Comparative Gene Prediction Based on Gene Structure Conservation

Shu Ju Hsieh[1], Chun Yuan Lin[2], Ning Han Liu[1], and Chuan Yi Tang[1]

[1] Department of Computer Science
[2] Institute of Molecular and Cellular Biology,
National Tsing-Hua University Hsinchu, Taiwan, ROC
hsieh@cs.hccvs.hc.edu.tw, cyulin@mx.nthu.edu.tw,
greg@cs.hccvs.hc.edu.tw, cytang@cs.nthu.edu.tw

Abstract. Identifying protein coding genes is one of most important task in newly sequenced genomes. With increasing numbers of gene annotations verified by experiments, it is feasible to identify genes in newly sequenced genomes by comparing with genes annotated on phylogenetically close organisms. Here, we propose a program, GeneAlign, which predicts the genes on one sequence by measuring the similarity between the predicted sequence and related genes annotated on another genome. The program applies CORAL, a heuristic linear time alignment tool, to determine whether the regions flanked by candidate signals are similar with the annotated exons or not. The approach, which employs the conservation of gene structures and sequence homologies between protein coding regions, increases the prediction accuracy. GeneAlign was tested on Projector data set of 449 human-mouse homologous sequence pairs. At the gene level, the sensitivity and specificity of GeneAlign are 80%, and larger than 96% at the exon level.

1 Introduction

Accurate prediction of gene structures, the exact exon-intron boundaries, is an important task in genomic sequence analysis, while it remains far from fully analyzed [9]. Numerous computational gene prediction programs have aided the identification of protein coding genes; however, no programs are accurate enough to predict all the protein coding genes perfectly. The best accuracy is achieved with spliced alignment of full-length cDNAs or comprehensive expressed sequences tags (ESTs) [8]. Sim4 [14], Spidey [28], BLAT [18], and GMAP [29] belong to this class. Nevertheless, to generate complete and accurate predictions of all genes is still an ongoing challenge because of the numerous genes lacking for the full-length cDNA. Single-genome predictors which predict gene structures by using one genomic sequence, e.g., GENSCAN [10], have been successfully used at prediction of newly sequenced genomes. With more and more organisms being sequenced, the comparative approaches provide more accuracy than the single-genome predictors. In addition to comparative analysis between genomes (e.g., ROSETTA [4], Pro-Gen [24], DOUBLESCAN [21], TWINSCAN [19], SGP2 [25], SLAM [1] and EXONALIGN [16], evidences from related organisms, such as cDNAs and ESTs of related

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 32 – 41, 2006.

organisms (e.g., GeneSeqer [8]), known proteins of related organisms (e.g., GeneWise [6], PROCRUSTES [15]) and known annotations of related organisms (e.g., Projector [20]), have been employed in the comparative approaches. Recently, several programs, Combiner [2], ExonHunter [7], and JIGSAW [3], devote to integrate multiple sources of information (e.g., multiple genomic sequences, cDNAs/ESTs and protein databases of related organisms, and the output of various gene predictors) to further increase accuracy for gene prediction.

Currently, the gene structures for complete genome sequences are generated by incorporating multiple computational approaches depending on the evidence available. The Ensemble gene prediction pipeline uses two streams of evidence, the direct placement of cDNAs and ESTs on the genome of the same organism and a related gene in another organism which is used as a template for the homologous gene [13]. Although cDNA and EST collections are far from comprehensive for most organisms, the abundance of valuable data provided by more than 1700 complete and ongoing genome projects [5] (Genomes Online Database http://www.genomesonline.org, January 2006) could help to locate the exon-intron boundaries for organisms which full-length cDNA sequences have not been generated. Moreover, the previous studies indicate that the known gene annotations coming from homologous genes are more powerful in aiding gene prediction than the evidence of homologous protein sequences [20].

In this study, we present a gene prediction tool - GeneAlign. The same as Projector, GeneAlign employs gene annotations of one organism to predict the homologous genes of another relative organism. GeneAlign integrates signal detectors with CORAL [16] to efficiently align annotated exons with predicted sequences. CORAL, a heuristic alignment program, aligns coding regions between two phylogenetically close organisms in linear time. The approach could identify the distinctive features, the high degree of conservation between protein coding sequences and gene structure conservation between phylogenetically close organisms. GeneAlign assumes the conservation of the exon-intron structures, but it can also align exons which differ by events of exon-splitting. GeneAlign can help gene structure prediction by a fairly diverged annotated genome that still shares a common gene structure. Here, we show that GeneAlign performs well in identifying coding exons; specifically, the rates of missing exons and wrong exons are both low.

