Genetic Algorithms and Tabu Search for Correcting Lanes in DNA Images

M.J. Angélica Pinninghoff, Q. Daniel Venegas, and A. Ricardo Contreras

Department of Computer Science
University of Concepción, Chile
{mpinning,rcontrer}@udec.cl

Abstract. This paper describes an experience that combines Genetic Algorithms and Tabu Search as a mechanism for correcting lanes in DNA images obtained through Random Amplified Polymorphism DNA (RAPD) technique. RAPDs images are affected by various factors; among these factors, the noise and distortion that impact the quality of images, and subsequently, accuracy in interpreting the data. This work proposes a hybrid method that uses genetic algorithms, for dealing with the highly combinatorial feature of this problem, and tabu search, for dealing with local optimum. The results obtained by using them in this particular problem show an improvement in both, fitness of individuals and execution time.

1 Introduction

Randomly Amplified Polymorphism DNA (RAPDs) [11] is a type of molecular marker which has been used in verifying genetic identity. During the past few years RAPDs have been used for studying philogenetic relationships [1,10], gene mapping [5], trait-associated markers [9], and genetic linkage mapping [2]. This technique has been used as support for many agricultural, forest and animal breeding programs [6].

In Figure 1, a photograph of a RAPD reaction is shown. In this case, 12 samples were loaded of which lanes 1 and 14 correspond to the molecular weight standards. In this case, four different genotypes of Eucalyptus globulus were studied, including three identical copies of each (known as ramets). If the ramets are identical, then quite similar band patterns should be expected when analyzed by the same primer. However, this is not always the case, due to, for example, mislabeling of samples.

The RAPD technique consists of amplifying random sequences of the genomic DNA by using primers, which are commonly 10 bp (base pairs) in length. This process is carried out by polymerase chain reaction (PCR) and generates a typical pattern for a single sample and different primers. The PCR products are separated in an agarose gel, under an electric field which allows smaller fragments of the PCR products to migrate faster, while larger ones much slower. The gel is stained with a dye (typically ethidium bromide) and photographed for further data analysis. One way of analyzing the picture obtained is simply

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2010, LNCS 6256, pp. 144–153, 2010.

by comparing visually the different bands obtained for each sample. However, this can be a tedious process when various samples with different primer combinations have to be analyzed. At the same time, since, in this case, the presence or absence of bands is to be scored, sometimes the band assessment can be very subjective and there is no reliable threshold level, since the intensities of the bands are affected by several factors (i.e staining, gel quality, PCR reaction, DNA quality, etc.).

During the process of generating the RAPD image, many physical-chemical factors affect the electrophoresis producing different kinds of noise, rotations, deformations and other abnormal distortions in the image. The effect of this problem is, unfortunately, propagated through the different stages in the posterior analysis, including visualization, background extraction, band detection, and clustering, can lead to erroneous biological conclusions. Thus, efficient image processing techniques will, on the other hand, have a positive impact on those biological conclusions.

Typical errors consider rotation in lanes, that is the problem we try to solve. This is the first step; once lanes are corrected, i.e., the complete image shows a minimum slope for each lane, it will be necessary to work in band correction, a difficult problem due to the nature of distortions, different to lane distortion. The second step, band correction, can be carried out in an analogous way; however, it is beyond the scope of this paper.

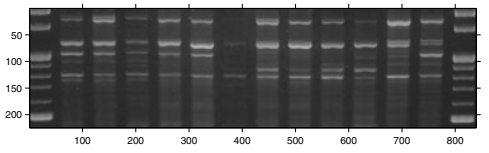

Fig. 1. A sample RAPD image with two reference lanes, and 12 lanes representing four ramets

The basis for this work is the experience described in [7], in which genetic algorithms are used to deal with a population of potential solutions, where solutions are intended as the best templates. A template is a set of lines representing lanes and therefore, the best template is one in which lines match closely to lanes in the original image. This work offered good solutions, although execution time is greater that expected. The other problem this approach presents, is the presence of local minimum.

The aim of this work is to correct distortions in lanes, by using genetic algorithms hybridized with *Tabu Search*. This allows a comparison of two strategies: the first one, already mentioned, that considers single genetic algorithms, and the second one, that uses genetic algorithms and Tabu Search as collaboration mechanisms.

This article is structured as follows; the first section is made up of the present introduction; the second section describes the specific problem to be faced; the

third section is devoted to genetic algorithms and tabu search considerations, while the fourth section shows the results we obtained with our approach, and the final section shows the conclusions of the work.

