DeTECTOR: Gene translation in Eukaryotes, a diagnostic tool

Massimiliano Orsini[1-2], Silvano Salaris[2] and Anna Tramontano[1-2]

[1] Department of Biochemical Sciences, University 'La Sapienza', P.le Aldo Moro, 5, 00185 Rome.
[2] Center for Advanced Studies, Research and Development in Sardinia (CRS4), Bioinformatics Unit, Parco Scientifico e Tecnologico, POLARIS, Edificio 1, 09010 PULA (CA).authors

Abstract

Gene prediction pipelines are often affected by a high false positive ratio. This is true in particular for splicing variant detection. We describe here a diagnostic tool (DeTECTOR: Detection Tool of Elements Controlling Translation of Open Reading frames), based on a neural network and aimed at classifying transcripts into three major classes: coding, pseudogenes and non-translated RNA. It evaluates the presence of transcriptional regulative elements, together with the properties of the putative translated products. It reaches good accuracy in detecting non-translated RNAs, while its performance in discriminating pseudogenes from coding transcripts is less satisfactory.

INTRODUCTION

Large-scale projects of genomes annotation often suffer by a high false positive ratio in detecting the correct genomic structure of a locus (Guigó R. et al, 2006). In addition, depending upon its function, a given transcript (even if correctly predicted by gene finding software) might not be translated in peptide. This latter case, which might be biologically relevant, depends upon transcript features such as the presence of a short polyA tail or a Non Mediated Decay (NMD) recognition site (Chang Y.F. et al, 2007) or the presence of specific secondary structures in its 5'UTR (Un-Translated Region) (Matlin A.J. et al, 2005). The translation machinery is a very complex system. Its complexity has increased in evolution (Gibson G. et Muse S., 2004). Regulation occurs at multiple levels: methylation of DNA, regulation of transcription, post transcriptional mechanisms, translational and post-translational mechanisms, etc. Many of these processes are mediated by protein-RNA interaction networks and depend upon specific RNA patterns (Abdul-Manan N. et, Williams K.R., 1996). Recognition of specific pattern by RNA binding protein can occur in UTRs or in coding regions (McCarthy J.E., 1998). Especially in the 5'UTR, many elements involved in the regulation of translational machinery are present. These elements include 5' cap 7-metil-guanosine (7mG), internal recognition sites, upstream open reading frames, secondary structure elements, patterns flanking the start codon, etc. (Maston G.A. et al, 2006). In a similar way other "regulative elements" are localized near the stop codon, the polyA sites or the 3'UTR and can act with different mechanisms (Munroe D. et Jacobson A., 1990). Several other classes of genes, with different features have been discovered. Whole genome studies have detected a large number of non-functional genes, called pseudogenes. With respect to their functional counterpart their sequence has "mistakes", for example they sometimes lack the start codon or have additional stop codons, do not present regulatory sequences, or derive from frame shifts that prevents them from producing functional peptides (Mighell, A.J. et al, 2000). According to their features or their origin they are classified in three

major classes: processed pseudogenes, non-processed pseudo genes or disabled pseudogenes (Brent M.R., 2005; Griffiths-Jones S., 2007). Due to their nature and intrinsic features, pseudogenes are often a problem for gene prediction algorithms and some authors claim that progress in pseudogene identification will improve accuracy of gene finding methods (van Baren, M.J. et Brent M.R., 2006).
In the last decade, many new RNA components have been discovered. In addition to the well known coding RNAs and the traditional group of non coding RNAs, such as rRNAs and tRNAs, the presence of RNAs with regulatory function has been established. Small RNAs molecules seems to have a crucial role in the regulation of gene expression through a process called RNA interference (Dykxhoorn D.M. et al, 2003). According to their functions, the molecules responsible for this mechanism are called miRNA (micro RNA), siRNA (short interfering RNA), snoRNA (small nucleolar RNA) and snRNA (small nuclear RNA).

THE DeTECTOR WEB SERVER.

