GENETIC ALGORITHM FOR MAPPING EPISTASIS IN CONTROLLED CROSSES

Paulo Tadeu Meira e Silva de OLIVEIRA[1]
Mariza de ANDRADE[2]
José Eduardo KRIEGER[3]
Júlia Pavan SOLER[1]

- *ABSTRACT: The identification of epistasis or interactions among genes plays an important step for understanding the genetic regulatory network of complex diseases. When there are more than one genetic locus influencing the phenotype, interaction effects among loci are probably involved.* ***Nevertheless***, *despite of the advances in genetic mapping studies the analytical detection of epistasis is still considered a challenge. In this work, we apply the genetic algorithm (GA) jointly with different criteria for model selection to search for molecular markers associated with multiple QTL´s (Quantitative Trait Loci) and their interaction effects. GA represents a more efficient alternative for searching high dimensional spaces and it is less affected to general problems of identification of epistatic genes. We use simulation studies to compare the performance of GA with the classic search procedures, exhaustive and conditional, under different configurations. Finally, we analyze data from a F2 rats design and the AG found more optimal results when compared to conditional procedure. Two QTL´s with epistatic effect on systolic blood pressure were identified, located in chromosomes 5 and 9 of the rat genome.*

- *KEYWORDS: Interval mapping, gene interaction, model selection, quantitative trait loci.*

1 Introduction

Epistasis is a phenomenon theoretically known and very important in genetic studies, which describes how interaction among genes can affect phenotypes. However, empirically it is difficult to investigate this phenomenon, possibly due the limitations on the methodological tools that have been used. Many studies have focused on the epistasis detection problem in complex diseases, such as hypertension, asthma, diabetes and multiple sclerosis (Cordell, 2002; Carlborg and Haley, 2004; Moore and Williams, 2005; Gao et al., 2010), but the analytical identification of epistatic genes continues to be a challenge.

The term epistasis was first used by Batenson (1909) to describe a biological phenomenon in which the expression of a gene depends on the presence of one or more

[1] University of São Paulo - USP, Department of Statistics, Postal Code 05508-900, São Paulo, São Paulo, Brazil. Email*: poliver@usp.br / pavan@ime.usp.br*.

[2] Division of Biomedical Statistics and Bioinformatics, Mayo Clinic, Rochester, Postal Code 55905, MN, USA. Email: *mandrade@mayo.edu*

[3] Laboratory of Genetics and Molecular Cardiology, Heart Institute, SP, Brazil. Email: *krieger@incor.usp.br*

genes. Later, Fisher (1918) proposed a statistical interpretation through the linear regression framework allowing to model epistasis as an interaction effect among predictor variables. In this paper predictor variables will be represented by molecular markers that will take three possible values 0, 1 and 2 corresponding to their aa, Aa and AA genotypes, respectively. One of the goals of the genetic analysis is to identify quantitative trait locus (QTL) that is a genetic locus associated to a quantitative trait. Inbred populations, such as F2 design, are one of the most used for QTL analysis due to the direct calculation of the genetic values of a series of observed or putative loci (Haley and Knott, 1992).

Genetic mapping consists of experimental and statistical procedures for detection, localization and effect estimation of genes associated with etiology and regulation of diseases. By considering design of experiments involving controlled crossing of animals or plants, different formulations of regression models can be used to identify QTL's, including their major effects and possible interaction effects (epistasis). The challenge in these studies lies on the comparison of models or likelihood functions that, in general, show only a small variation around an optimum point and involve a high dimensional search space.

For epistatic QTL identification the exhaustive and conditional search are the mostly common used methods (Carlborg et al. 2000; Sen and Churchill, 2001; Goldberg, 1996; Holland, 1998). Considering a set of finite points, the exhaustive procedure assess all possible solutions to the combinatory problem, but for high dimensional maps the computational costs, in both time and memory, are severe. The conditional search method is much faster, but it has limited power for interaction effect detection, since it takes no account of all possible locus combinations and the selection of a locus depends on the previous selections (Churchill, 2001; Jannink and Jansen, 2001; Kao and Zeng, 2002).

Furthermore, for epistatic mapping problem the use of Genetic Algorithm (GA) was first proposed by Carlborg et al. (2000) and Nakamichi et al. (2001), who showed the efficiency and applicability of GA when searching dense markers maps and using residual sum of squares as objective-function. Genetic algorithm (GA) is a general and flexible method for searching optimum solutions in complex spaces (Tsutsui and Gosh, 1986; Carlborg et al., 2000). For mapping of epistasis, GA represents an analytically useful tool that can increase the computational efficiency for searching bigger genomes and compact maps of molecular markers. Further, GA can be adaptable to the apparently non-linear optimization problem involved on epistasis detection.

For quantitative responses the regression models have been applied for epistasis mapping through interaction effects of predictor variables. The same approaches have been used in the context of logistic regression models for binary responses associated to diseases. Recently, Satagopan and Elston (2013) proposed a test statistic for detecting the presence of removable interaction in order to fit parsimonious additive models for searching epistasis. Despite of all efforts, epistasis still remains an opened research field and our limitation for identification of epistatic loci has been one factor responsible for the failure of many genetic studies, including hypertension genetic mapping using experimental designs as F2 inbred (Garret and Rapp, 2002; Levy et al., 2010; Newton-Cheh et al., 2009; Krieger, 2010; Gopalakrishnan et al., 2010).

