A Multi-Agent System for Exon Prediction in Human Sequences

Laurence Vignal [1] **Frédérique Lisacek** [2]
vignal@lirmm.fr Frederique.Lisacek@genetique.uvsq.fr

[1] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier
161 rue Ada, 34392 Montpellier Cedex 5, France

[2] Université de Versailles Saint-Quentin,
45 avenue des États-Unis, 78035 Versailles, France

Abstract

Given the problem of identifying exons in new genomic DNA, the sketch of a resolution process was drawn using sequence data and models of site/signal recognition. A multi-agent architecture is used to validate these models and test hypotheses on the chronology of events involved in gene splicing. Information is channelled through a hierarchy of agents. Each type of agent is the result of a successful step in the resolution process. The system does not rely on the compositional bias of coding sequences which is a key feature of current computer methods.

1 Introduction

The splicing of human genes is known to take place in a large molecular complex called spliceosome, containing close to one hundred proteins as well as five small RNA species. It involves a collection of site recognition mechanisms. Such mechanisms are not yet fully understood, if known at all, which makes the precise chronology of events resulting in the splicing of introns and the ligation of consecutive exons, quite difficult to determine.

Even though each known mechanism involved in gene splicing is rather easy to describe, its relative contribution to the overall process is still unclear. Consequently, the problem can be partitioned into sub-problems but, the resolution process is certainly not straightforward.

The automatic identification of exons is usually based on the coding properties of exonic sequences. Current computer methods (see [4] for review) rely on the definition of a measure of *codingness* of a sequence, to distinguish between coding and non-coding regions in genomic DNA. Statistical regularities found in exons depending on codon usage, compositional bias, periodicity, etc, are used to characterise sequences. But, these methods cannot account for a number of recently observed phenomena such as alternative splicing or non-coding exons [8].

In contrast to previous approaches, the method presented here, is designed only to deal with the available information when splicing occurs. It is performing as well as current programs but with minimal information, that is, sequence patterns. Preliminary results were encouraging [16].

Gene finding methods are often seen as splicing simulation devices. They include a training phase and require re-training when the testing phase is not satisfactory. The capacity of integrating new or modified information without re-training, is the main characteristic of distributed AI systems [5]. Besides, decisions reached can be traced and explained, while missing or erroneous information can be dealt with.

The multi-agent architecture is used here as a mean to validate models and test hypotheses on the chronology of events involved in gene splicing. Two models have been introduced to simulate splicing:

- The *scanning model* of the recognition of an acceptor site [11] states that no AG dinucleotide occurs between the branch site and the invariant AG of the acceptor site. Indeed, experiments show that if an AG dinucleotide is introduced between the branch site and the invariant AG of the acceptor site, it is used as a new acceptor site.

- The *exon definition model* [1] states that exons are recognised as a whole and serve as a reference for defining a gene. It shows that an acceptor site is first recognised, followed by a search for a donor site within the next 300 nucleotides. The occurrence of both a 3' and a 5' splice site in the correct orientation within these boundaries is a requisite for the formation of stable exon complexes.

The principles of the method are first presented, then the detection of splice junctions and e exons is detailed. Some results are given and commented.

2 Multi-agent architecture

The multi-agent architecture is organised in layers and sensitive to chronology. Information is channelled through a hierarchy of agents. Raw information reaches higher level agents only after being processed by so-called *basic agents*. Each type of agent is in fact, the result of a successful step in the resolution process.

Given the problem of identifying exons in new genomic DNA, the sketch of a resolution process was first drawn in collaboration with biologists. Originally, the challenge was to check whether sequence data and models of site/signal recognition, provided enough information to simulate splicing.

An agent is described by a list of characteristics. Basic agents were set to be donor and acceptor agents. For instance, a basic donor agent is characterised by a position in the sequence and a specific pattern surrounding the invariant GT. Basic agents represent the initial knowledge of the system.

