Genetic Neutrality in Naïve Gene Expression Programming

Abstract—Gene expression programming (GEP) is a genotype-phenotype system with non-coding regions in genome where possible neutral mutations would be accumulated. In this paper, a novel concept named naïve gene expression programming (NGEP) is introduced. The main contributions include: (a) proposing a novel algorithm named NGEP, which decoding method based on the complete tree, (b) discussing the role of genetic neutrality in GEP and NGEP, and (c) performing experiments to show that NGEP is faster than traditional GEP, and the success rate of NGEP is higher than that in traditional GEP.

Keywords—genetic neutrality; gene expression programming; naïve gene expression programming; evolutionary computation

I. INTRODUCTION

Most recently, the research of neutral theory attracts much attention not only from biological scientists, but also from computer researchers. The study results deriving from biology field provide more opportunity for computer scientists. To the best of our knowledge, the neutral theory of molecular evolution[1] is firstly invented by Kimura, and it asserts that the accumulation of neutral mutations is a key problem in evolutionary methodology. In neutral theory, "most of the large genetic mutations are neutral", i.e., the phase of accumulations by neutral mutations instead of natural selection plays an essential role in evolutions.

In the study of biological evolutions, more and more evidences show that the molecular or some part of molecular may bring about an important influence in evolutions, which are thought not to be important in functions.

GEP is a typical research practice of a full-fledged genotype-phenotype system which the chromosome/expression trees form a truly functional, indivisible whole[2]. It is a newly proposed data mining technique, and widely applied in the field of data mining in recent years. For the special structure of genes in GEP, the genetic representation of GEP has both the fragmentation of genes in the genome and the existence of junk sequences or non-coding regions in the genome.

This study is highlighted by the phenomena of the neutral mutations in terms of chromosomes, focusing on analyzing the efficient of neutral genetic in evolution. The literature[3] studied neutral regions in the GEP genome, and it showed that the existence of neutral regions in the genome and the accumulation of neutral mutations in the genome play a key role in evolutions. In this paper, we proposed a new concept called Naïve GEP (NGEP), proposed a decoding method which based on completed tree, and by our decoding method, there are more free neutral regions in genome than the traditional GEP.

II. PROBLEM STATEMENT

In this section, we mainly describe the decoding gene by binary tree and a new concept called NGEP.

A. Gene and Complete Tree

The gene representation using gene expression programming is composed of two distinct domains, i.e., head and tail[2]. The head domain contains symbols that represent both functions and terminals, whereas the tail is only composed of terminals.

For each problem, the length of the gene head denoted as $h$ is chosen, and the tail $t$ is a function of $h$ and the number of arguments of the function with more arguments $n_{\text{max}}$ (called maximum arity). They satisfied with the Equation (1).

$$t = h^*(n_{\text{max}}-1) + 1 \tag{1}$$

Consider a gene with the maximum argument is $n_{\text{max}}$, the length of gene $g = h + t$, constructing a $n_{\text{max}}$-branch complete tree including $g$ nodes. Thus, this tree has $h$ non-leaves, as well as $t$ leaf nodes, the $h$, $t$, $n$ must satisfied with formula (1). Therefore, there exists a one to one relationship between the lengths of a gene and a completed $n_{\text{max}}$-branch tree.

B. The Decoding Rules Based on Binary Tree

First, we give an important concept.

Definition 1. The special case of traditional GEP satisfying the following two conditions is called Naïve GEP (abbreviated as NGEP):

(i) The maximum argument numbers of functions is not greater than 2, i.e., $n_{\text{max}} = 2$;

978-1-4244-2108-4/08/$25.00 © 2008 IEEE
(ii) Gene decoding rule is based on binary tree rule.

To decode a gene, there are mainly two steps in NGEP:

(a) Label valid locus of a gene. Scan each gene from left to right, label the first loca, if it is a terminal then terminates this phase, otherwise, according to its arguments, record the positions of its children nodes. Recursive the above phase till labeling a terminal symbol.

(b) Translate labeled valid genes into an expression tree. The labeled gene is an ORF (Open Reading Frame) of the gene. In this phase, first generate a binary tree according to tagged gene, and this tree is an expression tree that is translated based on binary trees.

