel designs are explained.

Serial classifiers

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. In the following section, we describe our serial classifier methods.

K-means – GA

Good initial population can greatly affect the results of genetic algorithm. Our first approach is based on classification SNP fragments by K-means to generate an individual with acceptable fitness. The result of K-means is used for generating initial population for genetic algorithm. The initial population of GA is consisted of K-means answers combined with a predetermined error rate in input and also random individuals to escape from local minima. Genetic algorithm starts to optimize the result of the K-means method. For generating the initial population, we used error information. This combination produces good initial population. The algorithm for MEC (the changes needed for MEC/GI model) is shown in Algorithm 1. The results of this algorithm are discussed in experimental results section.

GA - K-means

For serial MCS information fusion techniques, K-means properties were the center of attention. In haplotype assembly problem, K-means heuristic method needs two centers to start classification procedure. So good initial centers can help the algorithm to converge to better results. In our approach, genetic algorithm was used to find these centers and K-means tries to change the results location towards the real centers. The GA- K-means designed algorithm for MEC model is shown in Figure 2. Here, we designed an algorithm for MEC/GI model with the same fusion function along with some modifications. It was predictable that the result of GA- K-means would be better than the results obtained by applying them individually (Tables 1 and 2).

KGK and GKG

We defined two different information fusion functions in the last two sections (K-means to GA (KG) and GA to K-means (GK). In the same manner, two more algorithms were designed (KGK and GKG) and also their result is discussed in experimental results section (Algorithm 2).

Parallel classifier

Using a single classifier and designing efficient algorithms were not enough to solve MEC and MEC/GI models. The best classifiers are not necessarily the ideal choice in this problem due to noisy and incomplete inputs, so we implement parallel classifier (Figure 3).

K-means and GA

A parallel classifier is proposed which combines K-means and GA. Information fusion combiner function decides according to the majority of classifier decisions. Information fusion function was designed based on ignoring noisy input data. Here is the function:

$$Fragment_i = \begin{cases} 1^{st}\, Class & 1^{st}\, Classifier = 2^{nd}\, Classifier = 1^{st}\, Class \\ 2^{nd}\, Class & 1^{st}\, Classifier = 2^{nd}\, Classifier = 2^{nd}\, Class \\ omitted & otherwise \end{cases}$$

$$i = 1 \ldots m$$

Table 1. The comparison of reconstruction rate of multiple classifiers algorithms for MEC model.

Daly database –MEC model								
Gap rate	Error rate	Single classifier		Serial classifier		Parallel classifier		
		K-Means(k)	G.A. (G)	K.G	G.K	G.K.G	K.G.K	G.K.G and K.G.K
0.25	0.1	0.999	1.000	1.000	0.999	1.000	0.999	0.997
	0.2	0.993	0.993	0.993	0.994	0.994	0.994	0.989
	0.3	0.931	0.933	0.933	0.920	0.929	0.932	0.875
	0.4	0.716	0.718	0.718	0.712	0.713	0.717	0.696
0.5	0.1	0.998	0.998	0.998	0.997	0.997	0.998	0.995
	0.2	0.972	0.973	0.973	0.994	0.973	0.975	0.972
	0.3	0.861	0.869	0.869	0.867	0.868	0.869	0.819
	0.4	0.691	0.694	0.694	0.689	0.690	0.690	0.677
0.75	0.1	0.977	0.977	0.977	0.978	0.976	0.978	0.978
	0.2	0.896	0.898	0.898	0.889	0.892	0.901	0.897
	0.3	0.772	0.774	0.774	0.770	0.765	0.768	0.762
	0.4	0.663	0.665	0.665	0.660	0.653	0.661	0.678

Table 2. The comparison of reconstruction rate of multiple classifiers algorithms for MEC/GI model.

