Hypothesis testing approaches to the exon prediction problem

Mireia Vilardell\textsuperscript{1,2} and Alex Sánchez-Pla\textsuperscript{1,*}
\textsuperscript{1}Statistics Department, University of Barcelona, Barcelona, Spain and \textsuperscript{2}Genetics Unit, Department of Experimental and Health Sciences, Pompeu Fabra University, Barcelona, Spain

Received on May 18, 2006; revised and accepted October 18, 2006
Advance Access publication October 25, 2006

Abstract

Motivation: Many gene identification methods assign scores to gene elements prior to their assembly into predicted genes. The scoring system is often based on log-likelihood ratios. These methods usually perform well but it is difficult to interpret how significant a score is.

Results: We have developed several tests of significance for the scores: (1) a sum-of-scores test (SST), (2) an intersection-union test (IUT), based on a multiple hypothesis testing interpretation of an exon’s score and (3) a meta-analytical approach (MA), which combines several \( P \)-values, corresponding to the exon’s parts, to yield a global \( P \)-value. We performed simulation studies, which show that the MA has better sensitivity and specificity than other methods and is easier to interpret by non-expert users. This is an improvement over other methods and is especially relevant for users who would like to predict incomplete gene sequences.

Contact: asanchez@ub.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The gene identification problem can be formulated as the deduction of the amino acid sequences encoded in a given DNA genomic sequence (Guigo, 1997; http://genetics.mgh.harvard.edu/doc/genefinder.doc.html). This is an important but difficult problem, especially in eukaryotes, where genes are often split into exons separated by introns. The beginning or end of these fragments is specified in the genome sequence by a few types of signals encoded in the primary DNA sequence (see Table 1), which instruct the cellular machinery in the pathway from DNA to protein. Signals involved in gene specification are all encoded in the primary DNA sequence. However, they are ill-defined, lack generality and are highly unspecific. Using currently available detection methods, it is impossible to distinguish signals truly processed by the cellular machinery from those actually not functional. Attempting to predict gene structure by processing only DNA sequence signals often results in a computationally intractable combinatorial explosion of potential products. As a result, any gene prediction method that relies on these signals has to be able to distinguish between false exons, i.e. sequences delimited by the signals but not producing any transcript, and true exons, which have both the signals and the capacity to produce transcripts.

The main objective of this study was to derive and compare methods to improve the performance of some gene prediction methods, such as geneid (Parra et al., 2000), which are based on a log-likelihood ratio scoring system. These methods have some drawbacks: scores are not bounded and their values depend on various factors such as exon type or exon length (Zhang, 1998). As a consequence, a high score is often not related to a high significance.

All the methods discussed in this study view the exon prediction step as a problem of hypothesis testing. This leads to the possibility of introducing a probabilistic scoring system based on \( P \)-values, which could be used as an additional input in the dynamic programming algorithms used to build the final gene prediction from the candidate exons. Other dynamic algorithms, such as genetic algorithms, could also be implemented using \( P \)-values. These latter could also be useful in detecting alternative splicing.

2 METHODS

A commonly used (\textit{ab initio}) approach to gene-prediction is to construct a set of potential exons, which are scored using a statistical model. These potential exons are highly overlapping and only a few are assumed to form part of a gene. The subset of exons that will be assembled into predicted genes are selected using some type of dynamic programming algorithm, which seeks to maximize the total score assigned to the gene.

The geneid (Parra et al., 2000) or GeneFinder (Green, P., 2001 unpublished data, http://genetics.mgh.harvard.edu/doc/genefinder.doc.html) programs use this approach and score potential exons using log-likelihood ratio scores. HMM based programs, such as GENSCAN (Burge and Karlin, 1997) also rely on the use of this type of scores (see Barret et al., 1997).

In simple terms, a gene, \( G \), can be considered as a sequence formed by a series of one (\( G = e \)) or more exons separated by zero or more introns (\( G = e_1 i_1 e_2 i_2 \ldots i_{n-1} e_n \)). As shown in Table 1, any exon is formed, as indicated, by three parts: Begin (\( B \) site) (start or acceptor site), Coding (\( C \)) part and End (\( E \)) site (donor or terminal site):

\[
\begin{align*}
0.1\text{ To whom correspondence should be addressed.}
\end{align*}
\]
In a study by Guigo et al. (2000) and available at http://www1.imim.es/databases/gpeval2000, it contained 22,041 starts, 76,728 acceptors, 64,566 donors, 54,919 stops and 6,811,740 coding fragments. Guigo et al. (2000) also describe how the set of potential exons for these models is constructed.

