A multiobjective approach for gene structure prediction

Javier Pérez-Rodríguez, Alexis G. Arroyo-Peña, and Javier Martínez-Luna

Department of Computing and Numerical Analysis, University of Córdoba, Spain
javier.perez@uco.es, i52arpea@uco.es, i52maluj@uco.es
http://www.cibrg.org/

Abstract. Current advances in DNA sequencing technology have motivated the investigation of reliable gene identification methods, an open area currently present in bioinformatics. Gene recognition can be considered as a search problem, where many evidence sources should be combined in a scoring function that must be maximized to obtain the most likely and right gene structure in a genomic sequence.
In this article, we combine a support vector machine classifier to reduce the search space together with a multiobjective genetic algorithm as main search engine to deal with a set of prospect structures. We made use of various content statistics that are commonly employed to obtain evidences of coding regions in DNA sequences which will determine the probability of a certain structure to be an actual gene.
We use the human sequences located at the chromosomes 3, 19 and 21 as training set, and the chromosome 18 genes to check the performance of our system. Very promising results are obtained.

1 Introduction

The terms gene recognition, gene structure prediction or gene finding are used when determining which parts of a sequence are coding and constructing the whole gene from its start site to its stop codon [16, 1]. The remaining of this work is concerned only with eukaryotic gene recognition, as that is more important and difficult.

There are two basic approaches to gene structure prediction. Homology based approaches search for similar sequences in databases of known genes. These methods are usually called extrinsic methods. The growing number of sequenced genomes and known genes is increasing the potential of homology based methods. However, it is clear that only genes that are somewhat similar to known genes can be identified in this way. The second set of methods are usually known as intrinsic methods, and include two basic approaches [8]: *ab initio* and *de novo* methods. Both are based on obtaining the features that characterize a coding region and/or the functional sites, and using them to find the correct structure of the unknown genes. *Ab initio* methods use only the information of the genome to be annotated (the target genome), whereas *de novo* methods add information of one or more related genomes (the informant genomes).

The methods for obtaining and using that information are many, such as neural networks, hidden Markov models, conditional random fields, etc. These methods try to recognize genomic sequence patterns that are characteristic of splice donor and acceptor sites, translation initiation sites and termination sites, and/or specific features of coding regions. Some of them also try to find other important parts, such as promoters, untranslated terminal regions, polyadenilation sites, etc., although the recognition of these sites is a very difficult problem on its own.

One of the first significant improvements in *ab initio* prediction was GENSCAN, which achieved both accuracy and robustness using a single genome as input. No new significantly better methods were obtained until the development of dual-genome predictors. Dual-genome predictors use two genomes, the genome to be annotated and the genome of a related organism.

We consider gene recognition as a search problem where many sources of evidences are combined to obtain the structure of a probable gene. The approach presented in this paper links two basic methods. Firstly, we use support vector machines (SVMs) [3] to localize the functional sites along the genomic sequence. The second basic part of our approach is evolutionary computation (EC). We approach the problem as a multiobjective search for which we use different content statistics combined in pairs.

We have used for our evaluation five different content statistics measures: in-frame hexamers, average mutual information, position asymmetry, length distribution and local compositional complexity, all of which are tested in an evolutionary framework for gene recognition.

The remainder of the paper is organized as follows: Section 2 describes the framework that will be used and the content measures uses in the study, presenting the multiobjective algorithm that will be used as a gene finder system, Section 3 states the experimental setup, in Section 4 results are showed, and Section 5 provides the conclusions of the study.

2 Evolutionary gene recognition framework

Evolutionary computation [12, 7, 17] is a set of global optimization techniques that have been widely used in the last few years for almost every problem within the field of Artificial Intelligence. As we have explained, gene recognition task can be considered as a search problem, where the objective is to find the most likely right gene structure in a target genomic sequence given. To carry out this search in the possible solution space, it is easy to think in a evolutionary technique due to its attractive features. These methods have the ability to simultaneously search different regions of a solution space, making possible to find a diverse set of solutions for difficult problems with non-convex, discontinuous, and multi-modal solutions spaces. The application of evolutionary computation to gene structure prediction design is based on a two-step procedure. The first step consists of reducing the search space. In a second step, we develop the evolutionary algorithm to find the most likely gene structure.

