Mann-Whitney Test-based Segmentation of Yeast Genomic Information

Jose C. Riquelme
Dept. of Computer Science
University of Seville, Spain
Email: riquelme@lsi.us.es

Daniel Mateos
Dept. of Computer Science
University of Seville, Spain
Email: mateos@lsi.us.es

Jesus S. Aguilar–Ruiz
Dept. of Computer Science
University of Seville, Spain
Email: aguilar@lsi.us.es

Abstract— **Segmentation algorithms differ from clustering algorithms with regard to how to deal with the physical location of genes throughout the sequence. Therefore, segments have to keep the original positions of consecutive genes, which is not a constraint for clustering algorithms. In this paper, we present an evolutionary algorithm for genomic segmentation with a fitness function based on the Mann-Whitney test for the precise differentiation of adjacent segments. We have used the activity of meiotic recombination of yeast as the variable of study for the yeast genome instead of the DNA sequence, as many other researchers have dealt with. Experimental results reveal the suitability of evolutionary computation to this sort of complex problem, not only because of the quality of solutions but also because they allow the researcher to address many different conditions, providing flexibility to experimental analysis.**

I. Introduction

Budding yeast *Saccharomyces cerevisiae*, the first eukaryote for which an entire genomic sequence is available, is important to the research comunity because the analysis of this unicellular eukaryote may allow us to detect fundamental patterns that shape the more complex genomes of multicellular organisms [1], [2].

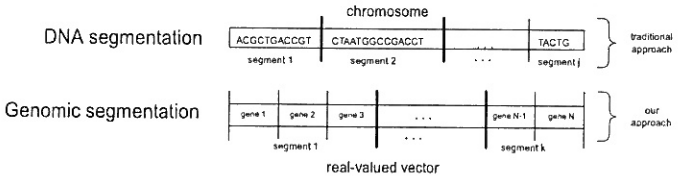

Fig. 1. Genomic segmentation.

Segmentation algorithms emerge observing fluctuations of DNA sequences in alternative homogeneous domains, which are named segments (see Figure 1). In 1974, Elton developed a series of theoretical models of increasing complexity and concluded that the data support a model in which the DNA consists of a sequence of "segments" with different underlying base compositions [3]. The key idea is that two genes that are controlled by a single regulatory system should have similar expression patterns in any data set. In [4] Kruglyak and Tang used the correlation coefficient as a measure of similarity to show that the expression patterns of adjacent genes are more often highly correlated than the expression patterns of randomly selected gene pairs. Traditionally, moving windows have been used, representing graphically the G+C (Guanine and Cytosine) percentages to observe the fluctuations [5]. A compositional heterogeneity measurement based on information theory to obtain segments with maximum compositional divergence has been introduced in [6]. In [7], the DNA sequence is viewed as a stochastic process with local compositional properties determined by the states of a hidden Markov chain. The model used is a discrete-state, discrete-outcome version of a general model for non-stationary time series. Another approach to DNA segmentation, based on Bayesian estimators, is presented in [8].

A critical issue about segmentation algorithms is the stopping criterion in segmenting homogeneous DNA sequences. W. Li [9] proposed a solution based on Bayesian Information Criterion which identified borders of biologically meaningful units (subtelomeric units, replication origin and terminus).

However, all of these approaches work directly with DNA sequences instead of numerical values associated to each gene. Researchers can study genomes under specific conditions and extract patterns from the level of expression of genes. From studies in the yeast Saccharomyces cerevisiae, it has become clear that recombination results from the induction of a meiosis–specific pathway for the formation and repair of DNA double-strand breaks. Therefore, a detailed understanding of meiotic recombination in yeast will give us a picture of how the process works in mammals. Meiotic recombination is the exchange of chromosomal segments between the paternal and maternal homologs during meiosis. In this paper, we present a new approach based on Evolutionary Algorithms (EAs) that differentiate segments of genes, which are represented by its level of meiotic recombination (see Figure 2). To our knowledge it is the first biological segmentation work that deals directly with real numbers instead of DNA sequences. To measure the relevance of segments the Mann–Whitney statistical test has been used.

