International Journal of Bio-Science and Bio-Technology
Vol.5, No.5 (2013), pp. 223-232
http://dx.doi.org/10.14257/ijbsbt.2013.5.5.23

Using Harmony Clustering for Haplotype Reconstruction from SNP fragments

Saman Poursiah Navi

Department of Computer Engineering, Islamic Azad University Quchan Branch, Quchan, Iran
samanpoursiah@gmail.com

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 differences. Haplotypes have more information for disease-associating than individual SNPs or genotypes; it is substantially more difficult to determine haplotypes through experiments. Hence, computational methods that can reduce the cost of determining haplotypes become attractive alternatives. MEC, as a standard model for haplotype reconstruction, is fed by fragments input to infer the best pair of haplotypes with minimum errors needing correction. It is proved that haplotype reconstruction in the MEC model is a NP-Hard problem. Thus, researchers' desire reduced running time and obtaining acceptable results. Heuristic algorithms and different clustering methods are employed to achieve these goals. In this paper, Harmony Search (HS) is considered a clustering approach. Extensive computational experiments indicate that the designed HS algorithm achieves a higher accuracy than the genetic algorithm (GA) or particle swarm optimization (PSO) to the MEC model in most cases.

Keywords: *Clustering, Bioinformatics, Evolutionary Optimization, Reconstruction Rate*

1. Introduction

Availability of the complete genome sequence for human beings makes it possible to investigate genetic differences and associate genetic variations with complex diseases [1]. It is generally accepted that all human beings share about 99% identity at the DNA level, with only some regions of differences in DNA sequences responsible for genetic diseases [4, 5]. 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 [16] and are found approximately every 1,000 base pairs in the human genome. They are promising tools for disease association studies. Every nucleotide in an 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 chromosome pair in a diploid genome is called a haplotype, which is a string over {'A', 'B'}. SNP fragments are composed of gaps and errors. One question arising from this discussion is how the distribution of gaps and errors in the input data affects computational complexity.

Some models discussed for haplotype reconstruction include Minimum Error Correction (MEC) [17], Longest Haplotype Reconstruction (LHR) [7], Minimum Error Correction with Genotype Information (MEC/GI) [12], and Minimum Conflict Individual Haplotyping (MCIH) [1]. Our research chose the standard MEC, a standard model for haplotype reconstruction that is fed by fragments as an input to infer the best pair of haplotypes with

ISSN: 2233-7849 IJBSBT

minimum error correction. For the MEC model, two different procedures can be employed to resolve the problem: First, partitioning 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 in this paper. The second approach is based on inferring haplotypes directly from SNP fragments and simultaneously correcting the errors.

It was proved that haplotype reconstruction in the MEC model is an NP-Hard problem [1]. Thus, researches desire reduced running time and obtaining acceptable results [18, 19].

A meta-heuristic algorithm, mimicking the improvisation process of music players, has been recently developed and named Harmony Search (HS) [10, 14]. In this paper, we propose an algorithm based on HS, for a haplotype reconstruction problem in a minimum-error-correction model. To demonstrate the effectiveness and speed of HS, we have applied HS algorithms on a standard SNP fragments database and received good results compared to GA and PSO. The evaluation of the HS experimental results showed considerable improvements and robustness.

In the next section, biological definitions such as SNP, SNP fragments, and haplotype are formulated. Next we introduce GA, PSO, and K-means as related works in the MEC model, two of these three approaches are considered as supplemental methods for our solution. In the next section, the proposed approach is discussed in detail. In this section, the HS algorithm (Harmony Search) and its properties, along with functions of the algorithm, are discussed for the MEC model. The final two sections are Results and Discussion regarding the different data-sets and Conclusion.