2 Methods

GeneAlign accepts two nucleotide sequences of homologous genes and known gene annotation of one of these two sequences as inputs and predicts the exon positions in another sequence according to the known gene annotation. The major components of GeneAlign for annotation-genome mapping and alignment include: (1) signal filtrations, (2) applying CORAL to measure sequence similarity following candidate signals for generating approximate gene structures.

2.1 Signal Filtrations

To model the conserved gene structures of homologous genes, GeneAlign measures similarities between annotated exons of one sequence and downstream/upstream to the potential splice acceptors/donors of another sequence. For the predicted sequence, GeneAlign firstly obtains a set of candidate signals, TISs(translation initiation sites), splice acceptors/donors, according to signal scores calculated by the signal prediction tool NetStart [26] and DGsplicer [12] respectively. The NetStart, the most popular and accessible program for TISs prediction [23], produces neural network predictions of translation start in nucleotide sequences. The DGsplicer employs a dependency graph model to score potential splice signals. The NetStart and DGsplicer could efficiently filter out many false TISs and splice signals but failed to remove false signals resulting from highly degenerate and unspecific nature. Integrating CORAL [16] could help to measure the similarity between annotated exons and potential regions marked by candidate TISs and splice signals.

2.2 COding Region ALignment – CORAL

CORAL is developed on the basis of the conservation of coding regions. Since most of coding regions among organisms are conserved at the amino acid level, suggesting that the hamming distance of two segments with an optimal alignment is low. Applying the idea of a random model, the codon mutations are supposed to occur randomly within a sequence. A probabilistic filtration method is built to efficiently find ill-positioned pairs, a less than optimal alignment which is supposed to result from a shifting mutation and could be solved by inserting a gap of a length of a multiple of three. A local optimal solution is used to obtain a significant alignment when an ill-positioned pair is detected and to determine the possible position and length for the inserted gap. Besides, considering that the nucleotide sequences of the translated regions are well conserved in the first and second positions of a codon and maybe less conserved in the third nucleotide of a codon, we utilized three nucleotides spread out in the pattern XXO (where the X indicated "absolute matching" and the O meant "don't care") to serve as the basis of alignment.

CORAL employs probabilistic analysis and local optimal solution to efficiently align sequences by sliding windows and, thus, obtains near optimal alignment in linear time. The detail for the concept of CORAL can refer to Hsieh *et al.* [16]. Additionally, CORAL is implemented another version to directly compare with amino acid instead of codon. An amino acid identity score is calculated by translating the codons according to the genetic code and comparing corresponding amino acids in the two compared regions.

2.3 Gene Structure Alignments – GeneAlign

After signal filtrations by NetStart and DGsplicer, the predicted sequences and annotated exons are aligned from 5' to 3'. GeneAlign is designed for detecting multi-exons genes. The coding exons are divided into three categories according to their location in the coding region, initial exon (ATG-GT, first coding exon of a gene), internal exon (AG-GT), and terminal exon (AG-stop codon, last coding exon of a gene). Splice sites are the most powerful signals for gene prediction, accurate modeling splice sites could improve gene prediction accuracy [9]. Thus, the alignments are processed from the splice acceptors, aligning the first annotated internal exons with regions following the candidate splice acceptors by CORAL. CORAL will stop aligning when the alignment score significantly drops. If the alignment score and aligned sequence length are greater than threshold, the aligned subsequence is predicted as a candidate exon. In general, the threshold is set at alignment score ≥ 50% and exon length ≥ 30 bp, which is determined empirically. Candidate splice acceptors and the following annotated exons are examined subsequently to search for meaningful alignments. For each aligned segment, the downstream boundary is delimited by an admissible candidate splice donor. A series of aligned segments is ended at the annotated terminal exon and delimited by a stop codon, e.g. TAG, TGA and TAA. The aforementioned process is repeated from 3' to 5', from the last internal exons aligning with the regions following the candidate splice donors, and ended with initial exon. TISs are selected according to the scores evaluated by NetStart. This procedure retrieves possible missing exons resulted from underestimation of splice acceptors by DGsplicer, a single intron insertion/deletion to one of the exon pair, and frameshift at the 5' end of exon pairs. Any annotated exon could not be mapped to the predicted sequence, the alignment score of threshold will be set lower, e.g., 30%, and the corresponding region is searched again.

