MBMEDA: An Application of Estimation of Distribution Algorithms to the Problem of Finding Biological Motifs

Carlos I. Jordán(✉) and Carlos. J. Jordán

Facultad de Ingeniería en Electricidad y Computación,
Escuela Superior Politécnica del Litoral (ESPOL), Guayaquil, Ecuador
cjordan@espol.edu.ec

Abstract. In this work we examine the problem of finding biological motifs in DNA databases. The problem was solved by applying MBMEDA, which is a evolutionary method based on the Estimation of Distribution Algorithm (EDA). Though it assumes statistical independence between the main variables of the problem, results were quite satisfactory when compared with those obtained by other methods; in some cases even better. Its performance was measured by using two metrics: precision and recall, both taken from the field of information retrieval. The comparison involved searching a motif on two types of DNA datasets: synthetic and real. On a set a five real databases the average values of precision and recall were 0.866 and 0.798, respectively.

Keywords: DNA dataset · Estimation of distribution algorithms · Molecular biology · Transcription factor · Motifs

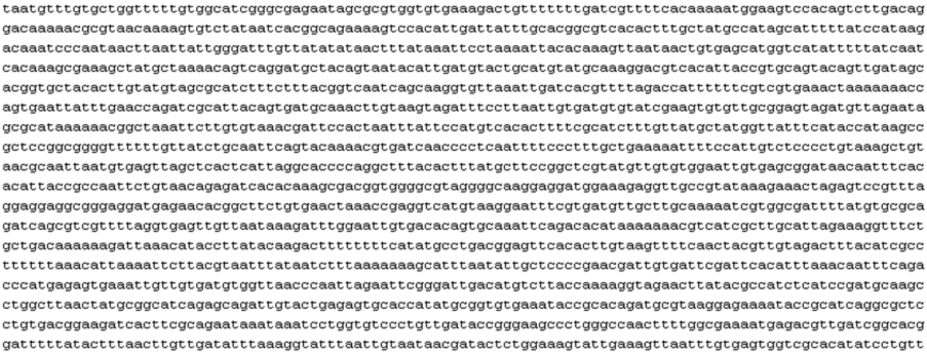
1 Introduction

The search for biological motifs is an important problem in molecular biology. A motif or transcription factor binding site (TFBS) is the sequence of nucleotides in the promoting zone of a gene, where a transcription factor (TF) binds and controls the process of transcription of that gene into an mRNA molecule [1]. This molecule eventually will be translated into a protein at a cells ribosome; all this happens according to the central dogma of molecular biology.

Basically the problem can be formulated as follows: given a DNA base consisting of n promoting zones of size m, with one TFBS per sequence, find a pattern of length l that constitutes a motif. No doubt this problem is rather difficult, because we dont know a priori the length of the motif or its location in the promoting zone, neither the specific sequence of nucleotides we are looking for. To make matters even worse, the TFBS may mutate from one instance to another. Fig. 1 shows how difficult is to find a pattern of nucleotides on a real DNA base.

There exists, however, a key to break this code: the motif is a sequence of nucleotides of length l that repeats with the highest frequency in the DNA

J.M. Ferrández Vicente et al. (Eds.): IWINAC 2015, Part I, LNCS 9107, pp. 39–46, 2015.
DOI: 10.1007/978-3-319-18914-7_5



```
taatgtttgtgctggtttttgtggcatcgggcgagaatagcgcgtggtgtgaaagactgtttttttgatcgttttcacaaaaatggaagtccacagtcttgacag
gacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgattatttgcacggcgtcacactttgctatgccatagcatttttatccataag
acaaatcccaataacttaattattgggatttgttatatataactttataaattcctaaaattacacaaagttaataactgtgagcatggtcatatttttatcaat
cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatgtatgcaaaggacgtcacattaccgtgcagtacagttgatagc
acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcaaggtgttaaattgatcacgttttagaccatttttcgtcgtgaaactaaaaaaacc
agtgaattatttgaaccagatcgcattacagtgatgcaaacttgtaagtagatttccttaattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata
gcgcataaaaaacggctaaattcttgtgtaaacgattccactaatttattccatgtcacacttttcgcatctttgttatgctatggttatttcataccataagcc
gctccggcggggtttttgttatctgcaattcagtacaaaacgtgatcaacccctcaatttccctttgctgaaaaatttccattgtctcccctgtaaagctgt
aacgcaattaatgtgagttagctcactcattaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcac
acattaccgccaattctgtaacagagatcacacaaagcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ggaggaggcgggaggatgagaacacggcttctgtgaactaaaccgaggtcatgtaaggaatttcgtgatgttgcttgcaaaaatcgtggcgattttatgtgcgca
gatcagcgtcgttttaggtgagttgttaataaagatttggaattgtgacacagtgcaaattcagacacataaaaaaacgtcatcgcttgcattagaaaggtttct
gctgacaaaaaagattaaacataccttatacaagactttttttcatatgcctgacggagttcacacttgtaagttttcaactacgttgtagactttacatcgcc
tttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaacgattgtgattcgattcacatttaaacaatttcaga
cccatgagagtgaaattgttgtgatgtggttaacccaattagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgcggtgtgaaataccgcacagatgcgtaaggagaaaataccgcatcaggcgctc
ctgtgacggaagatcacttcgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaaaatgagacgttgatcggcacg
gattttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagttaatttgtgagtggtcgcacatatcctgtt
```