In this article, we formalize the GA in the context of genetic mapping and applied it to identify epistatic QTLs using simulation data and F2 controlled crosses. In section 2, we describe the classic methods and our proposed GA. In section 3 we describe our simulation and real data. In Section 4 we show the results of our GA algorithm using simulated data under different scenarios, and analyzing F2 rat data (Schork et al., 1995). In section 5, we presented our final remarks.

2 Method

2.1 Epistatic regression model

The multiple interval mapping model (MIM) proposed by Kao et al. (1999) is based on a regression model with the predictor variable defined as genetic values associated with the effects of multiple QTLs and their interactions. Using this model to detect epistasis may increase the precision and power for identification of QTLs and their interaction effects. We assume that the QTLs are either observed or imputed molecular markers.

In this article, MIM is applied by considering two QTLs randomly sampled from the genome and analyzed as predictor variables and using three different search procedures (exhaustive, conditional and GA). We assume that all F2 rats have phenotype and genotype data available for analysis. By genotype data it means any molecular marker platform as microsatellite or SNPs (Single Nucleotide Polymorphisms). By setting two QTLs, Q1 and Q2, located on positions*pos1* and *pos2*in two different intervals, I_1 and I_2, respectively, the following regression model is used for epistasis detection between such QTLs,

$$y_i = m + a_1 X_{1i} + a_2 X_{2i} + i_{12} X_{1i} * X_{2i} + \varepsilon_i \qquad (1)$$

where y_i is a phenotype of interest (quantitative response) observed on the *i-th* individual; *m* is the general mean;a_1, a_2 andi_{12} represent additive and interaction effects for Q1and Q2, respectively;X_{1i} and X_{2i} are predictor variables that can be defined in terms of both observed or putative genotype values for loci in the I_1 and I_2 intervals, respectively. The error components, ε_i, are assumed uncorrelated, homoscedastic and following a Normal distribution (Haley and Knott, 1992) or, as proposed by Zeng (1994), be a Normal Mixture distribution.

Churchill (2001) considered regression models similar to equation (1) and proposed a test for interaction effect with one, two or three degrees of freedom (df), i.e., the interaction test can be assessed in the presence of both additive effects (H_0: $i_{12}=0 \wedge a_1 \neq 0 \wedge a_2 \neq 0$; using one df), or in the presence of only one additive effect (H_0: $i_{12}=0 \wedge a_1=0 \wedge a_2 \neq 0$; using two df) or without additive effects (H_0: $i_{12}=0 \wedge a_1=0 \wedge a_2=0$; using three df). In our article, we use the one df interaction test.

For search genetic maps using exhaustive method the interaction effect test is conducted for all possible pairs of positions extracted from the genomic map, for instance, using slices of 1cM (centi Morgan molecular unit) to cover the map. Under GA, an interaction test is accomplished only for pairs of positions randomly selected by the algorithm path. For the conditional search, first we test for one major additive effect (H_0: $a_1=0$) to identify a QTL with the greatest effect, say Q1; second we include Q1 in the model, and interaction tests are conducted to find the next QTL, Q2, considering all remaining loci (H_0: $i_{12}=0 \wedge a_1 \neq 0 \wedge a_2 \neq 0$).

For each pair of loci, Q1 and Q2,the model in equation (1) is fitted and statistics SSE (residual sum of squares), AIC (Akaike Information Criterion) and BIC (Bayesian

Information Criterion) are calculated and used as criteria for model selection (Paulino et al., 2003). In our implementation, SSE was calculated by the classical quadratic form given by

$$SSE = Y'\left[I - X(X'X)^{-1}X'\right]Y \tag{2}$$

where Y is the response variable vector; X is the predictor variables matrix, with the columns defined by the values corresponding the general mean effect,Q1 and Q2 additive effects and its interaction. AIC statistics was calculated by the following expression,

$$AIC = -2\ln(LR) + 2p \tag{3}$$

where *LR* is the likelihood ratio, *ln* is the natural logarithm and *p* is the difference between the number of the parameters of the two models under comparison (under H_0 and H_1).BIC statistics was calculated by (Raftery, 1995)

$$BIC = -2\ln(LR) - 2p\ln(n) \tag{4}$$

where *LR* and *p*are defined as in the expression (3) and *n* is the number of the rats. Equivalences among these criteria can be assessed (for instance, Sakamoto et al., 1986):

$$AIC - BIC = 2p(1 + \ln(n)),$$

$$AIC = -2n\ln\left(1 + \frac{SSR}{SSE}\right) + 2p,$$

and

$$BIC = -2n\ln\left(1 + \frac{SSR}{SSE}\right) - 2p\ln(n),$$

where *SSR* is the regression sum of square. The main difference between *AIC* and *BIC* is the penalization factor. BIC assumes that the “true” model is among the set of candidate models, while AIC searches for the best model among the available models. For model comparison with different number of parameters, the SSE criterion is easy to implement, ensures a good performance of the model selection procedure and can well discriminate the quality of adjusts between two or more competitive models (Carlborg et al., 2000; Wang, 2000).