Knowledge is updated and revised through processes of communication between agents. Two main processes are involved :

- Selection of agents (discarding irrelevant information)
- Co-operation between agents (merging information from distinct sources)

In the resolution process, assumptions are refined by modifying characteristics of agents and defining agents of higher levels. In the selection mode, new characteristics are added to the list associated with an agent. The knowledge of the system is enriched after selection. This filtering operation is meant to keep relevant agents. For instance, basic *donor* agents are selected upon similarity to consensus such that the refined *donor* agent is characterised by a position in the sequence, a specific pattern surrounding the invariant GT and a score of similarity of this signal to consensus.

In the co-operation mode, agents are merged and so are the associated lists of characteristics. For instance, when *donor* agents co-operate with *acceptor* agents, they give rise to *group of exons* agents characterised by relevant acceptors and donors.

The so-called breeding of agents representing the resolution process is summed up in figure 1. The order in which selection and co-operation operations are performed, reflects the sketch of resolution set earlier. It can be altered whenever priorities need be modified, but, a the same time, the relevance of the model is tested.

Depending on the instantiation of constraints, agents are more or less informative; information is quantified by scoring functions. Simple operations are implemented for score calculations.

The program is written in an actor language. The system was called AMELIE (acronym of Multi-agent Architecture for the Explanation and the Localisation Intron-Exon).

2.1 Learning

The LEGAL system [7] is used to identify regularities in sequences surrounding sites. The length of signals to be recognised was determined after a number of trials: *nnnGTnnnnnnn* for donors and

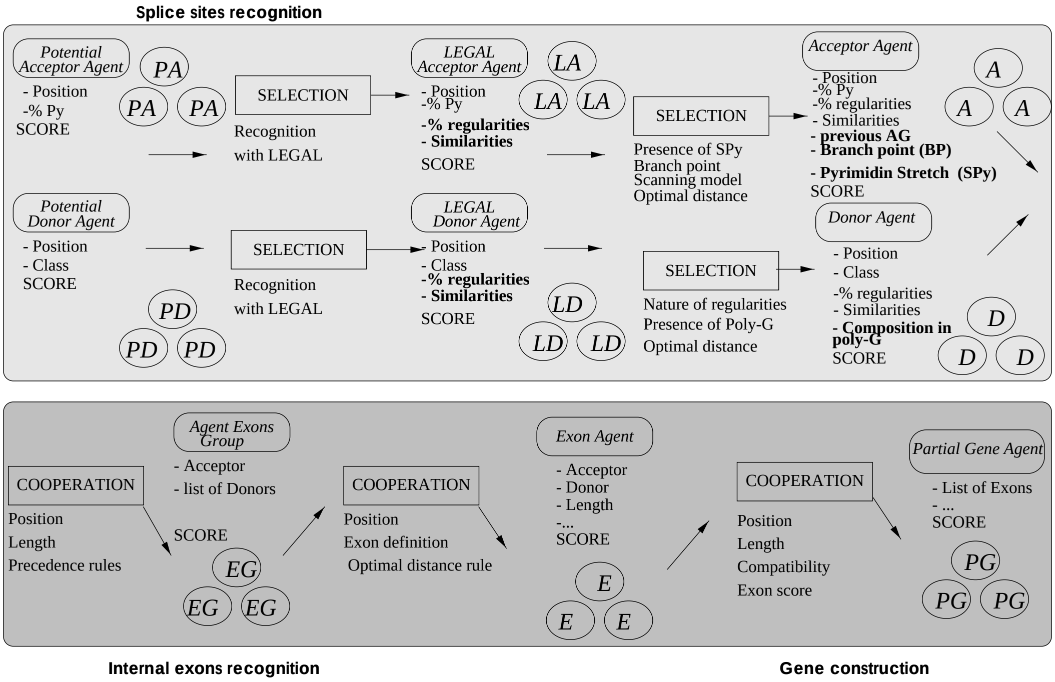

Figure 1: Multi-agent system instantiation

nnnnnnnAGn for acceptors were found to be optimal which is consistent with results found in [3]. Rules governing the occurrence of nucleotides in such signal sequences are generated and used for the definition of a similarity measure. A sequence is recognised as a signal depending on the number and the quality of rules which are satisfied. Thresholds are modifiable.

