# An Imporved Gene Expression Programming Based on Adaptive Correction Constant

Yonghong Yu[1, a], Li Wu[2, b] and Zhou Zhou[1, c]

[1]School of Management Science and Engineering, Anhui University of Finance & Economics, Bengbu 233030, China;

[2]School of Finance and Public Management, Anhui University of Finance & Economics, Bengbu 233030, China.

[a]ac120107@163.com, [b]bbwuli@163.com, [c]1057578619@qq.com

## Abstract

**This paper mainly focuses on the necessity of the presence of dominant chromosomes in the initialization population, and the importance of introducing reasonable constants during the genetic process. We proposed the rough multiple linear regression initialization_adaptive correction constant (RMLR_AC) method to obtain a rough set of full variable coefficients as the initialization constants and integrated these constants into the evaluation calculation process, which can ensure the number of variable of individuals and the number of reference constants be complete, and can increase fitness of population in early stage of evolution stably, not resulting in rapid local convergence. The experiment results show that the RMLR_AC accelerates evolution, increases the fitness of optimal chromosome compared to the result not using RMLR_AC.**

## Keywords

**Gene expression programming, Rough multiple linear regression, Adaptive correction constant.**

## 1. Introduction

In genetic algorithm[1], the way of organization and state of the population is one of the important factors that affect the evolutionary and the search efficiency for the target solution. Population is a collection of individuals in solution, the merits of the population is determined by the level of individual merits. The pros and cons of the population reflects the level of evolution, and the distance between the current generation and the generation that evolves a satisfactory result. As a new branch of genetic algorithm, Gene Expression Programming[2,3] (GEP) has the same features of genetic algorithm mentioned above.

The population of standard GEP consists of gene expression individuals generated by random initialization. These initialization individuals not only constitute the function of independent variables selected randomly, but also determine the constants of independent variables of the function. However, the function has some blindness, during the genetic process, it can not assure to be effective of genetic structural model. In addition, because of lacking prior knowledge, the constants existing in the form of coefficients or exponents in function are difficult to initialized, especially in structural model of random function. These constants decide the proportion of computing of each variable in the function, and changing the constants value often exerts a great influence on the result of the gene expression evaluation, and then affects the fitness computing. It is difficult to begin evolution for GEP if the fitness value of the best individuals in population generated randomly are low. Therefore, it is necessary to reduce the randomness of constants during individual initialization. In order to promote the formation of excellent gene fragments in the early stages of evolution, it is usual way to increase the fitness of gene expression by controlling constant of function.

There are three main approaches to introduce numerical constant into GEP: the GEP-RNC[4] with the addition of the DC domain to the tail of the gene, the MC[5,6] with direct constants as terminals, and

the constant creating by algorithm itself[3]. Both the GEP-RNC and the MC all focus on the way of integrating constants with gene expression, and the way of encoding and decoding of gene expression, but the constants of function are still introduced at random. In third method, the constants are generally derived from the operation of the same terminals or the constants, and are adjusted according to fitness function during genetic evolution. However, it does not provide definite method to determine the constants in the process of population individual initialization. If the initial value of constant is significantly different from the appropriate constant value, it may increases the swing amplitude of the parameter, and result in the dificulity of determination of the operator type in genetic process.

In order to improve the fitness of chromosome individuals in the initial population, this paper propose a novel approach to improving the average fitness in the early evolutionary stage effectively. The gene expression, generated by converting the multivariate linear regression of the training samples, will be used as the predominant individuals in the initial population. The coefficients of multivariate linear regression will be used as constants of gene expression, which will be involved in evolution under the adaptive correction of GEP.

## 2. Preliminary

Terminals and functions constitute the primitive elements of GEP. The terminals represent to the constant, input or no parameter functions in the mathematical expression, and correspond to the leaf nodes of the expression tree. The functions represent operators of espical application, components of programm and middle symbols of system, they correspond to the non leaf node of expression tree.

The expression tree(ET) is used as the phenotype of GEP individuals to represent an expression visually. In order to facilitate computer genetic manipulation, the nonlinear structure of the expression tree is translated into a linear structure which Ferreira called K-expression as chromosome of GEP, corresponding to the genotype of GEP individuals.

