GENE EXPRESSION PROGRAMMING

Dr. Sundaram Suresh
School of Computer Engineering
Nanyang Technological University
Singapore
Email: ssundaram@hotmail.com

Textbook

2

□ Candida Ferreira, Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence, Angra do Heroismo, Portugal. 2002

□ Weblink
- http://www.gene-expression-programming.com/
  - www.gepsoft.com

- GEP code can be found in
  - http://jgep.sourceforge.net/
  - http://www.gene-expression-programming.com/Downloads.asp

What is Gene Expression Programming?

3

- □ GEP is also an evolutionary based algorithm.
- □ Gene expression programming is developed by incorporating both the idea of simple, linear chromosomes of fixed length used in GAs and the ramified structures of different sizes and shapes used in GP.
- □ Genes - codes for a smaller program or sub-expression tree.
- □ The structure of chromosomes in GEP was designed to allow the creation of multiple genes.
- □ It is worth emphasizing that GEP is the only genetic algorithm with multiple genes.

GEP

4

- □ GA – does not represent the I/O relationship mathematically
- □ GP – complexity in genetic operators and increase in tree length due to genetic operation.
- □ GEP – is combination GA string representation and GP mathematical expression
- □ GEP uses genetic operators in GA to change the tree length. But, here the length of the string remains the same.
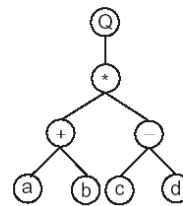- □ GEP is based on human genome…

GEP Representation

5

- □ We want to represent the arithmetic expression
- □ Chromosome made of genes
- □ Function set arguments (n)
- □ Gene – head and tails
- □ Heads (h) are specified for a given problem
- □ Tails are calculated based on number of heads and

  $t = h(n-1)+1$

$$\sqrt{(a+b)\times(c-d)},$$

which can also be represented as a diagram or ET:

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010

GEP Representation

6

- □ Arithmetic expression - Gene Equivalent – K-Expression

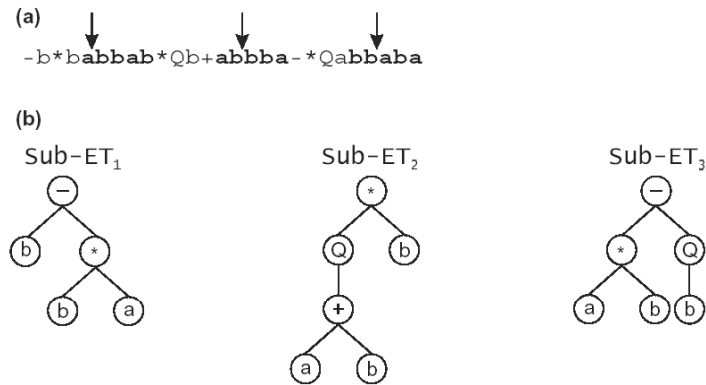| 0	1	2	3	4	5	6	7
Q	*	+	-	a	b	c	d

- □ Three Gene Representation

```
012345678012345678012345678
-b*babbab*Qb+abbba-*Qabbaba
```

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   **6**

Tree Construction of Three Gene

(a)

`-b*babbab*Qb+abbba-*Qabbaba`

(b)

Sub-ET₁   Sub-ET₂   Sub-ET₃

Figure 2. Expression of GEP genes as sub-ETs. **(a)** A three-genic chromosome with the tails shown in bold. The arrows show the termination point of each gene. **(b)** The sub-ETs codified by each gene.

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   **7**

Biological process in Gene

□ The main operations that occur in a Genome are:

◘ Genome Replications

◘ Genome restructuring

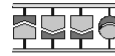◘ Transcriptions

◘ Translation and post-translation modifications

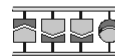Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   **8**
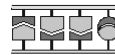
Replication

9

- □ Replication of DNA molecules.
  - ◘ The strands acts as a template for a new, complementary strand.
  - ◘ When copying is complete, there will be two daughter DNA molecules, each identical in sequence to the mother molecule.

