Determination of the SNP-SNP Interaction between Breast Cancer Related Genes to Analyze the Disease Susceptibility

Mei-Lee Hwang, Yu-Da Lin, Li-Yeh Chuang, and Cheng-Hong Yang

Abstract—Investigation of the single nucleotide polymorphism (SNP)-SNP interaction model can facilitate the analysis of the susceptibility to disease. The model explains the risk of association between the genotypes and the disease in case-control study. Thus, many mathematic methods are widely applied to identify the statistically significant model such as odds ratio (OR), chi-square test, and error rate. However, a huge number of data sets have been found to limit the statistical methods to identify the significant model. In this study, we propose a novel statistical method, complementary-logic particle swarm optimization (CLPSO), to increase the efficiency of significant model identification in case-control study. The complementary-logic is implemented to improve the PSO search ability and identify a better SNP-SNP interaction model. Six important breast cancer genes including 23 SNPs and simulated huge number of data sets were selected as the test data sets. The methods of PSO and CLPSO were applied on the identification of SNP-SNP interactions in the two-way to five-way. In results, the OR evaluates the breast cancer risk of the identified SNP-SNP interaction model. Compared to the corresponding non-interaction model, if the OR value is greater than 1 that indicates the model is significant risk between cases and controls. The results showed that CLPSO is able to identify the significant models for specific SNP-SNP interaction of two-way to five-way (OR value: 1.153-1.391; confidence interval (CI): 1.05-1.79; p-value: 0.01-0.003). The model suggests that the genes ESR1, PGR, and SHBG may be an important role in the interactive effects to breast cancer. In addition, we compared the search abilities of PSO and CLPSO for identification of the significant model. Results revealed that CLPSO can identify better model with difference values between cases and controls than PSO; it suggests CLPSO can be used to identify a better SNP-SNP interaction models.

Index Terms—Single nucleotide polymorphism (SNP), particle swarm optimization (PSO), breast cancer.

I. INTRODUCTION

SNP is an common bio-marker in genomes, and it has widely used in the investigation of association analysis of diseases [1], cancers [2], and pharmacogenomics [3]. These analyses reported that SNPs have specific associations with the risk of certain diseases. However, most of the SNP analyses were focused on a single SNP. The low or marginal significant SNPs could be excluded, but these SNPs may have significant associations with disease when they are combined as SNP-SNP interaction model. Thus, the identification of appropriate interaction model is an important issue for SNP analysis.

A SNP-SNP interaction model includes the SNPs and their corresponding genotypes (AA, Aa, and aa). Therefore, the possible models are rapidly increased by the number of SNPs and their corresponding genotypes. The huge potential SNP-SNP interaction model makes the statistical method difficulty identify the significant models. Currently, machine learning have applied to help statistical method on the identification of appropriate interaction models, such as particle swarm optimization (PSO) [4]-[6] and genetic algorithm (GA) [7], [8]. PSO and GA have the properties of randomized search and are an optimization technique that derives its working principles from simulations of the organism behavior. They provide a fast identification in high-dimension problem, e.g., biomarker selection [9]. Thus, they can be employed to identify an optimal SNP combination from the huge possible combinations. Although they had overcome the excessive computational time to identify the significant models, the search ability remains a challenge. Thus, an improved method is required to explore the SNP-SNP interaction.

Here, we propose a complementary-logic PSO, named CLPSO, to identify the significant model associated with SNP-SNP interactions. In this study, we hypothesize that the interactions between polymorphisms of genes may have a synergistic or non-additive effect on the pathogenesis of a disease. This interaction may explain differences between cases and controls in the disease risk. Six breast cancer related genes (COMT, CYP19A1, ESR1, PGR, SHBG, and STS) including 23 SNPs were selected to simulate huge number of data sets. Results indicate that CLPSO can identify the appropriate interaction models in breast cancer from the huge number of simulated data sets, and the results provide the significant information for determining the SNP-SNP interaction model with maximal difference between the cases and controls.