Daly database –MEC model								
Gap rate	Error rate	Single classifier		Serial classifier		Parallel classifier		
		K-Means(k)	G.A. (G)	K.G	G.K	G.K.G	K.G.K	G.K.G and K.G.K
0.25	0.1	1.000	1.000	1.000	1.000	1.000	1.000	0.924
	0.2	0.999	0.999	0.999	1.000	1.000	1.000	0.928
	0.3	0.993	0.974	0.993	0.994	0.994	0.992	0.925
	0.4	0.904	0.897	0.908	0.896	0.903	0.896	0.894
0.5	0.1	1.000	0.999	1.000	1.000	1.000	1.000	0.934
	0.2	0.999	0.993	0.999	0.999	1.000	0.999	0.926
	0.3	0.972	0.956	0.977	0.975	0.978	0.977	0.926
	0.4	0.881	0.882	0.884	0.881	0.882	0.884	0.876
0.75	0.1	0.998	0.992	0.998	0.998	0.998	0. 997	0.925
	0.2	0.977	0.973	0.981	0.978	0.983	0.980	0.928
	0.3	0.914	0.901	0.925	0.907	0.921	0.915	0.898
	0.4	0.872	0.869	0.871	0.868	0.873	0.870	0.875

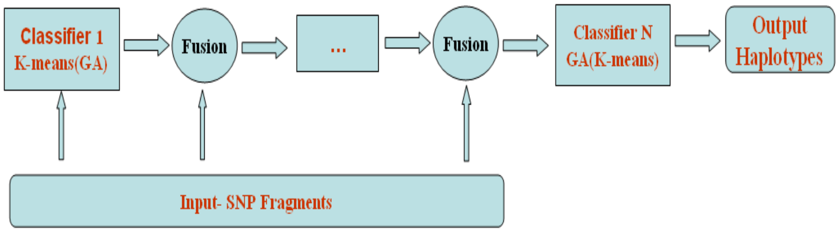


Figure 3. Serial classifier.

Table 3. Information fusion for parallel classifier.

SNP F.	1st classifier	2nd classifier	Parallel fusion function
1	Class 1	Class 1	Class 1
2	Class 1	Class 2	Eliminated
3	Class 2	Class 2	Class 2
-	-	-	-
-	-	-	-
-	-	-	-
n	Class 2	Class 1	Eliminated

For example, as can be seen in Table 3, 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 Algorithm 3 was designed. The results of this multiple parallel classifier were worse than serial ones due to probable elimination of some useful haplotypes. To modify the fusion function to avoid elimination of haplotypes, we must increase the number of different classifiers. It is predictable that if other classifiers be added to our MCS, a new fusion function be designed with three or more classifiers, we can improve the results. By doing this, a more number of haplotypes are kept and this can lead us to better results.

EXPERIMENTAL RESULTS

There are real biological datasets and also simulation datasets available for haplotyping problem like ACE, DALY, SIM0 and SIM50. We chose DALY dataset which includes 4 different subsets. Each subset has a different error rate (10, 20, 30 and 40%) and includes 384 different test cases. The results of the experiments on DALY set for MEC model, is shown in Table 1. In Figures 4a, b and c, the comparison of result methodology is focused in each model. Also we implemented all algorithms for MEC/GI model and the results are shown in Table 2 and methods are compared in Figure 4d, e and f.

Conclusion

In this paper, we focus on MCS (multiple classifier system) to solve MEC and MEC/GI model. The components which were used in our research were genetic and K-means algorithms. First, for MEC model, we designed four serial classifiers (GK, KG, KGK and GKG). Then one parallel classifier which is a combination of GA and K-means is designed. The information fusion function proposed for our MCS is described. Then all of the aforementioned methods are implemented and tested on MEC/GI model problem which are intended to infer haplotypes with high accuracy by employing genotype information. We compare the results of all methods in terms of RR. In both MEC and MEC/GI models, KGK and GK outperform the other approaches due to the fact that GA finds near optimal solutions in search space and K-means acts as a local heuristic classifier to find the real answer.