- Template for replication

Daughter - 1

Daughter 2

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010

Genome Restructuring

10

- □ This operation modify the gene structure.
  - ◘ Introduce genetic diversity
- □ Similar to GA and GP, in GEP also, populations of individuals (computer programs) evolve by developing new abilities and becoming better adapted to the environment due to the genetic modifications accumulated over a certain number of generations.
  - ◘ Mutation
  - ◘ Recombination
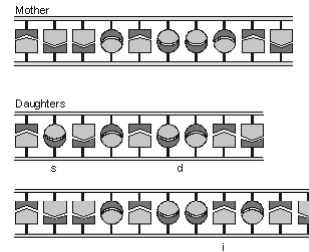 - ◘ Transposition
  - ◘ Gene duplication

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   **10**

Mutation Operation

11

- □ Basic mutation operations are:
  - ◘ Substitution (s)
  - ◘ Deletion (d)
  - ◘ Insertion (i)
- □ The base strand (mother) is modified using these three operations to generate daughters

Recombination

12

- □ Different recombination operators:
  - ◘ Homologous
  - ◘ Site specific
  - ◘ Non-homologous
- □ Transposition and genetic duplication
  - ◘ Transposition - genetic elements consist of genes that can move from place to place within the genome.
  - ◘ Genetic duplication - a gene is copied twice during replication.
    - ■ duplicated through the combined effects of gene transposition and recombination

Solution Representation

13

- □ Like in GP, GEP the chromosomes (solutions) are represented using function set and terminal set.
- □ In GEP – chromosomes are represented using genes
  - ◘ Genes – heads and tails
    - ■ Heads are coded using function and terminal set
    - ■ Tails are coded using only terminal set.
- □ Let F be function set, F = {*,+,-,Q}, where
  - ■ Q – square root
- □ Let T be terminal set, T = {a,b,c,d}

Solution Representation

14

- □ We want to represent the arithmetic expression
- □ Chromosome made of genes
- □ Max. of arguments for the elements in the Function set (n)
- □ Gene – head and tails
- □ Heads (h) are specified for a given problem
- □ Tails are calculated based on number of heads and

  t = h(n-1)+1
- □ Arithmetic expression - Gene Equivalent – K-Expression

$$\sqrt{(a+b)\times(c-d)},$$

which can also be represented as a diagram or ET:
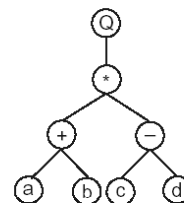
0	1	2	3	4	5	6	7
Q	*	+	-	a	b	c	d

Example

15

- K – expression

```
01234567890
Q*+*a*Qaaba
```

- Equivalent Tree is
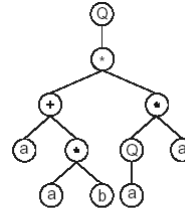
Representation…

16

- The structural organization of GEP genes is better understood in terms of open reading frames (ORFs).
- In biology, an ORF, or coding sequence of a gene, begins with the "start" codon, continues with the amino acid codons, and ends at a termination codon.
- However, a gene is more than the respective ORF, with sequences upstream from the start codon and sequences downstream from the stop codon.
- **Although in GEP the start site is always the first position of a gene, the termination point does not always coincide with the last position of a gene.**
- It is common for GEP genes to have non-coding regions downstream from the termination point.

Use of non-coding region

□ They are, in fact, the essence of GEP and evolvability, for they allow modification of the genome using any genetic operator without restrictions, always producing syntactically correct programs without the need for a complicated editing process or highly constrained ways of implementing genetic operators.

