Statistical Test-Based Evolutionary Segmentation of Yeast Genome

Jesus S. Aguilar–Ruiz, Daniel Mateos, Raul Giraldez, and Jose C. Riquelme

Dept. of Computer Science, University of Seville, Spain
{aguilar,mateos,giraldez,riquelme}@lsi.us.es

Segmentation algorithms emerge observing fluctuations of DNA sequences in alternative homogeneous domains, which are named segments [1]. 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 this work, we present a new approach based on Evolutionary Algorithms (EAs) that differentiate segments of genes, which are represented by its level of meiotic recombination[1]. We have tested the algorithm with the yeast genome [2][3] 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 (N). Each gene is a row of the file. Each column of file represents a genomic characteristic under specific conditions (in this case, 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. To measure the relevance of segments the Mann–Whitney statistical test has been used.

Each individual of the population is an array of natural numbers with size C, 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 (FC=31) cannot be moved, they have been included in all of the individuals, making easier the computational process. 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. We have chosen the Mann-Whitney test as the fitness function. 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. If we choose the mean as

[1] Meiotic recombination is the exchange of chromosomal segments between the paternal and maternal homologs during meiosis

K. Deb et al. (Eds.): GECCO 2004, LNCS 3102, pp. 493–494, 2004.

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"[2]. The fitness function returns the sum of all tests (for all the cutpoints of an individual).

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.

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.

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. 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 bench mark test based on dynamic programming to avoid the computationally unapproachable exhaustive search.

References

1. Elton, R.: Theoretical models for heterogeneity of base composition in DNA. Journal of Theoretical Biology **45** (1974) 533–553
2. Goffeau, A., et al.: The yeast genome directory. Nature **387** (1997) 5–105
3. http://www.yeastgenome.org/

[2] An apparent experimental result that is not actually real but is due to the experimental methods