Sinusoidal Map Based Particle Swarm Optimization Detect the SNP Barcode in Breast Cancer to Disease Susceptibility

Li-Yeh Chuang[1], Cheng-Han Wu[2], Yu-Da Lin[3], Cheng-Hong Yang[4.]*

[1]Department of Chemical Engineering, I-Shou University, Kaohsiung, Taiwan

[2]Department of Computer Science and Information Engineering, National Kaohsiung
University of Applied Sciences, Kaohsiung, Taiwan

[3]Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan

[4]Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
[*]*Email: chyang {at} cc.kuas.edu.tw*

ABSTRACT---- ***Single nucleotide polymorphisms (SNPs) are the most common type of DNA sequence variation in the human genome and are widely used to investigate the association analysis of diseases. SNP barcode is a combination of SNPs with genotypes (AA, Aa, and aa for an SNP) to find the difference between case data set and control data set for analyzing the disease association amongst SNPs. Currently, the computational time of statistical method becomes the weak to analyze the big data to find the significant SNP barcode. Here, we applied a sinusoidal particle swarm optimization (SPSO) algorithm facilitate the statistical methods to analyze the associated SNPs. We systematically evaluated the synergistic effect of 26 SNPs from eight epigenetic modifier-related genes in breast cancer. The 2- to 5-order SNP barcodes were found to determine the risk effects in breast cancer. We found that five of eight genes (BAT8, DNMT3A, EHMT1, DNMT3A, and BAT8) were statistically significant to breast cancer and play the important role in the SNP barcode. In addition, we compared the search ability between PSO and SPSO in the 2- to 5-order SNP barcodes. The results indicated that SPSO can find the better SNP barcode than PSO. In conclusion, SPSO is a precise algorithm for finding a significant model of SNP barcode.***

Keyword--- Sinusoidal map, Particle Swarm Optimization, SNP barcode

1. INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most common type of DNA sequence variation in the human genome [1]. SNPs have been widely used in genome-wide association studies (GWAS) to analyze the genetic susceptibility to disease [2-4]. The combination of SNPs can identify the potential associations in which top associated SNPs with disease may influence the genetic variation. Therefore, the selection of SNPs from gene sequences becomes the necessary for recognizing whether SNPs has marginally significant associations with disease or not.

SNP barcode is a combination of SNPs with genotypes which include three genotypes AA, Aa, and aa. SNP barcode is able to find the difference between cases and controls for identifying the disease association amongst SNPs. The massive combinations amongst all possible SNP barcodes make the computation difficult by traditional statistical analysis. Many computational approaches have been proposed to examine epistasis in family-based and case-control association studies [5-6]. However, currently the presented methods are still not able to simultaneously evaluate the complex interactions between all SNPs within several genes. The algorithms for reducing the number of search items among SNP combinations (e.g., PSO and GA) do not ensure that the predicted model has statistical significant and high statistical power.

Based on Pharoah's study, the 26 SNPs from eight growth factors-related genes (EGF, IGF1, IGF1R, IGF2, IGFBP3, IL10, TGFB1, and VEGF) were selected to detect their association with breast cancer [7]. In this study, the SNP barcode association analysis for disease susceptibility were computed by our proposed improved algorithm and provided further insight into genetic susceptibility to breast cancer.

We proposed a method called Sinusoidal map based particle swarm optimization (SPSO) to generate the SNP barcode and analyze the risk of disease susceptibility. The highest associated SNP barcode can be evaluated by the odds ratio (*OR*) and 95% confidence intervals (95% CI). We systematically evaluated the synergistic effect of 26 SNPs from eight epigenetic modifier-related genes in breast cancer. The results revealed that our prposed method effectively provided the statistically significant SNP barcode, and this barcode possessed better differences value than the original PSO algorithm.

2. METHODS

2.1 Particle swarm optimization

Particle swarm optimization (PSO) was proposed by Kennedy and Eberhart [8]. PSO belongs to a heuristic algorithm based on swarm intelligence to search the optimal resolution on a complex problem. Each individual (particle) in PSO is represented a possible solution of the problem. Swarm intelligence simulates social behavior for information sharing; it makes that particles can save its previous experience and shares the common knowledge in population. Thus, a significant knowledge in swarm can be exchanged to offer a direction and leads particles toward a good solution.

