Journal of Theoretical and Applied Information Technology
15th November 2011. Vol. 33 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SEARCH FOR CYTOLOGY PATTERNS USING THE GENETIC ISLAND MODEL

JESÚS ALVAREZ-CEDILLO[1], ISRAEL RIVERA-ZARATE[1],JUAN C. HERRERA-LOZADA[1]
[1] Instituto Politécnico Nacional, Centro de Innovación y Desarrollo Tecnológico en Cómputo, Dpto. de Posgrado, U. P. Adolfo López Mateos, Av. Juan de Dios Bátiz s/n casi esq. Miguel Othón de Mendizábal, Edif. del CIDETEC, Col. Nva. Industrial Vallejo, Del. Gustavo A. Madero, 07700, México, D. F.

E-mail: {jaalvarez, irivera, jlozada}@ipn.mx

ABSTRACT

Island model deals with a species which is subdivided into a number of discrete finite populations, races or subspecies, between which some migration process occurs. If the number of populations is small, an assumption of equal rates of migration between each pair of populations may be reasonable approximation. In this paper we shown a general solution is given for the process of population divergence under this model following subdivision of a single parental population, expressed in terms of the observed average frequency of cells and explains how to find a zone in an image that is representative and small enough to optimize the later analysis of the image in general whether manual or automatic. This diagnostic method will analyze the behavior of a group of special cells (could be Cancer).

Keywords: *Cytology Patterns, Genetic Island Model, Population Divergence, Diagnostic Method*

1. INTRODUCTION

The techniques of image analysis are distinguished from others by its low cost and the low impact caused to the patient not having to suffer physical body surgery. Imageology or image study has various specialized disciplines to obtain analysis: tomography, x-rays, ultrasound, cytology and others [1, 2].

Cytology is a technique used to study the anatomical phenomenon of cells, normally through a microscope [3]. An image can have so many cells that the quality of the analysis can be affected without a strategy to direct analysis force to a specific zone of the image.

The possibility to use genetic algorithm techniques to find this zone is explained, also parallelization techniques to give best results in time and quality.

After analyzing some development platforms, one is chosen and simple functional prototype is implemented to search this zone and compare the results with both sequence and parallel methods.

2. BASIC CONCEPTS.

One of the greatest challenges of medical science has been the early detection and diagnosis of grave illnesses such as cancer. Throughout medical history many technological elements have been incorporated to comply with this and other challenges. The technology information or it have been, in the last decades, the principal providers of new solutions to old problems; furthermore creating new study fields where informatics, medicine and genetics converge to establish innovative lines of investigation and development. Theoretical and practical applications to improve the quality of life of the people.

The use of information technology for the detection and diagnosis of cancer has as the principal objective to identify as early as possible the appearance of a cancerous manifestation, still more the possibility of a person has the potential to develop a certain type of cancer at some the stage of his life is examined.

Cytology or study of cell characteristics of sample is a very versatile technique as it can detect and diagnose different types of cells based on a simple

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

principal: morphological differences that are visible under a microscope.

Cells are complex elements composed of various corpuscles and sections that can be distinguished and should have certain characteristics of size, form, density and proportion as show in figure 1. [4, 5]

Figure 1 Normal cell cytology view

Cytology experts estimate that a sample can have from only hundreds of cells up to 5000005. These large quantities are not processed in an adequate form by specialist or by computation given the large concentration of cell features shows that a cytology could analyze various hundreds per second, turning this type of analysis in a form of luck more than a predictable.

Specialist show this is like finding the difference between cars whilst in an airplane at a height of 984 meters traveling at twice the speed of sound visualizing an area of 228 x 0.72 square kilometers.

This is where an opportunity for genetic algorithms and parallel processing is found: helping to find the most representative zone of an image to be processed afterwards by a specialist or by other automatic software.

2.1 GENETIC ALGORITHMS.

The sequential search for the most representative rectangle of an image increases as the image and the number of cells in it increase. If an image of 256 x 256 could contain up to 65535 originals and 65535 sizes within which fit 65536 cells, the solution space is 655353 [6].

However genetic algorithms are a good alternative to find in an adequate range of iterations a reasonable solution to this problem thanks to its capacity to resolve search problems, it's flexibility to adapt to any type of problem however unknown.