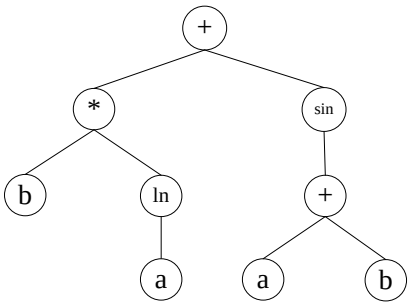
GEP genes are composed of a head and a tail[3,7]. The head contains symbols that represent both functions and terminals, whereas the tail contains only terminals. Therefore two different alphabets occur at different regions within a gene. For each problem, the length of the head h is chosen, whereas the length of the tail t is a function of h and the number of arguments of the function with the most arguments n, and is evaluated by the equation:

$$t = h^* (n-1) + 1 \tag{1}$$

Consider a gene composed of {+,*,ln,sin, a, b}. In this case n = 2. For instance, for h = 7 and t = 8, the length of the gene is 7+8=15. One such gene is shown below:

$$\begin{array}{c} 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 0\ 1\ 2\ 3\ 4 \\ +\ *\ \sin\ b\ \ln\ +\ a\ \ a\ \ b\ \ b\ \ a\ a\ b\ a\ b \end{array} \tag{2}$$

and it codes for the following ET:




The corresponding meaning of each chromosome member in the initial population created by the GEP algorithm is different. The genetic operator acts on the genetic manipulation of the population, and the higher the degree of fitness of the resulting offspring individuals, the greater the probability of being selected. GEP adopts linear equal-length coding, so long as the length of the gene is constant and the tail consists only of terminals, a legitimate offspring chromosome can be obtained. Therefore,

the GEP genetic operator can be designed more simply and flexibly. The basic genetic operators of GEP mainly include selection, mutation, transposition of insertion sequence elements, root transposition, gene transposition, one-point recombination, two-point recombination, and gene recombination.

Fitness is an indicator of biological ability to measure the ability of a species to adapt to the environment, it can also apply to the GEP algorithm. The fitness function is used to calculate the individual fitness to evaluate the pros and cons. By evaluating the chromosomes, expression trees can be obtained, then the objective function values are calculated from the phenotypes, and finally, according to the conversion rules, the phenotypic (individual) fitness values can be obtained using the objective function values. Ferreira proposed two fitness evaluation models: one is that the fitness fj of an individual chromosome j is calculated based on the absolute error, and the other is calculated based on the relative error. Let T be a dataset containing n samples, Ti is the i-th input dataset, vi is the observed value of Ti, and $\hat{v}_i$ is the estimated value. the fitness equations are:

$$f_j = \sum_{i=1}^{n} (K - | v_i - \hat{v}_i |) \tag{3}$$

$$f_j = \sum_{i=1}^{n} (K - | \frac{v_i - \hat{v}_i}{v_i} |) \tag{4}$$

Through the constraint of the fitness function, the population retains the high-adaptation chromosomes during the evolution process. When the individual chromosomes do not have relative or absolute errors in the sample estimates, the accuracy is maximized and the fitness is maximized.

## 3. Adaptive Correction Constant Based on Rough Multiple Linear Regression Initialization

In the extended multivariate linear regression model, we first obtain the parameters of the model by least squares estimate(OLS) or maximum likelihood(ML) estimation; second, perform goodness of fit tests(R2) and significance test (F, t) on sample data to compare the fit of the models; then, conduct a series of tests and elimination such as heteroscedasticity, autocorrelation and multicollinearity; finally, we obtain a more accurate linear regression formula. If the accurate linear regression formula is introduced as a gene expression individual of GEP population initialization, a highly adaptive individual can be obtained and can be treated as a predominant individual in the initial population.

The gene expression corresponding to the multivariate linear expression contains excellent gene fragments, by hybridizing with other individual gene fragments through genetic operators, they can increase the average fitness of offspring population, and lay the foundation of generating individuals with more fitness for evolutionary breakthrough. In addition, the coefficients of the independent variables reflects the strength of the corresponding sample attributes in determining the prediction results in the regression model, and we suggest that these coefficients be used as the coefficient constant of argument terminals of the chromosome individual gene expression in the initial population. Under the guideline of the GEP fitness function, the coefficient constants of the terminals of the independent variables are constantly adjusted to be more accordant with the current gene model by continuously adapting the evolution process. Therefore, excellent gene fragments and suggested coefficient constants of multivariate linear regression expressions have a positive effect on the evolution of the gene model.