2 The Proposed Approach

The problem addressed in this paper can be formally stated as follows.

Consider an image (matrix) $A = \{a_{ij}\}, i = 1, \ldots, n$ and $j = 1, \ldots, m$, where $a_{ij} \in Z^+$, and A is a RAPD image. Usually, a_{ij} is in the range [0..255] in a grey scale image, and we use a a_{ij} to refer to an element $A(x, y)$, where x and y are the pixel coordinates.

To deal with lane distortions, a set of templates is used. These templates are randomly created images with different distortion degrees, having lines that are in a one-to-one correspondence with lanes in the original RAPD image. A good template is the one that reflects in a more precise degree the distortions that the RAPD image under consideration has.

The template we consider is a matrix L (lanes) where $L = \{l_{ij}\}, i = 1, \ldots, n$ and $j = 1 \ldots, m$, $l_{ij} = 0$ or $l_{ij} = 1$ (a binary image), with 1 meaning that l_{ij} belongs to a line and 0 otherwise. A procedure described in [8] is used to approximately detect the initial position of the lanes. In doing so, the generation of matrix L is limited to those regions that correspond to lanes in matrix A. Due to the rotation of the lanes, it is necessary to consider different alternate configurations. If we are dealing with an image with 12 lanes, and if for each lane we consider 14 possible rotations, we are considering 12^{14} different configurations to evaluate. This causes a combinatorial explosion, which justifies the use of genetic algorithms.

Genetic algorithms allow to manage a large number of templates, and those that are similar to the original image are chosen. Thus, it is necessary to seek for an objective function that reflects this similarity in a precise way. This function is used as a measure for the quality for the selected template.

When the lane correction procedure is applied, templates contain straight lines. Different templates will show different slopes for each line, as shown in Figure 2. A template contains non-intersecting vertical lines, which are not necessarily parallel.

Results obtained in a previous work are promising but not really good. In considering this, we decided to hybridize the solving strategy by adding a Tabu

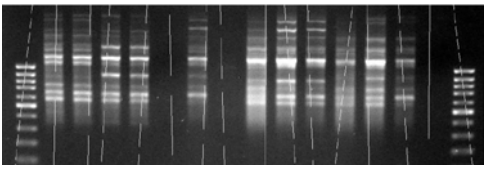

Fig. 2. A sample template for lane correction

Search component. Tabu Search is a mathematical optimization method belonging to the class of local search techniques. Tabu search enhances the performance of a local search method by using memory structures: once a potential solution has been determined, it is marked as "taboo" so that the algorithm does not visit that possibility repeatedly.

3 Genetic Algorithms and Tabu Search

Genetic algorithms (GA) are a particular class of evolutionary algorithms, used for finding optimal or good solutions by examining only a small fraction of the possible space of solutions. GAs are inspired by Darwin's theory about evolution. The basic concept of GAs is designed to simulate processes in natural system necessary for evolution, specifically those that follow the principles of survival of the fittest. As such they represent an intelligent exploitation of a random search within a defined search space to solve a problem.

The structure of a genetic algorithm consists of a simple iterative procedure on a population of genetically different individuals. The phenotypes are evaluated according to a predefined fitness function, the genotypes of the best individuals are copied several times and modified by genetic operators, and the newly obtained genotypes are inserted in the population in place of the old ones. This procedure is continued until a *good enough* solution is found [3].

In this work, the templates are the chromosomes, lines in a template are the genes, and a line having a particular slope represents the value (allele) that a gene has.

A good fitness means that a particular template (matrix L) fits better to the original RAPD image (matrix A). To evaluate a template, images corresponding to matrices A and L are put together, and a sum of intensities is obtained by considering neighborhood pixels within a range and for each line. This range is determined in this work considering the width of the brightest part of the lane. The aim of this is to gain precision in the fitness function. If a line in the template coincides with a lane, a higher value of the sum is obtained. In contrast, if they do not coincide, the value is lower than in the first case, because we are adding background pixel intensities (values close to zero).

Another issue added is that, the value obtained in the evaluation of each line is stored as part of the *gene*. In this way, the sum of intensities of pixels is only done when a new line is created; and this occurs in mutation and in tabu search. As a consequence of this issue, the execution time is reduced considerably compared to previous experiments.

Genetic operators: Different genetic operators were considered for this work. These genetic operators are briefly described below:

- Selection. Selection is accomplished by using the roulette wheel mechanism [3]. It means that individuals with a best fitness value will have a higher probability to be chosen as parents.