2.1.1 CROMOSOMATIC CODING.

As the intention is find the best rectangle imaginable than encloses the most representative zone of an image, from its position and size it is possible to evaluate its properties, so that these parameters are those that codify our individual genetics as is shown in the figure 2.

Figure 2. - Cromosomatic coding

Where each parameter measures 8 bits because the image has a measurement of 256 x 256 pixels size. The first 16 bits codify the origin of the rectangle whereas the second 16 bits codify the area of the rectangle.

At the moment of evaluating the fitness of a rectangle this returns the width and height to extract the biomarks of the enclosed cells.

So with the information provided by these biomarks, the rectangle size and the amount of cells that it contains, there is sufficient to resolve the fitness function.

2.1.2 FITNESS FUNCTION.

The named fitness function or objective function is that which allows the evaluation of the bonanza grade of the genetic algorithm throughout the evolution process, thus allowing determining which individuals are most or least apt genetically. They are then selected to continue participating in the evolution process or re-emplaced for their poor performance as the object of this work is to find the potentially most representative rectangular zone of the image that contains possible cancerous cells, the function will seek to evaluate positively those zones whose cells present a high grade of cancer (gc) on the other hand if the zones are very big and contain too many cells that embrace (ar) and for the number of cells contained (nc) thus from this it is proposed as a fitness function:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$f=(\sum_{i=1}^{nC} gCi)-(aR+nC)$$

Where the grade of cancer of a cells is defined as the sum of the characteristics that describe a normal cell so that the more positive it is, the more cancerous this cell is considered.

$$Gc = tC+cC+dN+tN+cN$$

Where tc represents the size of the cell, cC the circumference of the cell, dN the definition of the cell nucleus, tN the size if the cell nucleus and finally cN the circumference of the cell nucleus.

Thus it can be stated that this fitness function first encloses in a rectangle all the cells of the image, and closes and moves until the least amount of cells most representative of all the image is enclosed.

2.1.3 PARALLEL PROCEDURE.

The objective of the different techniques and forms of parallelization of a computerized procedure is to increase n times a computerized algorithm when n units of procedure and used, whether procedures within the same computer or independent computers within the same web.

There are many techniques and technologies for parallelization. However these works use a combination of messages and parallel data and RMI of Java technology for its implementation. The use of Java as a platform of software allows an independence in the hardware and at the same time the use of heterogeneous procedure units for the fulfillment of the algorithm.

2.2 THE ISLAND MODEL.

In genetic algorithms there are now some models for the parallelization of the evolution process. One such is the island model, whose idea is that various populations running evolution processes in parallel form are found. These populations are called sub-populations.

The point of this idea is that after a certain number of generations the asynchrony and separation that occurs in all the sub-populations is detained and some individuals are interchange between certain populations (depending on the topography of the interchange). Injecting foreign chromosomes in the sub-populations the evolution process is renewed. Thus each that passes a period of interchange (epoch) sufficient geographical diversification of the individuals is assured. The advantage of this model is the parallel architecture SIMD 8 present, allowing the procedure units to perform in a coordinated manner the evolution process, using its own memory and procedure resources. Only at the beginning and computer becomes leader to determine which teams to send the execution parameters and to synchronize the start of the evolution process. At the end this same team is in charge of receiving the results as well as storing them.

It is pointed out that it is not necessarily him that detains the evolution process, because if the best solution is found by any other of the teams they will be responsible for informing that the general process has fished.

This diversion of tasks, as well as representing an increase in the time necessary to find a solutions, allows finding better solutions because each procedure unit can explore different fragments of the total space of solutions as shown in figure 4.

This model has been used to solve complex procedure problems, given its strong parallelism, efficient scalability and the genetic diversity even after executing various generations of individuals its show in figure 3.

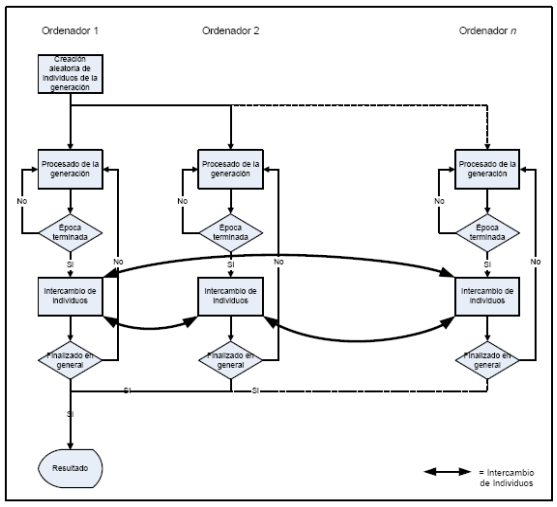


Figure 3. Various generations of individuals

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3 DEVELOPMENT PLATFORMS.