The first step is devoted to limit the search space. A gene is a structure delimited by two sites, the start and stop codons. Between these two boundaries we have two different substructures, exons and introns, flanked by donor and acceptor sites. If we consider no restrictions, the search space would be huge, and any method would very likely fail. The common approach for reducing the search space is to limit the putative start, splice and stops sites, to the most probable ones. In our system, we consider support vector machines (SVMs) using a string kernel function [6] for site recognition. String kernels are an appropriate and specific function kernel to deal with character sequences.

A second reduction of the search space is achieved by taking into account the constraints in the gene structure:

- The exons do not overlap.
- The gene starts and finishes with an exon.
- An intron must be flanked by two exons.
- A gene can be composed of only one exon.

Evolutionary methods are general purpose randomized optimization techniques which exploit principles inspired from biological systems [7]. A genetic optimization algorithm performs a search by evolving a population of candidate solutions (individuals) modeled with "chromosomes". From one generation to the next, the population is improved by mechanisms derived from genetics. Mathematicaly, an individual can be defined as:

$$I_p = \{x_T, [x_{D1}, x_{A1}, x_{D2}, \ldots, x_{An},]x_S\} \quad x \in \mathbb{N} \tag{1}$$

where each x is a specific functional site and $x_T < x_D < x_A ... < x_S$ (T, tis; D, donor; A, acceptor; S, stop-codon).

The most common form of EC involves the following steps. First, an initial population of chromosomes is randomly generated taking in account the list of splice sites from first step and the biological restrictions of a correct gene structure. Then, the goodness of each chromosome is evaluated according to a predefined fitness function representing the considered objective function. This fitness evaluation step allows one to keep the best chromosomes and reject the worst ones by using an appropriate selection rule based on the principle that the better the fitness, the higher the chance of being selected. Once the selection process is completed, the next step is devoted to reproducing the population. This is done by genetic operators such as crossover and mutation operators. The entire process is iterated until a user-defined convergence criterion is reached.

2.1 Multi-Objective Genetic Algorithms (MOGAs)

A multi-objective optimization problem has a number of objectives, each of them is to be either minimized or maximized. MOGAs use genetic algorithms (GA) to optimize these objectives simultaneously and result in a Pareto-optimal front (a solution set) for higher level analysis and decision. In order to find the Pareto

front, several multiobjective GA-based approaches have been proposed in the literature.

In this paper, we will adopt the nondominated sorting genetic algorithm (NSGA-II) for its low computational requirements and its ability to distribute uniformly the solutions along the Pareto front [4]. It is based on the creation of an initial random parent population. Individuals selected through a crowded tournament selection undergo crossover and mutation operations to form an offspring population. Both offspring and parent populations are then combined and sorted into fronts of decreasing dominance (rank). After the sorting process, the new population is filled with solutions of different fronts starting from the best. If a front can only partially fill the next generation, crowded tournament selection is used again to ensure diversity. Once the next-generation population has been filled, the algorithm loops back to create a new offspring population and the process continues up to convergence.

2.2 Content statistics

Several statistical features are often used for discriminating between coding and non-coding DNA regions. The measures are used in pairs within a multiobjective framework. A comparative study on the discriminating power of these features by themselves was previously conducted within an GA framework [18]. In that case, in a MOGA framework, a total of five content statistics have been studied: in-frame hexamers, local compositional complexity, position asymmetry, length distribution and average mutual information.