### Monte Carlo
The empirical distribution is approximated by scoring a large list of sequences generated using estimated PWM for Begin and End sites and 5th order Markov chain models for the Coding parts, as described above. Both PWM and Markov chain models were estimated from the same reference dataset used for resampling.

#### 2.2 Sum-of-scores tests (SST)
The SST is a straightforward approach that uses the scores as test statistics. It requires the distribution of the sum of the scores under the null hypothesis (i.e. in case the scored potential exon is false) to be known. This method yields approximate P-values of the significance test, allowing standard scores to be converted into probabilistic scores, and is easy to interpret. Although LLR depends on the length of the sequence, the SST works without exon length differentiation. We propose SST2, which assigns P-values accounting for exon length.

The hypothesis test can be stated as:

\[ H_0 : e \text{ is not an exon, versus } H_1 : e \text{ is an exon.} \]  

The test statistic is the exon score Equation (1). This test rejects the null hypothesis (i.e. it decides that \( e \) is an exon) if \( L_e(e) \), where \( L_e \) is the \( 1 - \alpha \)-quantile of the null (reference) distribution to be approximated by simulation.

The SST test does not consider the lengths of the exons. This problem has been overcome by introducing SST2 where null hypothesis and test statistics are the same as in SST. When simulating the null hypothesis distribution, exon length is accounted for by resampling each exon from a set of exons of similar length (see more details in section 2.3). The Monte Carlo method is not applied to SST because it builds exons with the same length as the candidate.

#### 2.2.2 Intersection-union tests (IUT)
In IUT, the components of the score are first considered separately and a global test is then performed using the intersection-union method [see Casella and Berger, 2002]. This seems to be the natural translation of the question ‘is this an exon?’ into a test, but has the drawback that a unique P-value cannot be obtained and therefore no probabilistic scoring is available. We consider an exon to have three components: Begin site, Coding part and End site. Given a three part sequence to be tested, the decision that it is an exon is taken if it is decided that the first part is a Begin site, the second part is Coding and the third part is an End site. Each decision can be seen as the alternative hypothesis for a null hypothesis of a random sequence, which gives:

\[ H_0^B : e_B \text{ is \textquoteleft \textquoteleft random–Begin\textquoteright \textquoteright, } H_1^B : e_B \text{ is a Begin site,} \]

\[ H_0^M : e_M \text{ is \textquoteleft \textquoteleft random–Coding\textquoteright \textquoteright, } H_1^M : e_M \text{ is Coding,} \]

\[ H_0^E : e_E \text{ is \textquoteleft \textquoteleft random–End\textquoteright \textquoteright, } H_1^E : e_E \text{ is an End site.} \]

Deciding that the sequence is an exon is equivalent to simultaneously rejecting the three null hypotheses. This suggests stating the (global) null hypothesis as a union of three (simple) null hypotheses. Applying the

### Table 1. Definition of exon types and relation with sites

<table>
<thead>
<tr>
<th>Exon type</th>
<th>Begin sites (B) Position</th>
<th>Type of site</th>
<th>End sites (E) Position</th>
<th>Type of site</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>Begin of the gene</td>
<td>Start (ATG)</td>
<td>Before first intron</td>
<td>Donor (GT)</td>
</tr>
<tr>
<td>Internal</td>
<td>After any intron but last</td>
<td>Acceptor (AG)</td>
<td>Before any intron but the first</td>
<td>Donor (GT)</td>
</tr>
<tr>
<td>Terminal</td>
<td>After last intron</td>
<td>Acceptor (AG)</td>
<td>End of the gene</td>
<td>Stop (TGA, TAA, TAG)</td>
</tr>
<tr>
<td>Single</td>
<td>Begin of the gene</td>
<td>Start (ATG)</td>
<td>End of the gene</td>
<td>Stop (TGA, TAA, TAG)</td>
</tr>
</tbody>
</table>

\[
L_E(e_i) = w_B \cdot L_B(e_i) + w_M \cdot L_M(e_i) + w_E \cdot L_E(e_i).
\]
In this case we have three test statistics: $L_d(e)$, $L_f(e)$ and $L_g(e)$. The null hypothesis will be rejected if there is simultaneous verification that:

$$|L_d| < f_{\alpha/2} \quad |L_f| < f_{\alpha/2} \quad |L_g| < f_{\alpha/2},$$

where $f_{\alpha/2}$, $f_{\alpha}$ and $f_{\alpha}$ are the respective 1-2, 1 and 1-alpha quantiles of the null hypothesis distributions, which can be approximated by simulation. As a consequence of the properties of intersection union-tests [see Casella and Berger, 2002] the global significance level of the IUT test, with every single thesis distributions, which can be approximated by simulation. As a consequence of the properties of intersection union-tests [see Casella and Berger, 2002] the global significance level of the IUT test, with every single thesis distributions, which can be approximated by simulation.

2.2.3 Meta-analysis based tests (MA) Appealing as the IUT test may be, it seems reasonable to assume that it may lack sensitivity (as shown in the next section) due to the strict requirements that must be fulfilled to reject the null hypothesis. An alternative approach consists of combining $P$-values obtained separately for each exon’s component using MA methods, which were originally developed to test the statistical significance of combined results [see Hedges and Olkin, 1985]. MA are based on the idea that relatively small $P$-values, not necessarily significant, do not appear at random. It is known that, under the null-hypothesis, $P$-values are distributed uniformly. If $P$-values are consistently small, even if they are not significant, the uniform law could be discarded and this could be interpreted as evidence favoring the rejection of the null-hypothesis.

Of the various methods for combining $P$-values those included in this study are among the most widely used. They are based on known reference distributions (implying a low computational cost as the distribution is tabulated) and all satisfy the monotonicity principle [see Hedges and Olkin, 1985] and will thus be optimal in most cases. In addition, all three methods account for exon length. The null hypothesis is the same for all three tests considered:

$$H_0 = H_0^a \cup H_0^b \cup H_0^c \Rightarrow H_1 = H_1^a \cap H_1^b \cap H_1^c.$$  

(6)

We propose using this schema to combine $P$-values, $p_i$, obtained from Begin sites, End sites or the Coding part, as follows:

(1) **MA1: Fisher’s Chi-square method**

The test statistic is defined as:

$$F = -2 \sum_{i \in \{B, M, E\}} \log p_i,$$

and the null hypothesis is rejected if:

$$F > \chi^2_{2k, \alpha},$$

where $\chi^2_{2k, \alpha}$ is the 1-alpha quantile of a chi-squared distribution with 2$k$ degrees of freedom.

(2) **MA2: Logit method**

The test statistic is defined as:

$$L = \sum_{i \in \{B, M, E\}} \log \left( \frac{p_i}{1 - p_i} \right),$$

where $L^* = \sqrt{c} L$ and $c = 3(5k + 4)\pi^2(5k + 2)$. The null hypothesis is rejected if:

$$|L^*| > t_{\alpha/2, k},$$

where $t_{\alpha/2, k}$ is the 1-alpha quantile of a Student’s $t$-distribution with $5k + 4$ degrees of freedom.

(3) **MA3: Inverse gaussian method**

The test statistic is defined as:

$$Z = \frac{1}{\sqrt{k}} \sum_{i \in \{B, M, E\}} \Phi^{-1}(p_i).$$

(11)

The null hypothesis is rejected if $Z > z_{\alpha}$, where $z_{\alpha}$ is the 1-alpha quantile of a standard Normal $\mathcal{N}(0,1)$ distribution.

All tests considered in this study were computed using an alpha value = 0.10, instead of the usual 0.05 in order to decrease the probability of false negatives, i.e. in order to avoid rejecting true exons unnecessarily.

2.3 Simulation experiments

Simulation experiments were designed to study and compare the performance of all tests. First, the number of simulation replicates was determined empirically, relying on the situation with the highest variability, which occurs when resampling coding regions between 5 and 104 bp long. The empirical procedure to determine the number of replicates consisted of:

1. Selecting a sample of (205) real exons with a length between 5 and 104 bp.
2. Repeatedly take 100 resamples of sizes 100, 200, 500, 1000, 2000, 5000, 10 000, 15 000, 20 000 and 50 000, from the false coding set having the same length as the corresponding real exon.
3. For each resample $R$ of size $N$ we approximated the $P$-value by the percentage of false exons scoring equal or more than the corresponding true exon.