For the implementation of this task are evaluated 3 areas of development/libraries directing computation problem solutions using parallel genetic algorithms and one is selected for the creation of a small applications. Each one of the areas has its advantage and disadvantage, but all have a Java in common as a platform base.

This characteristic is determined as previously explained with the aim to use heterogeneous computers as procedure units. The platforms are presented as follows.

3.1.1 GENETICA.

Genetica is a development area that allows genetic algorithms to be implemented without any particular focus, as it already has a varied group of classes that together allow construction of a genetic platform. As describes before, the island model is used as a parallelization technique.

This area as well as being based in Java, is developed with a focus orientated to agents, using the platforms of agents JADE, not only to solve computation tasks, but to better the process to0 perform the application of form of distribution, supporting failure tolerance and balancing dynamic loads.

3.1.2 JaGa.

JaGa is another group of libraries that permit the implementation of genetic algorithms that subdivide in 3 sub-groups to archive parallelization. They are:

1. JaGa. A group of libraries in java to implement simple evolution procedures. (similar to gallops but in Java).
2. Island Ev. A group of libraries in Java that implement the island model based on a model client-servant to co-ordinate the distribution of evolution processes as well as it's execution.
3. Distrit. Group of libraries in Java to distribute the applications between teams through a web, based on client and servant models that establish connections and execute methods under RMI (Remote Method Invocation).

3.1.3 JGDS.

Finally JGDS is a group of libraries in Java that permit the resolution of optimization problems using parallel genetic algorithms, combining the capacity of the evolution focuses with the distribution procedure to solve complex problems. Its parallelization architecture as well as the previous tools orientated also to the island model.

The design of this group of libraries is very simple and if it doesn't have too many classes, the ease to create a new problem and integrate it into the design has proved why it has been chosen in this task as a development platform.

Originally JGDS comes with half a dozen typical problems resolved, extending a subclass of the class problem. Similarly the class can be extended to implement any specific new problem as shown in figure 4.

JGDS reads the parameters of the execution of the evolution process from a very simple configuration archive, which makes it a very flexible tool during the parameter engagement to refine the quality of the obtained solution as it is now not necessary a recompilation of the class to vary the characteristics of our individual, of the population, of the operators of cross, mutation, selection, etc.

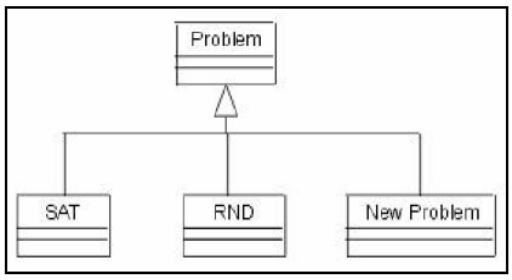


Figure 4. JDGS

4 INSTALATION AND FUNCTIONING.

Once all the Java source archives are compiled, including the subclass cancer Java it is enough to share the subdirectory where the class complied archives are found and where the archives for configuration of the cancer problem are stored as shown in figure 5.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

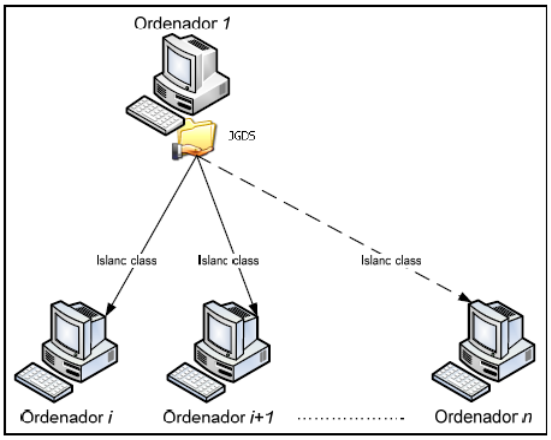


Figure 5.JDGS Network architecture

Now that all the teams can acceed to the classes it is necessary to approve them in the configuration archive, changing the following labels:

Num islands – number of computers.
Host = IP of computer.