In-frame hexamer statistics are related to codon usage bias. Position asymmetry is related to the asymmetric feature of the distribution of nucleotides at the three codon positions. Local compositional complexity is based on richness of exon information. Length distribution is based on the different average lengths of exons and introns. Average mutual information statistics are related to the correlation between nucleotides at a certain distance. A description of these statistics follows:

1. **In-frame hexamer frequency (IFH).** It has long been known that synonymous codons are not used with equal frequencies and that different organisms differ in their patterns of codon usage. The in-frame hexamer score for the interval starting at nucleotide i and ending at j, $IF_6(i,j)$ is calculated as follows:

$$IF_6(i,j) = max \begin{cases} \sum_{k=0,3,6,\ldots,j-6} & ln(\frac{f_k}{F_k}) \\ \sum_{k=1,4,7,\ldots,j-6} & ln(\frac{f_k}{F_k}) \\ \sum_{k=2,5,8,\ldots,j-6} & ln(\frac{f_k}{F_k}) \end{cases} \quad (2)$$

   where f_k is the frequency, in the table of in-frame hexamers in human coding sequences, of the hexamer starting at position k in the interval. In the calculation, F_k is the frequency of the same hexamer in a random population based on the base composition of the sequence. Hexamers with occurrences equal to those expected by composition have $IF_6 = 0$, those preferred have

a positive score and those avoided, a negative score. First exons, last exons, internal exons, unique exons and introns hexamers are evaluated from different frequency matrices.

2. **Local compositional complexity (LCC).** In non-coding regions of the eukaryotic genomes is typical to find large amounts of repetitive DNA sequences. In contrast, coding regions hold information richness. This property, quantified by the Shannon information [20], is a measure of the local redundancy of the sequence. We can define a local compositional complexity of a segment as a statistical property to distinguish between coding and non-coding sequences. This local entropy measure, LCC, using a segment of nucleotides of length L, is defined as:

$$LCC = -\sum_{k=\{A,C,G,T\}} (\frac{N_k}{L}) log_2 (\frac{N_k}{L}) \tag{3}$$

where N_k is the number of times base k occurs in the segment of nucleotides of length L [14].

3. **Position asymmetry (PA).** Let $f(b,r)$ be the relative frequency of nucleotide b at codon position r. Let $f(b) = \sum_{r=1}^{3}(f(b,r))/3$ be the average frequency of nucleotide b at the three codon positions, and define the asymmetry in the distribution of nucleotide b as the variance of this frequency, i.e., $asym(b) = \sum_{r=1}^{3}(f(b,i) - f(b))^2$, and the PA of the sequence is defined as follows [9]:

$$PA = asym(A) + asym(C) + asym(G) + asym(T) \tag{4}$$

4. **Length distribution (LD).** This statistic provides evidence that introns and exons have different extreme and average lengths [10]. Even within the class of exons, the length distributions of first, last and internal exons all differ significantly from one another. This information can be used as evidence that an interval is a member of a particular sequence type by looking up the frequency of the interval length in a table. A low score can be used as strong evidence that the interval is not part of the actual solution.
5. **Average mutual information (AMI).** The correlation ($\rho_{ij}(k)$) between nucleotide i and nucleotide j at a distance of k nucleotides can be calculated as $\rho_{ij}(k) = p_{ij}(k) - p_i p_j$, where p_i and p_j are the probabilities of nucleotides i and j in the sequence and $p_{ij}(k)$ is the probability in the sequence of the pair of nucleotides i and j at a distance of k nucleotides [11]. Thus, for each distance k, 16 different individual correlations can be calculated. A measure that summarizes all individual correlations at a given distance k is the mutual information function,

$$I(k) = \sum P_{i,j}(k) log_2 (\frac{P_{i,j}(k)}{P_i P_j}) \tag{5}$$

The mutual information $I(k)$ quantifies the amount of information that can be obtained from one nucleotide about another nucleotide at a distance k. In coding DNA, $I(k)$ oscillates between two values, whereas in non-coding DNA, $I(k)$ is rather flat. The two values between which $I(k)$ oscillates in coding DNA in the in-frame mutual information are called I_{in} at distances $k =$