2. Formulations and Problem Definitions

Suppose there are m SNP fragments from a pair of haplotypes. Each SNP fragment (here after "fragment") corresponds to one of the two target haplotypes. $M=m_{ij}$ is defined as a matrix of fragments, of which each entry m_{ij} has a value ‘A’, ‘B’ or ‘-’ (‘-’ is a missing or skipped SNP site, which is called a "gap"). The rows and column of the matrix $M_{n\times m}$ 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 $P(C_1,C_2)$ (C_1 and C_2 are two classes) to formulate the problem. P is an exact algorithm or clustering method that divides fragments into C_1 and C_2 (Figure 1).

```
---010—00   Class1:
110111—00   Class2:
1-0-111—1   Class2:
00111011-   Class1:
00-0110-1   Class1:
00011-1--   Class2:
110011001   Class1:
--00-11--   Class2:
-1-0--1--   Class2:
-1011-110   Class2:
1111-0011   Class1:
```

Figure 1. Classifying SNP Fragments from M

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 to decide the values on the corresponding SNP site of related haplotypes. The function is so defined: ($N^i_A(M)$ (or $N^i_B(M)$) denoting the number of 'A's (or 'B's) in j^{th} 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 (*RR*) is a simple, popular means 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 ($h = (h_1, h2)$) and reconstructed ones ($h' = (h'_1, h'_2)$). The formula $d(x,y)$ is defined as the difference of two alleles in one SNP site. HD of two fragments $HD(f_i, f_j)$ and $RR(h, h')$ are formulated as:

$$d(h_{ij}, h_{kj}) = \begin{cases} +1 & (m_{ij} \ne m_{kj} \ne -) \\ 0 & Otherwise \end{cases}$$
$$HD(h_i, h_k) = \sum_{j=1}^{n} d(h_{ij}, h_{kj})$$
$$r_{ij} = HD(h_i, h'_j) \qquad i, j = \{1, 2\}$$
$$RR(h, h') = 1 - \frac{\min(r_{11} + r_{22}, r_{12} + r_{21})}{2n}$$

HD_1 and HD_2 are considered the two distances obtained from comparison of f_i and the two other fragments (f_1 and f_2).

3. Related Works

- ***Genetic Algorithm (GA)***

To resolve haplotype assembly, Wang and colleagues proposed a genetic algorithm to cluster the fragments [2]. The chromosomes are defined as binary string of length *m* (number of fragments). When Ch_i is equal to 0 (or 1), it means that the i^{th} 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 are no cluster's centers obtained yet. There is a fitness function recommended for evaluating the individuals by Wang and colleagues that computes the distance of all fragments with their class centers (the class centers in this problem are computed by the voting function method) [2].

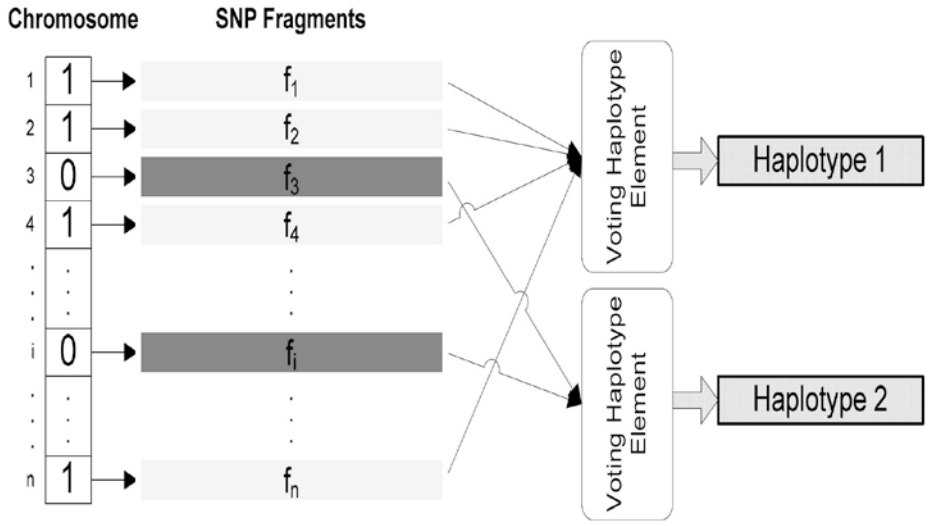


Figure 2. GA Chromosome and Inferring Haplotypes from One Partitioning

- ***Particle Swarm Optimization (PSO)***

The particle swarm optimization (PSO) method is much like GA. In this method, first a population of random solutions is generated and each of these solutions moves in the search space to get optimized. PSO has no evolutionary operators such as crossover and mutation, but the particles share their information of the visited areas and the best solutions met. Qian and colleagues used this method for haplotype reconstruction problem in a MEC model in which the particles are coded the same as the described genetic algorithm [6].