The criterion was to choose a size such that $P$-values did not change appreciably from this size on and this led us to use 15 000 as the appropriate resampling size for this study.

Ideally any gene identification method should have a type II error as low as possible in order to yield high sensitivity. In a real application, there are many more negative samples than real exons and this was dealt with by using 10 times as many false exons as real ones in all comparisons.

2.3.1 Experimental design

Test performance was evaluated by considering the following factors for each method: test (SST1, SST2, IUT, MA), exon type (First, Internal, Terminal, Single), exon length (arbitrarily split into: <5, 5–104, 105–504, >504), P-value generation method (Resampling and Monte Carlo) and testing data (known false exons and known true exons). More details of the simulations can be found in Vilardell and Sanchez-Pla (2004), http://www.imub.ub.es/publications/preprints/pdf/preprint353.pdf.

The results are presented as tables containing the percentage of times that the null hypothesis that a given sequence is not an exon is rejected during each simulation and in each combination of factors, i.e. the percentage of times that the sequence is identified as an exon. This percentage corresponds to the sensitivity for real exons and to specificity for false exons. A good test should have a high percentage in tables for real exons (i.e. high-sensitivity) and for false exons (i.e. high-specificity).

The general appropriateness of all tests was evaluated using receiver operator characteristic (ROC) curves. ROC were calculated to compare resampling and Monte Carlo methods. Sensitivity and specificity values for detecting sites and the coding part at the same alpha level were calculated (see Table 6). ROC curves corresponding to Tables 2–6 can be found in the Supplementary material. We also compared the resampling distribution of sites and the coding part with the simulated ones using variability graphs, which are shown in the Supplementary material.

All the methods considered here can be applied to any genome. This study used only Drosophila sequences. However, a wider study using other genomes is underway.
3 RESULTS

The Monte Carlo simulation as a method of generating \( P \)-values is similar to the resampling method for site models but differences appeared in coding part models (see Supplementary material).

Table 2. Sensitivity values for all tests and different exon lengths using the resampling method. Taking two medium-size groups, 5–104 and 105–504, decreases computer-memory requirements

<table>
<thead>
<tr>
<th>Exon length</th>
<th>Number of exons</th>
<th>SST (%</th>
<th>IUT (%)</th>
<th>SST2 (%)</th>
<th>FISHER MA1 (%)</th>
<th>LOGIT MA2 (%)</th>
<th>GAUSS MA3 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5</td>
<td>12</td>
<td>100</td>
<td>67</td>
<td>92</td>
<td>92</td>
<td>92</td>
<td>100</td>
</tr>
<tr>
<td>5–104</td>
<td>205</td>
<td>99</td>
<td>41</td>
<td>92</td>
<td>94</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>105–504</td>
<td>572</td>
<td>99</td>
<td>7</td>
<td>88</td>
<td>91</td>
<td>97</td>
<td>95</td>
</tr>
<tr>
<td>&gt;504</td>
<td>370</td>
<td>97</td>
<td>37</td>
<td>96</td>
<td>85</td>
<td>90</td>
<td>89</td>
</tr>
<tr>
<td>All lengths</td>
<td>1159</td>
<td>98</td>
<td>48</td>
<td>91</td>
<td>93</td>
<td>95</td>
<td>94</td>
</tr>
</tbody>
</table>

Table 3. Specificity values for all tests and different exon lengths using the resampling method

<table>
<thead>
<tr>
<th>Exon length</th>
<th>Number of exons</th>
<th>SST (%</th>
<th>IUT (%)</th>
<th>SST2 (%)</th>
<th>FISHER MA1 (%)</th>
<th>LOGIT MA2 (%)</th>
<th>GAUSS MA3 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5</td>
<td>246</td>
<td>100</td>
<td>99</td>
<td>90</td>
<td>93</td>
<td>91</td>
<td>78</td>
</tr>
<tr>
<td>5–104</td>
<td>10000</td>
<td>88</td>
<td>100</td>
<td>97</td>
<td>93</td>
<td>80</td>
<td>85</td>
</tr>
<tr>
<td>105–504</td>
<td>2500</td>
<td>79</td>
<td>100</td>
<td>95</td>
<td>90</td>
<td>79</td>
<td>81</td>
</tr>
<tr>
<td>&gt;504</td>
<td>100</td>
<td>52</td>
<td>99</td>
<td>53</td>
<td>80</td>
<td>69</td>
<td>70</td>
</tr>
<tr>
<td>All lengths</td>
<td>12 846</td>
<td>85</td>
<td>100</td>
<td>96</td>
<td>92</td>
<td>79</td>
<td>84</td>
</tr>
</tbody>
</table>