$2, 5, 8, \ldots$, and the out-of-frame mutual information I_{out} at $k = 4, 7, 10, \ldots$. To reduce the pair of numbers I_{in} and I_{out} to a single quantity, we compute the average mutual information (AMI) as follows:

$$AMI = \frac{I_{in} + 2I_{out}}{3} \tag{6}$$

3 Methodology

Four different SVM models were trained to predict the potential translation initiation sites, donor splice sites, acceptor splice sites and stop codons in a sequence. Using the potential site lists, we performed the evolutionary process described in Section 2.1. The evolution to obtain the final gene structures population of each test sequence was performed for a number of generations. Taking into account the results of the studies carried out on previous works, four different setups of NSGA-II were executed, each of one with using as objective a combination of two of the mentioned measures:

- In-frame hexamer frequency and Local compositional complexity
- In-frame hexamer frequency and Length distribution
- In-frame hexamer frequency and Position Asymmetry
- In-frame hexamer frequency and Average Mutual Information

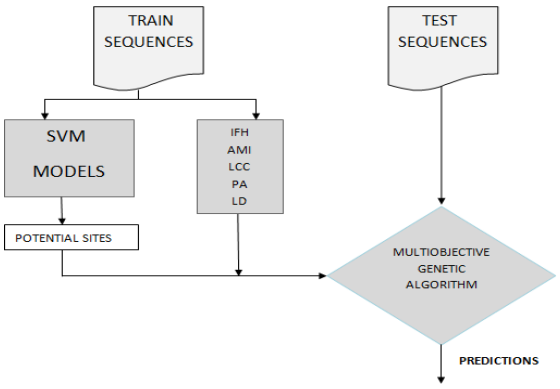


Fig. 1: System architecture.

3.1 Decision Maker

Once the stop criterion of the algorithm has been reached, a final evolved population has been obtained. All the individuals of this population are non-dominated individuals, in accordance with the sense of the strict concept of dominance.

Several actions can be conducted once this point is reached. If the population is relatively small, this can be examined by a human expert. However, if we follow a completely automatic system, it is necessary the implementation of a decision-maker (DM). This DM will be responsible of choosing one of the individuals of the Pareto front, by examining its features. In our experiments, the G-mean of the value of the considered objectives has been used like DM, despite the fact that it represents a field to be researched by itself.

3.2 Evaluation measures

Accuracy is not a useful measure for imbalanced data. In the prediction of gene structure, the ratio of coding against no coding regions is heavily imbalanced, and therefore other measures must be used. Several measures have been developed that consider the imbalanced nature of the problems. Given the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), we can define the following two basic measures: sensitivity $Sn = \frac{TP}{TP+FN}$ and specificity $Sp = \frac{TN}{TN+FP}$.

These are common measures in any class-imbalance problem. There are also specific measures to the gene recognition task. One of the most commonly used measures of this type is the correlation coefficient, CC:

$$CC = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{(PP)(PN)(AP)(AN)}} \tag{7}$$

where PP are the predicted positives, AP the actual positives, PN the predicted negatives and AN the actual negatives. CC will be our main measure of the performance of the method.

The source code used for all methods, in C and licensed under the GNU General Public License, as well as the partitions of the datasets, are freely available on request from the authors.

4 Experimental setup and results

The system was tested on the chromosome 18 of the human genome, and trained with the chromosomes 3, 19 and 21. Chromosome 3 has 4 contigs where 1497 genes are distributed, chromosome 19 has 1767 genes distributed on 4 contigs and chromosome 21 has 312 genes on 8 contigs. The size of the whole dataset is more than 200 million nucleotides. Figure 2 plots in bps the genes sequences lengths in chromosome 18.

We used the training dataset to obtain our content statistics. These content statistics are the objectives that guide our algorithm multiobjective search.

For setting the parameters of NSGA-II, we used k-fold cross-validation, where k is the total number of training contigs. Thus, we obtained a population size of 250 individuals, $p_{\mathrm{CRU}} = 0.2$ crossover probability, $p_{\mathrm{MUT}} = 0.01$ mutation probability and a window length value 11bps for LCC statistic. The maximum number of generations was set to 5000 in case the stop criterion is not reached.

