Optimization of multi-classifiers using a fuzzy logic approach: an application to the gene prediction problem

Rocío Romero-Zaliz[1] **Coral del Val**[1] **Igor Zwir**[1,2]

[1] DECSAI, UGR, Granada, Spain, {rocio,delval,igor}@decsai.ugr.es
[2] Howard Hughes Medical Institute, St. Louis, Missouri, USA

Abstract

Genomes of many organisms have been sequenced over the last few years. However, transforming such raw sequence data into knowledge remains a hard task. A great number of prediction programs have been developed to address part of this problem: the location of genes along a genome. We propose a multi-objective methodology using fuzzy logic to combine algorithms into an aggregation scheme in order to obtain optimal methods' aggregations. Results show improvements in specificity and sensitivity when our methodology is compared to the performance of individual methods for gene finding problems. The here proposed methodology is an automatic method generator, and a step forward to exploit all already existing methods, by providing optimal methods' aggregations to answer concrete queries for a certain biological problem with a maximized accuracy of the prediction. As more approaches are integrated, *de novo* accuracy can be expected to improve further.

1 INTRODUCTION

Genomes of many organisms have been sequenced over the last few years. However, transforming such raw sequence data into knowledge remains a hard task. A great number of prediction programs have been developed to address one part of this problem: the location of genes along a genome [2, 3, 1, 9]. Unfortunately, finding genes in a genomic sequence is far from being a trivial problem. Gene prediction is one of the most important problems in computational biology due to the inherent value of the set of protein-coding genes for other analysis.

Despite the advances in the gene finding problem, existing approaches to predicting genes have intrinsic advantages and limitations [11]. Furthermore, there is no program that can provide perfect predictions for any given input. Our methodology combines these approaches into an aggregation scheme to provide better predictions by taking advantage of the different methodologies' starknesses and avoiding their weaknesses. Moreover, we use a multi-objective approach to extract the best aggregation of methods by maximizing the specificity and sensitivity of their predictions.

We applied our methodology to a reference dataset in gene prediction containing 570 multi-species DNA sequences of known genes [5].

2 MATERIALS AND METHODS

The aggregation of methods is accomplished by using fuzzy union -∪- and fuzzy intersection -∩- operators [8, 14]. All potential aggregations conform a space of potential hypotheses, which can be represented as a lattice structure (Figure 1). We search for the best aggregation of methods, moving from hypothesis to hypothesis towards the most general (i.e., the union of all methods) and the most specific (i.e., the intersection of all methods) which are located at the top and the bottom of the lattice, respectively [12] (Figure 1). In the gene finding problem we explore three methods, n = 3, termed M1 to M3, conforming a total set of seven potential aggregations.

The aggregation of the different methods in the gene finding problem is performed at a nucleotide level. This aggregation joins two overlapping or adjacent exons into a new exon (Figure 2 amd 3) taking into account their exon probabilities.

Even though most *ab initio* gene finders develop a scoring scheme for exon prediction, many of them only report meaningless scores referring to the predicted

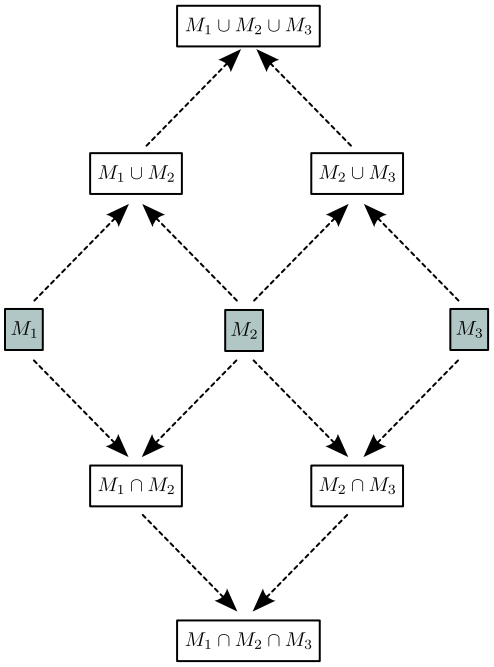


Figure 1: Lattice of potential hypothesis, methods' aggregations of $M_1, \cdots M_n$ using the -∪- and -∩- operators. The solid arrows show the direction of the search in the space of hypothesis.