2.2 Genetic algorithm (GA)

The major reason to use GA for epistasis search is its potential for optimization of high dimensional spaces. To search the genome for detection of epistatic QTLs, each pair of positions (pos1, pos2) extracted from the possible positions population defines a solution to be assessed and updated. GA reaches the pair of positions that represents the best solution taking into account the optimization of an objective-function. In our situation, the three model selection statistics, SSE, AIC, and BIC, are used as objective-functions for our proposed GA.

In this article, for the GA implementation we used a real codification where real values were assigned either for pairs of candidate positions and their correspondent objective-function. To cover the genome, we assume that the positions are fixed 1cM apart along the genome, and they can be observed or imputed positions. For implementation of the GA the following steps were considered, (see, Figure 1):

Step 1 – Evaluation: Assignment of a worse pair of positions (*pos1*, *pos2*) that reaches a high value for the objective-function assuming that the pair will be replaced during the process.

Step 2 – Initialization: Random selection, without replacement, of four different pairs of positions. Such pairs are sampled from the population of possible positions defined by the combination $\binom{K}{k}$, where K is the number of positions fixed into the genomic map and k is the number of positions selected at a time (we use $k = 2$ loci). Objective-functions are calculated for the four pairs and one is selected using two tournaments procedure, which has the advantage to provide diversity (Thierens and Goldberg, 1994; Blickle and Thiele, 1995; Carlborg et al., 2000; Nakamichi et al., 2001).

Step 3 – New Set of Positions: Generation of new positions through the following steps:

> **Step 3.1 – Selection:** In this stage, one pair of positions is selected from the group of four candidates for further recombination via a BLX-h or a mutation method. For decision between recombination (BLX-h)or mutation, a random variable is generated under uniform distribution, U[0,1]; if the result is lower than a fixed mutation probability (*pm* parameter, set to 0.1 or 0.4) the program will execute a mutation procedure; otherwise, it will execute recombination procedure.
>
> **Step 3.2 – Recombination:** Conduction of the *BLX-h* routine (Eshelman et al., 1997). This operator sustains the intuitive idea that information must be exchanged among different candidate solutions. The probabilities of recombination are set to 0.9 or 0.6 (corresponding to *pm* equal to 0.1 or 0.4, respectively).
>
> **Step 3.3 – Mutation:** A random operator is used that attribute equal probability for occurrence of a stronger mutation (limit mutation) or a lighter mutation (uniform mutation).

Step 4– Updating: In the initialization matrix including the four candidate pairs, the pair of positions with the worse adjustment (using SSE, BIC or AIC criteria) is excluded in favor of the new pair of positions obtained from recombination or mutation steps. This pair is included in the updating if its adjustment is “better” than the one previously inserted.

Step 5 – Stopping rule: The algorithm is run for a fixed number of generations (*ng*) and the best solution is obtained. Then, the algorithm is run for a number of solutions (*ns*) and global best solution is finally obtained by comparing all *ns* solutions.

Step 6 – Loop: If the stopping rule, *ns*, for one solution is not attained, the algorithm returns to step 2 for the next GA generation, until a given number of generations is completed, and a solution is achieved.