The use of counter-examples was subject to various trials, depending on the definition of *non-site.Non-sites* were automatically generated as sequences not verifying the acquired rules, or given more of a biological meaning, as sequences found inside exons. In practise, the two strategies yielded similar results.

Introducing counter-examples turned out to be an efficient mean of reducing the rate of false-positive without damaging the quality of recognition. In the case of donor sites, for instance, the number of false-positive drops from 12 to 6 per kb. Sequences in the training set were divided into classes depending on the $G + C$ content. Four classes are defined :

- class 1 : sequences containing from 20 to 39% of $G + C$
- class 2 : sequences containing from 40 to 49% of $G + C$
- class 3 : sequences containing from 50 to 59% of $G + C$
- class 4 : sequences containing over 59% of $G + C$

Results and detailed values given below correspond to the case of sequences of class 2.

2.2 Detection of Donor sites

GT is the invariant sequence of the donor site. Conserved nucleotides occur downstream GT, on the side of the intron and upstream GT, on the side of the exon. Different types of constraints are acting on each side but, these sequences are covariant: nucleotides may be conserved on both sides and if not, either the intronic sequence is weakly conserved while the exonic sequence is strongly conserved, or conversely.

Practically, the different cases depend on the presence of G at specific positions. Three distinct environments of the consensus were observed:

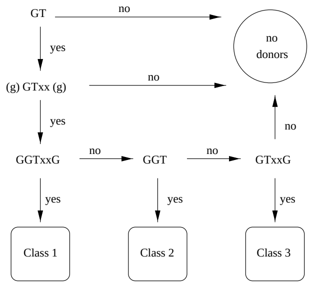

Figure 2: Donor sites: class-dependent learning and identification

(1) G **GT** NNG in 60% cases approximately
(2) G **GT** in 20% cases approximately
(3) **GT** NNG in 20% cases approximately

The LEGAL system is used to identify regularities in distinct sets of examples of these three types of donor sites (116 examples for type (1), 21 for type (2) and 29 for type(3)). Three sets of rules are obtained (denoted R_1, R_2, R_3). Figure 2 shows how the identification of donor sites is split into three sub-problems.

As mentioned above and seen on Figure 1, when scanning a new sequence, a *potential donor agent PD* is simply determined by the position in the sequence and the type of a consensus (1 to 3). Then, a *LEGAL donor agent LD* is selected if enough (threshold value) rules of the corresponding R_i are verified.

Donor agents D are selected among *LD* agents, upon the existence of a G-rich sequence to be found 7 to 50 nucleotides downstream the donor site.

Score calculations are performed for each Donor agent, as linear functions of the percentage of similarity to learned signals and the percentage of G downstream the donor site.

2.3 Detection of Acceptor sites: implementing and validating the Scanning Model

AG is the invariant sequence of the acceptor site and each occurrence of such dinucleotide in the sequence, is tested. Acceptor agents are built in a multiple step fashion. Learning is performed with a set of examples of signal sequences and rules are derived from a single set of 167 examples. These rules are combined to other characteristics to achieve the selection of agents. These characteristics rely on the hypothesis demonstrated in the *scanning model*, namely: if an AG dinucleotide is introduced between the branch point and the invariant AG of the acceptor site, it used as a new acceptor site.

More effects on modifying sequences upstream AG between the branch point and the pyrimidine stretch, modifying the pyrimidine stretch, etc, are studied in [11], to establish other restrictions on sequence length and distances between signals. Figure 3 shows the details of the procedure based on these constraints.