2.4 Performance Evaluation

The standard performance measures on prediction accuracy defined by Burset and Guigó [11] are applied to compare the accuracy of gene prediction. The measures of sensitivity (*Sn*) and specificity (*Sp*) are respectively $Sn=TP/(TP+FN)$ and $Sp=TP/(TP+FP)$ where *TP* (true positives) is the number of correctly predicted genes, *FN* (false negatives) is the number of true genes missed in the prediction, *FP* (false positives) is the number of pseudo genes wrongly predicted, and *TN* (true negative) is the number of correctly predicted pseudo genes. At the exon level, the *TP*, *FP, FN* and, *TN* are the same as the definitions except that exons are compared. An exon is assumed to be correctly predicted only when both its boundaries are correct. *ME* (missing exons) is the proportion of annotated exons not overlapped by any predicted exon, whereas *WE* (wrong exons) is the proportion of predicted exons not overlapped by any annotated exons.

3 Results

3.1 Data Sets

GeneAlign applies CORAL based on codon identity to efficiently find the partner exons to those of related known genes. The other version, GeneAlign*, which applies

CORAL based on amino acid identity, is in comparison with GeneAlign. To optimize the parameters, GeneAlign was trained by the IMOG data set [25]. The IMOG data set contains 15 homologous gene pairs. The testing set is Projector data set which collects 491 homologous gene pairs. As we aim to test the capability of the splice alignment, intronless genes were discarded. The average number of exons per gene in the test set of the remaining 449 homologous gene pairs is 9.3 exons. 45% of these gene pairs (204 out of 449) have the identical number of coding exons and the identical coding sequence length. 50% of these gene pairs (224 out of 449) have identical exons number but differ in coding sequence length. 5% of these gene pairs (21 out of 449) have different number of exons.

3.2 Performance

The performance of GeneAlign on accuracy of gene prediction was compared on that of Projector [20] and GeneWise [6]. Projector predicts gene structures by using the annotated genes on a related organism, which is the same with GeneAlign. GeneWise predicts gene structures by using the known proteins of a related organism. The set of genes predicted by Projector and GeneWise were retrieved from Projector web sever (http://www.sanger.ac.uk/Software/analysis/projector). We measure the performance in terms of sensitivity and specificity not only at exon level but also at gene level. The results are summarized in Table 1. These results show that the predictions obtained by GeneAlign are accurate on both gene level and exon level. GeneAlign also predicts better when evaluated by *ME* and *WE*. Besides, GeneAlign* has the lower ratios of *ME* and *WE* than those of GeneAlign.

In order to study the effects of sequence similarity on the performance of prediction accuracy, 449 homologous pairs were stratified into five classes with amino acid identities between two encoded proteins ranging from <60%, 60%~70%, 70%~80%, 80%~90% and 90%~100% (Fig. 1).

There are respectively 10, 21, 51, 143, 224 pairs in each class. The performances evaluated by the measures all exhibit the strong dependence on the amino acid

Table 1. Prediction accuracy on the Projector data set

Program	Gene level (%)		Exon level (%)			
	Sn	*Sp*	*Sn*	*Sp*	*ME*	*WE*
GeneWise	62.36	62.50	92.49	93.61	1.56	0.31
Projector	49.22	49.33	93.57	87.05	0.88	8.54
GeneAlign	79.96	79.96	96.47	96.92	0.72	0.31
GeneAlign*	79.73	79.73	96.43	96.84	0.62	0.24