II. METHODS

A. Problem Definition

For SNP-SNP interaction problem, a vector like $X = [x_1, x_2$, ..., $x_n]$.
[47x707]i.e., vector objective function defined the difference between frequencies existence of non-linear objective functions with multiple local interaction models are refereed as the different SNP each parameter $x$ birds in a flock or fish in a school. Each individual (named technique; it simulates the social behavior of organisms as particle in the search space until a predefined number of conditions.

The complementary-logic aims to avoid the particles to be easily trapped in a local optimum by moving their position to a new region in the problem space. In traditional PSO, particles could be trapped in a local optimum due to the premature convergence of population. All particles are randomly generated in the problem space and updated by the $pbest$, which is the best fitness value in the particle itself has achieved so far, and $gbest$, which is the best fitness of all particles in the population so far. However, all particles based on update function can moves toward the $gbest$ and the distance between $gbest$ and the neighboring particles can be gradual decreased by increased generation. Here, we use complementary-logic to improve the PSO search ability for SNP-SNP interaction model, in which the complementary particles are introduced to avoid premature convergence. These complementary particles replace 50% of the randomly selected particles in the population. The procedure of CLPSO includes (1) initialization of the parameter vectors, (2) objective function, (3) identification of $pbest$ and $gbest$, (4) particle updating, (5) complementary-logic adaption of particles, and (6) judgment of termination condition.

D. Initialization of the Parameter Vectors

CLPSO searches for a global optimum point in a D-dimensional real parameter space $R^D$. It begins with a randomly initiated population of NP $D$ dimensional real-valued parameter vectors. Each vector, also known as genome/chromosome, forms a candidate solution to the multidimensional optimization problem. We shall denote subsequent generations in CLPSO by $G = 0, 1, ..., G_{max}$. Since the parameter vectors are likely to be changed over different generations, we may adopt the following notation for representing the $i^{th}$ vector of the population at the current generation:

$$X_{i,G} = [x_{1,i,G}, x_{2,i,G}, x_{3,i,G}, ..., x_{D,i,G}]$$

(1)

For each parameter of the problem, there may be a certain range within which the value of the parameter should be restricted, often because parameters are related to physical components or measures that have natural bounds (for example if one parameter is a length or mass, it cannot be negative). The initial population (at $G = 0$) should cover this range as much as possible by uniformly randomizing individuals within the search space constrained by the prescribed minimum and maximum bounds: $X_{min} = \{x_{1,min}, x_{2,min}, ..., x_{D,min}\}$ and $X_{max} = \{x_{1,max}, x_{2,max}, ..., x_{D,max}\}$. Therefore, we may initialize the $j^{th}$ component of the $i^{th}$ vector as:

$$x_{j,i,0} = x_{j,min} + rand_{i}[0,1]* (x_{j,max} - x_{j,min})$$

(2)

where $rand_{i}[0,1]$ is a uniformly distributed random number lying between 0 and 1 (actually 0 $\leq$ rand, $[0, 1]$ $\leq$ 1) and is instantiated independently for each component of the $i^{th}$ vector. In CLPSO, a chromosome in the population represents a solution group and can be divided into two parts: the number of selected SNPs, and the genotypes associated with the SNPs.
The chromosome encoding can thus be represented by:

\[ x_{j, \text{max}} = \begin{cases} \text{SNP}_{\text{max}} & j \leq D/2 \\ \text{Genotype}_{\text{max}} & j > D/2 \end{cases} \]

\[ x_{j, \text{min}} = \begin{cases} \text{SNP}_{\text{min}} & j \leq D/2 \\ \text{Genotype}_{\text{min}} & j > D/2 \end{cases} \]

Genotype = \begin{cases} 1 & \text{dominant allele} \\ 2 & \text{heterozygous allele} \\ 3 & \text{recessive allele} \end{cases}

where SNP_{\text{max}} and SNP_{\text{min}} represents a limited SNPs, Genotype_{\text{max}} and Genotype_{\text{min}} represents the limited possible genotypes. For example, let \( X_{i, 0} = (2, 4, 6, 3, 1, 3) \), thus represents \( i^{th} X \) in first generation (i.e., 0) chosen SNPs (2, 4, 6) and genotypes (3, 1, 3), and can be described by the SNPs associated with the genotypes as follows: (2, 3), (4, 1) and (6, 3).