- Cross-over. Cross-over is used to exchange genetic material, allowing part of the genetic information of one individual to be combined with part of the genetic information of a different individual. For example, if we have two templates each containing $r+s$ lines, after cross-over, the generated children result in: children 1 will have the first r lines that correspond to template 1, and the following s lines that correspond to template 2. For children 2, the process is slightly different, in which the order the templates are considered is modified.
- Mutation. By using this genetic operator, a slight variation is introduced into the population so that a new genetic material is created. In this work, mutation is accomplished by randomly replacing, with a low probability, a particular line in a template.

Tabu search (TS) is a meta-heuristic that guides a local heuristic search procedure to explore the solution space beyond local optimality. The local procedure is a search that uses an operation called *move* to define the neighborhood of any given solution. One of the main components of TS is its use of adaptive memory, which creates a more flexible search behavior. In a few words, this procedure iteratively moves from a solution x to a solution x' in the neighborhood of x, until some stopping criterion has been satisfied. In order to explore regions of the search space that would be left unexplored by the local search procedure, tabu search modifies the neighborhood structure of each solution as the search progresses [4].

A solution is a template representing a RAPD image, let us say the x solution; then to move from a solution x to a solution x' means that the template is modified. To modify a template we have chosen two possibilities: the first one is the change in the value of the slope for one or more lines in the template, i.e., a rotation movement; the second one is a shifting movement, the line is moved to the left or to the right, without changing the value of the slope for that line. In other words, if we call x_{inf} and x_{sup} the bottom and top points in a line respectively, a rotation movement is realized by changing these points to $x_{inf} - \delta$ and $x_{sup} + \delta$ (or changing to $x_{inf} + \delta$ and $x_{sup} - \delta$). In an analogous way, the shifting movement is accomplished by changing the original points to $x_{inf} + \delta$ and $x_{sup} + \delta$ (changing to minus if the movement is towards the opposite side of the image). Figure 3 illustrates the rotation movement and the shifting movement. The values allowed in both, shifting and rotation movements, are gradually diminished to avoid dramatic changes in the quality of the solutions.

To avoid repeated movements during a certain bounded period of time, TS stores each movement in a temporal memory, which is called *tabu list*. Each element in the tabu list contains one lane and its corresponding movement. A particular lane in the list may occur more than once, but the associated movement needs to be different. In this work, the size of the tabu list is bounded by the number of lanes in the particular image under treatment.

When it is not possible to find a better solution in the neighborhood of x, it is used the, so-called, *aspiration criterion*, which allows to search in the tabu list for a movement that improves the current state x. If that movement doesn't

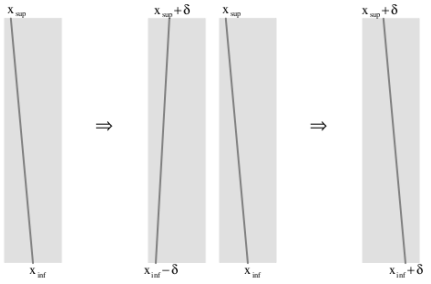

Fig. 3. The schema for rotating a line, and the schema for shifting a line

exist, then a movement with a higher residence time in the tabu list is chosen and, in this particular case, the inverse corresponding movement is applied to x; i.e., it is used to implement a backtracking strategy.

As previously mentioned, in spite of good results obtained by using genetic algorithms, the problem of local optimum is always present. By taking into account this issue, we decided to *hybridize* the procedure, that is to say to combine genetic algorithms with another strategy, in this specific work with tabu search, to let potential solutions avoid those local optimum points. The hybridization procedure considers the following strategy: the main objective of this process is to gradually improve individuals belonging to the population the genetic algorithm is working with. When the fitness measured during a certain number of iterations accomplished by the genetic algorithm doesn't vary; a reduced number of individuals is selected from current population. One of them is the best individual of the population, while the others are randomly selected. Each one of these individuals acts as an input for triggering a tabu search procedure.

Once the tabu search process is finished, the resulting individuals are reinserted into the genetic population, and the process continues with the genetic algorithm procedure, as before. The complete process is repeated several times depending on the quality of the genetic population and the stopping process condition. The latter is specified as a time condition (number of iterations) or as a specific fitness value.

4 Results

For testing, it is necessary to provide a set of images that consider most of the problematic situations: different slope in lanes, different bright, noise in images, missing lanes, and different number of lanes. According to this criteria, 13 images were chosen to carry out the final tests.