- ***Heuristic Method (K-means)***

A heuristic clustering method has been published by Wang and colleagues. First, two fragments are selected as the primitive centers. The other fragments are clustered according to their HD and the specified centers. In iterations, the centers are updated according to newly constructed clusters and voting function. Therefor 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.

4. The Proposed Framework

In this paper, we introduce Harmony Search to solve a MEC model. Pre-processing is used to make compatible input for the mentioned model. We stress the basic elements of this algorithm, as follows.

The Harmony Search (HS) algorithm was recently developed in an analogy with a music improvisation process whereby music players improvise the pitches of their instruments to obtain better harmony [11]. The steps in the Harmony Search procedure are as follows [11]:

Step 1. Initialize the problem and algorithm parameters.
Step 2. Initialize the harmony memory.
Step 3. Improvise a new harmony.
Step 4. Update the harmony memory.
Step 5. Check the stopping criterion.
These steps are described in the following subsections.

Step1: Initialize the Problem and Algorithm Parameters.

In Step 1, the optimization problem is specified as follows: In the light of the goal of the MEC model, the goodness and badness of an individual is dependent on the number of error corrections. Hence, we design the following objective function:

$$Max\ f\ (x_1,x_2,...,x_m)=\frac{m.n-E\left(P\{x_1,x_2,...,x_m\}\right)}{m.n},\ sb.\ to:x_i\in\{0,1\}$$

Where m is length of SNP fragment, n is number of SNP fragments, x_i is i^{th} SNP of current SNP fragment, P{x_1,x_2,…,x_m} is a partitioning of {x_1,x_2,…,x_m}, and E(P{x_1,x_2,…,x_m}) is the corresponding error correction in comparison with their own center, (i.e., the distance between center and each fragment).

Harmony memory (HM) is a memory location where all the solution vectors (sets of decision variables) are stored. HM is similar to the genetic pool in GA [13]. Here, harmony memory considering rate (HMCR) and pitch adjusting rate (PAR) are parameters used to improve the solution vector. Both are defined in Step.

Step2: Initialize the Harmony Memory

In Step 2, the HM matrix is filled as follows: The first half of harmony memory is generated randomly, and the rest of the harmony vector is produced by combining two fragments. For the second half, two different fragments are chosen from the list of fragments (random selection). These two fragments are considered as the centers of two classes. Then, the rest of the fragments are separated in two classes according to the hamming distance between the mentioned centers and the fragments:

$$h[center_1]=0$$

$$h[center_2]=1$$

$$h[i]=\begin{cases}0 & HD(M[i],M[Center_1])<HD(M[i],M[Center_2])\\ 1 & Otherwise\end{cases}$$

Where h[i] as one harmony_vector[i], HD as hamming distance, M[i] as i^{th} fragments and also center$_1$ and center$_2$ both as indexes of two fragments.

$$HM=\begin{bmatrix} x_1^1 & x_2^1 & \dots & x_{m-1}^1 & x_m^1\\ x_1^2 & x_2^2 & \dots & x_{m-1}^2 & x_m^2\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ x_1^{HMS-1} & x_2^{HMS-1} & \dots & x_{m-1}^{HMS-1} & x_m^{HMS-1}\\ x_1^{HMS} & x_2^{HMS} & \dots & x_{m-1}^{HMS} & x_m^{HMS}\end{bmatrix},$$

$$x_i^j\in\{0,1\},i=1,2,...,m,j=1,2,...,HMS$$

For construction of the hypothesis space, we use a binary string of {0, 1} to express a classification of SNP fragments (a feasible solution to the MEC model). The HMS is a number of solution vectors in HM, the length of the hypothesis space (m) is number of SNP fragments, and the value 0 or 1 on i^{th} site denotes i^{th} fragment's class-membership. For

example, if there are eight SNP fragments, a binary string of {10011100} denotes a partition: 1,4,5,6 are in a class and (left) 2,3,7,8 in another class. Thus, all binary strings with the length of *m* constitute the hypothesis space.