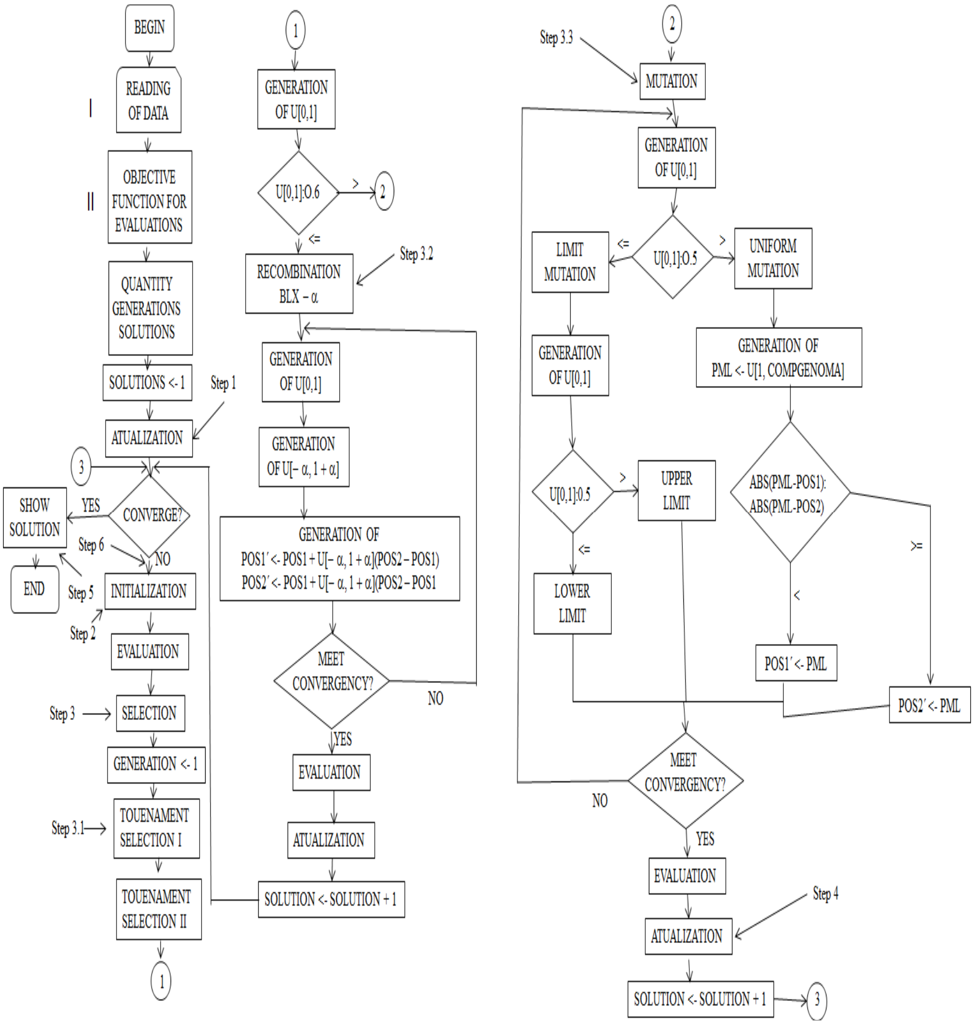

Figura 1 - Genetic Algorithm for epistasis search.

3 Data description

3.1 Simulation study

We consider eight different scenarios for data simulation based on the following parameters: dimension of the markers map (MAP), sample size (*n*), mutation probability (*pm*), number of solutions (*ns*) and number of generations (*ng*). The first parameter, MAP, refers to the number of chromosomes, number of observed markers per chromosome and the distance between markers for a map. The parameter *n* is the number of F2 rats and corresponds to the number of phenotype data simulated. The last three parameters are related to the **GA** application. Table 1 shows the parameter values used for simulation and data analysis.

Table 1. Characterization of simulated and real data

Data	MAP			Total	*n*	QTL1	QTL2	*pm*	*ns*	*ng*
1	2	4	4	24	50	5	26	0.1	100	100
2	2	4	4	24	50	5	26	0.4	100	100
3	2	4	4	24	50	5	26	0.4	100	1000
4	2	4	4	24	200	1	13	0.4	100	1000
5	2	4	4	24	200	5	22	0.4	100	1000
6	2	4	4	24	200	5	26	0.4	100	1000
7	10	4	4	120	200	104	113	0.4	100	1000
8	10	10	8	720	200	9	633	0.4	100	1000
9	21	8.7*	26.7*	6,436	221			0.4	100	100
10	21	8.7*	26.7*	6,436	221			0.4	100	1000

*: mean values for real data.

For the simulation of the markers map we consider two scenarios: first, smaller genome window (Data 1 to Data 6) with 2 chromosomes including 4 markers per chromosome equally 4cM spaced, totalizing 24cM for search; second, larger windows (Data 7 and Data 8) with 10 chromosomes, 4 markers per chromosome equally spaced by 4cM and by 8cM, totalizing 120cM and 720cM, respectively. Positions were imputed for covering the map in each 1cM. For each observed marker position, genotype data was generated, but for putative positions genotype data was calculated from the observed genotype of adjacent markers, as proposed by Haley and Knott (1992).

Data sets 1, 2 and 3 were generated using *n=50* observations and MAP defined by 24cM totalizing 276 possible pair of positions to search for epistasis. Data from 4 to 6 used *n=200* observations and the same MAP. Data 7 was generated using *n=200*and a MAP defined by 120cM totalizing 7,140 possible pair of positions to search for epistasis. Finally Data 8, also with *n=200* observations, but a MAP with 720cM, totalizing 259,560 possible pair of positions to search for epistasis.

For the genotype and phenotype data simulation we used Win QTL Cart (Wang et al., 2007) for a F2 population. We assume markers in linkage equilibrium. We generate the phenotype from a Normal distribution with mean and variance equal to 130 and 2, respectively. In each simulation scenario, two QTLs with epistatic effect were generated in fixed positions from the map (QTL1 and QTL2, indicated in Table 1). We use R package (R Development Core team, 2008) to implement our method.

3.2 Rats F2

We analyze data from Schork et al. (1995) involving 221 rats from a F2 design obtained from a crossing between a male Spontaneous Hypertensive Rat strain (SHR) with a female normotensive Brown-Norway rat (BN). Many cardiovascular variables were evaluated in each animal, such as systolic and diastolic blood pressure, measured before and after a salt diet exposition. Genotype data were also obtained for each animal considering a map with 182 molecular markers (microsatellite platform) distributed along the 21 rat chromosomes.

Schork et al. (1995) analyzed these data and identified 5 QTL's with additive effects associated with systolic blood pressure after salt loading (denoted by SBPS). They did not identify any locus with epistatic effect using conditional research. Thus, our goal is to use the same data and our proposed method to see if our GA method will be able to identify an epistatic effect.

