A new evolutionary computation technique for 2D Morphogenesis and information processing

ENRIQUE FERNANDEZ-BLANCO, JULIAN DORADO, JUAN RAMON RABUNAL, MARCOS GESTAL, NIEVES PEDREIRA
{efernandez, julian, juanra, mgestal, nieves}@udc.es sabia.tic.udc.es
Dept. Information and Communication Technologies, University of A Coruña
Facultade de Informatica Campus Elviña s/n 15071 A Coruña, Spain

Abstract: - Paper presents a new model that takes the behaviour of biological cells and tries to adapt some of their characteristics to the artificial cells in order to solve computational problems. This model can be related to the area known as Computational Embryology. Besides the theoretical approach, some of the tests that were performed as preliminary implementation of such model are also presented. This test consists into obtain simple forms are generated using the development of the cell model.

Key-Words: -Evolutionary Computation, Computational Embryology, Genetic Algorithm

1 Introduction

The biological cells of a given organism are able to build complex structures from a unique cell, known as zygote, with no need of a centralised control. The cells can perform such process due to existence of a general plan encoded at the DNA for the development as well as for the functioning of the system. Another interesting characteristic of natural cells involves the system that they build being tolerant to partial failures, small errors not to induce a global collapse of the system. Lastly, the tissues created by biological cells have, in every single cell of them, parallel processing of the information for the coordination of tissue functioning.

All the mentioned characteristics are quite interesting from the computational viewpoint. This paper presents the development of a model that tries to emulate the biological cells and to take advantage of some of their characteristics, trying to adapt them to artificial cells. This model can be set within a group of techniques known as *Computational Embryology* [10].

2 State of the art

The models into the area that was named as Computational Embryology by Kumar [10], have been emerging at the Evolutionary Computation (EC) area. Such group includes the whole of the models which try to adapt certain characteristics of biological embryonic cells to computer problem solving. These characteristics are self-organisation, failure tolerance and parallel processing of the information.

The work performed by the scientific community can be divided into two branches. The most theoretical branch deals with the proposal of a model that might work as a natural cell. It mainly focuses on the emulation of cell capabilities such as cellular differentiation and metabolism [8].The finally of this works is to do a better study of the biological model.

The other branch, considered the most practical one[1, 9, 10], focuses mainly on the development of a cell inspired-model that might be applicable to other problems such as generative encoding for building artificial organisms in simulated physical environments and real robots, robot control or development for evolutionary design of hardware and circuits [5, 11].

The model proposed in this paper is related to a few previous works. The first one was presented in 2004 by Kumar [10]. Other significant models were those developed by Eggenberger [4] and Dellaert & Beer [3]. These previous models, together with the one presented in this paper, should be not be mistaken for close concepts such as cellular automatons or genetic expression programming [2].

3 Main ideas

The cells of a biological system are mainly determined by the DNA strand, the genes and the proteins. The DNA is the structure that keeps the gene encoded information needed for the

development of the system. The genes are activated or transcribed after the protein shaped-information that exists at the cytoplasm. The genes have two main portions: one is the sequence, which identifies the protein that will be generated if the gene is transcribed; the promoter portion identifies the proteins needed into the cytoplasm for gene transcription.

Another remarkable aspect of biological genes is the difference between constitutive genes and regulating ones. The later ones are transcribed only when the proteins identified at the promoter region are present. The constitutive genes are always transcribed unless inhibited by the whole of the proteins of the proteins identified at the promoter portion, acting then as gene oppressors.

4 Proposed model

The model of artificial cellular system proposed in this article bases its functioning on the interaction of artificial cells using messages called proteins. The cells can divide themselves, die or generate proteins that will act as messages for the own cell and for the neighbourhood ones.

It is intended for the system to express a global performance for, initially, generate 2D structures. Such behaviour would emerge from the information encoded at a certain variables group of the cell, which analogically with the biological cells, would be named as genes.

4.1 Model structure

The central element of the model is the artificial cell. Every cell has a binary string encoded-information for the regulation of its functioning. Such string is named DNA, following the biological analogy. The cell also has a structure for the storage and management of the proteins generated by the own cell as well as the ones received from neighbourhood cells. This storage structure is called cytoplasm as the biological element.

4.1.1 DNA

The DNA of the artificial cell consists on functional units named genes. Every gene encodes a protein or message (which produces the gene).