Parameters are variables that maintain a fixed value during a particular processing. While they cannot be defined *a priori*, they have to be experimentally determined. We carried out different tests for both, the genetic algorithm and the hybrid algorithm processing.

Table 1. Parameters for testing

Parameter	Value(s)
Population Size	150
Num. Generations	2000
Cross-over %	70
Mutation %	2
Elitism %	10
Seed Value	10 different values

Parameter	Value(s)
Population Size	20, 50
Num. Generations	2000
Cross-over %	70, 60
Mutation %	2
Elitism %	10
Seed Value	10 different values
Triggering Cond.	30 generations
Num. Iterations	30
Tabu List Size	Number of lanes

The set of parameters used for the genetic algorithm approach is summarized in the left part of Table 1; and the set of parameters used for the hybrid approach (genetic algorithm plus tabu search), is summarized in the right part of the table. Figure 4 illustrates fitness improvement for 13 testing images.

The elements that have an important influence on the execution time, are the number of lanes the image has, the population size and the mutation probability.

The execution time for the hybrid algorithm approach, is higher than the genetic approach, because in this case, each movement implies to evaluate one or more lines for each new template created in the neighborhood. This is one of the reasons why we reduced the population size during tests. An increasing population size needs a larger execution time. A reduced number of tests considering a higher number for the population size didn't produce better results.

As shown in Figure 4, in all cases, the hybrid algorithm produced a better fitness than the genetic algorithm. In most of the test cases, considering the two different population sizes, the values obtained are similar.

If we consider the evolution of fitness, for early generations (below 300 genetic generations), the genetic algorithm offers better results that those obtained by using the hybrid algorithm. This is likely due to the bigger population size. For a medium evolution period (between 300 and 500 genetic generations) the fitness values for the hybrid algorithm increase dramatically. In this particular group, fitness is similar in both, the genetic algorithm and the hybrid algorithm. When

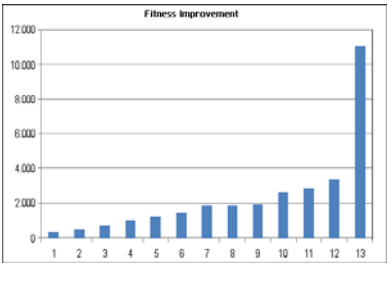

Fig. 4. Fitness improvement for different images

we consider a higher number of genetic generations (more that 500), values are clearly better when using the hybrid algorithm. On the other hand, as expected, the hybrid approach is much more time consuming than the genetic algorithm approach. Figure 5 shows the evolution while processing a particular image, the upper part (a) shows the original RAPD image, the middle part (b) shows the best individual, and the bottom part (c) shows the corrected image.

5 Analysis

The number of lanes and the population size, impact the execution time because of the increasing amount of data that needs to be processed. The mutation probability influence is related to the fact that the evaluation function evaluates a line each time a new line is created, the typical action when mutation is accomplished. In absence of mutation, each line keeps its evaluation while it is not modified.

The facts that the hybrid algorithm produced a better fitness and the similarity of results in spite of changing the population size, allow us to infer that the hybrid algorithm possibly reached an optimum value. For a particular image, the value obtained for the corresponding fitness is the same for the different population sizes and for the different seeds used for the random values that the genetic algorithm employs.

It is possible to say that when dealing with RAPD images, the hybrid algorithm works better than the genetic algorithm. One fact to remark is that this is true for both, large population sizes and small population sizes. This can be explained by considering that the hybrid approach acts from two different points

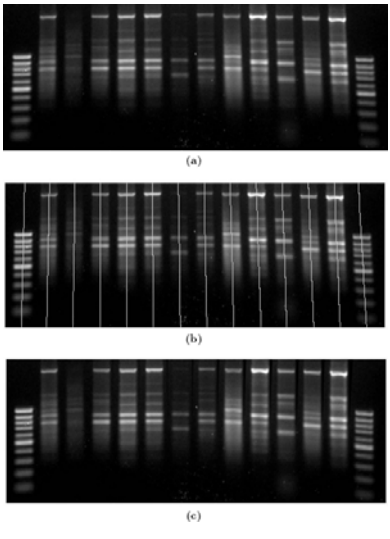

Fig. 5. The evolution in lane correction: a) the original image, b) the best individual, c) the corrected image

of view. From the first one, the genetic algorithm point of view acts as a global search process, while from the second point of view, tabu search acts as a local search process. Combining these two strategies leads to better results.