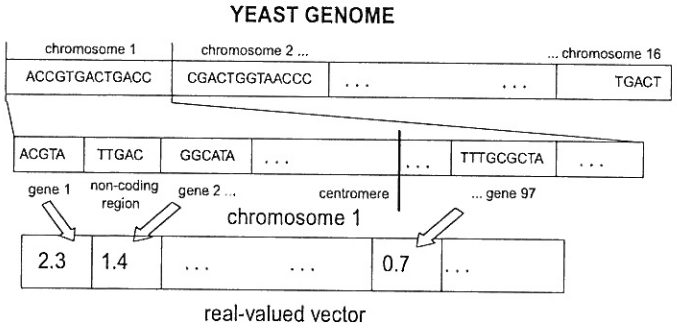

Fig. 2. Yeast genome.

From the field of evolutionary computation, there are many approaches to segment images. Bhanu et al. [10] addressed the adaptive segmentation of images. Yoshimura and Oe [11] proposed a segmentation algorithm for texture images using genetic algorithms that automatically determines the optimum number of segmentation areas. Cagnoni et al. [12] developed a method for evolving adaptive procedures for the contour–based segmentation of anatomical structures in 3D medical data sets. Applications of Genetic Algorithms in clustering and grouping problems are intensively described in

[13]. Recently, a new system [14] is presented for general symbol segmentation, which is applicable for segmentation of any connected string of symbols, including characters and line diagrams. Some interesting segmentation algorithms have been applied to financial times series [15].

The paper is organized as follows: in Section II the evolutionary algorithm is described, including the encoding, fitness function and genetic operators; experiments are discussed in Section III; finally, the main conclusions and future work are summarized in Sections IV and V, respectively.

II. Evolutionary Algorithm

A. Individual encoding

Each individual of the population is an array of natural numbers with size NC, and it represents a collection of cutpoints within the yeast genome. Fifteen of these cutpoints correspond to the boundaries of the sixteen chromosomes of the yeast genome, and they are permanent. The sixteen cutpoints corresponding to centromeres also are permanent, so we have 31 constant cutpoints. The centromere is approximately in the middle of a chromosome and separates it in two branches (L and R). Although these fixed cutpoints (FNC=31) cannot be moved, they have been included in all of the individuals, making easier the computational process (see Figure 3). For example, if a cutpoint array includes the values 34, 57, 7, 25 and 80, it means that there is a cutpoint between the 34^{th} and the 35^{th} genes, between the 57^{th} and the 58^{th} genes, between the 7^{th} and the 8^{th} genes, etc. Therefore, the segments comprise from the 1^{st} to the 7^{th} gene, from the 8^{th} to the 25^{th} gene, from the 26^{th} to the 34^{th} gene, from the 35^{th} to the 56^{th} gene, etc. We can also observe a priori that the cutpoint array might not be ordered. In fact, the algorithm begins with an initial population in which each individual comprises random cutpoints (except the fixed ones, which are in the first FNC positions of each individual). In order to facilitate the algorithmic process of the fitness function, the individuals are ordered right before calculating the fitness values, although the order is not preserved in the population.

Working with individuals which hold ordered cutpoints reduces the cost of the fitness function, but increases the cost of genetic operators. We have designed a specific crossover operator to reduce the aforementioned complexity.

B. Fitness function

We have chosen the Mann-Whitney test as the fitness function (see Figure 4). The Mann–Whitney test, also known as the Wilcoxon rank sum test, is a non–parametric test used to test for difference between the medians of two independent groups. This test is the non–parametric equivalent of the two–sample t–test. No distributional assumptions are required for this test, so the test does not assume that the populations follow Gaussian distributions.

The choice of this method is due to the necessity of differentiating adjacent segments clearly. Maximizing the difference between the fitness values of two consecutive segments can be a way for suitable differentiation of adjacent segments. If we choose the mean as representative statistical value for a segment, we can know when the mean of two adjacent segments is significantly different with the Mann–Whitney test. In order to verify the quality of the fitness function, we run the algorithm with the original data, and with randomized versions. We can understand that a fitness function is correct if the results obtained with the random data are lower in quality than those obtained from the original data. Otherwise, we can say that we have an "artifact" (an apparent experimental result that is not actually real but is due to the experimental methods). The implementation of the fitness function is very simple: for each pair of adjacent segments we run a Mann–Whitney test. If the null hypothesis is accepted (the mean of both segments is not significantly different) it returns zero, and one otherwise.