REFERENCES

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X,

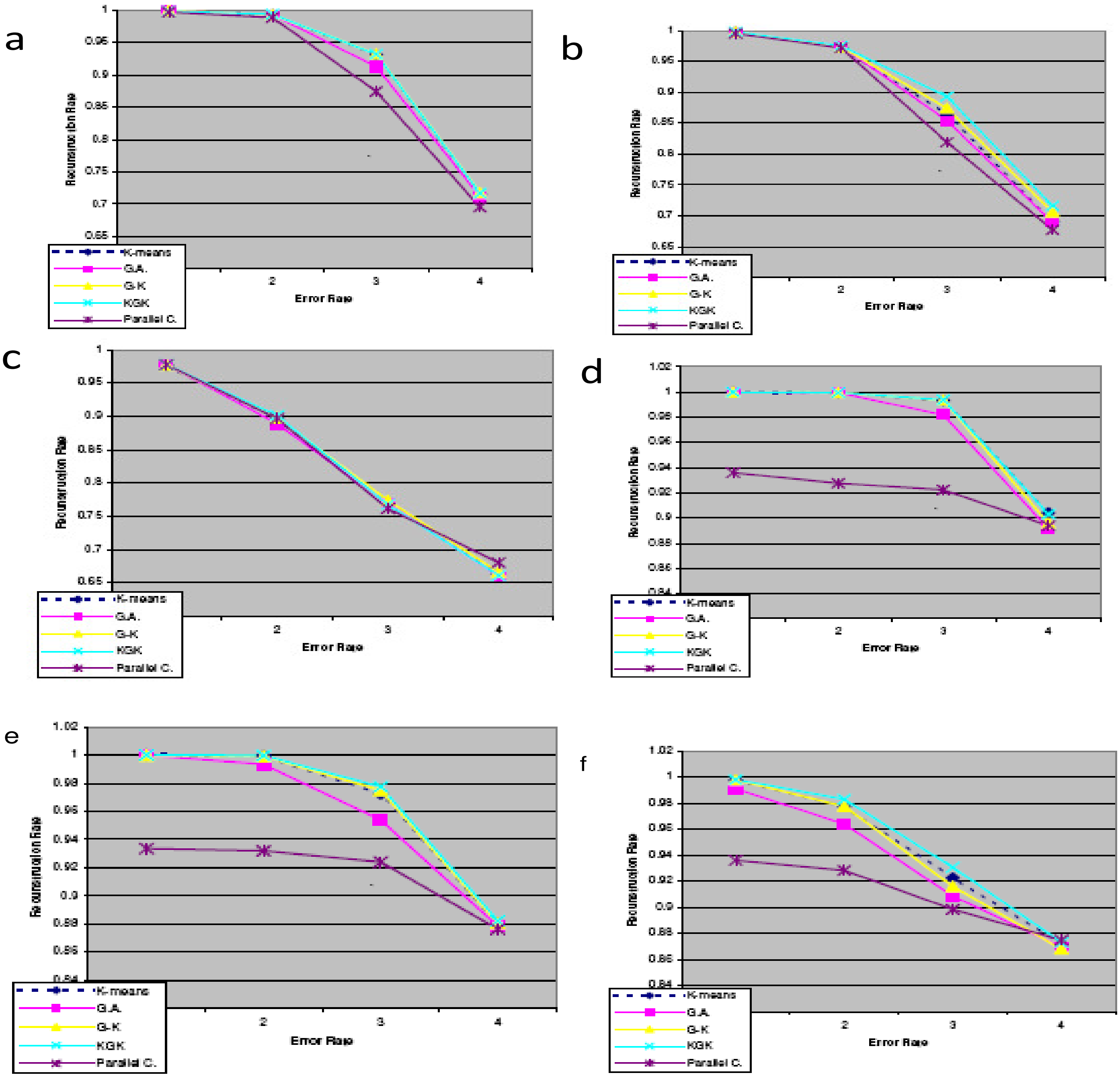


Figure 4. Comparison results of the algorithms using DALY database. a, b, c for MEC model and d, c, f for MEC/GI model . (a) DALY, g = 0.25; (b) DALY, g = 0.5; (c) DALY, g = 0.75; (d) DALY, g = 0.25; (e) DALY, g = 0.5; (f) DALY, g = 0.75.

Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001). The sequence of the human genome. Sci., 291(5507):1304–1351.

Terwilliger J, Weiss K (1998). Linkage disequilibrium mapping of complex disease: Fantasy and reality? Curr. Opin. Biotechnol., pp. 579–594.

Chakravarti A (1998). It's raining, hallelujah? Nat. Genet., 19: 216–217.

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

Zhang X, Wang R, Wu L, Zhang W (2006). Minimum conflict individual Haplotyping from SNP fragments and related Genotype. Bioinformatics, pp. 271-280.

Zhang X, Wang R, Wu L, Zhang W (2007). A clustering algorithm based on two distance functions for MEC model. Computational Bio. Chem., 148-150.

Moeinzadeh MH, Asgarian E, Mohammadzadeh J, Ghazinezhad A, Najafi-A A (2007). Three Heuristic Clustering Methods for Haplotype Reconstruction Problem with Genotype Information. The 4th IEEE International Conference on Innovations in Information Technology, pp. 404-406.

Bonizzoni P, Vedova GD, Dondi R, Li J (2003). The Haplotyping Problem: A review of Computational Models and Solutions. J. Comp. Sci. Tech., 18(6): 675-688.

Panconesi and Sozio M (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.

Greenberg HJ, Hart EW, Lancia G (2004). Opportunities for Combinatorial Optimization in Computational Biology. INFORMS J. Comp., 16(3): 211-231.

Rieder M, Taylor S, Clark A, Nickerson D (1999). Sequence variation n the human angiotensin converting enzyme. Nat. Genet., 22:59–62.

Wang Y, Feng E, Wang R, Zhang D (2007). The haplotype assembly model with genotype information and iterative local-exhaustive search algorithm. Comput. Bio. Chem., 31: 288–293.

Algorithm 1. Pseudo code serial KG.

Algorithm	Series k –means-GA for MEC (MEC/GI) with gap
Input	SNP fragments (Genotype)
Output	Two haplotypes
Step0	Initialize parameters
Step 1	Executing K –means for MEC (MEC/GI) to find two haplotypes as the class centre (c_1 and c_2).
Step2	Generating initial population for GA using k- means decision, making error them and random generation
Step3	Executing GA to find the two classes (c_1 and c_2).
Step4	Obtain two haplotypes from GA classes.

Algorithm 2. Pseudo code serial GK.

Alogorithm	Series GA k –means- for MEC (MEC/GI) with gap
Input	SNP fragments (Genotype)
Output	Two haplotypes
Step 0	Initialize parameters
Step 1	Executing GA for MEC (MEC/GI)to find the clusters (c_1 and c_2).
Step 2	Obtain to centers from the classes (c_1 and c_2).
Step 3	Set k-means intials centers (c_1 and c_2).
Step 4	Executing k- means for MEC (MEC/GI) to find two haplotypes as its centers

Algorithm 3. Pseudo code parallel classifier.

Algorithm	Series k –means-GA for MEC (MEC/GI) with gap
Input	SNP fragments (Genotype
Output	Two haplotypes
Step 0	Initialize parameters
Parallel steps	
Step 1	Executing GA for MEC (MEC/GI)to find the two classes (c1 and c2).
Step 1	Executing k-means for MEC (MEC/GI) to find the clusters fragments
Step 2	Voting on classifiers decision to eliminate noisy fragments
Step 3	Obtain two haplotypes from new classification