One hypothesis that can explain this fact is that the single genetic approach is good enough, and that there is no space for significant improvements concerning fitness. Despite the fact that the values of fitness have minor differences between both strategies (single and hybridized genetic algorithm), these differences are important for human experts, because they provide more reliability to their genetic conclusions.

The population size seems not to be important in terms of results. However there is a difference; a larger population size implies a higher genetic algorithm influence; on the contrary, a smaller population size reveals that the tabu search steps have a greater influence on the results. This is because the larger populations present a higher genetic variability, and genetic operations, consequently, possibly produce higher quality offsprings; while little size populations depend on tabu search movements quality to improve individuals, which tend to converge to local minimum values.

6 Conclusions

Experiments show that by using genetic algorithms and tabu search the final fitness value is slightly improved.

Another issue to remark is that the best performance for the hybrid algorithm is obtained after considering 500 genetic iterations. We believe that the reason behind this behavior is that at the beginning, the genetic algorithm can improve the population due to genetic operations, and so, the tabu search is not triggered until a local stability population is reached.

Results are slightly better with the hybrid approach, but it is necessary to pay a cost; that is the increasing execution time; in fact, the hybrid approach that considers 50 individuals takes an execution time similar to the single genetic approach that considers 150 individuals.

Some future work directions are mentioned in the following: the implemented solution depends strongly on a correct lane boundary detection. So, it is necessary to have a better lane detection method, probably automated. Currently, multilevel thresholding is being considered for band detection, but it could be appropriate for that purpose, as well. To increase the degree of parallelism when dealing with tabu search and to increase the population size are some pending issues that need to be taken into account.

Acknowledgment

This work has been partially supported by grant DIUC 209.093.014-1.0, University of Concepción, Chile.

References

1. Cao, W., Scoles, G., Hucl, P., Chibbar, R.: Philogenetic Relationships of Five Morphological Group of Hexaploid wheat Based on RAPD Analysis. Genome. 43, 724–727 (2000)
2. Casasoli, M., Mattioni, C., Cherubini, M., Villani, F. A Genetic Linkage Map of European Chestnut (Castanea Sativa Mill.) Nased on RAPD, ISSR and Isozime Markers. Theoretical Applied Genetics 102, 1190–1199 (2001)
3. Floreano, D., Mattiussi, C.: Bio-Inspired Artificial Intelligence. Theories, Methods, and Technologies. MIT Press, Cambridge (2008)
4. Glover, F., Laguna, M.: Tabu Search. Springer, Heidelberg (1997)
5. Groos, C., Gay, G., Perrenant, M., Gervais, L., Bernard, M., Dedryver., F., Charmet, G.: Study of the Relationships Between Pre-harvest Sprouting and Grain Color by Quantitative Trait Loci Analysis in the White X Red Grain Bread-wheat Cross. Theoretical Applied Genetics 104, 39–47 (2002)
6. Herrera, R., Cares, V., Wilkinson, M., Caligarip, D.: Characterization of Genetic Variations Between Vitis vinifera Cultivars from Central Chile Using RAPD and Inter Simple Sequence Repeat Markers. Euphytica 124, 139–145 (2002)
7. Pinninghoff, M.A., Contreras, R., Rueda, L.: An evolutionary approach for correcting random amplified polymorphism DNA images. In: Mira, J., Ferrández, J.M., Álvarez, J.R., de la Paz, F., Toledo, F.J. (eds.) IWINAC 2009. LNCS, vol. 5602, pp. 469–477. Springer, Heidelberg (2009)
8. Rueda, L., Uyarte, O., Valenzuela, S., Rodriguez, J.: Processing Random Amplified Polymorphism DNA Images Using the Radon Transform and Mathematical Morphology. In: Kamel, M.S., Campilho, A. (eds.) ICIAR 2007. LNCS, vol. 4633, pp. 1071–1081. Springer, Heidelberg (2007)
9. Saal, B., Struss, D.: RGA-and RAPD-derived SCAR Markers for a Brassica B-Genome Introgression Conferring Resistance to Blackleg Oil Seed in Oil Seed Rape. Theoretical Applied Genetics 111, 281–290 (2005)
10. Sudapak, M., Akkaya, M., Kence, A.: Analysis of Genetic Relationships Among Perennial and Annual Cicer Species Growing in Turkey Using RAPD Markers. Theoretical Applied Genetics 105, 1220–1228 (2002)
11. Tripathi, S., Mathish, N., Gurumurthi, K.: Use of Genetic Markers in the Management of Micropropagated Eucalyptus Germplasm. New Forests 31, 361–372 (2006)