□ Indeed, this is the paramount difference between GEP and previous GP implementations, with or without linear genomes

Example with Head and Tail
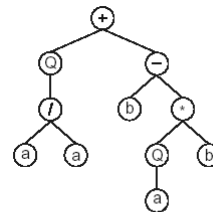
□ Function set : {Q,+,-,*,/}
□ Terminal set : {a,b}
□ Function arguments = 2
□ Head = 10
□ Tail = 10(2-1)+1=11
□ The K- expression

```
0123456789 01234567890
+Q-/b*aaQbaabaabbaaab
```

□ The bold face represent the tail.
□ Here ORF ends at 10, where as the gene end at 20.
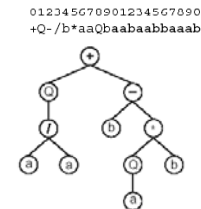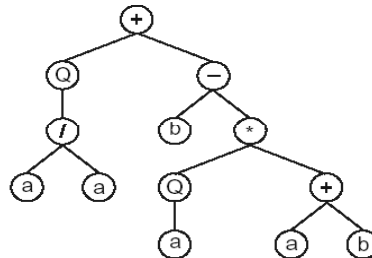□ ORF is phenotype representation

Use of Non-Coding Region

Suppose now a mutation occurred at position 9, changing the "b" into "+". Then the following gene is obtained:

```
0123456789012345 67890
+Q-/b*aaQ+aabaabbaaab
```

and its ET gives:

```
0123456709012345 67890
+Q-/b*aaQbaabaabbaaab
```


In this case, the termination point shifts two positions to the right (position 12).

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010

Use of Non-Coding Region

Suppose now that a more radical modification occurred, and the symbols at positions 6 and 7 in gene (3.5) change respectively into "+" and "*", creating the following gene:

```
0123456789012345 67890
+Q-/b*+*Qbaabaabbaaab        (3.7)
```

giving the ET:

In this case the termination point shifts several positions to the right (position 14).

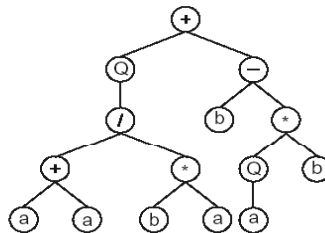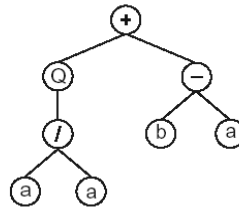Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010

Use of Non-Coding Region

21

Obviously the opposite also happens, and the ORF is shortened. For example, consider gene (3.5) and suppose a mutation occurred at position 5, changing the "*" into "a":

```
012345678901234567890
+Q-/baaaQbaabaabbaaab                    (3.8)
```

Its expression results in the following ET:

In this case, the ORF ends at position 7, shortening the original ET by 3 nodes.

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010

MultiGene Representation

22

□ We saw the representation for single gene representation.

□ Now, we discuss the multi-gene representation for the chromosome

□ Number of genes can be greater than one in a chromosome.

□ For all problem, number of genes and number of heads are fixed prior.

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   **22**

Three Gene Representation

23

- □ Three genes
- □ n = 2; h = 4
- □ K-Expressions for
  - ▫ Gene₁ :
  - ▫ Gene₂ :
  - ▫ Gene₃ :
- □ Position '0' is the start of the gene and position '8' is the end of the gene.
- □ The ORF ending of each gene can be calculated after tree construction.

```
012345678
-b*babbab
```

```
012345678
*Qb+abbba
```

```
012345678
-*Qabbaba
```

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010

Three Gene Representation
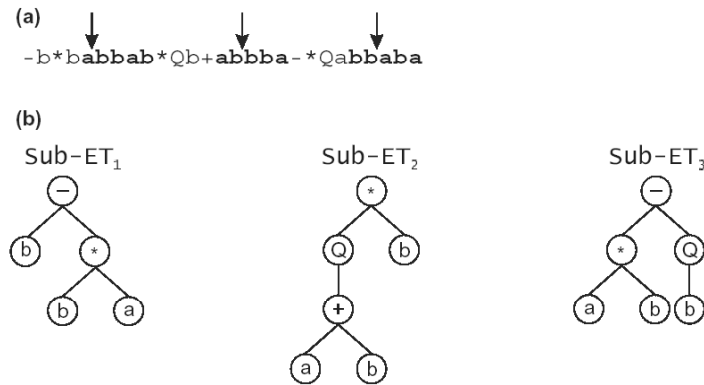
24

```
012345678012345678012345678
-b*babbab*Qb+abbba-*Qabbaba
```

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   **24**

Tree Construction

(a)

```
-b*babbab*Qb+abbba-*Qabbaba
```

(b)

Sub-ET₁   Sub-ET₂   Sub-ET₃

Figure 2. Expression of GEP genes as sub-ETs. **(a)** A three-genic chromosome with the tails shown in bold. The arrows show the termination point of each gene. **(b)** The sub-ETs codified by each gene.