The structure of a gene has four parts (Fig. 2):

- Sequence, is the binary string which corresponds to the protein that encodes the gene
- Promoters, is the gene area that indicates which are the proteins needed for its transcription.
- Constituent, this bit identifies if the gene is constituent and it is meaning to the model
- Activation percentage (binary value), is the percentage of minimal concentration inside the cell of promoters proteins that causes the transcription of the protein that the gene encodes.

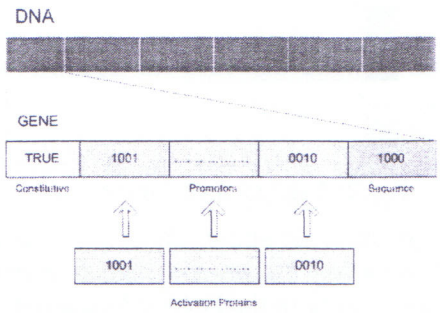

Fig. 1: Structure of a system gene

4.1.2 Cytoplasm

Other fundamental element for keeping and managing the proteins received or produced by the artificial cell is the cytoplasm. The stored proteins have a certain life time before being erased. At the cytoplasm it is checked which are the proteins, and their concentration, needed for the activation of the DNA genes; therefore the cytoplasm will answer all the cellular requirements for the concentration of a given type of protein among other ones. The cytoplasm will also extract the proteins from the structure in case they might be needed for a gene transcription.

4.2 Boolean algebra

Using this gene and protein structure, an operation with Boolean algebra-like structure might be defined. The space for the definition of the operations would be the presence or absence of certain proteins into the system, whereas the operation result would be the protein contained/encoded at the gene. The AND operation (see Fig.2) would be modelled with a gene that would need for its expression all the proteins of its

promoters. The OR operation would be modelled with two genes that, despite their different promoters, result in the same protein. Finally, the NOT operation would be modelled with the constituent part, which changes the performance of that gene. The presence of proteins belonging to the promoters would imply the absence of the gene resulting protein at the system. This behaviour is similar to the gene regulatory networks [7]

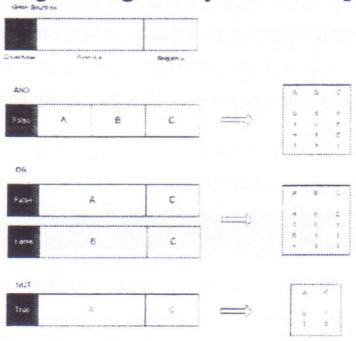

Fig. 2: Logical operators match

This operation set allows the cell to process the information, using the protein shaped stimuli that receives together with the DNA information.

4.3 Model functioning

The cellular cycles are established for the management of the different events occurring at a cellular tissue. These cycles will contain all the actions that might be performed by the cells and therefore restricting, in some cases, the occurrence of these actions. The model cells, during a cellular cycle, could be divided once and also every one of the genes could be expressed only once; however, and for instance, these cells could communicate several times with any of their neighbours.

The functioning of genes is determined by their type: constituent or non constituent. The transcription of the encoded protein occurs when the promoters of the non constituent genes appear in a certain rate at the cellular cytoplasm. Contrarily, the constituent genes are expressed during all the cycles until such expression is inhibited by certain rate of the promoter genes. It should be noticed that the proteins that connect with the promoter side might either be the same or bear certain resemblance. In order to control this similarity, the following equation was used:

$$\text{Pr}\,oteinConcentrationPercentage >= (Distance + 1) * ActivationPercentage \quad (1)$$

This activation plan is quite similar to the biological one and also provides the model with an increased flexibility. The gene transcription is then achieved if the condition expressed by Eq.1 is fulfilled, where *ProteinConcentrationPercentage* represents the cytoplasm concentration of the protein that is being considered; *Distance* stands for the Hamming distance between one promoter and the protein considered; finally, *ActivationPercentage* is the minimal percentage needed for the gene activation that is encoded at the gene. This equation is proved on each of the promoters and proteins. If the condition is fulfilled for all of the promoters, that gene is transcribed