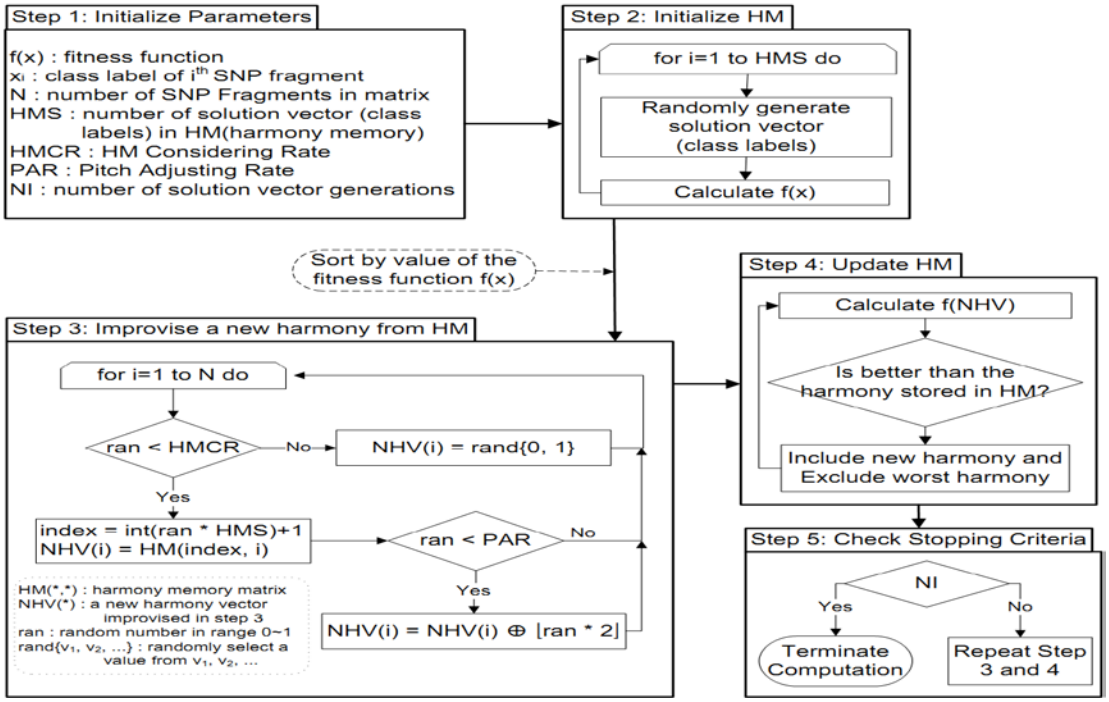


Figure 3. Harmony Search Approach

Step 3. Improvise a new harmony.

A new harmony vector, $x' = (x'_1, x'_2, ..., x'_m)$ is generated based on three rules: (1) memory consideration, (2) pitch adjustment and (3) random selection. Generating a new harmony is called "improvisation" [10].

In the memory consideration, the value of the first decision variable (x') for the new vector is chosen from any of the values in the specified HM range ($x'^1_1 - x'^{HMS}_1$). Values of the other decision variables $(x'_2, ..., x'_m)$ are chosen in the same manner. The HMCR, which varies between 0 and 1, is the rate of choosing one value from the historical values stored in the HM, while (1 – HMCR) is the rate of randomly selecting one value from the possible range of values.

$$x'_i \leftarrow \begin{cases} x'_i \in \{x_i^1, x_i^2, ..., x_i^{HMS}\} & \text{with probability} \quad HMCR, \\ x'_i \in X_i & \text{with probability} \quad (1 - HMCR). \end{cases}$$

For example, a HMCR of 0.85 indicates that the HS algorithm will choose the decision variable value from historically stored values in the HM with 90% probability or from the entire possible range with a (100–90) % probability. Every component obtained by the memory consideration is examined to determine whether it should be pitch-adjusted.

This operation uses the PAR parameter, which is the rate of pitch adjustment as follows:

$$x_i' \leftarrow \begin{cases} Yes & \text{with probability} \quad PAR, \\ No & \text{with probability} \quad (1-PAR). \end{cases}$$

The value of (1 – PAR) sets the rate of doing nothing. If the pitch adjustment decision for x' is YES, x' is replaced as follows:

$$x_i' \leftarrow x_i' \oplus \lfloor \text{rand}() * 2 \rfloor$$

Where rand() is a random number between 0 and 1. In Step 3, HM consideration, pitch adjustment or random selection are applied to each variable of the new harmony vector in turn.

Step 4. Update harmony memory.