First, a Y-rich sequence is searched upstream the AG. Practically, nucleotides in the segment defined by the 14th and the 5th nucleotide before AG should be a majority of pyrimidines.

A weight matrix is used to determine a potential branch point; this signal is supposed to be five nucleotides long as in [10] and [9]. Following the hypothesis of the *scanning model*, no AG dinucleotide occurs between the branch site and the invariant AG of the acceptor site. Consequently, contradicting sequences are discarded. In parallel, learning-based rules are used to refine the selection of signal sequences.

The final selection yields 84% success for exon prediction in short genes and 88% for long genes, with less than 4 false-positive per kb.

As mentioned above and seen on figure 1, when scanning a new sequence, a *potential acceptor agent*

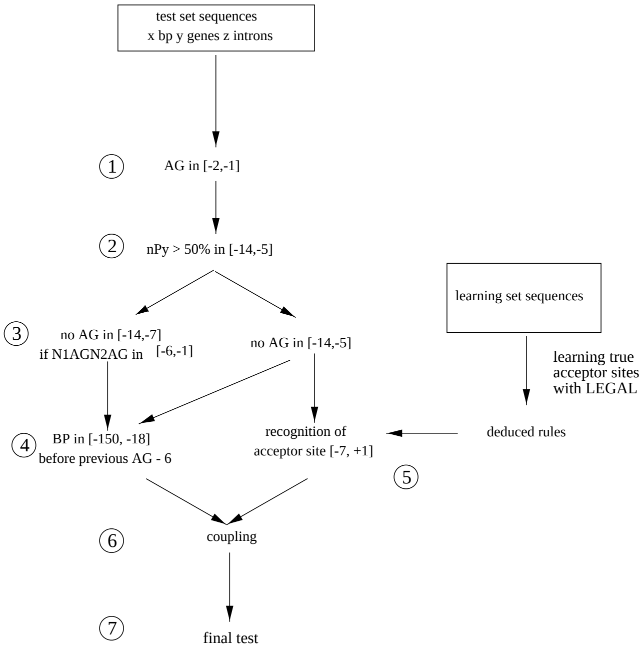

Figure 3: Acceptor site identification

PA is simply determined by the position in the sequence and the quality of the pyrimidin stretch. Then, a *LEGAL donor agent LA* is selected if enough (threshold value) rules are verified.
Acceptor agents A are selected among *LA* agents, upon the implementation of the procedure described above.
Score calculations are performed for each Acceptor agent, as linear functions of the percentage of similarity to learned signals, the percentage of pyrimidine upstream the acceptor site and the quality of the branch point determined by the weight matrix mentioned earlier.

2.4 Detection of Exons: implementing and validating the Exon Definition Model

One would think that since introns are to be removed, their sequence is first identified and then, spliced. But, in human sequences a donor site can be far apart from the next acceptor sites (1kb on average and up to 35kb) so that the ligation of exons is made difficult.
In contrast, exon sequences are short (130 bp on average and less than 300 bp with rare exceptions) which suggests that acceptor sites are first recognised. In fact, length is a determinant factor in site identification and the *exon definition* model [1] states further that exons represent units in the recognition process.
Two dependent factors are critical in the choice of a donor site to pair with a given acceptor site, namely, the distance between sites and the homology to consensus sequences. *In vitro*, sequences closer to consensus are usually chosen to favour a better binding to the U1-SnRNA, especially when the distance between sites is short [6]. On the contrary, if the distance is large, non consensus sequences can be chosen to the detriment of consensus ones.
Figure 4 shows how exons are put together. Exons are identified through a co- operation process between the previously built acceptor and donor agents. Figure 1 illustrates how groups of exons are defined as the result of merging one acceptor agent with corresponding donor agents.