\[ \text{Odds} \left( \frac{TP}{FP} \right) = \frac{TP \times TN}{FP \times FN} \]

The concept of the designed fitness value uses the intersection of set theory to evaluate the difference value between breast cancer cases and non-cancer cases. The intersection of two sets is the set that contains all elements found in both sets, but no other elements. For example: \( X = (\text{SNP}_{2, 3}, \text{Genotype}_{2, 3}) \) is used to evaluate the number of matching conditions in the breast cancer cases and non-cancer cases. Let’s suppose the number of independent matching SNP2 with genotype 3 and SNP3 with genotype 2 is 273 in the breast cancer cases, and the number of independent matching SNP2 with genotype 3 and SNP3 with genotype 2 is 51 in the non-cancer cases. According to Eq. (6), the fitness value is determined by subtracting 51 from 273, leaving 222, which represents a high risk.

\[ f(X_i) = \sum_{j=1}^{p} u(X_i, P_j) - \sum_{j=1}^{N} u(X_i, N_j) \]

\[ u(X_{i, A}) = \begin{cases} 1 & \forall x \in A \\ 0 & \forall x \notin A \end{cases} \]

where \( x_{\text{max}} \) and \( x_{\text{min}} \) are the maximum and minimum SNP number, respectively.

\[ x_{id}^{\text{Complement}} = (x_{\text{max}} + x_{\text{min}}) - x_{id}^{\text{selected}} \]

**E. Objective Function**

In the CLPSO process, the fitness function value measures the quality of chromosomes. The SNP-SNP interaction study focuses on the particular SNP combinations to detect the highest fitness value, i.e., the maximum difference value between breast cancer cases and non-cancer cases. This criterion divides the fitness function into three separate steps, and the relevant equation can be written as:

**G. Particle Updating**

Each particle moves its position by updating the velocity in the next generation via an evaluation of \( pbest \) and \( gbest \). The update functions of velocity and position are formulated in Eqs. (7) to (9):

\[ w_{LDW} = \left( w_{max} - w_{min} \right) \times \frac{\text{Iteration}_{\text{max}} - \text{Iteration}_{\text{current}}}{\text{Iteration}_{\text{max}}} + w_{min} \]

\[ v_{id}^{\text{new}} = w_{LDW} \times v_{id}^{\text{old}} + c_1 \times r_1 \times (pbest_{id} - x_{id}^{\text{old}}) + c_2 \times r_2 \times (gbest_{id} - x_{id}^{\text{old}}) \]

\[ x_{id}^{\text{new}} = x_{id}^{\text{old}} + v_{id}^{\text{new}} \]

In Eq. (7), \( w_{max} \) set to be 0.9, \( w_{min} \) set to be 0.4 and \( \text{Iteration}_{\text{max}} \) is the maximum number of generations. In Eq. (8), \( r_1 \) and \( r_2 \) are random values between (0, 1); \( c_1 \) and \( c_2 \) are acceleration coefficients equal to 2; these coefficients constantly influence a particle moves in a single generation.

**H. Complementary-Logic Adaption of Particles**

The complementary-logic aims to avoid the particles are trapped into a local optimum by adjusting particles to a new region in the problem space. Thus, complementary-logic can facilitate the ensuring global exploration. The position of a trapped particle is changed by the Eq. (10).

**I. Judgment of Termination Condition**

The CLPSO procedure is terminated if the maximum number of generation is reached. When the termination condition is reached, \( gbest \) is the optimal solution, i.e., best SNP-SNP interaction model. The maximum number of generation is explained in the parameter settings section.

**J. Parameter Settings**

We set the population size of parameters = 50, number of generations = 100. The termination condition is that the generations met the maximum number of allowed generations.