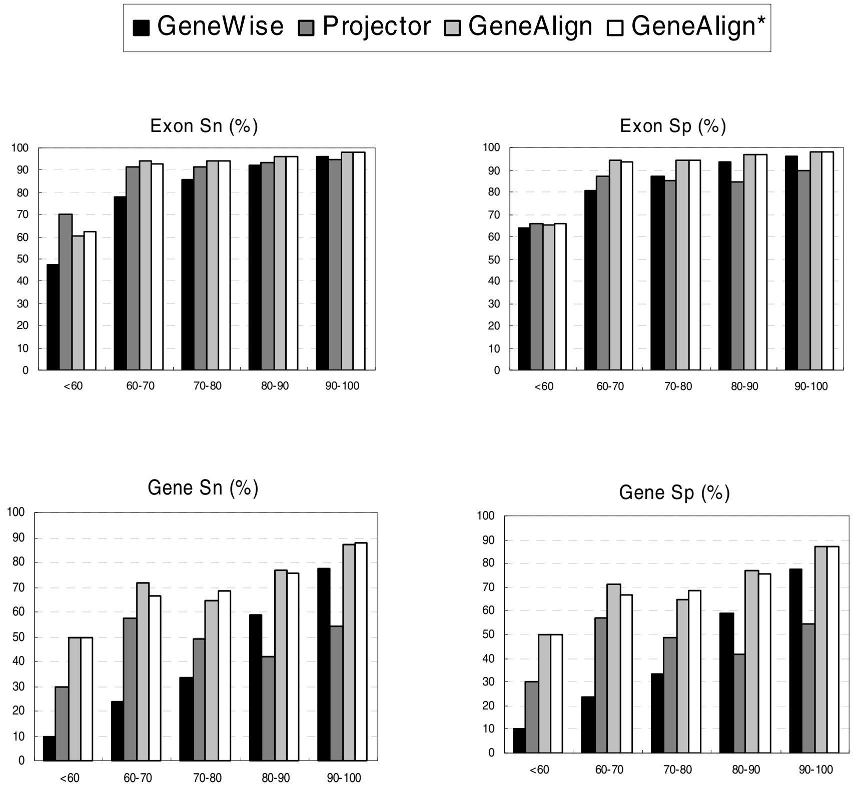


Fig. 1. Performance of the various gene predictors as a function of amino acid identities between two protein sequence encoded in each pair of homologous gene. We divided the data set by amino acid identities (<60%, 60%~70%, 70%~80%, 80%~90% and 90%~100%), and calculated the performance for each group.

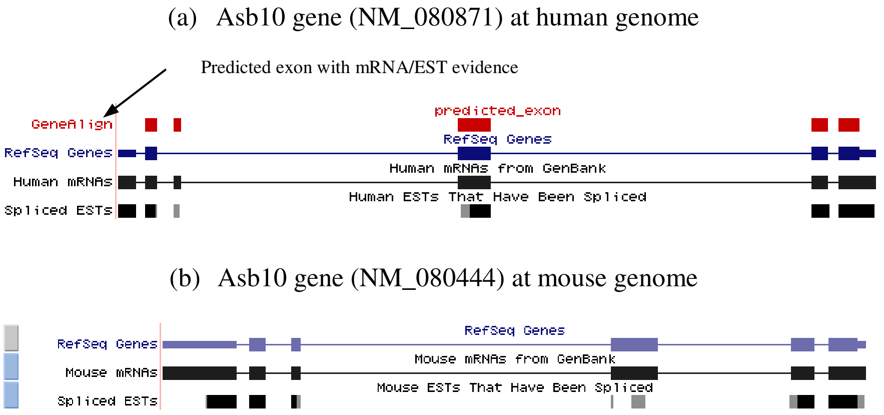
identities for three programs. GeneAlign outperforms the others in all classes of sequence similarity. GeneAlign and GeneAlign* have no significant difference on the gene level performance. At the performance of exon level, GeneAlign also displays high accuracy in sensitivity and specificity except for less well conserved gene pairs (percent identities less than 60). GeneAlign shows weaker performance than Projector whereas it can be slightly improved by GeneAlign*.

The behavior of the exon level performance can be further analyzed by the performance of rates of missing exons and wrong exons (Table 2). GeneAlign* shows the lower *ME* than those obtained by Projector and GeneWise in all class of sequence similarity, and performs better than GeneAlign. On the rate of wrong exons, GeneAlign* shows the very low rate of wrong exons except for the less well conserved gene pairs (percent identities less than 60), and GeneWise performs better in this class.