Similar to other evolutionary algorithm, PSO has a fitness function for evaluating the particle value. Thus, the fitness value can lead the convergence of particles towards a good solution through the particles adjusting their positions. Finally, the optimal solution of the problem can be found in the particles. The procedure steps of PSO are shown in the description below. First, each individual's position and velocity are randomly generated. Second, the fitness of each particle is evaluated by the fitness function. Third, current fitness for each particle is compared to the higher fitness value which is called $pbest_i$. Fourth, the common knowledge (global best value, *gbest*) was updated according to the best *pbest* amongst the population. Finally, the position of each particle was updated by velocity and position updating equations and they are formulated as following:

$$v_{id}^{new} = w \times v_{id}^{old} + c_1 \times r_1 \times (pbest_{id} - x_{id}^{old}) + c_2 \times r_2 \times (gbest_{id} - x_{id}^{old}) \tag{1}$$

$$x_{id}^{new} = x_{id}^{old} + v_{id}^{new} \tag{2}$$

r_1 and r_2 are random numbers between [0, 1], c_1 and c_2 control how far a particle move in a generation, v_{id}^{new} and v_{id}^{old} denoted the new and old velocities in i^{th} particle. x_{id}^{new} and x_{id}^{old} denoted the updated and old position in i^{th} particle. The velocity controls particle's movement in a generation. Particles can accelerate toward the best solution by moves. Velocities in each dimension are limited to within $[V_{\min,} V_{\max}]^D$, and particle positions are limited within $[X_{\min,} X_{\max}]^D$.

2.2 Sinusoidal particle swarm optimization (SPSO)

Sinusoidal map based particle swarm optimization used the chaotic variable instead of random variables. Sinusoidal map can carry out overall searches at higher speeds than stochastic searches due to the non-repetition and ergodicity of chaos. In this study, we used the sinusoidal map to improve the PSO for avoiding the local optima.

The random value r_1 and r_2 in velocity equation are important parameters that influence the particle towards the local optima or global optimal. Sinusoidal chaotic map generates chaotic sequence that is a good substitute for random value, and it does not store long random sequence. The random value r_1 and r_2 are modified on the equation below:

$$sr_{1i+1} = \sin(\pi sr_{1i}) \tag{3}$$

$$sr_{2i+1} = \sin(\pi sr_{2i}) \tag{4}$$

In the first generation, sr_1 and sr_2 are random values between [0, 1]. Value sr_{1i+1} and sr_{1i+1} denoted two chaotic values which replace random value r_1 and r_2 in velocity updating equation (Eq. 1) in i^{th} particle of PSO. The velocity updating equation for PSO based on the Sinusoidal chaotic map can be formulated as following:

$$v_{id}^{new} = w \times v_{id}^{old} + c_1 \times sr_{1i} \times (pbest_{id} - x_{id}^{old}) + c_2 \times sr_{2i} \times (gbest_{id} - x_{id}^{old}) \tag{5}$$

2.3 Application of the Sinusoidal PSO algorithm

a) Encoding population

In SPSO, a particle contains the number of selected SNPs and their corresponding genotypes, in which SNP selection cannot be repeated. A particle encoding is represented as:

$$Particle_i = (SNP_1, SNP_2, \ldots, SNP_j, Genotypes_1, Genotypes_2, \ldots, Genotypes_j) \quad (6)$$

where SNP is represented the selected SNP, Genotype is represented the three SNP genotypes (AA, Aa, aa), *i* is the size of the population, *j* is the number of SNP selected. For example, let 3-order SNP barcode = (4, 5, 20, 1, 2, 1), it means SNP_4 is selected and its genotype is AA, SNP_5 is selected and its genotype is Aa, and SNP_{20} is selected and its genotype is AA.