4 Results and discussion

Considering simulated data (Data 1 to Data 8) and real data (Data 9 and 10), we apply exhaustive, conditional and GA methodology to search the genomic map looking for epistatic QTLs. In each case, the objective-functions, SSE, AIC and BIC, were used for model selection. For Data 1 to Data 8, Table 2 shows the results obtained when GA is applied. In this table is indicated the objective function (*fc*); the global optimum value (*vog*), which is the *fc* value obtained by the exhaustive search; the percentage of convergence toward the global optimum *fc* value (*pc vog*), corresponding to the number of solutions in *ns* equal to *vog*; the minimum and maximum *fc* values(*lim_inf* and *lim_sup*, respectively) considering all (*ng* x *ns*) pair of positions searched; and the pair of the best global positions (pos1 and pos2), in centimorgan, for epistatic QTLs.

When the mutation probability (*pm*) changes from 0.1 to 0.4(Data sets 1 and 2) and the number of generations (*ng*) changes from 100 to 1000 (Data 2 and Data 3), we observe a reduction in the value of the difference between *lim_inf* and *lim_sup* values obtained by GA, and an increase in *pc vog*.

Table 3 shows the optimum value of the objective-functions for GA, exhaustive and conditional searches using Data sets 1 to 8. The results indicate that the best global solution was found for the same pair of positions using exhaustive search and GA, for the three objective-functions (SSE, AIC and BIC). The conditional search did not attain such optimum solutions.

Table 2. GA results obtained for simulated data analysis

Data	*fc*	*vog*	*pcvog*	*liminf*	*li*	pos1	pos2
1	SSE	47.89	14	47.89	6	5	26
1	AIC	0.04	4	0.04	8	5	26
1	BIC	-23.61	4	-23.61	-	5	26
2	SSE	47.89	20	47.89	6	5	26
2	AIC	0.04	6	0.04	7	5	26
2	BIC	-23.61	5	-23.61	-	5	26
3	SSE	47.89	36	47.89	6	5	26
3	AIC	0.04	18	0.04	7	5	26
3	BIC	-23.61	17	-23.61	-	5	26
4	SSE	197.66	2	197.66	2	1	13
4	AIC	-31.65	4	-31.65	1	1	13
4	BIC	-2.46	1	-2.46	-	1	13
5	SSE	57.50	28	57.50	6	5	22
5	AIC	0.72	18	0.72	7	5	22
5	BIC	-22.92	17	-22.92	-	5	22
6	SSE	64.39	32	64.39	7	5	26
6	AIC	-0.38	21	-0.38	1	5	26
6	BIC	-24.03	22	-24.03	-	5	26
7	SSE	185.84	1	185.84	3	104	113
7	AIC	-3.15	1	-3.15	1	104	113
7	BIC	-32.52	1	-32.52	-	104	113
8	SSE	192.05	1	192.05	3	9	633
8	AIC	-7.94	0	0.49	1	1	601
8	BIC	-32.13	0	-31.48	-	1	641

Table 3. Results of GA, exhaustive and conditional searches for the global optimum solution

DATA	GA			EXHAUSTIVE			CONDITIONAL		
	SSE	AIC	BIC	SSE	AIC	BIC	SSE	AIC	BIC
1	47.89	0.04	-23.61	47.89	0.04	-23.01	47.8	0,19	-
2	47.89	0.04	-23.61	47.89	0.04	-23.01	47.8	0,19	-
3	47.89	0.04	-23.61	47.89	0.04	-23.01	47.8	0,19	-
4	197.6	-2.46	-31.65	197.6	-2.46	-31.65	248.	9.30	-
5	57.50	0.72	-22.92	57.50	0.72	-22.92	57.5	0.72	-
6	64.39	-0.38	-24.03	64.39	-0.38	-24.03	57.5	0.72	-
7	185.8	-3.15	-32.52	185.8	-3.15	-32.52	334.	-3.15	-
8	192.0	0.49	-31.48	192.0	0.49	-31.48	305,	10.93	-
9	385.5	2.44	-2.31				404.	4.53	-2.87
10	170.8	-4.74	-34.33				404.	4.53	-2.87

When we increase the number of rats from 50 to 200 (Table 2, Data sets 3, 4, 5 and 6), we observe an increase in SSE values, a reduction in AIC and BIC values, and a strong reduction in the time that took GA search to attain the global optimum point (*pcvog*). These results are expected since AIC and BIC incorporate a penalty as a result of the sample size and SSE does not it.

For Data set 7 (Table 2), we observe that the simulated epistatic positions 104cM (chromosome 8) and 113cM (chromosome 9) were identified for three objective functions using exhaustive and GA searches. When using the conditional search this solution was attained for SSE and BIC methods, whereas a different epistatic pair on positions, 100cM (chromosome 8) and 113cM (chromosome 9), was identified using the AIC method.

By increasing the window size from 24cM to 120cM (Table 2, Data sets 6 and 7), we observe a decrease in tied solutions, i.e., solutions with equal objective-function values. Moreover, we also observed a reduction in the number of times that GA attains the global optimum point (*pcvog*).It occurred in a single solution among the hundreds researched (by assessing 1,000 generations in each solution). Empirically, it indicates that the greater the genome greater is the requirement to increase the number of generations (*ng*) to have a higher probability to attain the global optimum solution. Furthermore, when we increase the genome size from 120 to 720 (Table 2, Data sets 7 and 8) the computational time to run the exhaustive search increases significantly (Data 7, around 24 hours, and Data 8, around 13 days, using a 2 Gb RAM memory and a 3.4 Ghz clock computer) and GA failed to attain the global optimum point, suggesting that 1,000 generations per/solution might be insufficient due the great increase in the search space dimension, from 7,260 to 259,560 points (epistatic positions).