If the new harmony vector, is better than the worst harmony in the HM, judged in terms of the objective function value, the new harmony is included in the HM and the existing worst harmony is excluded from the HM.

Step 5. Check stopping criterion.

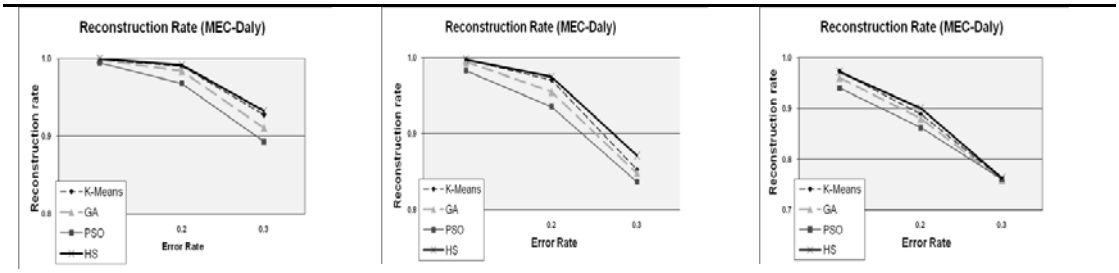
If the stopping criterion (maximum number of improvisations) is satisfied, computation is terminated. Otherwise, Steps 3 and 4 are repeated.

5. Results and Discussion

There are some simulation and real biological datasets available for haplotype reconstruction problems. In this paper, Daly, ACE, SIM0, and SIM50 were chosen. Our approaches were implemented using Visual C#.Net 4.0 and executed on all the datasets. All datasets have 12 different gap and error rates (Error Rate = 0.1, 0.2, 0.3 and 0.4 and Gap Rate = 0.25, 0.50 and 0.75).

- ***Simulation Datasets (SIM0 and SIM50)***

These two datasets are generated according to the similarity of the result haplotypes (or the percentage of heterozygous site in genotype). There is no similarity between the two obtained haplotypes in SIM0 datasets. Therefore, all positions are considered as heterozygous sites. In this dataset there are 30 test cases of 20 fragments with 50 SNP site lengths.



(a) g = 0.25 (b) g = 0.5 (c) g = 0.75

Figure 4. Comparison the Results Different Clustering Approaches for MEC Model on Daly Dataset

Table 1. Reconstruction Rate on Daly, ACE, SIM0 and SIM50 Datasets for Different Gap and Error Rate (MEC Model)

Gap Rate	Error Rate	Daly Database - MEC				ACE Database - MEC				SIM_0 Database - MEC				SIM_50 Database - MEC			
		K-Means	GA	PSO	HS	K-Means	GA	PSO	HS	K-Means	GA	PSO	HS	K-Means	GA	PSO	HS
0.25	0.1	0.999	0.999	0.995	1.000	0.998	0.998	0.987	0.998	0.999	0.987	0.979	0.996	0.996	0.998	0.985	0.999
	0.2	0.990	0.984	0.968	0.991	0.952	0.942	0.904	0.944	0.952	0.948	0.909	0.965	0.965	0.952	0.938	0.960
	0.3	0.928	0.911	0.893	0.930	0.814	0.840	0.820	0.846	0.817	0.862	0.803	0.870	0.849	0.820	0.800	0.835
	0.4	0.721	0.722	0.725	0.720	0.650	0.669	0.687	0.683	0.651	0.580	0.590	0.609	0.618	0.655	0.637	0.650
0.5	0.1	0.997	0.994	0.983	0.998	0.977	0.976	0.969	0.977	0.977	0.976	0.941	0.988	0.988	0.966	0.947	0.979
	0.2	0.971	0.955	0.936	0.974	0.898	0.901	0.876	0.900	0.910	0.901	0.863	0.920	0.919	0.888	0.878	0.924
	0.3	0.853	0.849	0.837	0.861	0.755	0.763	0.743	0.786	0.768	0.775	0.745	0.797	0.752	0.742	0.763	0.802
	0.4	0.693	0.688	0.695	0.694	0.637	0.633	0.659	0.637	0.640	0.553	0.595	0.555	0.556	0.624	0.618	0.643
0.75	0.1	0.973	0.960	0.940	0.979	0.887	0.892	0.898	0.913	0.908	0.882	0.863	0.914	0.912	0.915	0.886	0.913
	0.2	0.888	0.880	0.862	0.891	0.738	0.791	0.779	0.793	0.744	0.783	0.704	0.803	0.693	0.808	0.753	0.799
	0.3	0.762	0.759	0.761	0.765	0.680	0.673	0.673	0.672	0.675	0.610	0.607	0.657	0.628	0.658	0.674	0.670
	0.4	0.657	0.652	0.662	0.636	0.624	0.637	0.621	0.625	0.606	0.556	0.560	0.539	0.543	0.608	0.598	0.609