Table 2. The rates of missing exons and wrong exons

Program	<60%		60%~70%		70%~80%		80%~90%		90%~100%	
Missing exon rate (%)										
GeneWise	26.42	(14)	5.48	(8)	2.05	(9)	1.66	(22)	0.55	(12)
Projector	18.87	(10)	2.06	(3)	0.91	(4)	0.90	(12)	0.36	(8)
GeneAlign	20.75	(11)	0.68	(1)	0.68	(3)	0.68	(9)	0.27	(6)
GeneAlign*	16.98	(9)	0.68	(1)	0.46	(2)	0.60	(8)	0.27	(6)
Wrong exon rate (%)										
GeneWise	0.00	(0)	1.37	(2)	0.68	(3)	0.30	(4)	0.18	(4)
Projector	24.53	(13)	7.53	(11)	8.66	(38)	11.30	(150)	6.54	(144)
GeneAlign*	13.21	(7)	0.00	(0)	0.46	(2)	0.08	(1)	0.14	(3)
GeneAlign	11.32	(6)	0.00	(0)	0.22	(1)	0.00	(0)	0.14	(3)

* The numbers of missing exons and wrong exons are given in the brackets. The exon numbers in each interval, from <60% to 90%~100%, are 53, 146, 439, 1327 and 2203, respectively.


(a) Asb10 gene (NM_080871) at human genome

(b) Asb10 gene (NM_080444) at mouse genome

Fig. 2. Annotations of Asb10 at human and mouse genomes. (a) The human reference sequence is based on NCBI Build 35 and was produced by the International Human Genome Sequencing Consortium. (b) The mouse draft genome data was obtained from the Build 35 assembly by NCBI. The figures were produced with UCSC genome browser (http://genome.ucsc.edu/).

To analyze the missing and wrong exons in more detail, we found some wrong exons have corresponding exons at mouse annotation but are not annotated in human. They show high degree sequence conservation between human and mouse, the lengths are multiple of three. On the other hand, part of missing exons due to lose corresponding mouse annotations. These may be due to the differential representation of alternative splice forms. These can be shown by the example depicted in Figure 2.

According to the RefSeq [27] annotations, the ASB10 of Projector data set contains 4 exons in human but five exons in mouse. GeneAlign predict the 4 exons correctly and a wrong exon according to the mouse annotation. We found this predicted exon with the evidence of mRNAs from the UCSC browser [17].

4 Conclusion

A great number of gene prediction tools have been developed to identify genes, whereas accurate gene identification is still a challenging problem. Each gene prediction tool is applicable and performs well in different stage of gene prediction depending on available information. Incorporate with different gene prediction tools can acquire higher accuracy. Direct mapping the cDNA to the genomes acquire the most accurate gene structure, whereas current experimental and bioinformatics approaches for exploring transcriptomes are yet to catch up the speed of the growing genomic information. However, the rapid growth of sequence repositories, especially the experimentally verified genes, provides ongoing genome projects an extraordinarily valuable data. In contrast of comparative identifying genes with genome sequences or protein sequences from other organisms, the genes have been validated in one organism helps to identify homologous genes in another organism. Gene structure conservation had been employed in existing gene prediction tools (e.g. Projector), the incorporation with gene structure conservation showed greatly improvement of the accuracy of gene prediction.

GeneAlign, the same as Projector, predicts a gene by utilizing an annotated homologous one in the other organism. With the perspective of gene structure conservation aiding gene prediction, GeneAlign can provide highly accurate and sensitive predictions. Table 1 shows that GeneAlign performed well on gene prediction. The sensitivity and specificity are high at both the gene and the exon levels, while the rates of missing exons and wrong exons are both low. The applicability of CORAL results in low rate of missing exons, while aligning directly from the conserved splice signals results in low rate of wrong exons. However, there were still missing exons and wrong exons. Underestimating the splice signals is one of the reasons for missing exons. This problem can be solved by decreasing the cutoff score of splice signals. However, decreasing the cutoff score also results in longer running time. Lacking partner annotation is the other reason that GeneAlign fails to detect exons. Nevertheless, the true concordance of gene structure between the two organisms is probably higher, because of differences will be due to differential representation of alternative splice forms [22]. And some wrong exons are conserved between organisms. These wrongly predicted exons may be alternative spliced exons but due to lack of transcriptome evidence which may be biased with abundantly expressed splicing forms.

We have successfully applied GeneAlign to exon identification in human-mouse homologs. Extension the application to other relative distance organisms should be achieved. As GeneAlign only predicts exons with annotated similarity evidence, it will miss the exons without partner annotations in another genome. For such missing exons, EXONALIGN [16] can be applied to find the likely coding exons within putative introns.