exons. Although some gene finders, such as GENSCAN, give a probabilistic score to every predicted exon, the score does not respond to the likelihood correctly and is not reliable, especially when implementing in large DNA sequences [4]. Therefore, we applied the local polynomial regression method, a nonparametric regression model, to transform the raw scores to probabilistic ones as implemented in [10].

To perform the aggregation of exons using the fuzzy union and intersection operators, we first need to introduce some notation. We define the *exon* fuzzy set X as the a pair (A, m) where A is a set and $m : A \rightarrow [0, 1]$. For each $x \in A, m(x)$ is the grade of membership of x, where m corresponds to the probabilistic score calculated from the raw scores of each gene finder.

The fuzzy union operator joins two overlapped exons –exon x and exon y– when $m(x)$ and $m(y)$ are higher than a certain threshold. If $m(x) > \gamma$ while $m(y) < \lambda$, only exon x is kept (Figure 2 (c)). If $m(x) > \gamma$ and $m(y) > \gamma$, a new exon z is constructed by appending both exons (Figure 2 (b)) with $m(z) = max(m(x), m(y))$. If there is no overlap, only the exons with membership above threshold λ are kept (Figure 2 (a)).

The fuzzy intersection operator intersects two overlapped exons –exon x and exon y– when $m(x)$ and $m(y)$ are, again, higher than a certain threshold. If $m(x) > \lambda$ and $m(y) > \lambda$, a new exon z is constructed by taking only those nucleotides appearing in both exons (Figure 3 (b)) with $m(z) = min(m(x), m(y))$. If there is an overlap but $m(x) < \lambda$ or $m(y) < \lambda$, then no intersection is performed (Figure 3 (a)). If there is no overlap, neither exon x nor exon y is kept (Figure 3 (c)).

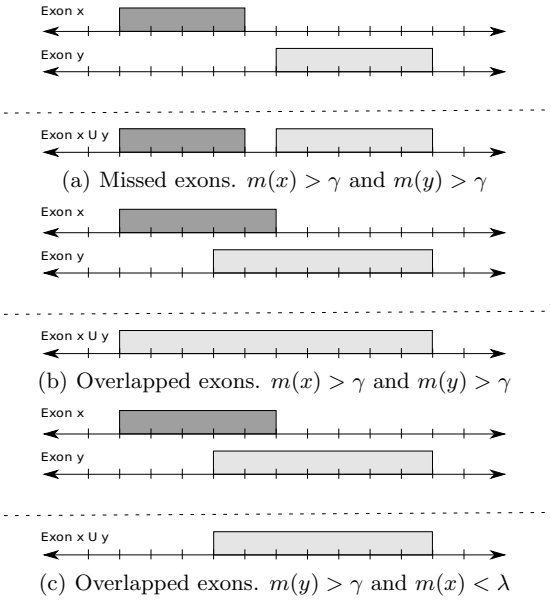
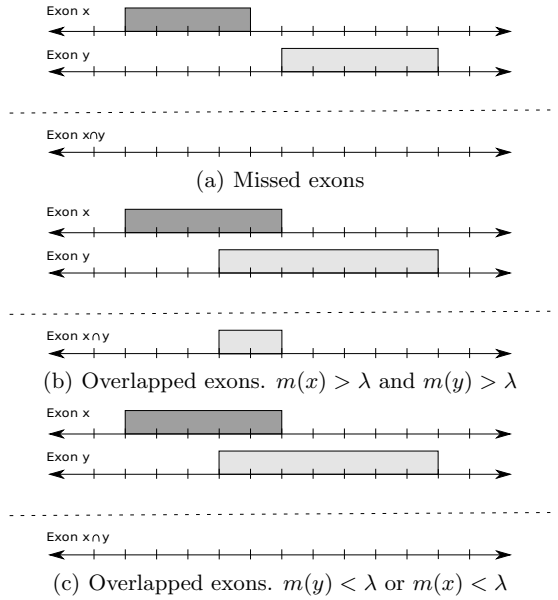


Figure 2: Example of exons aggregation by the fuzzy union operator.