b) Fitness evaluation

Fitness function is used to evaluate the difference value between the case and control data sets from the SNP barcode. The fitness value was calculated by the difference value between the total number of SNP barcode in the control data and the total number of SNP barcode in the case data. The fitness function is formulated as following:

$$F(particle_i)=\text{number}(control\cap particle_i)-\text{number}(case\cap particle_i) \quad (7)$$

where number($control \cap particle_i$) represents the total number of SNP in control data which have the case, such as $particle_i$. number($case\cap particle_i$) represents the total number of SNP in case data which have the case, such as $particle_i$.

c) Updating particle's experience and common knowledge

At each generation, every particle moves its position by computing the velocity to the current iteration. Particles can be adjusted by the $pbest_i$ and *gbest* in the velocity equation Eq. (5). The value of $pbest_i$ is computed by comparing each particle its current fitness value with its fitness value of *pbest*. If the current fitness of particle was better than its $pbest_i$, then position and fitness value of $pbest_i$ are updated to $particle_i$. Similar updating *pbest*, the *gbest* is updated by comparing all *pbest*s.

d) Statistical value

The odds ratio and *p*-value are widely used in epidemiology study for the criterion of the performance [1]. The odds ratio (*OR*) can be used to determine the best SNP barcode and quantitative measurement of the risk of disease; the *p*-value is used to demonstrate if the results are statistically significant for the difference between case and control group.

3. RESULTS AND DISCUSSION

3.1 Parameter settings

In this study, the population size is set to 50 and the iteration is set to 100. The value of the inertia weight *w* is set in the range between 0.9 and 0.4 [9]. Both c_1 and c_2 are set to 2 [10].

3.2 Data sets

The breast cancer data sets used in this study were proposed by Pharosh et al.[7]. The data sets consist of the epigenetic modifiers-related genes, included the genes of BAT8, DNMT1, DNMT3A, DNMT3B, EHMT1, HDAC2, MBD2, and SETDB1 with 27 SNPs. A total of 5000 sample size was simulated and normalized in the case data sets and the control data sets, in which the normalized data sets were randomly generated according the reported frequency of SNPs.

3.3 Identification of the best SNP barcode model with maximum difference value

The combinations of 2- to 5-order SNP barcode are shown in Tables 1 and 2. The PSO method and our proposed method (SPSO) were compared on the difference value between the case and control data sets of the predicted SNP barcode. The difference value of the 2-order SNP barcode, SNPs (10, 17) with genotype 2-1, showed 130 by SPSO method, and PSO also provided the same difference value. However, for the 3- to 5-order SNP barcode, SPSO provided better difference value than PSO. The difference value of 3- to 5-order SNP barcodes indicated that SPSO is robust approach to identify SNP barcode in high order interaction.

3.4 The odds ratios and its 95% CI analysis of effects of SNP barcode on breast cancer

The statistical values provided the evidence of the detection results. As shown in Tables 1 and 2, the statistical values (e.g., Odds Ratio (*OR*) and its 95% CI, *p*-value) present the estimation of the specific SNP barcode on the occurrence of breast cancer. If the *OR* value is bigger than 1, it indicates a higher risk association between the SNP barcode and the disease. Compared to both methods, SPSO provided greater *OR* value (1.103-1.823) than that of PSO (1.103-1.097). The results suggest that these genes with the combinations of SNPs and genotypes have the risk to breast cancer. For the analysis of *p*-value, SPSO revealed the good *p*-value ($P < 0.05$) in 2-order to 5-order SNP barcode, but PSO only identified a good *p*-value in 2-order SNP barcode. It indicated SPSO can provide a significant association (SNP barocde) amongst the SNPs.

3.5 Discussion

Detection of multi-SNP barcode association on disease susceptibility is widely used in genome-wide of case-control association studies. The massive computations of SNP barcode analysis remain a challenge, especially for the high-order SNP barcode. In this study, we proposed a novel method, SPSO, to identify the significant association of SNP barcode between case data set and control data set. The SPSO performed an outstanding result on the association detection. According to the statistical values (e.g., Odds Ratio (*OR*), *p-value* and 95% CI) obtained, the results demonstrated that SPSO provided higher reliability and a stronger ability thanPSO. The SPSO can identify the best difference value between case and control data. It also can be applied to detect the multi-SNP barcode amongst the huge quantitative of SNPs involved in GWAS.