However, the excellent gene fragments of multivariate linear regression expression is linear and the structural model of function is more single. The goodness of fit of the gene expression transformed by extended multivariate linear regression is much higher than that of other individuals in the initial population formed at random, so the obviously excellent may lead the direction of gene evolution, and the gene segments in other chromosome individuals close to the dominant gene segment, and are even assimilated by the linear gene segment even after a few evolutionary or several generations. Therefore, the single structure of extended multivariate linear regression gene expression is often too

excellent to evolute other individuals in the population, and to maintain the diversity of population.In addition, the extended multivariate linear regression test is more strict. In the process of testing the significance of explanatory variables, some insignificant variables are removed, and these variables may be useful for the final determination of the gene expression structure during genetic process. Similarly, in the solution of multicollinearity, the final result may still have missing arguments due to adopting the method of eliminating the secondary or alternative variables directly, stepwise regression or principal component regression. That is, it can not provide a complete parameter recommendations for GEP.

GEP hopes that the number of variables of individuals participating in population initialization and the number of reference constants shoud be complete, and there exist elite individuals to increase fitness of population in early stage of evolution stably. That is, GEP hopes that the predominant individuals will do not affect the gene diversity of future populations, do not result in rapid local convergence. This paper presents a simple constant initialization method RMLR_AC (Rough Multiple Linear Regression Initialization_Adaptive Correction Constant).

RMLR_AC requires GEP to make a multiple linear regression on all training samples before it is established. Regression uses only OLS to estimate the parameters of the model, minimizing the residual sum of squares of the multivariate linear regression model. By formalizing the normal equations to obtain the full variable coefficients, these full variable coefficients are provided to the GEP as constant reference values. This process gives up the series tests of statistical and economic that followed in the standard multiple regression and is therefore rough. The gene expression corresponding to this linear expression has some prioritized advantages in fitness, but its structural model has some loopholes and therefore is excellent in the initial population but does not occupy the absolute dominant position.

These constants are coefficients of all the variables that are linearly estimated, so they are simply connected to the terminal by a multiplication operator in the gene expression. In order to reduce the length of the gene expression and increase the computing speed, we do not use the coefficient constant when constructing the gene expression, but just multiply the sample value of the independent variable by the estimated coefficient as the new terminal value to participate in the operation. The estimated value of sample obtained by this indirect calculation is the same as the estimated value of sample obtained by directly calculating the gene expression with the coefficient constant.

The improved GEP is controlled by the evolution direction of the fitness during the genetic process, which means that the coefficient constant also participates in evolution indirectly. The GEP evolution process has the function of adaptive correction constants, so the coefficient constants is also involved in the adaptive correction.

For example, if the coefficient constant of an independent terminal in a certain generation is c, and it is inherited to the next generation, the front operation obtains a constant k, and k and c are connected by an operator to obtain a new coefficient of the terminal of the independent variable, the constant c is considered evolutionary after this genetic.

Fig.1 depicts the RMLR_AC approach to initialize the population flow. Fig.2 depicts the algorithm flow of introducing the coefficient constant into the chromosome evaluation:

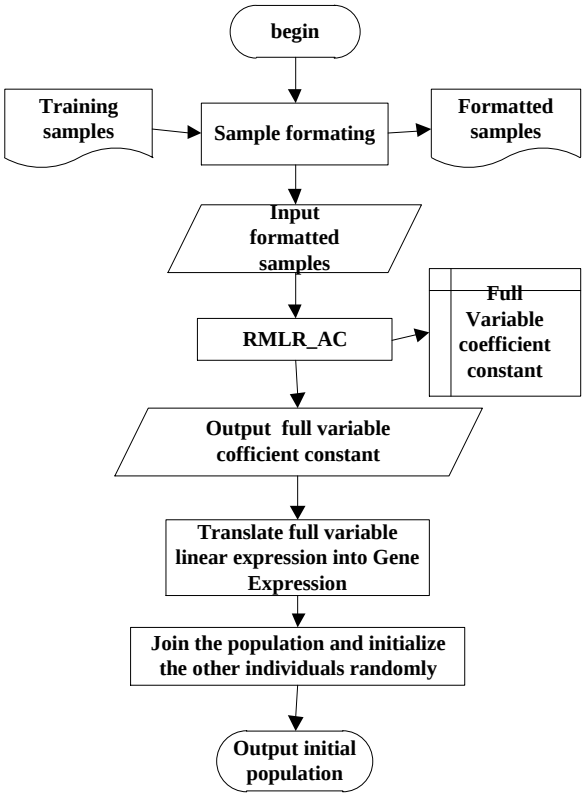



Fig 1. Algorithm flow of introducing RMLR_AC into initial population

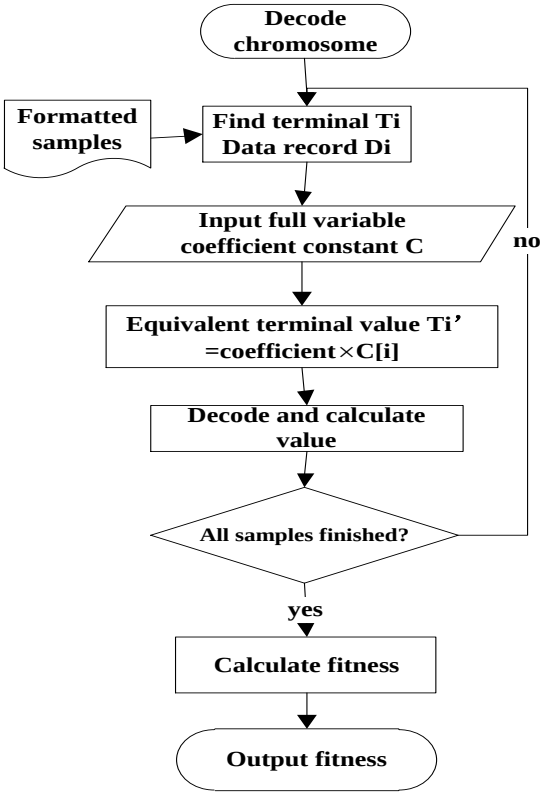



Fig 2. Algorithm flow of introducing coefficient constant into chromosome evaluation

## 4. Experiment and Analysis

A part of the Iris data test of UCI data set is used to prove the validity of the improved GEP algorithm compared to the standard GEP. The data contain three flower samples, setosa, versicolor and virginica. Each sample contains 4 attributes, including sepal length, sepal width, petal length and

petal width. The calyx length of the setosa sample was selected as the dependent variable y, and the other attributes were used as independent variables, which were expressed as a,b and c respectively. Setosa sample data, as shown in Table 1:

Table 1. Iris-Serosa Sample Dataset.

| Num | Sepal length y | Sepal width a | Petal length b | Petal width c |
|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 |
| 5 | 5.0 | 3.6 | 1.4 | 0.2 |
| 6 | 5.2 | 3.5 | 1.5 | 0.2 |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 |
| 8 | 5.0 | 3.4 | 1.5 | 0.2 |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 |

In order to improve the efficiency of genetic evaluation, we adopt reverse polish expression_sack decoding(RPE_SD) to evaluate gene expression. To compare the time required for genetic evaluation of GRCM[8] and RPE_SD in the same population size and the same generation of heredity, we set the population size to 40, the generation of heredity to 200, and use single gene to represent chromosomes. The length of the gene expression is set to 10, 20,... 90 respectively. The original size of the sample is 10, in order to verify the evaluation time, we set the experiment size of sample to 20,30,...,90 respectively by copying the original sample. Each group are executed 90 times repeatedly, and the average evaluating time is adopted as the result. The experiment results show in Fig. 3:

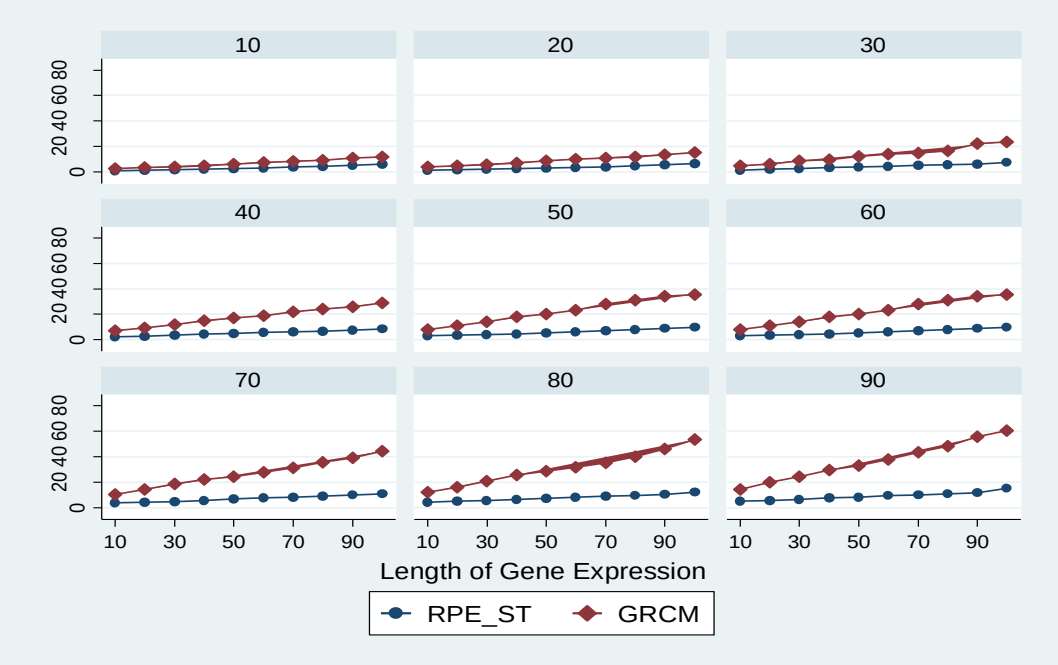



Fig 3. The decoding time of RPE_SD and GRCM

Fig.3 shows that under the same experimental conditions, regardless the size of sample space and the head length of gene expression, the time required for genetic evaluation using the RPE_SD is about 1/3~1/4 the time of GRCM, especially when the sample space is large.

In order to prove the effect of RMLR_AC constant on evolution, the experiment was divided into two groups, the experimental group and the control group, both groups use the same data set in Table 1, and they are executed based on RPE_SD.

The control group was introduced into the array {1.0,1.0,1.0,0.0}, the first three were corresponding to the sepals width $0, the petal length $1 and the petal width $2, and the last 0 was a constant

parameter. After RMLR_AC multiple linear regression, the full variable regression results in the experimental group were as follows (k means coefficient constant and v means partial correlation coefficients):

k(0)=0.858695652173916 v(0)=0.998931620843001

k(1)=0.34420289855079 v(1)=0.966804304896309

k(2)=-2.61594202898534 v(2)=0.971745634545923 k(3)=2.06304347826073

Sum of square: q=0.169565217391305 mean standard deviation: s=0.130217209842365

R-Square=0.709648600357355 Regression Sum of Squares: u=0.414434782608693

and we can get following equation:

$$\hat{y} = k(0) \times \mathrm{a} + k(1) \times b + k(2) \times c + k(3)$$

When the coefficient constant is not a terminal symbol, the corresponding ORF string is "+ $ 0 + $ 1 $ 2". The R-square value is 0.71, so we can see that the gene expression transformed by multivariate linear regression has a good fitness. Then an non-coding segment with appropriate length is added to the ORF string, and the extended ORF string is converted into a complete gene expression, which is added to initial population as the dominant chromosome individual. In addition, K = {k [0], ..., k [3]} is used as the coefficient constant of variables.