Translation

- The process of converting the K-expressions into tree (ET) and reducing it to mathematical form is called Translation.
- GEP chromosomes are composed of one or more ORFs, and obviously the encoded individuals have different degrees of complexity.
- The simplest individuals are encoded in a single gene, and the iorganismî is, in this case, the product of a single gene - an ET.
- In other cases, the organism is a multi-subunit ET, in which the different sub-ETs are linked together by a particular function.
- In other cases, the organism emerges from the spatial organization of different sub-ETs (e.g., in planning and problems with multiple outputs).
- And, in yet other cases, the organism emerges from the interactions of conventional sub-ETs with different domains (e.g., neural networks). However, in all cases, the whole organism is encoded in a linear genome.

Linking Function

- □ In case of multi-gene representation, the genes are linked together to form a mathematical expression.
- □ The link functions are defined apriori.
- □ This function will not appear in the chromosome representation.
- □ Now, we show an example for better understanding.

Example 1

(a)
012345678012345678
Q*Q+**bbaaa***-ba**baabb**
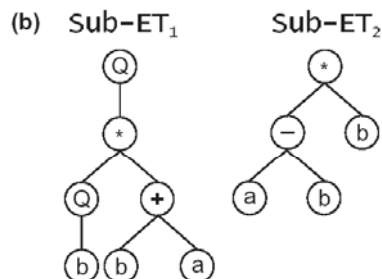
(b) Sub-ET₁   Sub-ET₂

Example with + Link

(c)   ET
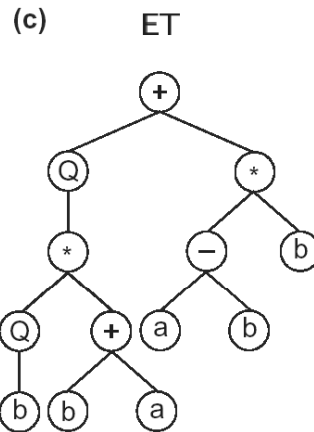
Analysis

- □ Can we represent the final tree with K-Expression?
  - ◘ Yes.
- □ What is the equivalent K-expression?

```
0123456789012
+Q**-bQ+abbba
```

- □ Issues?
  - ◘ It is difficult to use the genetic operators to evolve because of less number of tails
  - ◘ Multi-genes representation – faster convergence

Example 2 for Non-coding region in multi-gene

31

- Chromosome has two genes
- Head = 3 and tail = 4
- Operators: N – NOT, O – OR
- Fig a) represent the K expression
- Fig b) the first operator is the connecting operator
- In gene$_1$ – OOcacab – 'the last two character 'ab' belongs to non-coding region
- In gene$_2$ – NNNbbcb – the last three character 'bcb' belongs to non-coding region

a. 01234560123456
   OOcacabNNNbbcb

b.

Gene$_1$

Gene$_2$

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010
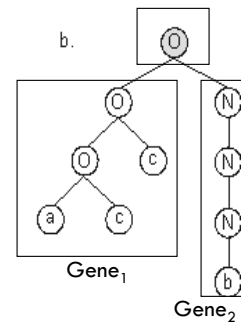
Example 3