The fitness function returns the sum of all tests (for all the cutpoints of an individual). In Figure 4 we show an easy procedure to calculate the Mann–Whitney test for two independent segments (A and B) with five values (in fact, the same procedure is applied to more than five values) for each group (if a segment has less than five genes, the procedure returns zero). Groups do not need to have the same size. The null hypothesis means that there is no difference between the means.

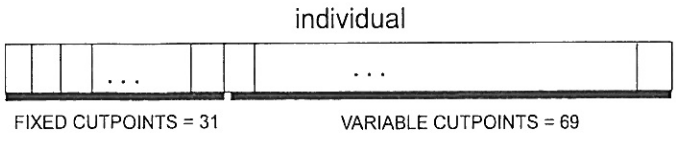

Fig. 3. Representation of an individual of the genetic population.

Procedure M-W

1. Rank all the data values. The smallest number gets the rank 1.
2. Give an average rank if two or more values are tied.
3. Sum the ranks of some segment (T=sum of ranks for one group).
4. Calculate for that group $\mu_a = \frac{n_a(N+1)}{2}$ and $\sigma_a = \sqrt{\frac{n_a n_b (N+1)}{12}}$.
5. Calculate $Z = \frac{T-\mu_a \pm 0.5}{\sigma_a}$ (if $T > \mu_a$ we choose -0.5, else +0.5).
6. If $|Z| > |Z_0|$ for a significance level, we reject the null hypothesis.

end M-W

Fig. 4. Mann-Whitney procedure (n_a and n_b are the sizes of A and B respectively, and $N = n_a + n_b$).

Let us see an example for the steps 1, 2 and 3. We can observe the average rank for tied values. If we choose the group A, the total rank results: $T = 1 + 3 + 5 + 7.5 + 9 = 25.5$. In this case, as $T > \mu_a = 27.5$ we choose +0.5 in line 5. Then, we calculate the value Z=-0.495 and if $|Z| > |Z_0|$ for a significance level, we reject the null hypothesis, i.e. both segments are statistically different. In this case, for 0.05 level of significance, $Z_0 = 1.96 > 0.495$, so the segments are statistically similar.

Group A
0.86
1.00
0.94
2.35
0.01

Group B
0.94
3.15
0.78
0.94
1.00

Values	Group	Rank
0.01	A	1
0.78	B	2
0.86	A	3
0.94	A	(4+5+6)/3
0.94	B	(4+5+6)/3
0.94	B	(4+5+6)/3
1.00	B	(7+8)/2
1.00	A	(7+8)/2
2.35	A	9
3.15	B	10

Fig. 5. Example of Mann-Whitney test.

C. Genetic Operators

1) Crossover: Due to the intrinsic characteristics of the problem, a variant of the uniform crossover has been chosen. That is, a new individual is built by randomly choosing cutpoints from both parents. Also, we tested other well–known operators (one–point and two–points crossovers), but they did not provide better results. This operator has to maintain the diversity, controlling values different than those assigned to the boundaries of chromosomes and centromeres.

2) Mutation: The mutation operator alters each cutpoint according to two probabilities: p_1 and p_2. The probability p_1 controls if a cutpoint is going to be modified; and the probability p_2 controls if the mutation will result in a random cutpoint within the range, or in a slight variation (currently, 5 genes to the left or to the right) of the cutpoint. Basically, these two options are: $indiv[i] := random(N)$ and $indiv[i] := indiv[i] + (-1)^{random(2)} * random(5)$, respectively. The choice of value 5 is not critical, other values around 5 can be used as well. However, if that value is high, consecutive populations will present greater diversity than its ancestors. Logically, the genes at the boundaries of chromosomes and centromeres are not mutated.

III. Experiments

We have tested the algorithm with the yeast genome because this organism is very interesting for the research community, as it preserves many biological properties from more complex organisms and it is simple enough to run experiments.