Table 4. Sensitivity values for all tests and different exon types using the resampling method

<table>
<thead>
<tr>
<th>Exon type</th>
<th>Number of exons</th>
<th>SST (%</th>
<th>IUT (%)</th>
<th>SST2 (%)</th>
<th>FISHER MA1 (%)</th>
<th>LOGIT MA2 (%)</th>
<th>GAUSS MA3 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>274</td>
<td>100</td>
<td>45</td>
<td>95</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>Internal</td>
<td>472</td>
<td>100</td>
<td>71</td>
<td>93</td>
<td>99</td>
<td>99</td>
<td>98</td>
</tr>
<tr>
<td>Terminal</td>
<td>275</td>
<td>99</td>
<td>29</td>
<td>90</td>
<td>93</td>
<td>95</td>
<td>92</td>
</tr>
<tr>
<td>Single</td>
<td>138</td>
<td>87</td>
<td>12</td>
<td>83</td>
<td>64</td>
<td>75</td>
<td>74</td>
</tr>
<tr>
<td>All</td>
<td>1159</td>
<td>98</td>
<td>48</td>
<td>91</td>
<td>93</td>
<td>95</td>
<td>94</td>
</tr>
</tbody>
</table>

Table 5. Specificity values for all tests and different exon types using the resampling method

<table>
<thead>
<tr>
<th>Exon type</th>
<th>Number of exons</th>
<th>SST (%</th>
<th>IUT (%)</th>
<th>SST2 (%)</th>
<th>FISHER MA1 (%)</th>
<th>LOGIT MA2 (%)</th>
<th>GAUSS MA3 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>1073</td>
<td>85</td>
<td>99</td>
<td>94</td>
<td>91</td>
<td>77</td>
<td>83</td>
</tr>
<tr>
<td>Internal</td>
<td>7139</td>
<td>85</td>
<td>100</td>
<td>96</td>
<td>91</td>
<td>79</td>
<td>82</td>
</tr>
<tr>
<td>Terminal</td>
<td>4420</td>
<td>87</td>
<td>100</td>
<td>96</td>
<td>94</td>
<td>81</td>
<td>87</td>
</tr>
<tr>
<td>Single</td>
<td>214</td>
<td>90</td>
<td>100</td>
<td>96</td>
<td>95</td>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td>All</td>
<td>12 846</td>
<td>85</td>
<td>100</td>
<td>96</td>
<td>92</td>
<td>79</td>
<td>84</td>
</tr>
</tbody>
</table>

Table 6. Sensitivity values for sites and coding part depend on exon type (RS: Resampling; MC: Monte Carlo)

<table>
<thead>
<tr>
<th>Exon type</th>
<th>Begin sites (B)</th>
<th>End sites (E)</th>
<th>Coding part (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RS (%)</td>
<td>MC (%)</td>
<td>RS (%)</td>
</tr>
<tr>
<td>First</td>
<td>274</td>
<td>58</td>
<td>61</td>
</tr>
<tr>
<td>Internal</td>
<td>472</td>
<td>92</td>
<td>94</td>
</tr>
<tr>
<td>Terminal</td>
<td>275</td>
<td>94</td>
<td>96</td>
</tr>
<tr>
<td>Single</td>
<td>138</td>
<td>46</td>
<td>47</td>
</tr>
</tbody>
</table>

3.1 3006

Fig. 1. ROC curve for resampling based simulation.