Precedence and length determine this correspondence: each potential acceptor site, A, is paired with potential donor sites, D_1, D_2,... D_n, within the next 300 nucleotides with a minimum of 15

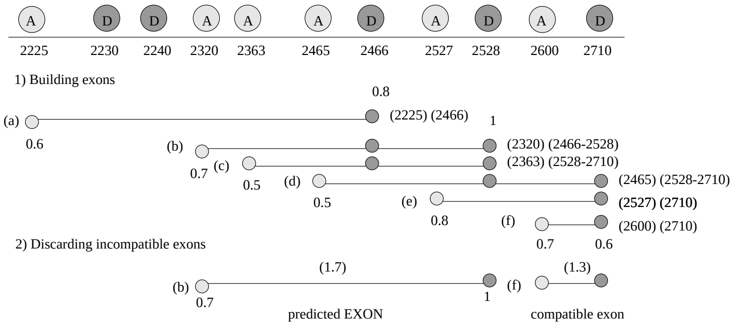
A D D A A A D A D A D
2225 2230 2240 2320 2363 2465 2466 2527 2528 2600 2710
1) Building exons
(a) 0.6 0.8 (2225) (2466) 1
(b) 0.7 (2320) (2466-2528)
(c) 0.5 (2363) (2528-2710)
(d) 0.5 (2465) (2528-2710)
(e) 0.8 (2527) (2710)
(f) 0.7 0.6 (2600) (2710)
2) Discarding incompatible exons
(b) 0.7 (1.7) 1 predicted EXON
(f) (1.3) compatible exon

Figure 4: Building exons with the exon definition model.

nucleotides separating sites A and any D_i. Since exons are being built, donor sites are downstream acceptor sites.

The score of an exon is the product of scores of the acceptor and donor sites modified by a log factor of the length of the exon.

Usual rules depending on the reading frame or the presence of a STOP codon, are irrelevant to determine how compatible exons are, in this scheme. Instead, an experimentally proven rule defines the compatibility between exons. Indeed, according to [12], there are at least 51 nucleotides separating the donor site from the branch point corresponding to acceptor site of the next exon. Such a rule is implemented (the minimal distance is set to 60). Figure 5 shows an example of how it is applied. Finally, preference is given to exons with higher scores.

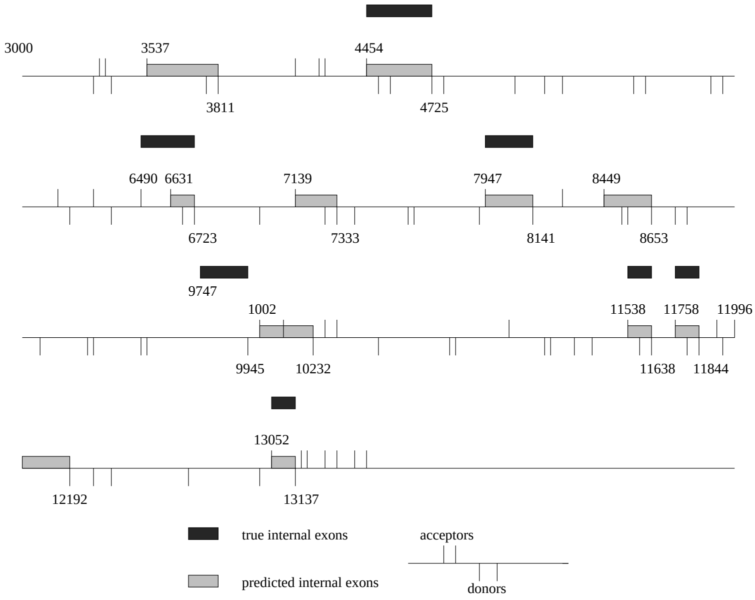

Figure 5: Prediction on *hspaia* gene : acceptor and donor sites, internal exons

Group of Exons agents are merged with respect to compatibility to give rise to *exon agents* and a *Partial Gene* is formed as a result of the co-operation between exon agents.