The aim of the first step, as a preparation for the second phase, can accelerate the speed of gene decoding, therefore improving performance of NGEP system.

In the following, Example 1 is illustrated to this idea.

**Example 1.** Given a gene $G_1$ that is represented as follow:

$$G_1: / +a*/ +aaaaaaaa$$

We label valid genes, and obtain the locus from 0 to 4 and from 7 to 10; these fragments form a valid string of that the corresponding gene. The locus that is not be tagged is a neutral region and the corresponding symbol is the useless gene or a neutral gene; these genes play a vital role in evolutions. We decode this gene and express it into expression tree, show it in Figure 1.

![Figure 1. The ET of $G_1$ is decoded by binary tree method.](image)

Map labeled genes into a binary tree. In Fig. 1, after decoding, we obtain a mathematic formula of this gene as $(a *a + a / a) / a$.

Figure 2 depicts gene $G_1$ decoded by Karva language. Its result formula is $((a + a) * a + 1) / a$.

![Figure 2. The ET of $G_1$ decoded using Karva language](image)

The binary tree-based encoding mode is straightforward, because the relationship among locus in a gene and the label of nodes in a binary tree are one-to-one relationship. In NGEP, encoding and decoding use a specific mode, thus its performance is higher than that of the traditional GEP.

### III. Analysis of Neutral Region in a Gene

In order to improve the evolution efficiency, genes in both traditional GEP and NGEP adopt the moderate redundancy strategy. The redundant genes form neutral genes, and the occupied region in a gene is called neutral region. In this section, we will further analysis the properties of neutral regions.

#### A. Neutral Region in a GEP Gene

The gene in traditional GEP consists of head domain and tail domain, the head is mainly for functions selected for the special problem, the tail works as a buffer or reservoir with terminals in order to guarantee the validity of structures. The head contains symbols both for functions and terminals and the tail is composed of terminals only.

For each problem, the length of a head $h$ is chosen beforehand, whereas the length of a tail $t$ is a function of $h$ and the number of arguments in a function with more arguments $n_{\max}$ and is evaluated by Equation 1. It has been proved in [4] that any modification made in genome, no matter how profound, always results in a valid expression tree. Obviously, the structural organization in terms of genes must be kept, i.e., maintaining boundaries between the head and the tail.

**Example 2.** Given a gene $G_1$ that is expressed as follow.

$$G_2: /a Q/b*a b/Qa*b*-abbaababbbabbbba$$

Its function set is $\text{FunSet} = \{Q, *, /, -,+\}$, the terminal set is $\text{TermSet} = \{a, b\}$, then $n_{\max} = 2$. Here, consider $h=15$, then $t=16$, $g=15+16=31$.

Decoding it by Karva language, we obtain its expression tree as Figure 3.

![Figure 3. The ET of $G_1$ by Karva language](image)

From Fig.3, we know that the length of $G_1$ is 31, but the ORF of $G_1$ ends at position 7. It implies that from positions 8 to 30 are not expressed in its ET. This region is called neutral region, and the gene fragments in this region are called neutral code. If genetic operations occur at neutral region, it does not affect an individual, i.e. its expression tree keeps unchanged, and hence called neutral operation.

#### B. Neutral Region in a NGEP Gene

The encoding approach in NGEP is an improvement of GEP. It produces more free neutral regions in a gene than traditional GEP method. Example 3 illustrates this feature.
Example 3. Given a corresponding gene \( G_2 \), we decode this gene by using binary tree rules. Its expression tree is depicted in Figure 4.

![Figure 4. ET of \( G_2 \) by binary tree techniques.](image)

Fig. 4 shows that in gene positions from 3 to 4, 6 to 10, 13 to 22, and 25 to 30 are all neutral regions. It shows that NGEP gene has more flexible neutral regions than that in a traditional GEP gene.

Generally, GEP gene has one unique neutral region and it always locates at the rear of a GEP gene, but NGEP gene may have multiple neutral regions, and it may reside in any location of a gene besides the root of the gene.

IV. EMPIRICAL STUDY AND DISCUSSION

This section studies the performance of neutral region and the neutral genetic modifications. The platform is follows.

CPU: AMD Athlon 1.53GHz PC; Memory: 512M RAM; Program Language: Java.