It systematically yielded lower values in most combination of factors affecting sensitivity and specificity. ROC curves (see Figs 1 and 2) illustrate this when different tests are compared. For this reason, only results obtained with the resampling method are presented. The main results of the simulation study are presented in Tables 2–6 and can be summarized as follows:

1. The IUT test performed worst, with a lower sensitivity than the SSTs (1,2) and MA tests (1,2,3).
2. SST seems to be less specific than SST2. SST tends to lose specificity with exon length (Table 3 or Fig. 1) tending to classify large potential exons as true exons. SST2 is more consistent and yields similar specificity values for all lengths (Table 3). Therefore, we propose using SST2 rather than SST.
3. The SST2 test performs very similarly to all three MA tests in terms of sensitivity. However globally it is less specific than any of the MA tests. SST2 has a significant loss of specificity in classes >504 bp. In general, we observed that, as the exon length increased, the sensitivity of SST2 tended to be slightly higher than that of any of the MA tests, while its specificity tended to be lower (cf. Tables 2 and 3).
The performance of the different tests in the various conditions analyzed showed some well-defined trends:

1. Most tests performed better for internal exons and worse for single ones. The internal exon was the exon type with the strongest signals and the single exon that with the weakest (see Table 6). In single exons, signals were globally consistent with random distribution in P-values. Here, MA test fails by definition. An alternative model for coding parts based on Markov Chain Models should be considered for single exons.

2. Fisher’s test (MA1) performed slightly better than MA2 and MA3 globally, although it was less robust to the lack of adjustment than MA2 and MA3 (see single exons in Table 4). If there were an alternative model for coding part in single exons MA1 would be more powerful and more differences among MA1, MA2 and MA3 could be found.

In summary, MA-based tests performed similarly to the SST test, with the advantage that no additional parameter, such as weights in SST tests, needed to be estimated. Fisher’s chi-square (MA1) test was slightly better than the Logit (MA2) or Gauss (MA3) tests in terms of higher sensitivity and specificity. It performed better than the other for most combinations of factors and would thus be the test of choice to obtain a probabilistic score.

4 DISCUSSION AND CONCLUSIONS

This study tries to formulate an approach to the meaning of the scores as test statistics (see Vilardell and Sanchez-Pla, 2004), which led us to introduce the IUT. This test allows the question ‘Is a sequence, $S$, an exon?’ to be recast as a test of hypothesis, which we have called a test of exonicity and seems to be a natural choice because the question is broken down into three individual tests: (1) ‘Is the initial fragment a Begin site?’ (2) ‘Is the internal sequence Coding?’ (3) ‘Is the final fragment an End site?’ The final decision comes from rejecting the three null hypotheses. Appealing as it may be, the test has been shown to perform badly in simulation studies. The nature of this test implies that all three hypotheses must be rejected in order to decide that the sequence is an exon, and the fact that the signals are so weak (see Table 6) suggests that it is very easy for this not to occur in at least one of the three tests.

Although a testing approach method based on IUT seemed reasonable, the weakness of the signals meant that it did not work well enough. One further possibility was to try to combine the different P-values instead of using them separately. Techniques for combining P-values have been developed by several authors and are often used in meta-analysis (Hedges and Olkin, (1985)). We chose three of these techniques and carried out three different tests based on combining P-values for tests (1), (2) and (3) of the previous paragraph. All three tests performed well high-sensitivity and high-specificity in simulation studies.

However, we would recommend Fisher’s method to improve the selection of exon candidates as it is easier to apply and performed slightly better than the other MA tests. Since exact P-values could be calculated from background models with low computational costs [see Huang et al. (2004)] and although better background models in the coding part are needed to avoid the use of resampling methods, Fisher’s method could be used to develop methods to assign exact P-values for each potential exon. In addition, background models take exon length into account, and are not biased towards large exons, as occurs with SST. Exons longer than 504 bp present a classification problem, which is directly related to the difficulty in classifying single type exons, as this class tends to have longer exons, as shown in Table 5. We suggest studying single exons in more detail in order to build specific models for them.

In summary, we have introduced a new method for detecting coding sequences. This method performs at least as well as LLR methods but does not require estimation of additional parameters, such as in Parra et al., (2000). In addition, this method is easier to interpret than LLR because it yields P-values, which can be interpreted as probabilistic scores ranging from 0 to 1, making its interpretation more intuitive than is the case with the arbitrary scales used in LLR scores.

This method could be extended in several directions. For example, the resulting P-values could be used with genetic algorithms to detect alternative splicing or to compare genomes.

ACKNOWLEDGEMENTS

The authors thank Roderic Guigo, Genis Parra and Enrique Blanco from Genome Bioinformatics for their support and helpful comments. The authors also thank Lauro Sumoy, Roberto Castello and Josep Lluis Mosquera for reviewing the manuscript.

Conflict of Interest: none declared.

REFERENCES