For the experimental section we used a threshold $\gamma = 0.8$ and a threshold $\lambda = 0.2$.

2.1 DATASET

We selected the dataset from Guigó et al. [5] which is a reference for assessing the quality of gene prediction programs. This set contains 570 sequences from vertebrate genomes 570, having only those sequences representing only one complete spliceable functional product of a gene in the forward strand. The programs used in this study are Genscan [1], GeneID [7] and Augustus [16]. Genscan uses a general probabilistic model for the gene structure of human genomic sequences. It has the capacity to predict multiple genes in a sequence, to deal with partial as well as complete genes, and to predict consistent sets of genes occurring on either or both DNA strands [1]. GeneID combines different algorithms using Position Weight Arrays to detect features such as splice sites, start and stop codons and Markov Models to score exons



(a) Missed exons

(b) Overlapped exons. $m(x) > \lambda$ and $m(y) > \lambda$

(c) Overlapped exons. $m(y) < \lambda$ or $m(x) < \lambda$

Figure 3: Example of exons aggregation by the fuzzy intersection operator.

and Dynamic Programming (DP) to assemble the gene structure [7]. Augustus is a gene predictor for eukaryotic genomic sequences that is based on a generalized hidden Markov model, a probabilistic model of a sequence and its gene structure [16].

2.2 MEASURE OF ACCURACY OF PREDICTIONS

We measured the accuracy of a prediction on a test sequence by comparing the predicted coding value (coding or non-coding) with the true coding value for each nucleotide along the test sequence. This has been one of the most widely used approaches in evaluating the accuracy of coding region identification and gene structure prediction methods. Nucleotide level accuracy is calculated as a comparison of the annotated nucleotides with the predicted nucleotides. Sensitivity (Sn) (Equation 1) is the proportion of annotated nucleotides (as being coding or part of an mRNA molecule) that is correctly predicted, and specificity (Sp) (Equation 2) the proportion of predicted nucleotides (as being coding or part of an mRNA molecule) that is so annotated. As a summary measure, we have computed the correlation coefficient (CC) (Equation 3) between the annotated and the predicted nucleotides [5].

$$Sn = \frac{TP}{TP + FN} \quad (1)$$

$$Sp = \frac{TP}{TP + FP} \quad (2)$$

$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN_FP) \times (TP + FP) \times (TN + FN)}} \quad (3)$$

3 RESULTS

Out of all gene prediction programs analyzed and all methods' aggregations, the union of all methods – Genscan ∪ GeneID ∪ Augustus– achieved the highest number of correctly predicted genes[1] (525 out of 570, over 92% of the dataset) and the highest average CC, 0.896 (Table 1). Whats more, these percentage of correctly predicted gene is increased by a 10% approximately when compared with the best individual predictor, GeneID. The Genscan ∪ GeneID methods' aggregation achieved the best specificity values while the union of all methods obtained the highest sensitivity value. All methods' aggregations using the fuzzy union operator obtained better sensitivity values compared to the individual methods. Moreover, some of these methods's aggregations' specificity values are also better than most of the individual gene predictors, while the others do not differentiate to much from them. If a crisp union operator is used, the sensitivity values are increase, but most of the time its specificity values decrease (data not shown) [15].

On the other hand, the fuzzy intersection operator proposed did not produce better results than individual methods (Table 1). This is mainly due to the fact that the fuzzy intersection greatly decreases the sensitivity of the results, and thus producing a very low CC.