**K. Statistical Analysis**

Odds ratio (OR) is used to evaluate the difference between cases and controls. It determines the combination of genotypes and quantitatively measures the association disease risk [10], and OR is defined as follows:

\[ \text{Odds ratio} = \frac{TP \times TN}{FP \times FN} \]
30%×5000=1500 for Aa and 20%×5000=1000 for aa. Thus, according to the above frequencies, the simulated data for 5000, and the SNP genotypes are randomly generated huge data set to test PSO and CLPSO. The sample size is shown in Table I. The SNP frequencies were obtained from 4

<table>
<thead>
<tr>
<th>SNP No.</th>
<th>Gene (SNP)</th>
<th>Genotype</th>
<th>AA (Control /Cases)</th>
<th>Aa (Control /Cases)</th>
<th>aa (Control /Cases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>COMT (rs6269)</td>
<td>1769</td>
<td>2390</td>
<td>841</td>
<td>917</td>
</tr>
<tr>
<td>2</td>
<td>COMT (rs46680)</td>
<td>1377</td>
<td>2417</td>
<td>1206</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>CYP19A1 (rs10046)</td>
<td>1430</td>
<td>2497</td>
<td>1073</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>CYP19A1 (rs3020314)</td>
<td>2147</td>
<td>2280</td>
<td>573</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>ESRI (rs2234693)</td>
<td>1450</td>
<td>2524</td>
<td>1026</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>ESRI (rs1543404)</td>
<td>1467</td>
<td>2441</td>
<td>1092</td>
<td>18</td>
</tr>
<tr>
<td>7</td>
<td>ESRI (rs3798577)</td>
<td>1413</td>
<td>2494</td>
<td>1093</td>
<td>19</td>
</tr>
<tr>
<td>8</td>
<td>ESRI (rs2747652)</td>
<td>1372</td>
<td>2447</td>
<td>1181</td>
<td>20</td>
</tr>
<tr>
<td>9</td>
<td>ESRI (rs2077647)</td>
<td>1383</td>
<td>2449</td>
<td>1093</td>
<td>21</td>
</tr>
<tr>
<td>10</td>
<td>ESRI (rs2175898)</td>
<td>1350</td>
<td>2507</td>
<td>1143</td>
<td>22</td>
</tr>
<tr>
<td>11</td>
<td>ESRI (rs9340799)</td>
<td>2107</td>
<td>2302</td>
<td>591</td>
<td>23</td>
</tr>
<tr>
<td>12</td>
<td>ESRI (rs1709182)</td>
<td>1988</td>
<td>2341</td>
<td>671</td>
<td>24</td>
</tr>
</tbody>
</table>

III. RESULT AND DISCUSSION

A. Data Set

We selected six important breast cancer genes (COMT, CYP19A1, ESRI, PGR, SHBG, and STS) as the test data sets, including 23 SNPs. The SNPs and genotype distributions are shown in Table I. The SNP frequencies were obtained from the reference [11]. These frequencies were used to simulate a huge data set to test PSO and CLPSO. The sample size is 5000, and the SNP genotypes are randomly generated according to the above frequencies. The simulated data for SNP, is based on its percentage which is obtained by multiplication of the percentage with the amount of the complete data set. For example, 50%×5000=2500 for AA, 30%×5000=1500 for Aa and 20%×5000=1000 for aa. Thus, the data for SNP, has generated to sample 5000 (2500+1500+1000=5000). All original data are generated by the same number in this manner. The simulation procedure is shown in the Fig. 2.

Pseudo-code for randomly generated data

01: begin
02: Set size = 5000
03: Set number of genotype = 3
04: Calculate amount of three genotypes
05: while (all SNPs are not normalized)
06: Calculate amount of each genotype
07: Calculate numbers of each normalized genotype
08: for n = 1 to number of genotype
09: Randomly create numbers of each normalized genotype
10: next n
11: end

Fig. 2. Pseudo-code for randomly generated data.
B. Analysis of Breast Cancer Susceptibility from 23 SNPs in Six Genes

The odds ratio and its 95% CI for all SNPs of six genes (COMT, CYP19A1, ESR1, PGR, SHBG, and STS) show that 13 SNPs, including COMT-rs6269, ESR1-rs3020314, ESR1-rs2175898, ESR1-rs1709182, ESR1-rs9478249, ESR1-rs1514348, ESR1-rs532010, PGR-rs660149, PGR-rs500760, SHBG-rs858518, SHBG-rs272428, SHBG-rs858524, and STS-rs2017591, display the statistically significant OR (p < 0.05) for breast cancer; the OR values range between 1.268 to 0.846.