Host n = IP of the computer N approving the computer, it is enough to establish the type of problem of the label problem type = 6 that corresponds to the new problem created (JGDS comes with 5 defect problems) so the execution can begin. If there are many parameters in the configuration archive, only the most important are described to execute a sample process. The rest of the parameter relationed with the same algorithm computer can be explored to find its best configuration combination for the cancer problem.

4.1 EXECUTION OF TESTS AND RESULTS.

The gC take values of 1 to 9 being the most positive the most cancerous. In table 1 we can see the location of the supposed cells and it's gC, while in figure 6 a graphical representation of the simulated image is shown, with the location of the supposed cells and with the representation of gC for the size of the figure 8 cancer, as shown in sphere.

That invokes the island class with the parameter of the confcancer.txt archive and retains the exit archive txt. As this test is made with only one computer, number 1 indicates that this computers leader and shouldn't wait for any other to begin the execution proposed. As the class is a subclass of abstract class abstract problem, it has tended to be implemented in 2 important functions: (evaluation and print solution) that respectively calculate the fitness and print the results. When island calls this class to begin an evolution process, the application begins a bi dimensional arrangement with values to simulate an image to process and calculates the fitness until the processes stop and then the result is printed on the screen.

x	y	gC
1, 21		1
2, 24		3
70, 26		8
75, 28		2
80, 30		1
85, 33		1
90, 40		5
95, 80		6
100, 90		7
105, 100		6
150, 110		5
170, 120		4
180, 130		2
190, 1		1
200, 30		8
210, 60		9
220, 90		6
230, 120		9
240, 150		7
250, 180		4

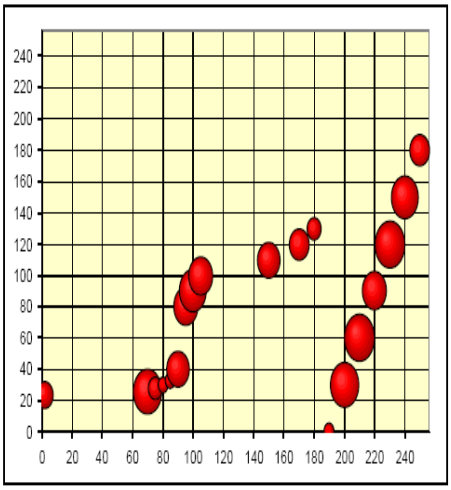


Figure 6. - Representation of gC

During its execution the valuation of fitness depends on the value of the individual.

4.2 TWO PROCEDURE UNITS.

To execute the process in two computer it is sufficient to configure the confcancertxt archive as previously indicated and execute in the second computer the command:

JGDS > Java island 2 o exit 2txt-c
Confcancertxt.

It is important to show that the last computer to execute the command should be the team leader; otherwise the process will throw out an error and will not begin.

4.3 THREE PROCEDURE UNITS.

To execute the process with three computers it is enough to follow the previous example to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

configure the confcancertxt archive and execute in the third computer the command:

JGDS > Java island 3 o exit 3txt-c
Confcancertxt.

Also it is important to show that the last computer to execute the command should be the team leader otherwise the process will throw out an error and not begin.

4.4 ISLAND METHOD PSEUDOCODE.

```
==========
for all members of population
   sum += fitness of this individual
end for

for all members of population
   probability = sum of probabilities + (fitness / sum)
   sum of probabilities += probability
end for

loop until new population is full
    do this twice
       number = Random between 0 and 1
      for all members of population
         if number > probability but less than next probability
              then you have been selected
      end for
    end
    create offspring
end loop
==========
```

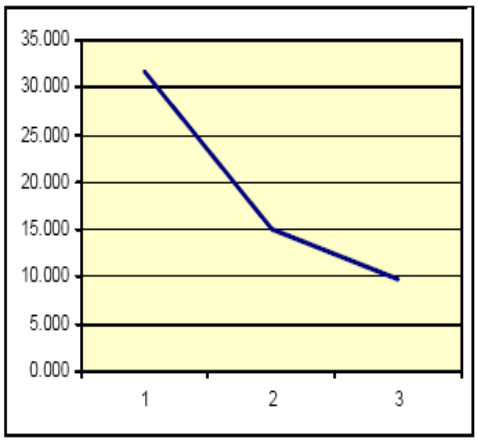
With this information it is possible to analyze the data calculating speed up, in other words the bonanza of adding procedure units, for the execution time as well as for the evaluation necessary to find a solution. Also the efficiency can be calculated to determine the grade to increase the performance such that the procedure units are also increased.