After the activation of one of the genes, there are four possibilities. Generated protein might be either to be stored at the cell cytoplasm or to be communicated to the neighbour cells, or it might induce cellular division (mitosis) or death (apoptosis). Three threshold values were defined at the cellular system proposed here in order to modulate these four possibilities. Two of those values will manage the communications, and the third one will regulate the system apoptosis. When a cellular gene produces a protein, the concentration of the protein is checked to be less that the value of the first communication threshold (lower communication threshold). If that is case, the protein will be kept within the cell cytoplasm. If the concentration is higher than the first value but lower than the second communication value (upper communication threshold) the cells sends that protein to one of its randomly selected neighbour ones. If concentration exceeds the second threshold the protein is sent to all the neighbour cells. The third implemented threshold is the apoptosis threshold and it determines the protein concentration needed for inducing cellular death. Finally, it should be noted that the mitosis and apoptosis procedures are executed when the cell generates certain types of proteins that have been identified as special ones. Five special proteins have been identified at the system: apoptosis, upwards mitosis, downwards mitosis, leftwards mitosis and rightwards mitosis.

The cells will express the performance encoded at the DNA. During the first stage of the cellular functioning, it is checked the existence of cytoplasm proteins that might affect to the promoters of the different DNA genes. Afterwards, and bearing in mind the previous information, it should be determined which are the genes that are going to be transcribed. The transcription will be performed after the recovery of needed proteins from the cytoplasm and then, the proteins encoded

at the selected genes will be generated. Once the genes are transcribed with the generated proteins, the cellular reactions to be performed will be also determined; this means that every protein can either be communicated to a neighbour cell, stored at the cytoplasm or to perform a special action (mitosis or apoptosis). The cytoplasm performance during every cycle of the system can be observed at the following pseudo-code, where P_x identifies the number of x type-proteins and TTL_y represents the lifetime of y type-protein:

```
FOR every Protein at the cytoplasm do
  TTL_j = TTL_j - 1;
  IF (TTL_j = 0)
    P_i = P_i - 1;     protein j ∈ Type i
  ENDIF
ENDFOR

WHILE no New Cycle DO

  FOR each genes promoter
    Return %P_a = (P_a / ∑ P_b) * 100
     a, b ∈ 1,N
  ENDFOR

  IF Requirement a type protein AND P_a > 0
    P_a = P_a - 1;
  ENDIF
ENDDO
```

According this pseudo-code, the cytoplasm firstly actualises the lifetime of its stored proteins and it later returns the information about the different concentrations of the N proteins of the system. It should be borne in mind that the concentration is used for the calculation of genes activation. The cytoplasm concentration of proteins decreases when they are required for genes transcription. When all the possible actions of the cellular cycle finish, the process will restart by means of the actualisation of the cytoplasm with those proteins identified for storage and the ones received after the communications with neighbour cells during the previous cycle.

4.4 Genome evaluation

After the establishment of the functioning mode for the artificial cell, the following step is to find a method for achieving the best genes configuration for the solution of every problem. A classical approach of EC proposes the use of GA [6] for optimisation, in this case, of the values of the DNA genes (binary strands). Every individual of the GA population will represent a possible DNA strand for problem solving.

For the calculation of the fitness value for every individual in the GA or the DNA, the strand will be introduced into an initial cell or zygote. After simulation during certain number cycles, the contained information will be then expressed and the characteristics of the resulting tissue will be evaluated using different criteria according to the goal to be achieved.

The encoding of the individual genes follows a structure as the described below:

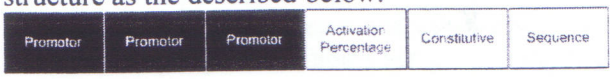

Fig. 3: GA gene structure

This structure presents all the parts previously enumerated and explained at the paragraph 4.1.1. As it can be seen, the gene activation percentage, a value that is needed to be adjusted for gene performance at the system (Eq. 1), has been also specified. All in all, the DNA strands will be strands with encoded genes, structured as previously seen. At the system used, the number of promoters of every gene might vary but the white and indivisible section of the Fig. 3, known as section "Activation Percentage – Constituent – Sequence" (PCS), must always be present. Being indivisible is a characteristic that is quite relevant during the crossover operation that is following described. The PCS sections determine the genes of the individual and the promoter sections get associated with the PCS themselves, as Fig. 4 shows. Such fact provides the GA with a wide range of genes combinations in order to achieve the solution.

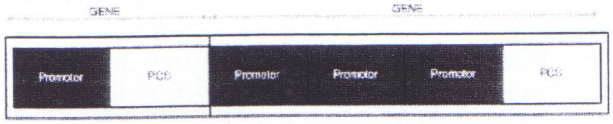

Fig. 4: Example of GA genes for encoding cellular genes

The search of a set of structures similar to those shown at Fig. 4 was performed by adapting the crossover and mutation GA operations to this specific problem.