C. Analysis of SNP-SNP Interaction Model with Difference between the Cases And Controls

Table II shows the results of SNP-SNP interaction models by 2 to 5-way. The left side in Table II represents the two to five SNP combinations. In these combinations, the 2-way model with SNP combinations and corresponding genotype values range between 1.268 to 0.846. The difference column indicates the difference between the cases and controls. In CLPSO, the difference values are explained as above mention. In CLPSO, the difference values between cases and controls are reduced from 126 to 41 in the two-way to five-way SNP-SNP interaction models. In PSO, the difference values between cases and controls are reduced from 126 to 15 in the two-way to five-way SNP-SNP interaction models. The larger difference value between cases and controls represents the better model.

D. Estimation of the Best Interaction Model Generated by PSO and CLPSO Using or and 95% CI in Breast Cancer

Table II shows the best interaction model in the 2-way to 5-way SNP-SNP interaction models. These results reveal that the total number of cases exceeds the total number of controls; it means that all models are a risk association in breast cancer. The right side in Table II shows the evaluated effects using odds ratio, 95% CI, and p-value. In PSO, the OR values in 2-way to 5-way SNP-SNP interaction models show the range of 1.05 to 1.79. All of the SNPs-SNP interaction models (2-way to 5-way) show significant OR values (p-value < 0.05). In CLPSO, the OR values in 2-way to 5-way SNP-SNP interaction models show the range of 0.89 to 2.28. Only the 2-way SNP-SNP interaction model shows the statistically significant (p-value < 0.05). However, in CLPSO, the OR values in 2-way to 5-way SNP-SNP interaction models show the range of 1.109 to 1.459, and the 95% CI of OR is in the range of 0.89 to 2.28. The 2-way SNP-SNP interaction model shows the statistically significant (p-value < 0.05). However, in CLPSO, the OR values in 2-way to 5-way SNP-SNP interaction models show the range of 1.153 to 1.391 and the 95% CI of OR is in the range of 1.05 to 1.79. All of the SNP-SNP interaction models (2-way to 5-way) show significant OR values (p-value < 0.050).

E. Comparison of PSO and CLPSO for the Interaction Model of Breast Cancer

The results represent the SNP-SNP interaction model identified by CLPSO has a better p-value than the model identified by PSO. PSO seems to provide a better OR value in the four-way SNP-SNP interaction model, however, the p-value shows the model does not statistically significant for breast cancer. The computational complexity of CLPSO is evaluated by objective function computation. Let n generation is implemented in a test, the computational complexity of PSO is O(n) which is the big-O in complexity analysis. The complementary-logic only adds an updated function, i.e., equation 10. Thus, the computational complexity between PSO and CLPSO is the same, but CLPSO is superior to PSO in terms of identifying the best SNP-SNP interaction model.

IV. Conclusion

In this study, a novel method, CLPSO, is proposed to identify the statistically significant SNP-SNP interaction models between related genes of breast cancer. These models can be used to analyze disease susceptibility and provide the information of SNPs located in the genes and their associated pathways. We used the huge number of simulated data to test the methods of PSO and CLPSO, the results indicate CLPSO can robust to search the statistically significant models with the difference value between SNPs of genes amongst the huge number of SNPs involved in real data sets.

REFERENCES

Yu-Da Lin received the MS degree from the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan, in 2011. He is currently working toward the PhD degree in the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan. He has rich experience in computer programming, database design and management, and systems programming and design. His main areas of research are bioinformatics and computational biology.

Li-Yeh Chuang received the MS degree from the Department of Chemistry, University of North Carolina in 1989 and the PhD degree from the Department of Biochemistry, North Dakota State University in 1994. She is a professor in the Department of Chemical Engineering and Institute of Biotechnology and Chemical Engineering at I-Shou University, Kaohsung, Taiwan. Her main areas of research are bioinformatics, biochemistry, and genetic engineering.

Cheng-Hong Yang received the MS and PhD degrees in computer engineering from North Dakota State University in 1988 and 1992, respectively. He is a professor in the Department of Electronic Engineering at the National Kaohsiung University of Applied Sciences, Taiwan. His main areas of research are evolutionary computation, bioinformatics, and assistive tool implementation.