32

(a)
IIAIca3aa2acuNNAOab2u3c31cAu12ua3112cac

(b) Sub-ET$_1$ Sub-ET$_2$ Sub-ET$_3$

(c)   ET
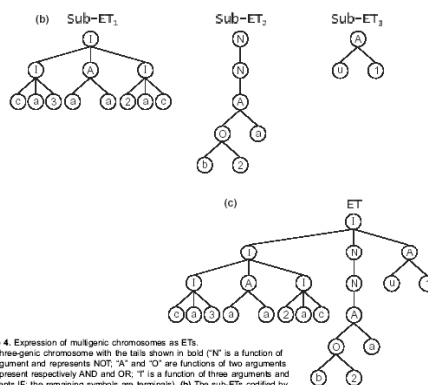
Figure 4. Expression of multigenic chromosomes as ETs.
(a) A three-genic chromosome with the tails shown in bold ("N" is a function of one argument and represents NOT; "A" and "O" are functions of two arguments and represent respectively AND and OR; "I" is a function of three arguments and represents IF; the remaining symbols are terminals). **(b)** The sub-ETs codified by each gene. **(c)** The result of posttranslational linking with IF.

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   **32**

Example 4

(a)
```
131u3ab2ubab23c3ua31a333au3
```
(b)
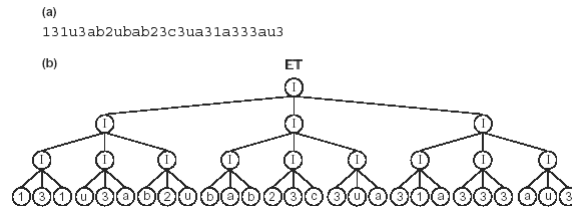
Figure 5. Expression of multigenic chromosomes as ETs. **(a)** A 27-genic chromosome composed of one-element genes. **(b)** The result of posttranslational linking with IF.

Points to Remember

- The type of linking function, as well as the number of genes and the length of each gene, are *a priori* chosen for each problem.
- So, we can always start by using a single gene chromosome, gradually increasing the length of the head; if it becomes very large, we can increase the number of genes and of course choose a function to link them.
- We can start with addition or OR, but in other cases another linking function might be more appropriate.
- The idea, of course, is to find a good solution, and GEP provides the means of finding one.

Mutation

35

- □ Mutations can occur anywhere in the chromosome. However, the structural organization of chromosomes must remain intact.
- □ In the heads any symbol can change into another (function or terminal); in the tails terminals can only change into terminals.
- □ This way, the structural organization of chromosomes is maintained, and all the new individuals produced by mutation are structurally correct programs.

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   35

Mutation – Mother genome

36

K – Expression – Equation 3.5
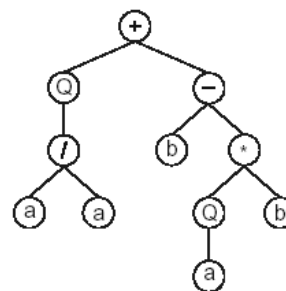
```
01234567890123 4567890
+Q-/b*aaQbaabaabbaaab
```

Equivalent ET-tree

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   36

18

Daughter Genome

Suppose now that a more radical modification occurred, and the symbols at positions 6 and 7 in gene (3.5) change respectively into "+" and "*", creating the following gene:

```
0123456789 01234567890
+Q-/b*+*Qbaabaabbaaab           (3.7)
```

giving the ET:
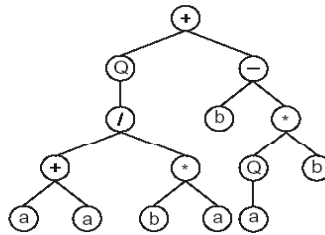
Mutation…

Typically, a mutation rate (p_m) equivalent to two point mutations per chromosome is used. Consider the following 3-genic chromosome:

```
012345678012345678012345678
-+-+abaaa/bb/ababb*Q*+aaaba
```

Suppose a mutation changed the element in position 0 in gene 1 to "Q"; the element in position 3 in gene 2 to "Q"; and the element in position 1 in gene 3 to "b", obtaining:

```
012345678012345678012345678
Q+-+abaaa/bbQababb*b*+aaaba
```

Neutral Mutation

□ Mutation occur in non-coding region is called neutral mutation. This mutation does not affect the ET of mother and daughter.

□ ORF ends at position 7 of the head

□ Suppose, mutation occur at tail position 9.