Fig. 1. DNA base for searching the TFBS of CRP in Escherichia Coli

dataset. This clue reduces the problem to a mathematical one, i.e., an optimization problem. To solve it a number of different methods have been devised; among others: MEME (Multiple Expectation Maximization for Motif Elicitation) and BioProspector [2].

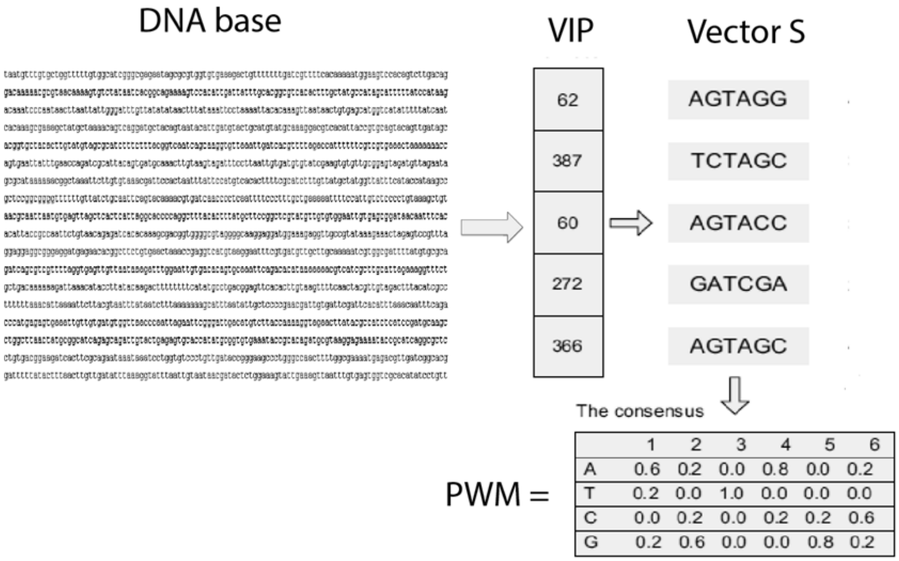
It is well known that optimization problems can be solved efficiently by evolutionary methods [3]. For instance: genetic algorithms are a good option; but in this case we are required to guess appropriate values for the rates of crossover and mutation, which are its classical operators [4]. We could avoid guessing these values if we use the method Estimation of Distribution Algorithm (EDA) [5]. However, in this case the challenge is to construct a good estimator. For the problem of finding biological motifs, it has been proposed in [6] to use a multivariate Gaussian estimator in order to capture possible correlations among the positions in the motif instances.

However, looking for simplicity and better processing times, we assume here that the nucleotides on a motif instance are statistically independent. Then, four univariate Gaussian Estimators (GE) will be required instead of a multivariate one to generate a new individual, where each estimator represents the distribution of a particular nucleotide estimated from the best individuals in the population. Our method will be called MBMEDA (Mtodo de Bsqueda de Motivos con base en un Algoritmo por Estimacin de Distribuciones) and its results will be compared systematically with those of EDAMD (Estimation of Distribution Algorithms for Motifs Discovery) published in [6]; this will allow us to explore two questions: 1) whether our method gives better or similar results compared with those of the multivariate approach, and 2) whether an EDA based motif search algorithm is more efficient than other computational motif search methods.

2 Materials and Methods

To test MBMEDA we used two types of DNA datasets: synthetic and real. A synthetic base is generated artificially following criteria used in other similar works: the length of the motif, the size of the promoting zones and the presence of noise [7]; in this bases a motif is implanted at known sites. On the other hand, in the real or biological DNA datasets, the sequences of nucleotides of the motif were determined experimentally by analyzing a number of promoting regions for each particular organism. Each biological dataset is labeled with the name of the TF that binds on its motif. Here we use five databases: CRP, E2F, ERF, ME2F and MYOD.