A graphical representation of the methods' aggregations performance can also be seen in Figure 3. Specificity and sensitivity values are plotted for all methods' aggregations, both using the fuzzy union or intersection fuzzy operators. Methods' aggregations belonging to the Pareto set are highlighted in red, i.e., those methods that are both better in specificity and sensitivity than the rest. We can therefore infer that Genscan ∪ GeneID ∪ Augustus and Genscan ∪ GeneID methods' aggregations are better in both specificity and sensitivity than individual methods.

If we take a closer look into the results we can extract many specific genes where individual methods fail, while the aggregation of methods produced better results (e.g., MMU12565, HUMSEMIIB, MMIL5G).

[1] We express the accuracy of the method aggregation by considering a gene correctly retrieved when its CC > 0.7.

Method	Sp	Sn	CC	Correctly predicted %
Genscan	0.885	0.753	0.753	78.42%
GeneID	0.899	0.808	0.830	82.28%
Augustus	0.829	0.715	0.796	73.33%
Genscan ∪ GeneID	*0.902*	0.903	0.881	90.35%
Genscan ∪ Augustus	0.882	0.841	0.847	84.56%
Augustus ∪ GeneID	0.900	0.894	0.886	90.00%
Genscan ∪ GeneID ∪ Augustus	0.893	*0.928*	*0.896*	*92.11%*
Genscan ∩ GeneID	0.836	0.622	0.680	64.91%
Genscan ∩ Augustus	0.783	0.586	0.657	61.93%
Augustus ∩ GeneID	0.809	0.613	0.696	63.16%
Genscan ∩ GeneID ∩ Augustus	0.757	0.517	0.601	52.98%

Table 1: Results obtained by all methods' aggregation using both the fuzzy union and the fuzzy intersection operators. The best result for each column is highlighted in italic and color-coded in blue.

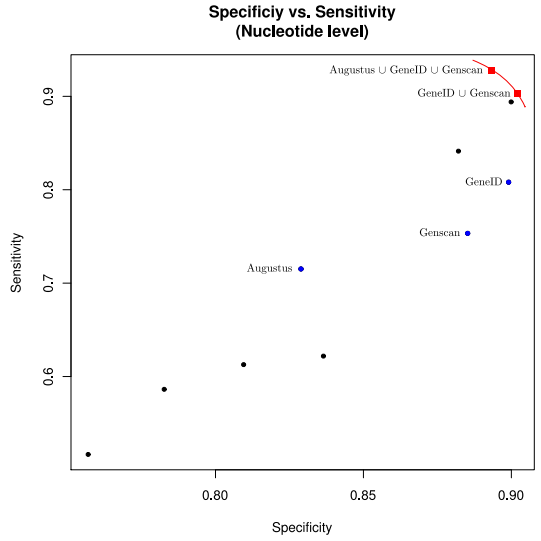


Figure 4: Graphical representation of the specificity and sensitivity values obtained by the methods' aggregations using the fuzzy union and fuzzy intersection operators to predict genes. Methods' aggregations belonging to the Pareto set are highlighted in red.

4 DISCUSSION

We propose a methodology to combine programs into a aggregation scheme using fuzzy logic operators. This idea provides better predictions by combining the advantages of the different methodologies used in each program. We introduced the use of a multi-objective approach to extract the best aggregation of methods by maximizing the specificity and sensitivity of their predictions. This way we avoid redundant and overlapping predictions that might be produced depending on the methodologies and the aggregation scheme used. The application of the proposed methodology to the gene finding problem to obtain optimal methods' aggregations showed an improvement in both sensitivity and specificity when compared to the performance of individual methods. The specificity levels obtained by the aggregation of gene finding methods improved or decreased depending on the methods used in the aggregation. When determining which aggregation of methods was the best one for the gene prediction problem, sensitivity and specificity were in contradiction. Nevertheless, the calculation of the correlation coefficient helped in the selection of the best methods' aggregation. The best aggregations include methods employing different algorithmic strategies that predict correctly different subset of the genes in the dataset. Although the statistical properties of coding regions allow for a good discrimination between large coding and non-coding regions, the exact identification of the limits of exons or of gene boundaries remains difficult.