Figure 2 shows the dispersion of the SSE, *AIC* and *BIC* values using GA for one solution with 10,000 generations. The horizontal axis represents the number of generations (index) and the vertical axis represents the values of the objective-functions. These values tend to decrease as the generations increase until they attain the global optimum solution. The lower value corresponds to the best pair of positions that detects the epistatic loci.

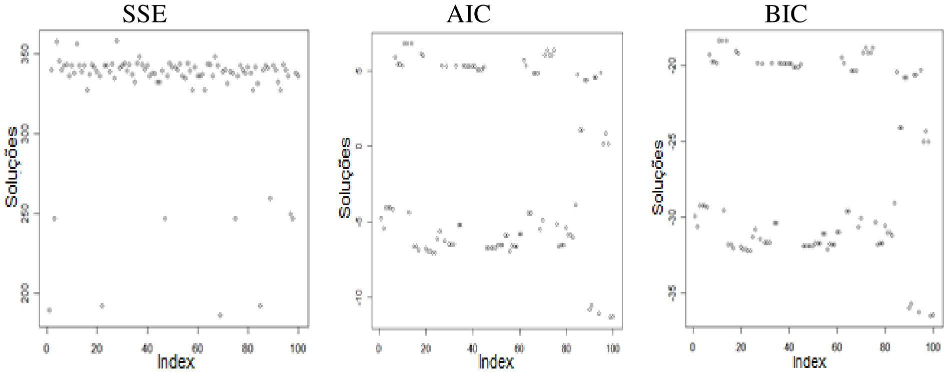

Figura 2 - Dispersion of *SSE*, *AIC* and *BIC* values using GA (Data 8).

Considering the analysis of F2 rats data (Data sets 9 and 10), Table 3 depicts the optimum value of the objective-functions (*SSE*, *AIC* and *BIC*) obtained via GA and conditional searches. For both procedures, the optimum pair of positions (in cM) is located in putative positions on the map. AG found more optimal results when compared to conditional procedure. We found the following markers, R1335and R5175, located on chromosomes5 and 9, respectively, using the optimum pair of position identified by GA (considering 100 solutions with 1,000 generations/solution). Figure 3 shows the interaction effect of these markers for the systolic blood pressure mean values after salt diet. We assumed that alleles B and S are segregating from normotensive (BN) and spontaneously hypertensive (SHR) parents for genotypes BB, BS and SS in F2 animals.

The graph depicted in Figure 3 strongly suggests the presence of an epistatic effect since the three mean lines are not parallel.

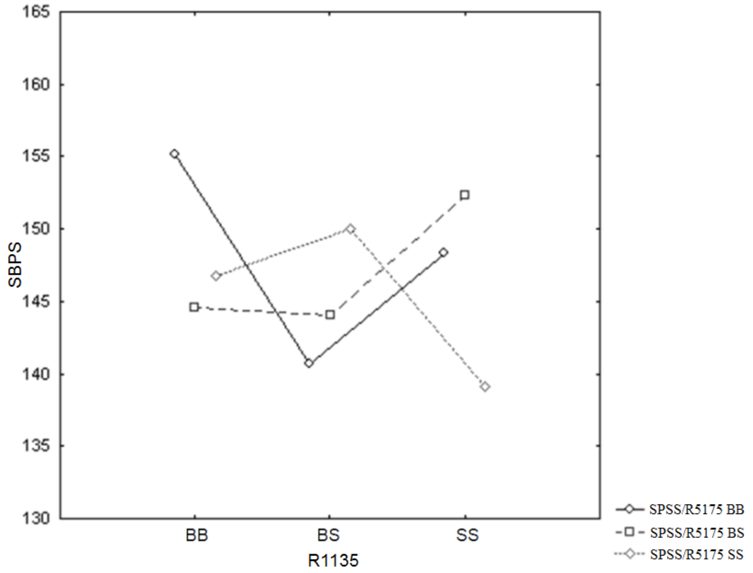

Figura 3 - Systolic blood pressure means for markers R1135 and R5175.

Figure 4 shows contour graphs for SSE, AIC and BIC values, obtained from the regression model in equation (1), considering all possible pairs of positions located precisely on 182 observed markers on the F2 rats map (a total of 16,471 possibilities). Lower values attained to SSE (SSE lower than 200), AIC (AIC lower than -1) and BIC (BIC lower than -10) are represented in dark blue, which correspond to the best pairs of positions for epistatic markers, whereas higher values of these objective-functions (SSE higher than 1,000, AIC between 0 and 2, and BIC between (-10,-2]) are represented in light blue and green. The optimum pair of positions found via GA is indicated in the graphs as a red asterisk.