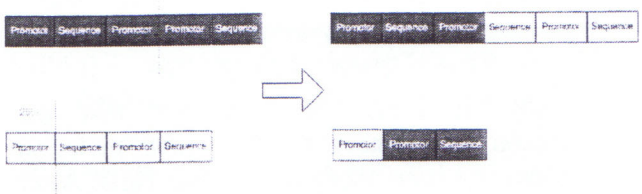

Fig. 5: Crossover example

Due to the variable length of the individuals, the crossover had to be performed according to these lengths. When an individual is selected, a random

percentage is generated for determining the crossover point of that individual; for instance, at Fig. 5 the 60% of length has been selected as crossover point, defining length as the number of sections that compose the strand. The indivisible PCS section accounts as length 1, as well as the promoter sections that cannot be divided by intermediate points of the strands. Once this has been done, the crossover point selection process is also repeated at the second selected parent (25% at the figure). From this stage, the descendants will be composed using the traditional way, by mixing the characteristics of the parents. With this type of crossover, the number of promoters of the genes might vary drastically and therefore, an alteration of the subsequent performance might be noticed when evaluating the new individuals. Moreover, one of the effects of this type of crossover is the absence of promoters for a given gene. If such fact occurs, the gene neither would be ever expressed nor, if it is constituent, it could be inhibited.

With regards to mutation, it should be mentioned that the types of the promoter or PCS sections are identified according to the value of the first strand bit. Bearing that in mind, together with the variable length of individuals, the mutation operation had to be adapted to for modifying , not only the number of these sections, but also the value of a given section. At the 20% of chances, one of the sections (promoter or PCS) will be added to a certain selected position. At other 20% an existing section placed at a selected position will be discarded. The remaining 60% mutation changes, the value of one bit of a randomly selected section will be varied at a previously selected position. The later, might induce the change, not only of one of the values, but also the type of section if the varied bit is the one that identifies such type. For instance, if a promoter section is changed into a PCS section, the promoter sequence will be the gene sequence; besides, they are also generated values for the constituent and for the activation percentage, as it shows Fig.6.

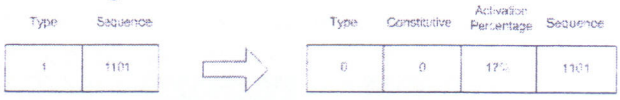

Fig. 6: Example of change type

Finally, it should be mentioned that the distribution of chances for the three mutation operations has been obtained on a trial basis. After several tests, the most suitable values were determined.

5 Results

In the present section some of the tests performed with the proposed model will be described. Some of the problems for model evaluation are related to generation of simple structures. In every test the reasons and the parameters configuration, as well as the results obtained, are described.

During the test, the divisions of the cells were restricted to the Cartesian axis: up (North), down (South), left (West) and right (East). The proteins that induce these divisions were established in 1000, 1100, 1010 and 1110 respectively.

Experiments were performed for the selection of the sequences that were identified as division or apoptosis ones. Both types of sequences, randomly chosen and with the larger Hamming distance possible, were tested. The results proved that the specification of those sequences has no influence on the developments of the tests.

The protein lifetime parameter was fixed in 3 cycles. Such value was obtained after diverse empirical tests developed during 50 cycles, as this is time enough for protein to be communicated and used but avoiding the saturation of the system if the protein is not used eventually.

Also, after several tests, it was determined that the most suitable GA operators were roulette selection, parent replacement and one point crossover.

5.1 5x5 Square

In this problem it is intended to achieve, from a unique cell or zygote, a square cellular structure, whose cell population might be kept stable after reaching that shape.

After several trials, the system parameters which provided the best solution were:

- The environment (20x20 matrix), doesn't insert messages into the developing cells.
- The protein 0000 was chosen as apoptosis signal. Apoptosis threshold: 10%.
- The lower-upper communication thresholds were fixed in 5%-10% respectively.
- The GA parameters for a population of 100 individuals were 90% crossover and 5% mutation on a 100-individual population.
- Finally, 50 cycles were allowed to develop before the tissue evaluation.