References

1. Alexandersson, M., Cawley, S., Pachter, L.: SLAM: cross-organisms gene finding and alignment with a generalized pair hidden Markov model. Genome Res. 13 (2003) 496-502
2. Allen, J.E., Pertea, M., Salzberg, S.L.: Computational gene prediction using multiple sources of evidence. Genome Res. 14 (2004) 142-148
3. Allen, J.E., Salzberg, S.L.: JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics 21 (2005) 3596-3606
4. Batzoglou, S., Pachter, L., Mesirovi, J.P., Berger, B., Lander, E.S.: Human and mouse gene structure: comparative analysis and application to exon prediction. Genome Res. 10 (2000) 950-958
5. Bernal, A., Ear, U., Kyrpides, N.: Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. Nucleic Acids Res. 29 (2001) 126-127
6. Birney, E., Clamp, M., Durbin, R.: GeneWise and Genomewise. Genome Res. 14 (2004) 988-995
7. Brejova, B., Brown, D.G., Li, M., Vinar, T.: ExonHunter: a comprehensive approach to gene finding. Bioinformatics 21 (2005) 57-65
8. Brendel, V., Xing, L., Zhu, W.: Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinformatics 20 (2004) 1157-1169
9. Brent, M.R., Buigo, R.: Recent advances in gene structure prediction. Curr. Opin. Struct. Biol. 14 (2004) 264-272
10. Burge, C., Karlin, S.: Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268 (1997) 78-94
11. Burset, M., Guigó, R.: Evaluation of gene structure prediction programs. Genomics 34 (1996) 353-367
12. Chen, T.M., Lu, C.C., Li, W.H.: Prediction of splice sites with dependency graphs and their expanded bayesian networks. Bioinformatics 21 (2005) 471-482
13. Curwen, V. Eyras, E. Andrews, T.D. Clarke, L. Mongin, E. Searle, S.M., Clamp, M.: The Ensembl automatic gene annotation system. Genome Res. 14 (2004) 942-950
14. Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., Miller, W.: A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. 8 (1998) 967-974
15. Gelfand, M.S., Mironov, A.A., Pevzner, P.A.: Gene recognition via spliced sequence alignment. Proc. Natl. Acad. Sci. 93 (1996) 9061-9066
16. Hsieh, S.J., Lin, C.Y., Chung, Y.S., Tang, C.Y.: Comparative exon prediction based on heuristic coding region alignment. Proceedings of the International Symposium on Parallel Architectures, Algorithms, and Networks (2005) 14-19
17. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J. et al.: The UCSC Genome Browser Database. Nucleic Acids Res. 31 (2003) 51-54
18. Kent, W.J., Zahler, A.M.: Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. Genome Res. 10 (2000) 1115-1125
19. Korf, I., Flicek, P., Duan, D., Brent, M.R.: Integrating genomic homology into gene structure prediction. Bioinformatics 17 (2001) 140-148

20. Meyer, I.M., Durbin, R.: Gene structure conservation aids similarity based gene prediction. Nucleic Acids Res. 32 (2004) 776-783
21. Meyer,I.M., Durbin,R.: Comparative ab initio prediction of gene structures using pair HMMs. Bioinformatics 18 (2002) 1309-1318
22. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature 420 (2002) 520-562
23. Nadershahi, A., Fahrenkrug, S.C., Ellis, L.B.: Comparison of computational method for identifying translation initiation sites in EST data. BMC Bioinformatics 5 (2004) 14
24. Novichkov, P.S., Gelfand, M.S., Mironov, A.A.: Gene recognition in eukaryotic DNA by comparison of genomic sequences. Bioinformatics 17 (2001) 1011-1018
25. Parra, G., Agarwal, P., Abril, J.F., Wiehe, T., Fickett, J.W., Guigó,R.: Comparative gene prediction in human and mouse. Genome Res. 13 (2003) 108-117
26. Pedersen, A.G., Nielen, H.: Neural Network Prediction of Translation Initiation Sites in Eukaryotes: Perspectives for EST and Genome analysis. Proc. Int. Conf. Intell. Syst. Mol. Biol. 5 (1997) 226-233
27. Pruitt, K.D., Tatusova, T., Maglott, D.R.: NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 33 (2005) 501-504
28. Wheelan, S.J., Church, D.M., Ostell, J.M.: Spidey: a tool for mRNA-to-genomic alignments. Genome Res. 11 (2001) 1952-1957
29. Wu, T.D., Watanabe, C.K.: GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21 (2005) 1859-1875