A GENETIC ALGORITHM FOR IDENTIFICATION OF REGULATORY SIGNALS

* Stavrovskaya E.D., Mironov A.A.

State Scientific Center GosNIIGenetika, 113545, Moscow, Russia
Integrated Genomics, POBox 348, 117333, Moscow, Russia
e-mail: esta191@fromru.com
*COrresponding author

Key words: regulation signal, genetic algorithms

Resume

There exist numerous algorithms for identification of regulatory signals in unaligned DNA fragments. Here we present a genetic algorithm for signal identification, describe its implementation and testing on simulated data. It is the first application of genetic algorithms in this area.

Introduction

The existing algorithms for identification of regulatory sites can be divided into optimization and combinatorial ones. The former class includes greedy algorithms, expectation-maximization, DMS, MEME; and also stochastic algorithms: simulated annealing and the Gibbs sampler. The combinatorial algorithms are ConsInd and MatInd, WORDUP, CONSENSUS, WINNOWER pattern, graphs, and numerous other algorithms.

The genetic algorithm suggested here can be considered as the optimization one. We believe that it works faster than other stochastic algorithms of comparable recognition power.

Description of Algorithm

The following abstract concepts will be used (to avoid confusion with standard biological terms, they will be italicized): genome, gene, allele, quality of genome, population, crossing, selection and mutation. Consider set of DNA fragments. Each fragment corresponds to a gene, and each position, specifying a candidate site, is an allele. Thus a set of candidate sites, one in each fragment, generates a set of alleles, that is, a genome. Each genome is characterized by its quality, defined as the information content of the respective set of sites. At each step the algorithm processes population, that is a set of genomes, and performs the following operations:

Crossing: select at random a pair of genomes and generate new a one:

Genome1: \[ S_1, S_2, \ldots S_k S_{k+1} \ldots S_n \]
Genome2: \[ T_1, T_2, \ldots T_k T_{k+1} \ldots T_n \]
New genome: \[ S_1, S_2, \ldots S_k T_{k+1} \ldots T_n \]
Position \( k \) of the cut is given by the random uniform distribution.

Selection: Delete the genome with the lowest quality.

Mutation: Select a random gene in a random genome and change the current allele to a random one. It is equivalent to selecting a random site in the corresponding fragment.

These steps are iterated for some fixed time.

Results and Discussion

Each test file contained ten fragments of length 200. The signal was a fixed word of twenty nucleotides. The sites were modeled by introducing some mismatches into the signal word and then inserting the resulting word into the sequence fragments at random positions. The number of mismatches varied from one to seven. To model corrupted samples, some fragments did not contain the signal.

The results are presented in the Table.
Table. Probability of correct signal identification.

<table>
<thead>
<tr>
<th>Number of mismatches</th>
<th>10</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.97</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>1</td>
<td>0.99</td>
<td>0.96</td>
<td>0.90</td>
<td>1.0</td>
<td>0.90</td>
<td>0.70</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>0.99</td>
<td>0.93</td>
<td>0.97</td>
<td>0.90</td>
<td>0.40</td>
</tr>
<tr>
<td>3</td>
<td>0.84</td>
<td>0.74</td>
<td>0.81</td>
<td>0.81</td>
<td>0.90</td>
<td>0.66</td>
</tr>
<tr>
<td>4</td>
<td>0.92</td>
<td>0.77</td>
<td>0.73</td>
<td>0.79</td>
<td>0.68</td>
<td>0.40</td>
</tr>
<tr>
<td>5</td>
<td>0.71</td>
<td>0.53</td>
<td>0.44</td>
<td>0.44</td>
<td>0.43</td>
<td>0.54</td>
</tr>
<tr>
<td>6</td>
<td>0.56</td>
<td>0.70</td>
<td>0.20</td>
<td>0.44</td>
<td>0.55</td>
<td>0.44</td>
</tr>
<tr>
<td>7</td>
<td>0.17</td>
<td>0.27</td>
<td>0.23</td>
<td>0.36</td>
<td>0.22</td>
<td>0.0</td>
</tr>
</tbody>
</table>

The dependency of the mean quality and its standard deviation on the number of iterations for different population sizes (500 and 1000) are presented in the diagrams. It can be seen that in larger populations the mean is larger, the standard deviation is smaller, although the stationary values are reached later. The iteration process should be stopped when the mean becomes practically constant and standard deviation is close to zero.

If two signals are introduced simultaneously, only one of them is found (an arbitrary one).

In these tests the best results were obtained with the population of 1000 genomes at 200000 iterations.

References