Also the genes were restricted to have up to 16 different proteins and 2 at the promoters, therefore 4 bit strings are used for protein encoding

Lastly, the minimisation of the Eq. 2 was fixed as evaluation criterion:

$$|cell\_number - 25| + |(X\,maximum - X\,minimum) - (Y\,maximum - Y\,minimum)| + |4 - X\,maximum + X\,minimum| + |4 - Y\,maximum - Y\,minimum| \quad (2)$$

Every one of the terms of the evaluation function tries to support those individuals that might fulfil the specified requirements. The first addend intends that the cell number might be 25. The following two, try to achieve the square shape and penalise the fact that length and the width might not be equal. The two last addends try that the cells might be concentrated into a 5x5 space.

With the previously described configuration and after 100 GA generations, the best DNA achieved is shown at Table 1(see at the end of the article). In that table, and for a better understanding, the Active Portion, Protein Sequence and Activation Percentage columns are shown using decimal values instead of the binary ones that encode them at the DNA. This best DNA has 2.0 error or fitness value, achieving a stable population close to the desired square shape (see Fig 7 with the growing from 10 to 30 cycles). After this point, the individual was allowed to remain more cycles, resulting in a stable tissue.

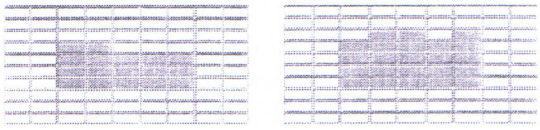

Fig. 7: The tissue after 10 cycles (left) and its stabilisation after 30 cycles (right)

The reason because the configuration shown at the previously mentioned table solves the problem is extracted from a detailed analysis of Table 1, as it is shown as follows.

The first achieve is that genes 7, 8, 11, 13 and 16 contain information that will not be ever expressed. The later is due to the fact that they, not only are not constitutive genes but also do not have promoters to be activated.

The gene 14 is quite difficult to activate, as it needs a very high protein concentration at its promoters, being such circumstance almost impossible.

The genes 2, 4 and 9 are identified as the constituent of the system, so they will express until the proteins of their promoters inhibit this expression. Genes 2 and 4 are really the same but they have different activation level; therefore, in absence of protein 3, both genes will be expressed. If the later protein is present, it is quite probable that only gene 2 might be expressed, however, none of them both will be expressed if the presence of the protein is highly remarked. The production of protein 9 by genes 2 and 4 can induce the deactivation of gene 4 as, although hamming distance between proteins 3 (that inhibits gene 4) and 9 is large, the activation threshold is low and therefore the activation will be enabled by means of the previously presented equation.

If protein 9 and 2 are present at the system, also can be activated the genes 1 (directly), 5,6,10 and 15 (by promoters similar to the ones of proteins 2, 0, 1 and 3). Nevertheless, the threshold of gene 15 is so high that will not allow it to activate. The expression of gene 5 will induce cellular division towards west, and the activation of gene 3 will do the same towards east. The growth towards north is induced by activation of gene 12 due to proteins similar to 14, which is not directly generated.

The growth of the cellular system stops at certain stage because, as cycles progress, it is more difficult to achieve high protein concentrations of the promoters for the activation of growing genes.

5.1.1 5x5 Square with obstacles

Starting from the satisfactory results of the previous test, other tests are performed in order to evaluate the flexibility of the DNA solution that was found. To achieve some obstacles are placed in the environment. Such obstacles are certain boxes that can not be filled by the cells. The results obtained are shown at Fig. 8.

Fig. 8: Development with a large blockade and development with partial blockades

As it can be observed in this figure, in both types of blockades, the tissue development is affected by the obstacles placed in the environment.

The cells try to solve the problem and tend to build the square trying to avoid the obstacles. The two obtained solutions take up an approximately 5x5 box square area, although they do not use all the possible boxes of such area and they only use 14 cells. During the GA DNA search no obstacles were used to reach the DNA fitness so this proves that flexibility is inherent in the model used.

Also it should be mentioned that different simulations performed with the same individual

generate slightly different results. This fact is due to the communication pattern, as the proteins produced between the two activation thresholds are expelled in a randomly selected direction.

5.2 Cross

Another test consisted on starting from a single cell in order to achieve a cross formed by cells on the centre. The parameters used at this test were the following ones:

- The GA population contained 1000 individuals
- 1000 generations were performed.
- GA evaluation criterion:

$$|9 - cell\_number| + |4 - height| + |4 - width| + |height - width| + |center - centroid| \quad (3)$$

This criterion is penalized adding one value for every cell not well-position. This evaluation criterion was added too, a 200 penalisation factor that misevaluates the solutions with no-cell. This penalisation tries to remove the solutions that might induce an excessive cell loss.