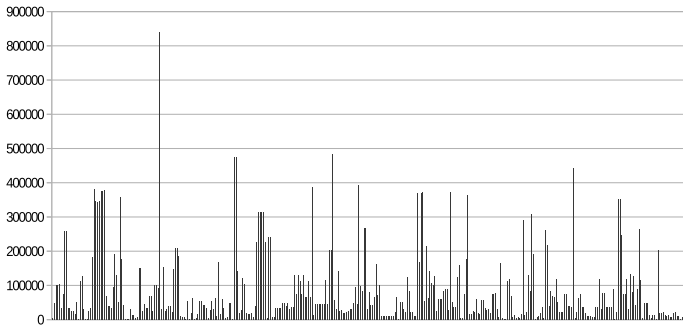


Fig. 2: Lengths of the test sequences.

Figure 3 shows the performance of SVM classifiers. Receiver operating characteristic (ROC) curve is a graphical plot which illustrates the behavior of a classifier when its discrimination threshold is varied. Its corresponding area under the curve (AUC) give a numerical measure of its performance. We can notice AUC has a value over 0.9 in the four cases.

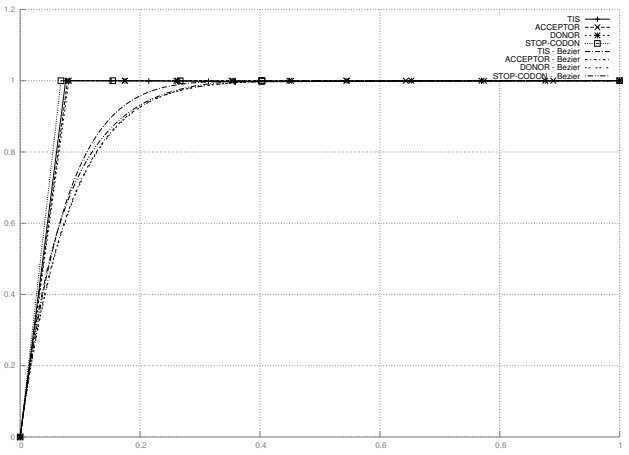


Fig. 3: ROC curves that illustrate the site SVMs performances.

The results obtained in the experimental process are shown in table 1. It shows the mean of the achieved accuracies in each of the 382 genes of chromosome 18.

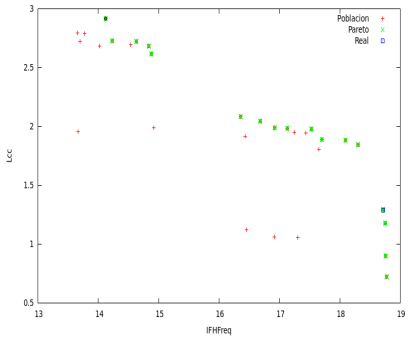
Table 1 shows the specificity, sensitivity and correlation coefficient at the nucleotide level. The combination that uses IFH combined with LCC as objec-

Objetives	Specificity	Sensitivity	CC
IFH+LCC	0.712	0.475	0.597
IFH+LD	0.667	0.444	0.558
IFH+PA	0.638	0.456	0.536
IFH+AMI	0.600	0.427	0.519

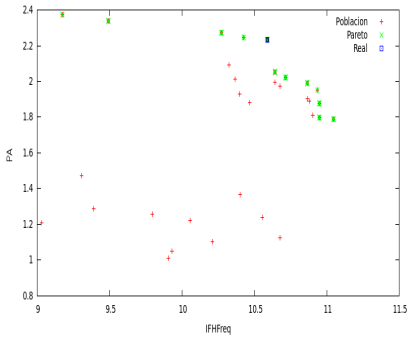
Table 1: Comparative statistical results for genes in chromosome 18 at nucleotide level.

tives is clearly the most discriminant one, achieving better accuracy than all of the remaining at the nucleotide level. IFH+LD and IFH+LCC show a medium performance at this level.

To analyze the MOGA final population and individuals that composed it, could be another interesting topic. In that sense, we can state that the actual solution, not always but often, is found inside the final population. However, that individual is chosen by the DM with much lower frequency. Figure 4 shows two cases where the actual solution is in the Pareto front and it is chosen by the DM system.