Set the size of population to 60, the head length of gene expression to 15, the probability of selection to 0.3, the probability of recombination to 0.75, and the mutation probability to 0.1. Comparing the optimal fitness of the two individuals obtained by the RMLR_AC and non-AC algorithm after the 2000 generation,the individual with high fitness wins. Fig.4 shows the results of the experiment:

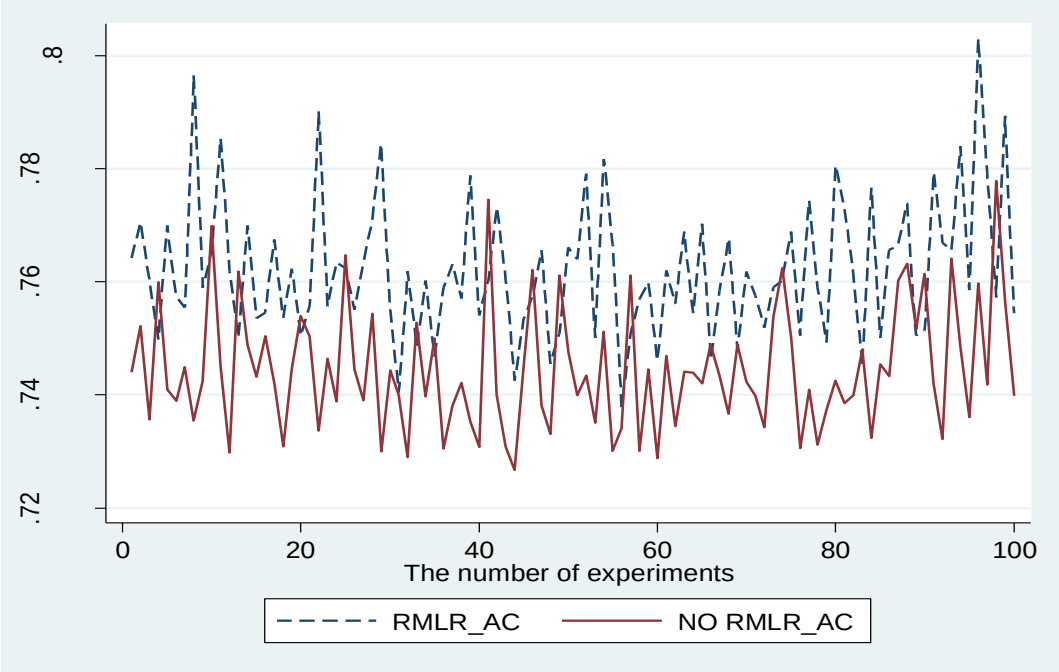



Fig 4. Fitness value of RMLR_AC and non RMLR_AC

Fig.4 shows that in the comparison of the 100 experimental results, the optimal fitness of the chromosomes obtained by the RMLR_AC algorithm after the 2000 generation is 81 times greater than that of the ProGEP without RMLR_AC. The result shows that the RMLR_AC algorithm can improve the evolution speed. One group uses RMLR_AC algorithm and the other group does not use RMLR_AC algorithm, each group is executed 100 times respectively, and after 2000 generations of evolution, the average fitness of optimal chromosome is 0.76243, compared to 0.740124 of not using RMLR_AC, increased by 3.01388%.

## 5. Conclusion

The constant selection in the process of population individual initialization of GEP plays an important role during the genetic process. An improved GEP algorithm based on RMLR_AC dealing with the constants selection in the process of population individual initialization is presented. The RMLR_AC can ensure the number of variable of individuals and can ensure the number of reference constants be complete, it can also improve evolution efficiency stably. The experiment results show that the RMLR_AC accelerates evolution, increases the fitness of optimal chromosome compared to the result not using RMLR_AC.

## References

[1] Koza J R. Genetic Programming 2 (The MIT Press, MA 1994).
[2] Information on https://www.gene-expression-programming.com
[3] Ferreira Candida. Gene Expression Programming: A New Adaptive Algorithm for Solving Problems. Complex Systems, vol. 13(2001), p.87-129.
[4] Ferreira Candida. Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence (Springer- Verlarg, Germany 2006).
[5] Jie Zuo: Researches on The Core Technology of Gene Expression Programming(Ph.D., Sichuan University, China 2004).
[6] Yuan Rao: Improved Algorithm of Gene Expression and Its Application in Function Optimization (MS., Guangxi Normal University, China 2007).
[7] Ferreira Candida. Genetic Representation and Genetic Neutrality in Gene Expression Programming. Advances in Complex Systems, vol. 4(2002), p.389-408.
[8] Dazhi Jiang. New Method Used in Gene Expression Programming: GRCM. Journal of System Simulation, vol. 18(2006), p.1466-1468. (In Chinese).