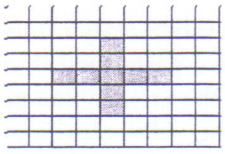

Fig 9: Tissue after 50 development cycles

6 Conclusions

Taking into account the model developed here, it can be stated that the use of certain properties of the biological cellular systems is feasible for the creation of artificial structures that might be used in order to solve certain computational problems.

Once simple tests were performed using the implemented model, the difficulty for developing GA evaluation functions is quite obvious. Such functions are the key for the development of solutions to different problems, being their creation crucial for the applicability of the system. With all that has been said, it can be stated that the results obtained are very promising but a lot of work still remains to be done for this type of systems to be developed in the next future.

7 Future developments

There are certain improvements that should be applied to the developed model and whose introduction could be attempted. Firstly, it would be interesting to develop at the cells a specialisation operator. Such operator, at certain level of development, would block certain genes of the cell and also of its future descendants. In that way, a complex structure could be approached by means of the growth of different simpler structures.

The promoter portion should be also approached by modifying the current protein sequence-functioning and using a more general proposal where proteins are overlapping sub-sequences of the whole promoter portion. The expression of a gene would occur after proteins available at the cell have completely covered the promoter portion by overlapping it.

8 Acknowledgement

This work was supported in part by the Spanish Ministry of Education and Culture (Ref TIC2003-07593, TIN2006-13274), the IMBIOMED network (Ref G03/160, PI052048) financed by the Carlos III Health Institute, the Regional European Development Funds (FEDER) program (Ref. PROLIT/SP1.E.199/03)and grants from the General Directorate of Research of the Xunta de Galicia (Ref. PGIDIT03-PXIC10504PN PGIDIT04-PXIC10503PN, PGIDIT04-PXIC10504PN)

References:

[1] Bentley, P.J. Digital Biology. Simon and Schuster, New York, 2002

[2]Ferreira C., Gene Expression Programming: A new adaptive algorithm for solving problems. CoRR 2001, 2001

[3] Dellaert F., Beer R.D., A Developmental Model for the Evolution of Complete Autonomous Agent, From animals to animats: Proceedings of the Forth International Conference on Simulation of Adaptive Behavior, Massachusetts, 1996, pp. 394-401, MIT Press.

[4] Eggenberger P., Cell Interactions as a Control Tool of Developmental Processes for Evolutionary Robotics, From animals to animats: Proceedings of the Forth International Conference on Simulation of Adaptive

Behavior, Massachusetts, 1996, pp. 440-448, MIT Press.

[5] Endo K., Maeno T., Kitano H: Co-evolution of morphology and walking pattern of biped humanoid robot using evolutionary computation -designing the real robot. ICRA, 2003, pp. 1362-1367

[6] Holland, J.H., Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor, MA, USA, 1975.

[7] Kauffman, S.A. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets, Journal of Theoretical Biology 22: 437-467

[8] Kitano, H. et al., Using process diagrams for the graphical representation of biological networks, Nature Biotechnology,Vol.8, No. 23, 2005, pp. 961 - 966

[9] Kumar, S. and Bentley P.J. (Eds), On Growth, Form and Computer, Academic Press, London UK, 2003.

[10] Kumar, S., Investigating Computational Models of Development for the Construction of Shape and Form. PhD Thesis. Department of Computer Science, University College London, 2004.

[11] Tufte, G. and Haddow, P. C., Towards Development on a Silicon-based Cellular Computing Machine, Natural Computing, Vol. 4, No. 4, 2005, pp.387-416

Gen Number	Promoter	Protein Sequence	Activation Percentage	Constitutive	Special Behaviour
1	2	6	13.5990	False	
2	3	9	59.2441	True	
3	7	7	36.8929	False	Grow E
4	3	9	1.4519	True	
5	1	3	39.8667	False	Grow W
6	1	6	16.6105	False	
7		12	67.7347	False	
8		1	14.8576	False	Grow S
9	13	2	58.6701	True	
10	0	6	42.2078	False	
11		4	0.3101	False	
12	14	5	23.8562	False	Grow N
13		10	56.1970	False	
14	5	12	91.6035	False	
15	3	3	65.6439	False	Grow W
16		6	53.9346	False	

Table 1: Square Problem Gene Configuration.