(a) IFH and LCC scores on OSBPL1A gene predictions.



(b) IFH and PA scores on C18orf1 gene predictions.

Fig. 4: Examples of Pareto front.

.

5 Conclusions

In this paper, we have presented the first attempt of using evolutionary multi-objetive computation as the main tool for gene structure prediction reported as yet. A simple system is presented where no other search paradigm is used. The system achieves good results at the nucleotide level using efficiently a very small amount of information.

This study has concluded that the combination between in-frame hexamer frequency and local compositional complexity as the only two objectives to maximize is the best performing pair, improving over the performance of each measure isolated [18] and the remaining of combinations. The superior performance of this system has been confirmed at nucleotide level. These promising results (in comparison with [8]) and the flexibility of the methodology will provide a tool that can deal with more complex problems, as alternative splicing, non-canonical functional sites, ignored stop codons, etc.

In future research, it would be interesting, although computationally quite expensive, to perform the same study by grouping the measures in three or more set of objectives. The proposed methodology opens a new field of application of genetic algorithms to gene structure prediction. Many new sources of evidence can be added to the system, as well as more sophisticated evolutionary methods. In the same sense, a more in depth decision maker study is required to improve the final solution selection.

References

1. Brent, M.R., Guigó, R.: Recent advances in gene structure prediction. Current Opinion in Structural Biology 14, 264–272 (2004)
2. Claverie, J., Sauvaget, I., Bougueleret, L.: k-tuple frequency analysis from intron/exon discrimination to t-cell epitope mapping. Methods Enzymology 183, 237–252 (1990)
3. C. Cortés and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
4. K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.
5. García-Pedrajas, N., Pérez-Rodríguez, J., García-Pedrajas, M.D., Ortiz-Boyer, D., Fyfe, C.: Class imbalance methods for translation initiation site recognition in dna sequences. Knowledge-Based Systems 25, 22–34 (2012)
6. H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Christiani, and C. Watkins, "Text classification using string kernels," Journal of Machine Learning Research, vol. 2, pp. 419–444, 2002.
7. D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison–Wesley, 1989.
8. Gross, S.S., Do, C.B., Sirota, M., Batzoglou, S.: CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. Genome Biology 8, R269.1–R269.16 (2007)
9. Guigó, R.: DNA composition, codon usage and exon prediction. In: Bishop, M. (ed.) Genetic Databases, pp. 53–80. Academic Press (1999)
10. Hawkins, J.D.: A survey of intron and exon lengths. Nucleic Acids Research 16, 9893–9908 (1988)
11. Herzel, H., Große, I.: Measuring correlations in symbolic sequences. Physica A 216, 518–542 (1995)
12. J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: The University of Michigan Press, 1975.

13. Japkowicz, N.: The class imbalance problem: significance and strategies. In: Proceedings of the 200 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning. vol. 1, pp. 111–117. Las Vegas, USA (2000)
14. Konopka, A.K., Owens, J.: Complexity charts can be used to map functional domains in DNA. Genetic analysis, techniques and applications 7(2), 35–38 (1990)
15. X. Li, Y. Yan, and Y. Peng, "The method of text categorization on imbalanced datasets," in *Proceedings of the 2009 International Conference on Communication Software and Networks*, 2009, pp. 650–653.
16. Mathé, C., Sagot, M.F., Schiex, T., Rouzé, P.: Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research 30(19), 4103–4117 (2002)
17. Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*. New York: Springer–Verlag, 1994.
18. J. Pérez-Rodríguez, A.G. Arroyo-Peña, N. García-Pedrajas: A Comparative Study of Content Statistics of Coding Regions in an Evolutionary Computation Framework for Gene Prediction. IEA/AIE 2012: 206-215
19. Pérez-Rodríguez, J., García-Pedrajas, N.: An evolutionary algorithm for gene structure prediction. IEA/AIE 2011: 386-395
20. Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. The University of Illinois Press, Urbana, IL (1964)