5 RESULTS.

In table 1 is presented the time results in second needed to find a reasonable answer. Five executions are made and the results averaged. In graphic 1 is shown the curve that describes time as procedure units are added.

Also the evaluation needed to complete the evolution process is quantified, so table 3 shows the results of evaluation quantity, while graphic 2 shows the reduction curve of the number the evaluations required to find a reasonable solution.

Table 1. Results of evaluation quantity



# Computers		1	2	3
processing time	1	27.860	11.344	7.813
	2	10.886	14.375	19.985
	3	55.069	8.532	6.940
	4	34.139	4.547	4.688
	5	30.253	36.422	9.243
average		31.641	15.044	9.734

6 CONCLUSIONS.

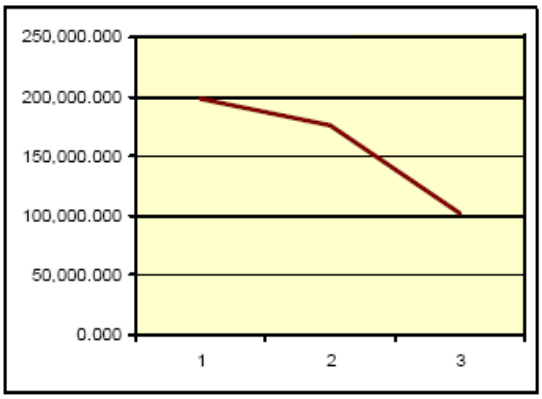
Add an computer to the evolution process proportions a speed up time = 2103, while adding two computers shows a speed up time = 3.251.

On the other hand, in the evolution aspect we find that one computer more gives a speed up evaluation = 1.123, while adding a third computer proportions speed up evaluation = 1.937.

Concerning efficiency we find the execution time for an additional computer is beneficial efficiency time = 1.052 while adding a third computer improves efficiency time = 1.084.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2. Results of evaluation quantity



# computers		1	2	3
sampling data	1	150,782.000	133,734.000	101,104.000
	2	66,436.000	173,011.000	232,315.000
	3	368,409.000	62,254.000	55,635.000
	4	188,572.000	43,859.000	51,978.000
	5	217,504.000	469,840.000	71,056.000
average		198,340.600	176,539.600	102,417.600

Mean while in the evaluation aspect, we find that an additional computer gives an efficiency evaluation = 0.562 and adding a third computer gives efficiency evaluation = 0.646.

The tests appear to demonstrate that it is possible to find a representative zone in a image with the proposed solution and demonstrate that the island model allows an efficiency of process.

Even when it is supposed that it isn't possible to have an efficiency better than 1, in this case the genetic algorithms not only find a solution proportionally quicker to increase the procedure units, but having a better geographical diversity within the solution space it can be found in less evaluations.

These two combined factors are what allows an efficiency to be better than 1. The genetic algorithms and the parallelization techniques are useful tools and capable of giving real solutions to practical problems.

REFERENCES.

[1] Altschul, S.F. and Gish, W. (1996) Local alignment statistics. Methods Enzymol, 266, 460–480[Web of Science][Medline].

[2] Altschul, S.F., et al. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403–410[CrossRef][Web of Science][Medline].

[3] Altschul, S.F., et al. (2001) The estimation of statistical parameters for local alignment score distributions. Nucleic Acids Res., 29, 351–361[Abstract/Free Full Text].

[4] Ginalski, K., et al. (2003) Detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res., 31, 3804–3807[Abstract/Free Full Text].

[5] Mood, A.M., Graybill, F.A., Boes, D.C. Introduction to the Theory of Statistics, (1974) 3rd ed. McGraw-Hill.

[6] Olsen, R., Bundschuh, R., Hwa, T. (1999) Rapid assessment of extremal statistics for gapped local alignment. In Lengauer, T., Schneider, R., Bork, P., Brutlag, D., Glasgow, J., Mewes, H.-W., Zimmer, R. (Eds.). Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology., , Menlo Park, CA AAAI Press, pp. 211–222.