	GeneFinder	*GRAIL3*	*GeneParser*
Codingness	Preference of triplets Preference of oligonucleotides	Preference of hexamers and frame-dependent hexamers Markov model	Preference of hexamers and frame-dependent hexamers Preference of octanucleotides
Splice sites	yes and no	Distinct neural nets	Statistical tests using search matrix
Branch point	Fixed	no	no
Pyrimidin stretch	Fixed	Some variability	no
poly-G	Number of G, GG and GGG	no	no
$G + C$ content	no	Of isochore Of potential exon	3 classes (rich, medium and low $G + C$)
Exon length	no	Exon length distribution	Exon and intron length distributions
Other features			Similarity score using BLAST
Scoring	Discriminant function	Neural net	Neural net
Gene construction constraints	Compatible reading frames Minimal intron length No STOP codon in frame	Compatible reading frames Minimal intron length No STOP codon in frame	No overlap between intron and exon
Restrictions	No alternative splicing No STOP codon in frame AG and GT consensus	No alternative splicing No STOP codon in frame YAG and GT Consensus	No alternative splicing No genes exceding 15000 bp

Table 1: Features of the main gene structure prediction systems

3 Results

The accuracy of results differs according to the $G + C$ content of sequences. As remarked upon for other methods, the success rate of recognition is lower with G+C-rich sequences. In other cases, the success rate ranges between 80 and 90% with less than two false positives per kilobase [17].
Our understanding of a better prediction for long genes relies on a simple observation: the size of introns is, on average, larger in long genes than in short ones and conversely the size of exons is relatively constant in both cases.
Figure 5 shows how the method performed on the *hspaia* gene.

4 Discussion

The development of large sequencing projects and more specifically, efforts put into sequencing the human genome, stress the need for the automatic identification of exons.
In the most popular methods, the so-called *codingness* of a sequence is estimated with the compositional bias of coding sequences. This bias is observed in words composing exon sequences. Hexamers are considered in the Markov model defined in GenMark [2] and in GRAIL3 [15], while both hexamers and octamers are tested in the discriminant analysis used in [14], and the neural net of [13]. No such feature was taken into account in the method presented here. More generally, table 1 and 3 show which of the main features of current methods are common or distinct to those used in AMELIE.

As far as the quality of results is concerned, AMELIE is comparable to most methods. Considering it does not use coding properties, AMELIE does as well with less but more targeted information. In fact, it is improved if coding properties are added (data not shown). Performances are compared in table 2.

Method	Exons	All	False per Kb	Exons	All	False per Kb
GRAIL	42%	73%	0.3	48%	81%	0.25
GeneFinder	71%	79%	0.1	62%	84%	0.15
AMELIE	48%	71%	1	54%	90%	1.3

Table 2: Comparison with GRAIL and GeneFinder. Results on respectively short genes and long genes. *Exons* gives the percentage of exons totally predicted and *All* gives the percentage of exons totally or partially predicted.

Moreover, the specificity of the multi-agent scheme is that the performance is explainable. Assumptions are explicitly confirmed or contradicted by AMELIE whereas neural networks such as [15] [13] do simulate splicing but cannot easily argue a decision.
The *scanning model* provided a good basis for the detection of acceptor sites. Improvement is always possible but, results are rather satisfactory.
On the other hand, the *exon definition model* is incomplete [1]. Splicing enhancers are involved when signals are weak. These sequences are located either in exons or in introns and provide a binding site for proteins stimulating or stabilising the spliceosome. These intermediary stages of the splicing process are not yet elucidated. Such a lack of information can partly justify some errors made by the system. Furthermore, the compatibility between exons is determined by one rule only which turns out to be insufficient and still yields too many false-positive.

5 Conclusion

A novel approach to the problem of gene identification was presented. Further work is needed in refining the computer method and mostly in gathering information on splicing mechanisms. The direction for future improvement is set by the performance of the system as weaknesses of the decision process can always be explained.

Acknowledgments

This work was made possible by the precious collaboration of Yves d'Aubenton-Carafa and Claude Thermes. The AMELIE project was supported by the Groupement de Recherches pour l'Etude des Genomes (GREG).