- ***Daly and ACE Datasets***

Daly dataset includes 383 different test cases for each error rate (1532 for all error rates). Each test case consists of 40 fragments of 53 SNP sites. The experimental results of new (HS) and previous (K-means, GA and PSO) approaches for the MEC model are shown in Figure 4, A-C. These diagrams are the reconstruction rate comparisons of K-means, GA, UWNN, and GKM approaches in Daly datasets. ACE (Angiotensin Converting Enzyme) as real dataset, includes 24 different test cases for each error rate.

6. Conclusions

In this paper, two new approaches were proposed to solve the haplotype reconstruction problem in the MEC model. HS was used to cluster data of our problem. Other improvements were performed in different parts of HS. In this approach, HS was used to cover almost all solution space and improve the accuracy of the solutions. The results of the Harmony Search (HS) implementation were obtained from different real and simulation datasets (Daly, ACE, SIM0, and SIM50). It was proved by experiences that the proposed methods outperform all previous related works.

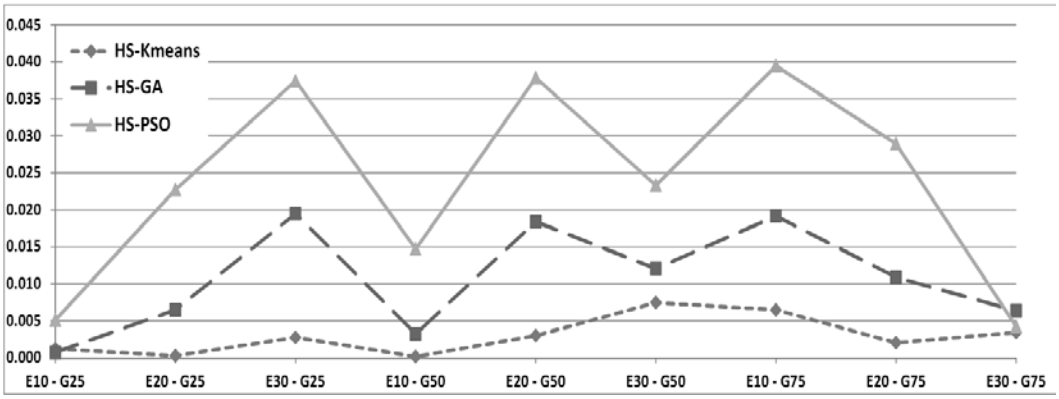


Figure 5. Reconstruction Rate Subtraction (Improvement of HS from K-means, GA and PSO) on Daly Dataset