◘ Change 'a' to 'b'

◘ The 'phenotype' of daughter genome is same as mother.

```
012345678901234567890
+Q-/baaaQbaabaabbaaab
```

Its expression results in the following ET:

Workshop on bio-inspired Computing,
VTU, Mysore, 7-10, June, 2010

39
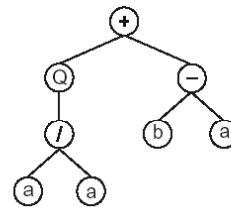
Comments on mutation

40

□ If a function is mutated into a terminal or *vice versa*, or a function of one argument is mutated into a function of two arguments or *vice versa*, the ET is modified drastically.

□ The change in tree size take place with-out increasing the computational complexity.

□ It is worth noticing that in GEP there are no constraints neither in the kind of mutation nor the number of mutations in a chromosome: in all cases the newly created individuals are syntactically correct programs.

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010

40

Comments on mutation…

41

- □ In nature, a point mutation in the sequence of a gene can slightly change the structure of the protein or not change it at all, as neutral mutations are fairly frequent (e.g., mutations in introns, mutations that result in the same amino acid due to the redundancy of the genetic code, etc.).
- □ In GEP also neutral mutations exist (e.g., mutations in the non-coding regions).
- □ A mutation in the coding sequence of a gene has a much more profound effect: it usually drastically reshapes the ET.

Transposition and insertion sequence elements

42

- □ The transposable elements of GEP are fragments of the genome that can be activated and jump to another place in the chromosome.
  - ◘ A) Short fragments with a function or terminal in the first position that transpose to the head of genes, except to the root (insertion sequence elements or IS elements).
  - ◘ B) Short fragments with a function in the first position that transpose to the root of genes (root IS elements or RIS elements).
  - ◘ C) Entire genes that transpose to the beginning of chromosomes.
- □ This operator will be useful only when the chromosome has more than one gene.

A) Transposition…

43

- Any sequence in the genome might become an IS element, therefore these elements are randomly selected throughout the chromosome.
- A copy of the transposition is made and inserted at any position in the head of a gene, except at the start position.
- Typically, an IS transposition rate (p_{is}) of 0.1 and a set of three IS elements of different length are used.
- The transposition operator randomly chooses the chromosome, the start of the IS element, the target site, and the length of the transposition.

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   **43**

A) Transposition…

- Suppose that the sequence "bba" in gene 2 (positions 12 through 14) was chosen to be an IS element, and the target site was bond 6 in gene 1 (between positions 5 and 6).
- Then, a cut is made in bond 6 and the block "bba" is copied into the site of insertion.
- During transposition, the sequence upstream from the insertion site stays unchanged, whereas the sequence downstream from the copied IS element loses, at the end of the head, as many symbols as the length of the IS element (in this case the sequence "a*b" was deleted).

```
0123456789012345678900123456789012345678
*-+*a-+a*bbabbaabababQ**+abQbb*aabbaaaabba   --Mother
```

```
0123456789012345678900123456789012345678
*-+*a-bba+babbaabababQ**+abQbb*aabbaaaabba   -- Daughter
```

Workshop on bio-inspired Computing,
VTU, Mysore, 7-10, June, 2010   **44**

B) Root Transposition

45

- □ All RIS elements start with a function, and thus are chosen among the sequences of the heads. For that, a point is randomly chosen in the head and the gene is scanned downstream until a function is found.
- □ This function becomes the start position of the RIS element.
- □ If no functions are found, it does nothing.
- □ Typically a root transposition rate (*pris*) of 0.1 and a set of three RIS elements of different sizes are used.
- □ This operator randomly chooses the chromosome, the gene to be modified, the start of the RIS element, and its length.

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   **45**

Root...