MBMEDA is a method that does a global search on the problem space of possible solutions, where a solution -also known as an individual- is defined by a vector VIP of the initial positions of a candidate motif on the n rows of the DNA dataset. Therefore, with each individual we associate a vector S of n sequences of length l that starts at the initial positions specified in VIP; we also associate with each individual a matrix of positional weights, denoted as PWM m x l, where l is the length of the motif sequence and m the cardinality of the nucleotide alphabet, in this case m = 4. The PWM has one row for each symbol of the alphabet: 4 rows in our case; it also has one column for each position in the pattern of sequences [8]. Each entry on the PWM represents the relative frequency of a nucleotide on its correspondent column in S. See Fig. 2.



	1	2	3	4	5	6
A	0.6	0.2	0.0	0.8	0.0	0.2
T	0.2	0.0	1.0	0.0	0.0	0.0
C	0.0	0.2	0.0	0.2	0.2	0.6
G	0.2	0.6	0.0	0.0	0.8	0.2

Fig. 2. Representation of an individual or candidate solution in MBMEDA

The initial population of individuals is usually generated randomly. The quality of a solution is evaluated by the fitness function, which in this case is the information content (IC) of the individual as defined by expression (1) [9] where fb is the frequency that nucleotide b appears at position i on the PWM and pb is the frequency of b on the entire DNA base. At each iteration, a number of the best individuals in the current population will be chosen by a tournament selection operator, in order to model with them how solutions will distribute in the next generation.

$$IC = \sum_{i=1}^{L} \sum_{b} f_b\left(i\right) log\left(\frac{f_b(i)}{p_b}\right) \tag{1}$$

The information content of an individual is a measure of the difference between the distribution of the nucleotides in the PWM -which represents a solution- and the distribution of the nucleotides on the entire DNA dataset. The larger is this difference, the more information content the solution has and, therefore, the larger is the possibility that it to be a motif. This concept of IC is crucial to the process of getting a subset of the best individuals in a population [10]; with them well estimate the four univariate gaussian models that will be used to calculate the next population.

Since we work here with the assumption of statistical independence of nucleotides on the motif instances, we have to estimate a set of four Univariate Gaussian distributions, one for each nucleotide, by calculating their corresponding values of mean and variance [11]. Then, by sampling the frequencies of these distributions using expression (2), well get the components for the new individuals in the next generation.

$$I_b = \mu_b + Z * \sigma_b^2. \tag{2}$$

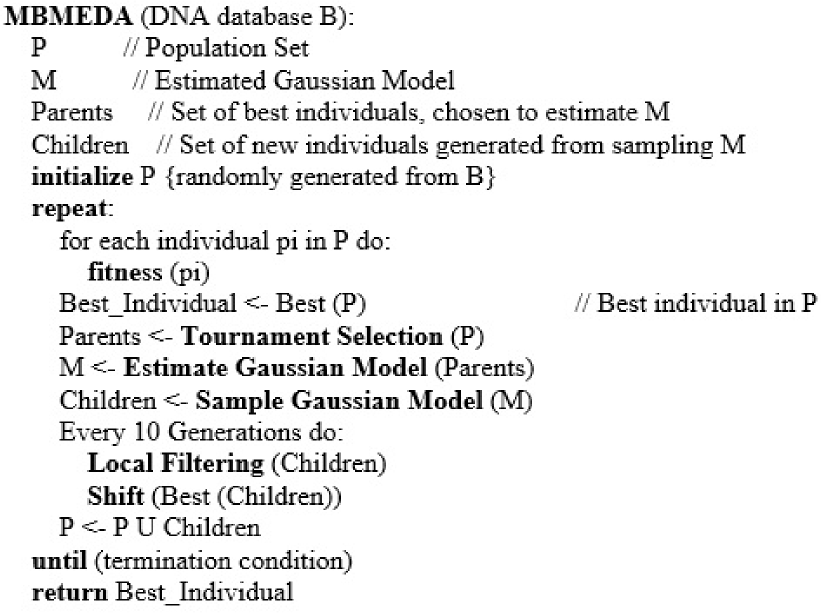
Where I_b represents the component of nucleotide b for a sampled individual, mu_b represents the mean of the distribution for nucleotide b, $sigma_b^2$ represents its variance and Z is a vector of random values obtained by the Box Muller Transformation.

The EDA algorithm iterates until appropriate termination conditions are satisfied; in our case, the value of the fitness function for the best individual remains constant through at least 10 generations [12]. To avoid being trapped on local minima, at each iteration two operators unique to this method are applied after sampling: the Shift and the Local Filtering operators [6]. Fig. 3 presents a pseudo-code for the MBMEDA algorithm.