There are several previous works combining gene finding programs [13, 17], but they fail to obtain good results as they use simultaneously all programs instead of optimize their aggregation. *De novo* gene prediction for compact eukaryotic genomes is already quite accurate, although mammalian gene prediction lags way behind in accuracy. One future scope would be the application of this approach to identify ways to quickly combine many or all-existing programs trained for the same organism, and determine the upper limit of predictive power by aggregations of programs genome wide [6].

In the last ten years, the existing competitive spirit has increased the number of programs/algorithms created, updated and adapted for the two biological problems here presented [11, 2, 9]. On one side the development of a new algorithm always implies the sacrifice of an objective in favor of another, which makes very difficult for novel approaches to improve in absolute terms

the quality of the existing ones. On the other side, the impressive amount of alternative algorithms available for different biological problems is confusing the users, who wonder what makes the programs different, which one should be used in which situation and which level of prediction confidence to expect. Finally, users also wonder whether current programs can answer all their questions. The answer is most probably no, and will remain to be negative as it is unrealistic to imagine that such complex biological processes can be explained merely by looking at one objective. The here proposed methodology is an automatic method generator, and a step forward to exploit all already existing methods, by providing optimal methods' aggregations to answer concrete queries for a certain biological problem with a maximized accuracy of the prediction.

References

[1] C. Burge and S. Karlin. Finding the genes in genomic dna. *Struct. Biol.*, 8:346–354, 1998.

[2] J. M. Claverie. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.*, 6:1735–1744, 1997.

[3] R. Guigó. Computational gene identification: an open problem. *Comput. Chem.*, 21:215–222., 1997.

[4] R Guigó, P. Agrawal, J.F. Abril, M. Burset, and J.W. Fickett. An assessment of gene prediction accuracy in large dna sequences. *Genome Res.*, 10:16311642, 2000.

[5] R. Guigó and M. Burset. Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–367, 1996.

[6] R. Guigó et al. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447, 2007.

[7] R. Guigó, S. Knudsen, et al. Prediction of gene structure. *J. Mol. Biol.*, 226:141–157, 1992.

[8] P. Halmos. *Naive Set Theory*. D. Van Nostrand Company., Princeton, NJ, 1960.

[9] D. Haussler. Computational genefinding. *Trends Biochem. Sci.*, pages 12–15., 1998.

[10] Xiao Li, Qingan Ren, Yang Weng, Haoyang Cai, Yunmin Zhu, and Yizheng Zhang. Scgpred: A score-based method for gene structure prediction by combining multiple sources of evidence. *Genomics, Proteomics & Bioinformatics*, 6(Issues 3-4):175–185, 2008.

[11] C. Mathé, M. F. Sagot, et al. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.*, 30(19):4103–4117, 2002.

[12] T. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[13] K. Murakami and T. Takagi. Gene recognition by combination of several gene-finding programs. *Bioinformatics*, 14:665–675, 1998.

[14] W. Pedrycz, P. Bonissone, and E. Ruspini. *Handbook of fuzzy computation*. Institute of Physics, 1998.

[15] R. Romero-Zaliz, C. Rubio-Escudero, I. Zwir, and C. del Val. Optimization of multi-classifiers for computational biology: application to gene finding and expression. *Theor Chem Acc*, 125, 2010. In Press.

[16] Mario Stanke and Stephan Waack. Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics*, 19, Suppl. 2:ii215–ii225, 2003.

[17] M. Tech and R. Merkl. Yacop: Enhanced gene prediction obtained by a combination of existing methods. *Silico Biology*, 3(4):441–451, 2003.