EXONSCAN: EXON Prediction with Signal Detection and Coding Region Alignment in Homologous Sequences

Shu Ju Hsieh\textsuperscript{a,1}, Yun Sheng Chung\textsuperscript{a,2}, Chun Yuan Lin\textsuperscript{b,3}, Chuan Yi Tang\textsuperscript{a,4}

\textsuperscript{a}Department of Computer Science National Tsing-Hua University Hsinchu, Taiwan, ROC
\textsuperscript{b}Institute of Molecular and Cellular Biology National Tsing-Hua University Hsinchu, Taiwan, ROC

ABSTRACT
Identifying the protein coding genes in the genomic sequences is a very important application and challenging work. A great number of computational gene prediction programs have been proposed with satisfied sensitivity and specificity at nucleotide level. However, their sensitivity and specificity at exon level are often low. Here, we propose EXONSCAN, a novel exon prediction program that combines signal detection and CORAL (COding Region ALignment) between homologous genomic sequences with the conservation of protein coding regions. EXONSCAN first uses the signal detection and CORAL to find candidate exons. Then EXONSCAN predicts the gene structures by assembling predicted exons. In the experimental test, our program was tested on ROSETTA data set of 117 human-mouse sequence pairs. The experiment results show that the sensitivity and specificity of EXONSCAN are both 98\% at nucleotide level; they are 87\% and 89\% at exon level, respectively. These results are superior to those of all existing programs.

Categories and Subject Descriptors
J.3 [Computer Applications]: Life & Medical Sciences.

General Terms
Algorithms

Keywords
Gene prediction, Sequence alignment and Comparative genomics.

1. INTRODUCTION
A comparative analysis of the mouse and human genomes [4] shows that approximately 99\% of the mouse genes have a homologue in the human genome and protein coding sequences exhibit the highest degree for conservation with on average 85\% sequence identity. Moreover, concerning the conservation of gene structure, 86\% of the homologous gene pairs are estimated to have the same number of coding exons. Accordingly, our method utilizes the properties of the conserved coding regions and gene structures in human and mouse genome.

2. METHODS
2.1 Exon Prediction – EXONSCAN
EXONSCAN accepts two genomic sequences containing homologous genes as input and predicts the gene structures in both sequences. Essentially, it is a novel method to integrate the signal prediction programs with CORAL, the sequence similarity measurement program. The aim of this method is to find a chain of exons with acceptable similarity. Our approach bases on the following idea. Given two genomic sequences, we first find a set of candidate signals (start codons, stop codons, splice acceptors and splice donors) that contain most part of true signals for each sequence. The experiment results show that the sensitivity and specificity of EXONSCAN are both 98\% at nucleotide level; they are 87\% and 89\% at exon level, respectively. These results are superior to those of all existing programs.

2.2 COding Region ALignment - CORAL
Most genes are subject to much stronger selective constraints on nonsynonymous changes than on synonymous. Specifically, coding regions of nucleotide sequences are fairly well conserved
in the first and second codon positions, the third position tends to diverge freely. To cope with this we develop an algorithm that is conceptually similar to that used in WABA [3]. Instead of aligning at the nucleotide level or the amino acid level, we utilize three nucleotides spread out in the pattern XOX (where the X’s indicates match and the O’s means “don’t cares”) to serve as the base. And in order to reduce computation complexity without sacrificing too much accuracy, referring to SPA [7], a more efficient filtration probability model is built with considering the conservation of coding regions. We call this algorithm CORAL which combines the methods of probabilistic and combinatorial analysis. Especially, it’s computation complexity is reduced to as low as O(n).

CORAL identifies the approximate positions of gaps by measuring the local similarity in a window of length w. When the hamming distance increase over a threshold, we assume that an ill-positioned pair has occurred, specifically, a gap could exist there. A greedy algorithm is applied to determine the possible position and length for each predict gap. By aligning a pair of sliding window with length w to determine where the ill-positioned pair occurs, and then we shift the sliding window with inserting a gap for finding the right-positioned pair, and then continue aligning until another ill-positioned pair occurrences is touched. The process is repeated until either of the sequences is completely aligned.