We have a file with about 6100 genes, divided into sixteen yeast chromosomes (NC). Each gene is a row of the file. Each column of file represents a genomic characteristic under specific conditions (in this experiment, only the activity of meiotic recombination). The goal is to group consecutive genes properly differentiated from adjacent segments. Each group will be a segment of genes, as it will maintain the physical location within the genome.

The evolutionary algorithm used the values 0.4 and 0.2 for p_1 and p_2, respectively. The value 0.4 indicates that approximately 40% of individuals are mutated. From these candidates to be mutated, 20% (p_2) are randomly generated and 80% are mutated slightly, shifting each selected cutpoint only 5 genes to the right or to the left. The significance level for the Mann–Whitney test was 0.05 in all the experiments.

A. Results

In the biological literature there is no reference about what is the best number of segments for the yeast genome. Before running our experiments we have consulted a biologist, who we are collaborating with, and for further biological research a number around 100 is satisfactory. For that reason, this is the number set in all the experiments. However, those cutpoints should not be distributed uniformly through the genome, so the evolutionary algorithm could find 2 cutpoints in the chromosome 1 and 20 cutpoints in the chromosome 2.

If we set 100 cutpoints, we have 100 pairs of segments to test (or 101 independent segments). We know that the fitness function returns the sum of all of the Mann–Whitney tests, and that each test returns 1, if the means of the adjacent segments are significantly different, and 0 otherwise. Therefore, the maximum value that the fitness function can return is 100 (all the segments are "different" in pairs). The number of initial segments is not a constraint for the evolutionary algorithm, because it is only the upper bound for the number of valid segments. In this example, for 101 possible segments, the number of valid ones was about two thirds.

The evolutionary algorithm found 62 cutpoints that differentiated two consecutive segments statistically. In addition, the best result from a random version of data (genes were randomly relocated in different chromosomes) was zero (zero different segments statistically). Therefore, the fitness function based on the Mann–Whitney test is not an artifact. The computation time was about 25 minutes for each run with 400 individuals and 200 generations on a Pentium IV, 2.4 GH, with 512 MB of RAM. It is worth to note that every evaluation needs to order the values of genes within segments to assign a value in the ranking, so a quicksort algorithm is called many times. The best result is graphically shown in Figure 6. Vertical lines (in black) are statistically valid cutpoints and segments are alternated in yellow and in blue. As we can see, not all the pairs of consecutive segments have a vertical black line separating them from each other, so those pairs are not statistically different. For each chromosome, the number of segments and the number of statistically valid segments are reported. For example, chromosome 4 has 13 segments, but only 5 cutpoints were statistically valid, generating 6 segments. These new 6 segments were also tested on its adjacent segments to guarantee the quality of results.

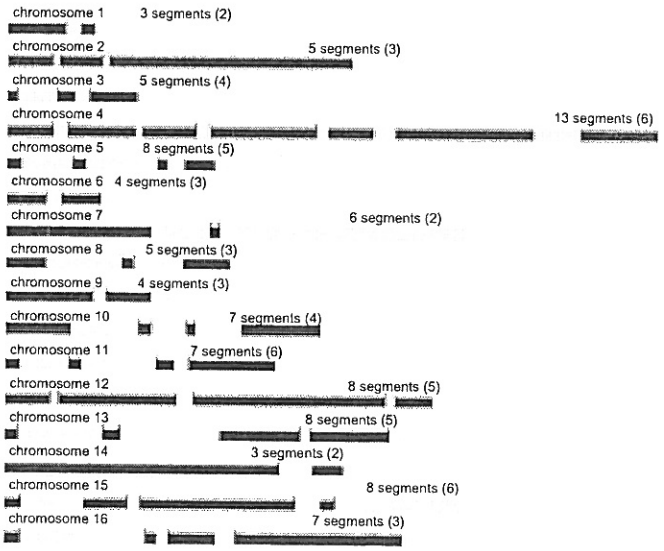

Fig. 6. Segmentation of yeast genome.