To measure the performance of EDA so that we are able to compare its results with those obtained by other methods, two metrics were used: Precision and Recall; both were taken from the field of Information Retrieval [13] and calculated by the following expressions (3) and (4), respectively; where N_c represents the correct number of motif instances found by the algorithm, N_p the number of promoter regions in the DNA database and N_t represents the total number of real instances of the motif.

$$Precision = \frac{N_c}{N_p}. \tag{3}$$

$$Recall = \frac{N_c}{N_t}. \tag{4}$$



```
MBMEDA (DNA database B):
  P          // Population Set
  M          // Estimated Gaussian Model
  Parents    // Set of best individuals, chosen to estimate M
  Children   // Set of new individuals generated from sampling M
  initialize P {randomly generated from B}
  repeat:
    for each individual pi in P do:
      fitness (pi)
    Best_Individual <- Best (P)                    // Best individual in P
    Parents <- Tournament Selection (P)
    M <- Estimate Gaussian Model (Parents)
    Children <- Sample Gaussian Model (M)
    Every 10 Generations do:
      Local Filtering (Children)
      Shift (Best (Children))
    P <- P U Children
  until (termination condition)
  return Best_Individual
```

Fig. 3. MBMEDA algorithm

Table 1. Results of MBMEDA applied on synthetic bases

Number of Sequences	**Motif Size**	**Noise**			
		Noiseless		**With-Noise**	
		Pr	**Rc**	**Pr**	**Rc**
100	16	1.00	1.00	0.99	0.97
20	16	0.99	0.99	0.98	0.95
100	8	1.00	1.00	0.99	0.93
20	8	0.99	0.99	0.98	0.92
100	16	0.97	0.97	0.84	0.79
20	16	0.95	0.95	0.93	0.86

3 Results

Table 1 shows the MBMEDAs performance with different synthetic DNA bases that corresponds to each row in the table. When we include noise for each base -which represents a more realistic situation-, the average values for both metrics were above 0.90; this is certainly promising

Table 2. Results of applying MBMEDA and EDAMD on real DNA bases

	MBMEDA		**EDAMD**	
Base	**Pr**	**Rc**	**Pr**	**Rc**
CRP	0.83	0.65	0.94	0.74
ERE	0.80	0.80	0.76	0.76
E2F	0.80	0.74	0.71	0.80
MYOD	1.00	0.80	0.86	0.90
ME2F	1.00	1.00	1.00	1.00

Table 3. Results when MBMEDA and other methods are applied on real DNA bases

	MBMEDA		**MBMAG**		**MEME**		**BioProspector**	
Base	**Pr**	**Rc**	**Pr**	**Rc**	**Pr**	**Rc**	**Pr**	**Rc**
CRP	0.83	0.65	0.88	0.69	0.92	0.52	1.00	0.35
E2F	0.80	0.75	0.76	0.70	0.80	0.70	0.52	0.41
ERE	0.80	0.80	0.76	0.76	0.88	0.60	0.30	0.56
ME2F	1.00	1.00	0.94	0.94	0.93	0.82	0.71	0.71
MYOD	1.00	1.00	0.94	0.76	0.00	0.00	0.00	0.00

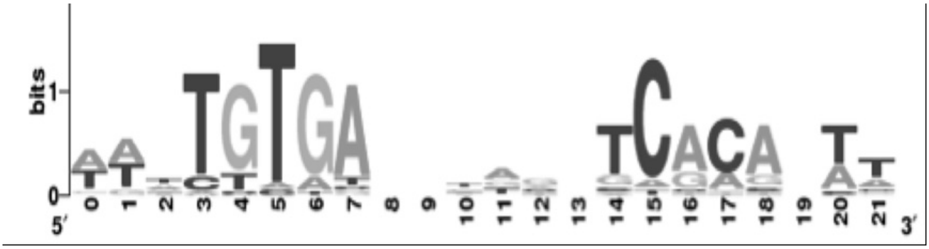


Fig. 4. Sequence logo of CRP motif consensus found experimentally

Table 2 shows the performance of MBMEDA and EDAMD on real datasets; this methods are both based on estimation of distribution algorithms. The average value for Precision for MBMEDA on these bases was 0.886, better than 0.854 for the reference method EDAMD, while for the other metric, Recall, the average value for MBMEDA was 0.798, a bit smaller than the 0.846 obtained for the reference method.

Table 3 presents results for the same real DNA bases as those of Table 2, obtained by applying our method MBMEDA and three others: MBMAG (Mtodo de Bsqueda basado en Algoritmos Genticos), which is a method based on genetic algorithms [9], and two non-evolutionary ones: MEME and BioProspector. The average values for Precision and Recall were 0.886 and 0.798 respectively, higher for the method we propose than for the other three.