References

[1] Berget, S. M. (1995). Exon definition in vertebrate splicing. *J. Biol. Chem.*, 270(6):2411–2414.

[2] Borodovsky, M., and McIninch, J. (1993). GenMark : parallel gene recognition for both DNA strands. *Computers and Chemistry*, 17(2):123–133.

[3] Brunak, S., Engelbrecht, J., and Knudsen, S. (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol.Biol.*, (220):49–65.

	AMELIE
Codingness	No analysis of codingness
Splice sites	Learning-based identification of acceptor sites Class-dependent learning of donor sites
Branch point	Variable
Pyrimidin stretch	Parametred
Poly-G	Content in G
$G+C$ content	4 classes according to $G+C$ content
Exon length	yes and no
Other features	Scanning model (previous AG)
Scoring	
Gene construction constraints	Distance rule between the donor site and the branch point of the next exon Minimal and maximal exon length Exon definition model
Restrictions	AG consensus (acceptor site) Donor site belongs to one of the 3 classes ($GGTxxG$, GGT or $GTxxG$ consensus)

Table 3: Features of the AMELIE system

[4] Fickett, J. (1996). The gene identification problem : an overview for developers. *Computers and Chemistry*, 20(1):103–118.

[5] Huhns, M. N. (1987) *Distributed Artificial Intelligence.* Morgan Kaufmann.

[6] Lear, A. L., Eperon, L. P., Wheatley, I. M., and Eperon, I. C. (1990). Hierarchy for 5' splice site preference determined in vivo. *Journal of Molecular Biology*, 211:103–115.

[7] Mephu Nguifo, E. and Sallantin, J. (1993). Prediction of primate splice junction gene sequences with a cooperative knowledge acquisition system. In *First International Conference on Intelligent Systems for Molecular Biology.*

[8] Mount, S. M., Peng, X., and Meier, E. (1995). Some nasty little facts to bear in mind when predicting splice sites. In *Workshop on Gene-Finding and Gene Structure Prediction, 1995*, Philadelphia, Pennsylvania.

[9] Penotti, F. E. (1991). Human pre-mRNA splicing signals. *Journal of Theoretical biology*, (150):385–420.

[10] Senapathy, P., Shapiro, M. B., and Harris, N. L. (1990). *Splice junctions, branch point sites, and exons*, volume 183, pages 252–278.

[11] Smith, C. W. J., Chu, T. T., and Nadal-Ginard, B. (1993). Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Molecular and Cellular Biology*, 13(8):4939–4952.

[12] Smith, C. W. J. and Nadal-Girard, B. (1989). Mutually exclusive splicing of a α-tropomyosin exons enforced by an unusual lariat branch point location : implications for constitutive splicing. *Cell*, 56:749–758.

[13] Snyder, E. E. and Stormo, G. D. (1995). Identification of protein coding regions in genomic DNA. *Journal of Molecular Biology*, 248:1–18.

[14] Solovyev, V. V., Salamov, A. A., and Lawrence, C. B. (1994). The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. In *2nd International Conference on Intelligent Systems for Molecular Biology.*

[15] Uberbacher, E. C., Xu, Y., shah, M., Matis, S., Guan, X., and Mural, R. J. (1995). DNA sequence pattern recognition methods in GRAIL. In *Workshop on Gene-Finding and Gene Structure Prediction, 1995*, Philadelphia, Pennsylvania.

[16] Vignal, L., d'Aubenton Carafa, Y., Lisacek, F., Mephu Nguifo, E., Rouzé, P., Quinqueton, J., and Thermes, C. (1996). Exon prediction in eucaryotic genomes. *Biochimie*, (78).

[17] Vignal, L. (1996). *Aide à la validation de modèles par une architecture multi-agents. Application à la localisation des introns exons.* PhD thesis, Université Montpellier II.