References

[1] X. Zhang, R. Wang, L. Wu and W. Zhang, "Minimum conflict individual Haplotyping from SNP fragments and related Genotype", Bioinformatics Oxford Journal, **(2006)**, pp. 271-280.
[2] Y. Wang, E. Feng and R. Wang, "A clustering algorithm based on two distance functions for MEC model", Computational Biology and Chemistry, vol. 31, no. 2, **(2007)**, pp. 148-150.
[3] R.-S. Wang, L.-Y. Wu, Z.-P. Li and X.-S. Zhang, "Haplotype reconstruction from SNP fragments by Minimum Error Correction", Bioinformatics, vol. 21, no. 10, **(2005)**, pp. 2456-2462.
[4] J. C. Venter and M. D. Adams, "The sequence of the human genome", Science, vol. 291, no. 5507, **(2001)**, pp. 1304-1351.
[5] J. Terwilliger and K. Weiss, "Linkage disquilibrium mapping of complex disease: Fantasy and reality?", Current Opinion in Biotechnology, vol. 9, no. 6, **(1998)**, pp. 578-594.
[6] W. Qian, Y. Yang, N. Yang and C. Li, "Particle swarm optimization for SNP haplotype reconstruction problem", Applied Mathematics and Computation, vol. 196, **(2007)**, pp. 266-272.
[7] M. Sozio Panconesi, "Fast Hare: A Fast Heuristic for Single Individual SNP Haplotype Reconstruction", Algorithms in Bioinformatics, **(2004)**, pp. 266-277.
[8] M. H. Moeinzadeh, E. Asgarian, A. Najafi-Ardabi, S. Sharifian-R, M. Sheikhaei and J. Mohammadzad, "Three Heuristic Clustering Methods for Haplotype Reconstruction Problem with Genotype Information", Innovations in Information Technology, **(2007)**, pp. 402 - 406.
[9] M. H. Moeinzadeh, E. Asgarian, S. Sharifian-R, A. Najafi-Ardabili and J. Mohammadzadeh, "Neural Network Based Approaches, Solving Haplotype Reconstruction in MEC and MEC/GI Models", Second Asia International Conference on Modelling and Simulation, **(2008)**, pp. 934-939.
[10] M. Mahdavi, M. Fesanghary and E. Damangir, "An improved harmony search algorithm for solving optimization problems", Applied Mathematics and Computation, vol. 188, no. 2, **(2007)**, pp. 1567-1579.
[11] K. S. Lee and Z. W. Geem, "A new meta-heuristic algorithm for continues engineering optimization: harmony search theory and practice", Computer Methods in Applied Mechanics and Engineering, vol. 194, **(2005)**, pp. 3902-3933.
[12] H. J. Greenberg, W. E. Hart and G. Lancia, "Opportunities for Combinatorial Optimization in Computational Biology", INFORMS Journal on Computing, vol. 16, no. 3, **(2004)**, pp. 211-231.
[13] Z. W. Geem, J. H. Kim and G. V. Loganathan, "Harmony search optimization: application to pipe network design", International Journal of Modelling and Simulation, vol. 22, no. 2, **(2002)**, pp. 125-133.
[14] Z. W. Geem, C. Tseng and Y. Park, "Harmony search for generalized orienteering problem: best touring in China", Lect Notes Comput Sci, **(2005)**, pp. 741-750.
[15] C. A. C. Coello, "Constraint-Handling using an Evolutionary Multi-objective Optimization Technique", Civil Engineering and Environmental Systems, vol. 17, **(2000)**, pp. 319-346.
[16] Chakravarti, "It's raining SNPs, hallelujah?", Nature Genetics, vol. 19, **(1998)**, pp. 216-217.
[17] P. Bonizzoni, G. D. Vedova, R. Dondi and J. Li, "The Haplotyping problem: An overview of computational models and solutions", Journal of Computer Science and Technology, vol. 18, no. 6, **(2003)**, pp. 675-688.
[18] E. Asgarian, M. H. Moeinzadeh, S. Sharifian-R, A. Najafi-A, A. Ramezani and J. Habibi, "Solving MEC model of haplotype reconstruction using information fusion, single greedy and parallel clustering approaches", International Conference on Computer Systems and Applications, **(2008)**, pp. 15-19.
[19] E. Asgarian, M. H. Moeinzadeh, J. Mohammadzadeh, A. Ghazinezhad and J. Habibi, "Solving MEC and MEC/GI Problem Models, Using Information Fusion and Multiple Classifiers", Innovations in Information Technology, **(2007)**; pp. 397-401.
[20] E. Asgarian, M. H. Moeinzadeh, A. Rasooli and S. Moaven, "Solving Haplotype Reconstruction Problem in MEC Model with Hybrid Information Fusion", Australian Journal of Basic and Applied Sciences, vol. 3, no. 1, **(2009)**, pp. 277-282.
[21] H. Bohnenkamp, H. Hermanns, I. P. Katoen and R. Klaren, "The MoDeST Modeling Tool and its implementation", Proc. of the Computer Performance Evaluation Modelling Techniques and Tools (TOOLS'03), Lecture Notes in Computer Science, Springer-Verlag, vol. 2794, **(2003)**, pp. 116-133.

Author

Saman Poursiah Navi is currently Senior Lecturer at Computer Engineering Department of Islamic Azad University Quchan Branch. He received his M.S. degree in computer engineering from Iran University of Science and Technology in 2007.