46

Consider the following 2-genic chromosome:

```
012345678901234567890 012345678901234567890
-ba*+-+-Q/abababbbaaa Q*b/+bbabbaaaaaaaabbb
```

Suppose that the sequence "+bb" in gene 2 was chosen to be an RIS element. Then, a copy of the transposon is made into the root of the gene, obtaining:

```
012345678901234567890 012345678901234567890
-ba*+-+-Q/abababbbaaa +bbQ*b/+bbaaaaaaaabbb
```

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   **46**

Root…

47

- □ During root transposition, the whole head shifts to accommodate the RIS element, losing, at the same time, the last symbols of the head (as many as the transposon length).
- □ As with IS elements, the tail of the gene subjected to transposition and all nearby genes stay unchanged.
- □ Note, again, that the newly created programs are syntactically correct because the structural organization of the chromosome is maintained.
- □ The modifications caused by root transposition are extremely radical, because the root itself is modified.
- □ In nature, if a transposable element is inserted at the beginning of the coding sequence of a gene, causing a frameshift mutation, it radically changes the encoded protein.
- □ Like mutation and IS transposition, root insertion has a tremendous transforming power and is excellent for creating genetic variation.

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   **47**

c) Gene Transposition

48

- □ In gene transposition an entire gene functions as a transposon and transposes itself to the beginning of the chromosome.

- □ In contrast to the other forms of transposition, in gene transposition the transposition (the gene) is deleted in the place of origin.

- □ This way, the length of the chromosome is maintained.

- □ The chromosome to undergo gene transposition is randomly chosen, and one of its genes (except the first, obviously) is randomly chosen to transpose.

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   **48**

Gene…

Consider the following chromosome composed of 3 genes:

```
012345678012345678012345678
*a-*abbab-QQ/aaabbQ+abababb
```

Suppose gene 2 was chosen to undergo gene transposition.
Then the following chromosome is obtained:

```
012345678012345678012345678
-QQ/aaabb*a-*abbabQ+abababb
```

Recombination operator

□ Three type of recombination operator
- ◘ One-point
- ◘ Two-point
- ◘ Gene

□ Two parents are randomly chosen and paired to exchange the genetic material between them.

One-point

51

During one-point recombination, the chromosomes cross over a randomly chosen point to form two daughter chromosomes. Consider the following parent chromosomes:

```
012345678012345678
-b+Qbbabb/aQbbbaab
/-a/ababb-ba-abaaa
```

Suppose bond 3 in gene 1 (between positions 2 and 3) was randomly chosen as the crossover point. Then, the paired chromosomes are cut at this bond, and exchange between them the material downstream from the crossover point, forming the offspring below:

```
012345678012345678
-b+/ababb-ba-abaaa
/-aQbbabb/aQbbbaab
```

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   51

Two-point

52

□ In two-point recombination the chromosomes are paired and the two points of recombination are randomly chosen.

□ The material between the recombination points is afterwards exchanged between the two chromosomes, forming two new daughter chromosomes.

Workshop on bio-inspired Computing, VTU, Mysore, 7-10, June, 2010   52

Two-point…

53

Consider the following parent chromosomes

```
0123456789001234567890
+*a*bbcccac*baQ*acabab- [1]
*cbb+cccbcc++**bacbaab- [2]
```

Suppose bond 7 in gene 1 (between positions 6 and 7) and bond 3 in gene 2 (between positions 2 and 3) were chosen as the crossover points. Then, the paired chromosomes are cut at these bonds, and exchange the material between the crossover points, forming the offspring below:

```
0123456789001234567890
+*a*bbccbcc++*Q*acabab- [3]
*cbb+ccccac*ba*bacbaab- [4]
```

Gene Recombination

54

In gene recombination an entire gene is exchanged during crossover. The exchanged genes are randomly chosen and occupy the same position in the parent chromosomes. Consider the following parent chromosomes:

```
012345678012345678012345678
/aa-abaaa/a*bbaaab/Q*+aaaab
/-*/abbabQ+aQbabaa-Q/Qbaaba
```

Suppose gene 2 was chosen to be exchanged. In this case the following offspring is formed:

```
012345678012345678012345678
/aa-abaaaQ+aQbabaa/Q*+aaaab
/-*/abbab/a*bbaaab-Q/Qbaaba
```

Observation

55