Table 1. Estimation of the best SNP barcode model on the occurrence of breast cancer as determined by SPSO

Combined SNP	**SNP genotypes**	**Cases No.**	**Controls No.**	**Difference**	**Odds Ratio**	**95% CI**	***p*-value**
SNPs(10,17)	2-1	1245	1156	89	1.103	1.01-1.21	0.004
	other	3755	3844				
SNPs(7,11,21)	3-2-1	93	57	36	1.644	1.18-2.29	0.004
	other	4907	4943				
SNPs(1,7,11,21)	1-3-2-1	87	53	34	1.653	1.17-2.33	0.005
	other	4913	4947				
SNPs(1,2,7,11,21)	1-2-1-1-1	49	27	22	1.823	1.14-2.92	0.011
	other	4951	4973				

Table 2. Estimation of the best SNP barcode model on the occurrence of breast cancer as determined by PSO

Combined SNP	**SNP genotypes**	**Cases No.**	**Controls No.**	**Difference**	**Odds Ratio**	**95% CI**	***p*-value**
SNPs(10,17)	2-1	1245	1156	89	1.103	1.01-1.21	0.004
	other	3755	3844				
SNPs(1,10,11)	1-1-2	1096	1020	76	1.095	1.00-1.21	0.063
	other	3904	3980				
SNPs(1,10,11,26)	1-1-2-1	824	763	61	1.096	0.98-1.22	0.095
	other	4176	4237				
SNPs(1,10,11,19,26)	1-1-2-1-1	500	460	40	1.097	0.96-0.1.25	0.186
	other	4500	4540				

4. ACKNOWLEDGMENT

This work is partly supported by the National Science Counci in Taiwan under grants 101-2622-E-151-027-CC3,100-2221-E-151-049-MY3, and 100-2221-E-151-051-MY2.

5. REFERENCE

[1] L. E. Mechanic, B. T. Luke, J. E. Goodman, S. J. Chanock, and C. C. Harris, "Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions," *BMC bioinformatics,* 2008, pp. 146-146.

[2] P. Kraft and C. A. Haiman, "GWAS identifies a common breast cancer risk allele among BRCA1 carriers," *Nature genetics,* vol. 42, 2010.

[3] J.-C. Yu, C.-N. Hsiung, H.-M. Hsu, B.-Y. Bao, S.-T. Chen, G.-C. Hsu, W.-C. Chou, L.-Y. Hu, S.-L. Ding, and C.-W. Cheng, "Genetic variation in the genome-wide predicted estrogen response element-related sequences is associated with breast cancer development," *Breast Cancer Res,* 2011, pp. R13-R13.

[4] X. Li, H. Chen, J. Li, and Z. Zhang, "Gene function prediction with gene interaction networks: a context graph kernel approach," *Information Technology in Biomedicine, IEEE Transactions on,* 2010, pp. 119-128.

[5] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics,* 2010, pp. 445-455.

[6] C.-H. Yang, L.-Y. Chuang, Y.-J. Chen, H.-F. Tseng, and H.-W. Chang, "Computational analysis of simulated SNP interactions between 26 growth factor-related genes in a breast cancer association study," *Omics: a journal of integrative biology,* 2011, pp. 399-407.

[7] P. D. Pharoah, J. Tyrer, A. M. Dunning, D. F. Easton, B. A. Ponder, and S. Investigators, "Association between common variation in 120 candidate genes and breast cancer risk," *PLoS genetics,* 2007, pp. e42-e42.

[8] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of IEEE international conference on neural networks*, 1995, pp. 1942-1948.

[9] Y. Shi and R. C. Eberhart, "Empirical study of particle swarm optimization," in *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, 1999.

[10] A. Ratnaweera, S. Halgamuge, and H. C. Watson, "Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients," *Evolutionary Computation, IEEE Transactions on*, 2004, pp. 240-255.