Section IV.A and Section IV.B, we test the performance of both NGEP and GEP, using single genetic and multi-genetic system, respectively. By controlling the length of genes, we achieve controlling the neutral region proportion for single genetic system. For multi-genetic system, we control the neutral region by changing the number of genes in a chromosome.

A. Neutral Genetic Operations in Single genetic System

In our experiments, the parameter settings in are showed in Table 1.

For analyzing the traditional GEP and on the neutral region effect of GEP on evolutions, the problem chosen a more complex function finding problem, that is, \( f(x) = x^2/2 + 3x \), a set of 10 random fitness cases chosen from the range between [-10, 10].

We counted the number of evolution system reach the best solution, and then obtain the success rate. We evaluate performance of both GEP and NGEP system by this success rate.

The neutral region in single genetic system can be adjusted by changing the length of chromosome. For the function finding problem, we change the length of chromosomes, running NGEP and GEP 100 times, respectively. Figure 5 shows that relationship between success rate and the length of chromosome.

Fig. 5 shows that, most compact organizations may not be the most efficient. For instance, when the length of gene \( g < 17 \) (corresponding to the head length of gene \( h < 8 \)), both traditional GEP and NGEP are unsuccessful.

![Figure 5. Success rate with varying length of genes](image)

Note that, the notation \( * \) represents the operation that is not used in single genetic system.

When \( g = 17 \) (correspondingly, \( h = 8 \)), success rate is just 3% for both of them. With the length of gene increasing, the size and/or the quantity of neutral regions increase, the probability of neutral genetic operation increases, while the success rate ascending. Therefore, a certain amount of redundancy is important for evolution to occur efficiently. Straightforwardly, excessive redundancy in a gene of chromosome, cause success rate decrease.

Note that the highly redundant systems works, nonetheless, considerably better than the compact systems. It further shows that evolutionary systems, for both GEP and NGEP, can cope fairly well with genetic redundancy.

B. Neutral Genetic Operation in Multi-genetic System

Now we study the performance of neutral regions in multi-genetic system. Adjust the quantity and the size of neutral regions by adjusting the number of gene in chromosomes. For multi-genetic system, each gene length in chromosome is equal. Chosen the length of gene in chromosome in multi-genetic system is the shortest length of gene which can be success in the single genetic system. The length of gene in multi-genetic chromosome is considerably compact. Thus we can compute the chromosome redundancy rate in the multi-genetic system by formula (2).

rateRedundancy = 1 - lengthOfGene / lengthOfChrom  \( (2) \)

Note that here, \( g = 17 \) (responding to \( h = 8 \)), i.e. \( g = 17 \) is the shortest chromosome in our study. The numbers of gene in a chromosome range from 1 to 10. For each the length of chromosomes, we are running 100 times traditional GEP and NGEP, respectively. The success rate is shown in Figure 6.
Fig. 6 shows that: (a) when the size and quantity of neutral regions increasing, the probabilities of neutral genetics also increase, the success rate of both traditional GEP and NGEP increases as well as, and (b) under a moderate redundancy (the corresponding to number of gene is 2~5, i.e. the redundancy rate is 50~80%), the success rate is higher than that other cases.

V. CONCLUSIONS

This study proposed a new concept named naïve gene expression programming (NGEP), and analyzed the neutral genetic in NGEP. Our extensive experiments show that neutral genetic mechanism is an important property in evolutionary system. The structural genes of NGEP enriches the size and quantity of neutral region, therefore, thus the existence of neutral regions in NGEP accelerates the evolution. The authors also analyzed the role and the ability of neutral genetic in gene. Compared with traditional GEP, NGEP has the following characters: (a) the most compact organizations may not be the most efficient, whereas a certain amount of redundancy is important for evolution; (b) multi-genetic system is more efficient than single genetic system, and (c) the average success rate of NGEP is higher than that traditional GEP under the same redundancy rate at single genetic and multi-genetic system respectively.

ACKNOWLEDGMENT

The authors would like to thanks the anonymous reviewers for their invaluable feedback. This research is supported by Grant of National Science Foundation of China (60773169), supported by the 11th Five Years Key Programs for Sci. &Tech. Development of China under grant No. 2006BAJ05A01 and the Software Innovation Project of Sichuan Youth (AA0807).

REFERENCES