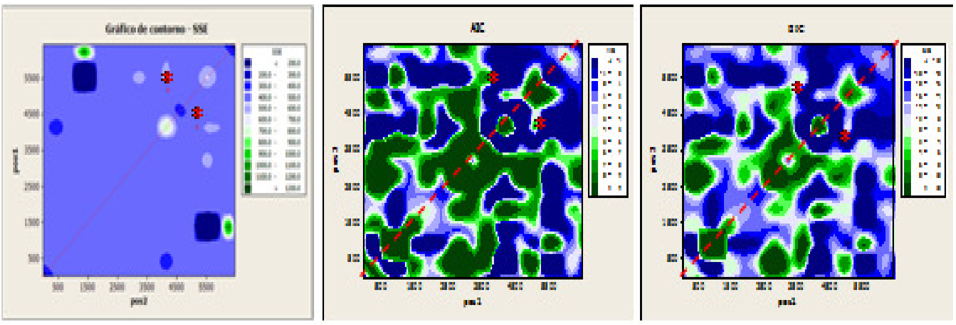

Figura 4 - Contour graphs for SSE, AIC and BIC values considering all pair of positions of the observed markers in the F2 rat map.

Figure 4 showed only results from observed marker data, without assessing imputed positions along the map (putative markers). Since the inclusion of imputed positions in the analysis might improve precision and power for the identification of epistatic loci, we have also evaluated our method using imputed positions. The, optimum pair of positions identified using GA is indicated in the graphs as a red asterisk (chromosomes 5 and 9, the solution by GA, Data set 10). When we combine the observed and putative positions, the optimum SSE, AIC, and BIC values using GA was 170.83, -4.74, and -34.33, respectively (Table 3, Data set 10). It is worth noting that regions in the contour graph include points that do not correspond to the best solutions, indicating that the optimization process used to identify pairs of epistatic loci is non-linear and the optimum solution found depends on the gradient used to refine the search space.

Conclusions

In this paper, we applied the genetic algorithm to find epistatic loci in molecular markers maps considering data from F2 inbred designs. For epistatic QTL search exhaustive and conditional methods were also used, but they had their own limitations. In our application we consider GA with three different objective-functions, SSE, AIC and BIC. Our results using simulated data showed that the conditional search method may lose power by disregarding pairs of positions, including optimum pairs of positions found using GA. Furthermore, GA not only showed to be as powerful as the exhaustive method for epistasis detection, since both methods found the same solutions when analyzing moderate size data set (10 Chr, 4 M\Chr and Genome 120 cM), but also GA is feasible to analyze dense maps while the other two methods cannot.

The GA search had a fast convergence process toward a global optimum point (with a more significant epistatic effect) regardless of the mutation probability values, the number of generations, and the number of solutions by generations. We also observed that for established given computing time it was more efficient to increase the number of generations than to increase the number of solutions. Our simulation results showed that for GA to converge it was important to keep a moderate number of solutions (around 100).

By applying GA to analyze F2 rat data a pair of QTLs, located in chromosome 5 and 9, was identified with epistatic effect in the regulation of the blood pressure. These loci were identified considering imputed positions on the marker map with 1cM apart and using GA with mutation probability equal 0.4 (recombination probability equal 0.6), 1,000 generations and 100 solutions/generation. Nevertheless, the chromosomal region that contains such optimum solution was not identified when exhaustive search was applied considering only positions of observed markers on the map. This indicates that the optimum solution depends on the gradient adopted to search the marker map, and the GA is flexible enough to move towards this solution.

Our map with 182 markers available was refined in distances from 1 to 1 cM, but by exploring the flexibility of GA to work on high dimensional search spaces additional epistasis studies can be conducted by using more severe genotype imputation or adopting more dense maps. In this context, surveys of the genetic variation based on SNP (Single Nucleotide Polymorphism) platforms can be used for mapping in rat strains (STAR Consortium, 2008). Thus, based on our QTLs currently identified, such chromosome regions can be refined by covering them with SNP data and finding significant QTNs (Quantitative Trait Nucleotides) associated to blood pressure.

An R source code considering the GA implementation is available on the webpage www.ime.usp.br/~poliveir and also it can be obtained from the authors upon request.

Acknowledgements

The authors thank the Laboratory of Genetics and Molecular Cardiology, Heart Institute – INCOR, Clinics Hospital of the University of São Paulo for granting the actual database of F2 rats used in this study.

OLIVEIRA, P. T. M. S; DE ANDRADE, M; KRIEGER, J. E.; SOLER, J. M. P. Algoritmo Genético no mapeamento de epistasia em delineamentos com cruzamentos controlados. *Rev. Bras. Biom.*, São Paulo, v.32, n.1, p.143-157, 2014.

- RESUMO: A identificação de epistasia ou interação entre genes é um passo importante para a compreensão da rede de regulação genética de doenças complexas. No entanto, apesar dos avanços em estudos de mapeamento genético a detecção analítica de epistasia ainda é considerada um desafio. Neste trabalho, aplicamos o algoritmo genético (GA), em conjunto com diferentes critérios para a seleção de modelos, para pesquisar o espaço de marcadores moleculares em busca de QTLs (do inglês, Quantitative Trait Loci) epistáticos (com efeitos de interação). Estudos de simulação sob diferentes configurações são usados para comparar o desempenho do GA com procedimentos de busca clássicos, como a exaustiva e condicional. Para o mapeamento epistático, dados de um projeto com ratos F2 são analisados sob o procedimento de busca condicional e GA. Por meio deste último fomos capazes de identificar dois QTL's com efeito epistático sobre a pressão arterial sistólica, localizados nos cromossomos 5 e 9 do genoma do rato. OGA representa uma alternativa eficiente para pesquisar espaços de alta dimensão, sendo menos afetado por problemas gerais de identificação de genes epistáticos. Apesar do GA ser suficientemente flexível para se mover para uma solução ótima, o procedimento de busca depende do gradiente (locos imputados) adotado para pesquisar o mapa de marcadores. Foi observado que para um tempo computacional de cálculo fixado é mais eficiente aumentar o número de gerações/solução do que aumentar o número de soluções, mas para garantir a convergência do GA é importante manter um número moderado de soluções (cerca de 100, em nossa aplicação).