IV. Conclusions

An evolutionary algorithm to address the problem of genomic segmentation is presented. The aim is to find groups of consecutive genes with common properties of the variables in study. In this case, we have been focused in the activity of meiotic recombination in yeast genome. The population of this algorithm is composed of vectors of natural numbers that represent a set of cutpoints within the genome. In this sort of problem it is not clear what is the appropriate fitness function to use, so after choosing the Mann–Whitney test, the results show that the fitness function is not an artifact. Also, experiments show that the genomic distribution in yeast genome is not random under the perspective of the activity of meiotic recombination. The evolutionary algorithm has a very satisfactory performance from the biologist point of view, as it can find a high percentage of valid adjacent segments, which can add knowledge to the biological research of functional properties of groups of genes.

V. Future Work

The results reported in this work are not comparable, as we have not found any other system that addresses the segmentation problem by using numerical information. We are now designing a benchmark test based on dynamic programming to avoid the computationally unapproachable exhaustive search.

From now on, the improvement of the algorithm should be focused on not restricting the number of cutpoints of each individual. This way, the evolutionary algorithm should be able to find the correct number of segments. Currently, it finds a number of valid segments less than a constant (100 in this paper). Also, the analysis of other statistic–based fitness functions is now under research. In general, Evolutionary Algorithms are very flexible in order to experiment with different initial and environmental conditions. Another future research direction is the adaptation of our approach to multi–attribute segmentation, i.e. including more relevant features from genes (like G+C content, as it has been proven that genes are richer in G+C than non–coding regions) in addition to the activity of meiotic recombination.

References

[1] Goffeau, A., Barrell, B., Bussey, H., Davis, R., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J., Jacq, C., Johnston, M., Louis, E., Mewes, H., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.: Life with 6000 genes. Science **275** (1997) 1051–1052

[2] Goffeau, A., et al.: The yeast genome directory. Nature **387** (1997) 5–105

[3] Elton, R.: Theoretical models for heterogeneity of base composition in DNA. Journal of Theoretical Biology **45** (1974) 533–553

[4] Kruglyak, S., Tang, H.: Regulation of adjacent yeast genes. Trends in Genetics **16** (2000) 109–111

[5] Bradnam, K., Seoighe, C., Sharp, P., Wolfe, K.: G+C content variation along and among saccharomyces cerevisiae chromosomes. Mol. Biol. Evol. **16** (1999) 666–675

[6] Oliver, J., Roman-Roldan, R., Perez, J., Bernaola-Galvan, P.: SEGMENT: Identifying compositional domains in DNA sequences. Bioinformatics **15** (1999) 974–979

[7] Churchill, G.A.: Stochastic models for heterogeneous DNA sequences. Bull. Math. Biol **51** (1989) 79–94

[8] Ramensky, V., Makeev, V., Roytberg, M., Tumanyan, V.: DNA segmentation through the bayesian approach. Journal of Computational Biology **7** (2000) 215–231

[9] Li, W.: New stopping criteria for segmenting DNA sequences. Physical Review Letters **86** (2001) 5815–5818

[10] Bhanu, B., Lee, S., Ming, J.: Adaptive image segmentation using a genetic algorithm. IEEE Transactions on Systems, Man and Cybernetics **25** (1995) 1543–1567

[11] Yoshimura, M., Oe, S.: Evolutionary segmentation of texture image using genetic algorithms towards automatic decision of optimum number of segmentation areas. Pattern Recognition **32** (1999) 2041–2051

[12] Cagnoni, A., Dobrzeniecki, A., Poli, R., Yanch, J.: Genetic algorithm-based interactive segmentation of 3D medical images. Image and Vision Computing **17** (1999) 881–895

[13] Falkenauer, E.: Genetic Algorithms and Grouping Problems. John Wiley & Sons (1998)

[14] Pearce, S., Ahmed, M.: An evolutionary algorithm for general symbol segmentation. In: Seventh International Conference on Document Analysis and Recognition. (2003) 726–730

[15] Fu, T., Chung, F., Ng, V., Luk, R.: Evolutionary segmentation of financial time series into subsequences. In: Proceedings of the 2001 Congress on Evolutionary Computation CEC2001, IEEE Press (2001) 426–430