3. RESULTS
We measure the accuracy of prediction at both the nucleotide and exon levels by the quantities Sn (sensitivity) and Sp (specificity). At the nucleotide level, the measures of sensitivity Sn=TP/(TP+FN) and specificity Sp=TP/(TP+FP) where TP (true positives) is the number of coding nucleotides predicted as coding, FP (false positives) is the number of non-coding nucleotides which are predicted as coding, FN (false negatives) is the number of coding nucleotides that are predicted as non-coding. We use the same measures at the exon level and an exon is assumed to be correct predicted only when both its boundaries are predicted exactly. The approximate correlation AC summarizes the overall nucleotide sensitivity and specificity by one number where TN (true negatives) is the number of non-coding nucleotides which are predicted as non-coding. AC is defined as follows:

\[ AC = 1 - \left( \frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FP} + \frac{TN}{TN+FN} \right)^2 \]

At exon level, the average of Sn and Sp is used instead. ME (missing exons) is the proportion of annotate exons not overlapped by any predicted exon, and WE (wrong exons) is the proportion of predicted exons not overlapped by any annotated exons.

EXONSCAN was tested on ROSETTA data set of 117 homologous gene pairs [2] and was compared to ROSETTA, SGP-1, SGP2, SLAM, TWINSCAN and GENSCAN. The TWINSCAN.p is the modifier TWINSCAN to run on a syntenic nucleotide sequence pair. SGP2 allows the external specification of a gene model which can be used to predict a single complete gene and multiple genes separately. The results of the programs are summarized in Table 1. The results for GENSCAN, TWINSCAN, ROSETTA, SGP-1 and SLAM were retrieved from [1]. SGP2 results were obtained from [5].

4. DISCUSSION
The main components in the EXONSCAN are currently the signal generator and CORAL. An important difference from the other homologous based methods is that EXONSCAN generates signals with signal detectors and content alignments at first instead of applying local or global alignments to input sequences. This could explicitly model both the conservation of exon boundaries and the encoded amino acid sequences between two related genes and thus obtain additional information compared with existing methods. This should help cross annotation of related genomes, as none of the existing homologous based methods explicitly models the conservation of exon boundaries. Experiment results demonstrate that EXONSCAN outperforms all the others at both the nucleotide and the exon levels. However, taking advantage of the conservation of exon boundaries results in some limitations of this method. One is that the gene structures of both input sequences are demanded to be similar. In other words, it is unsuitable for distant organisms. And the other is losing short exons, such as the exon length is only three. This drawback is the same with most of the previous methods.

<table>
<thead>
<tr>
<th>Program</th>
<th>Nucleotide level</th>
<th>Exon level</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sn</td>
<td>Sp</td>
<td>AC</td>
<td>Sn</td>
<td>Sp</td>
</tr>
<tr>
<td>GENSCAN</td>
<td>0.98</td>
<td>0.91</td>
<td>0.93</td>
<td>0.82</td>
<td>0.77</td>
</tr>
<tr>
<td>TWINSCAN.p</td>
<td>0.96</td>
<td>0.94</td>
<td>0.94</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>TWINSCAN</td>
<td>0.96</td>
<td>0.89</td>
<td>0.92</td>
<td>0.85</td>
<td>0.77</td>
</tr>
<tr>
<td>SGP2(single)</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
<td>0.84</td>
<td>0.85</td>
</tr>
<tr>
<td>SGP2(multiple)</td>
<td>0.96</td>
<td>0.97</td>
<td>0.96</td>
<td>0.71</td>
<td>0.79</td>
</tr>
<tr>
<td>ROSETTA</td>
<td>0.94</td>
<td>0.98</td>
<td>0.95</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>SGP-1</td>
<td>0.94</td>
<td>0.96</td>
<td>0.94</td>
<td>0.70</td>
<td>0.76</td>
</tr>
<tr>
<td>SLAM</td>
<td>0.95</td>
<td>0.98</td>
<td>0.96</td>
<td>0.78</td>
<td>0.76</td>
</tr>
<tr>
<td>EXONSCAN</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.87</td>
<td>0.89</td>
</tr>
</tbody>
</table>

5. REFERENCES