- PALAVRAS-CHAVE: Mapeamento intervalar; interação gênica; seleção de modelos; QTL, genes da hipertensão.

References

BATENSON, W. *Mendel´s Principles of Heritability*. Cambridge University Press, 1909.

BLICKLE, T.; THICLE, L. A Mathematical Analysis of Tournament Selection. In: PROCEEDINGS OF THE 6TH INTERNATIONAL CONFERENCE MORGAN KAUFMANN, San Francisco, USA, 1995.

CARLBORG, O.; ANDERSON, L.; KINGBORN, B. The Use of a Genetic Algorithm for Simultaneous Mapping of Multiple Interacting Quantitative Trait Loci. *Genetics,*v.155, p.2003-2010, 2000.

CARLBORG, O.; HARLEY, C. S. Epistasis: Too Often Neglected in Complex Traits Studies? *Nature Reviews Genetics*, v.5, p.618-625, 2004.

CORDELL, H. J. Epitasis: what it means, what it doesn´t mean, and methods to detected it in humans. *Human Molecular Genetics,* v.11, n.20, p.2463-2468, 2002.

ESHELMAM, L.; MATHIAS, K.; SCHAFFER, J. Cross-over operator biases: Exploring the population distribution. In: *PROCEEDINGS OF THE SEVENTH INTERNATIONAL CONFERENCE ON GENETIC ALGORITHMS AND THEIR APPLICATIONS,* Michigan, T. Baeck Ed., 2007, p.354-361.

FISHER, R. The correlations between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.*, v.52, p.399-433, 1918.

GAO, H.; GRANKA, J.M.; FELDMAN, M. W. On the classification of epistatic interactions. *Genetics*, v.184, p.827-837, 2010.

GARRET, M. R.; RAPP, J. P. Two closely linked interactive blood pressure QTL on rat chromosome 5 defined using congenic Dahl rats. *Physiol. Genomics*, v.8, p.81-86, 2002.

GOLDBERG, D. G. *Genetic Algorithms in Search, Optimization and Machine*. Learning Addison and Wealey, Reading, MA, 1996.

GOPALAKRISHNAN, K.; SAIKUMAR, J.; PETERS, C. G.; KUMARASAMY, S. FARMS, P.; YERGA-WOOLWINE, S.; TOLAND, E. J.; SCHNACKEL, W. GIOVANNUCCI, D. R.; JOE, B. Defining a rat blood pressure quantitative trait locus to a <81.8 kb congenic segment: comprehensive sequencing and renal transcriptome analysis. *Physiol. Genomics*, v.42A, p.153-161, 2010.

HALEY, C.; KNOTT, S. A Simple regression method for mapping quantitative trait loci in line flanking markers. *Heredity*, v.69, p.315-324, 1992.

HOLLAND, J. *Adaption in Natural and Artificial System.* The University of Michigan Press, Ann Arbor, Mich, 1998.

JANNINK, JL;JANSEN, R. Mapping Epistatic Quantitative Trait Loci with One-Dimensional Genome Searches. *Genetics*, v.157, p.445-454, 2001.

KAO, C.H.; ZENG, Z. B.;TEASDALE, R.D. Multiple intervals mapping for quantitative trait loci. *Genetics*, v.52, p.1203-1216, 1999.

KAO, C. H.;ZENG, Z. B. Modeling Epistasis of Quantitative Trait Loci Using Cockerham's Model. *Genetics*, v.160, p.1243-1261, 2002.

KRIEGER, J. E. Mapping genes for hypertension using experimental models: a challenging and unanticipated very long journey. *Physiol. Genomics*, v.43, p.99-100, 2010.

LEVY, D.; EHRET, G. B.; RICE, K.; VERWOERT, G. C.; LAUNER, L. J. Genome-wide association study of blood pressure and hypertension. *Nature genetics*, v.41, n.6, p.677-687, 2009.

MOORE, J. H. A global view of epistasis. *Nature Genetics*, v.37, p.13-14, 2005.

MOORE, J. H.; WILLIAMS, S. M. Traveling the conceptual divide between biological and statistical epistasis system's biology and a more modern synthesis. *Problems and paradigms, Bio Esays,* v.27, p.637-646, 2005.

NAKAMICHI, R.; UKAI, Y.; KISHINO, H. Detection of Closely Linked Multiple Quantitative Trait Loci Using a Genetic Algorithm. *Genetics*, v.158, p.463-475, 2001.

Received in 25.10.2013

Approved after revised in 21.03.2014