Figure 4 shows a sequence logo [14], which is a graphic representation of the consensus word for the TFBS of transcription factor CRP, found experimentally

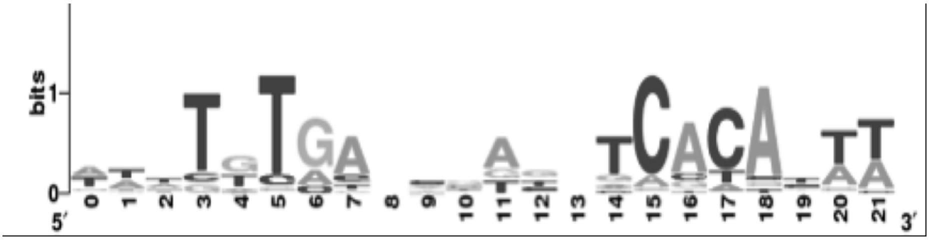


Fig. 5. Sequence logo of CRP motif consensus found by MBMEDA

[6]. Figure 5 on the other hand presents the sequence logo for the same motif as it was found by MBMEDA, the method proposed in this work.

4 Discussion and Conclusion

From the tables and figures above, its clear that the results of applying MBMEDA on DNA synthetic and real databases are quite satisfactory; they are similar and in some cases better than those obtained by other methods, like EDAMD for example.

Comparing Fig. 4 and Fig. 5, we observe that logo sequences for the motif consensus of the TFBS of protein CRP resemble each other quite well, which confirms the good results obtained with MBMEDA when searching for a motif. All this would imply that the assumption of statistical independence among the positions of the nucleotides in the motif instances is a reasonable one. However, we still consider necessary to make a more rigorous analysis of this assumption, which is fundamental to the performance of EDA based methods, since it simplifies the modeling of distributions and the process of sampling new individual for the next population.

Acknowledgements. The authors would like to thank Dr. Daniel Ochoa, Director of the Artificial Vision and Robotics Laboratory at ESPOL, for its constructive comments that helped shape the present form of this work.

References

1. Stormo, G.: DNA binding sites: representation and discovery. Bioinformatics 16(1), 16–23 (2000)
2. Liu, X.: Bioprospector: Discovering Conserved DNa Motifs in Upstream Regulatory Regions of Co-expressed Genes. In: Pacific Symposium on Biocomputing, vol. 6, pp. 127–138 (2001)
3. Hertz, Z., Stormo, G.: Identifying DNA and Protein Patterns with Statistically Significant Aligments of Multiple Sequences. Bioinformatics 15(7), 563–577 (1999)
4. Eiben, E. , Smith, J. : What Is an Evolutionary Algorithm. Introduction to Evolutionary Computing. Springer, New York (2003)

5. Endika, B., Larrañaga, P., Bloch, I., Perchant, A.: Estimation of Distribution Algorithms: a New Evolutionary Computation Approach for Graph Matching Problems. Energy Minimization Methods in Computer Vision and Pattern Recognition, 454–469 (2001)
6. Gang, L., Chan, T., Leung, K., Hong, K.: An Estimation of Distribution Algorithm for Motif Discovery. Evolutionary Computation, 2411–2418 (2008)
7. Wei, Z.: GAME: Detecting Cis-regulatory Elements Using a Genetic Algorithm. Bioinformatics 22(13), 1577–1584 (2006)
8. Sinha, S.: On counting position weight matrix matches in a sequence, with application to discriminative motif finding. Bioinformatics 22(14), 454–463 (2006)
9. Schneider, T., Stormo, G., Gold, L., Ehrenfeucht, A.: Information Content of Binding Sites on Nucleotide Sequences. Journal of Molecular Biology 188(3), 415–431 (1986)
10. Shannon, C.: A Mathematical Theory of Communication. Bell Syst., Techn. J. 27, 379–423 (1948)
11. Jordán, I., Jordán, C.: Aplicación de Algoritmos Evolutivos a la búsqueda de motivos biológicos en bases de regiones promotoras de ADN. Revista Matemática ICM, 33–42 (2012)
12. Fogel, D.: Evolutionary Computation: Toward a new Philosophy in Machine Intelligence. IEEE Press (1995)
13. Manning, D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval, pp. 151–158. Cambridge UP, New York (2008)
14. Schneider, T., Stephens, R.: Sequence Logos: A New Way to Display Consensus Sequences. Nucleic Acids Res. 18(20), 6097–6100 (1990)