We developed a public server which returns the propensity of a given transcript to be translated into a functional peptide. This tool consists of a neural network, which is trained on different classes of cDNAs (coding cDNA, pseudogenes cDNAs and a miscellaneous group of non-translated RNAs). Three classes of algorithms have been developed for scoring the input sequences. The first two add scores from different parameters, the third uses a decision tree where sequences whose parameters are below a given threshold are excluded from the final list. Once the score has been produced, the input sequence is classified in one of the three groups or labelled as undetermined.
Scoring parameters are determined on the basis of the presence of a given set of translational regulative elements and of the intrinsic features of putative open reading frames. This set of features has been chosen after a careful survey of the literature. They include: length of the transcript, GC composition, occurrence of additional start codons (AUG), relative occurrence of the three stop codons (UAG, UAA, UGA), number and size of polyG islands in the 5'UTR, polyA sites, ribosomal binding sites (McCarthy J.E., 1998), regulatory elements, RNA protein binding sites (van Helden J. et al, 1998; Abdul-Manan N. et al, 1996) and properties of all putative open reading frames.
The training set was built using cDNA sequences obtained by Ensembl (www.ensembl.org) and grouped in three main categories: protein coding, pseudogenes, and a miscellaneous group of untranslated RNAs (rRNA, miRNA, snRNA, siRNA, snoRNA).
Users can retrain the method using other specie-specific datasets (Saccharomyces, Drosophila, Danio Rerio, Mus, Rattus) available online The whole pipeline was implemented in Python. A preliminary versions of a web interface to access the database and run the pipeline with user's dataset is available at the address: https://detector.crs4.it (Figure 1). It uses the APACHE/PHP/HTML technology and includes a system to prevent automatic access and a pre-loaded example.
Preliminary tests on lower eukaryotes, based on regulative elements only, showed interesting results in classifying RNAs versus coding or pseudogenes transcripts, but rather poor accuracy in discriminating coding genes versus pseudogenes. To improve the method, we tested additional features such as the amino acid composition and length of the putative product. This improved the performance (Table 1) although pseudogenes still remain difficult to predict.
As an example, we report here the results of the analysis of the entire chimpanzee transcriptome (from Ensembl database; Flicek P. et al. 2008). The three algorithms, trained on human

sequences, exhibited different performance. They showed high specificity (true negatives/(true negative + false positive)) and a quite good sensitivity (true positive /(true positive + false negative)) on the chimpanzee dataset. With all methods, coding peptides are rarely mistakenly assigned as RNAs, confirming the good classification ability of the tool for these class of transcripts (Table 1).

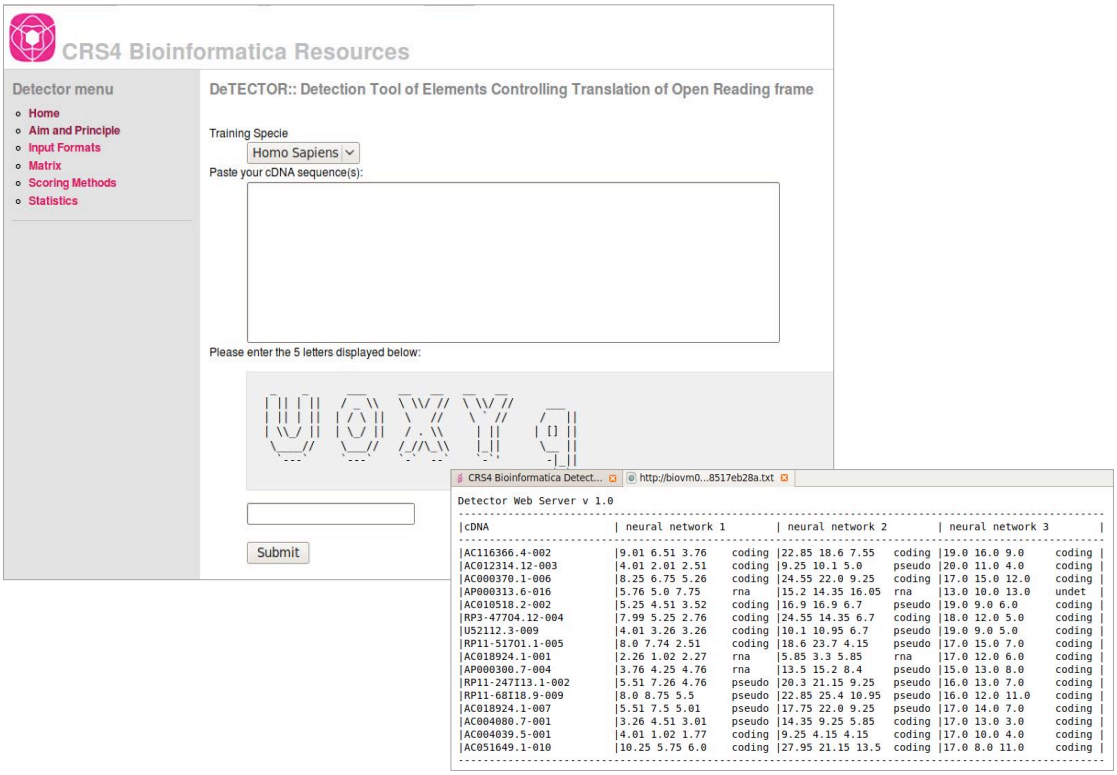

Figure 1. DeTECTOR Web Page. Results can be displayed in a separate page or downloaded.

Table 1. Chimpanzee transcriptome results. SAM: Scoring Assessing Methods

	coding vs non-coding (pseudogene + RNAs)			Mis classified pseudogenes	
SAM	*undetermined*	Sensitivity	Specificity		*undetermined*
A	0.8%	0.86	0.97	47.1%	1.7%
B	3.7%	0.80	0.98	21.1%	8.4%
C	3.0%	0.83	0.99	24.4%	2.0%

The pseudogenes subset remains the more difficult class to predict correctly. A consistent fraction of pseudogenes in the chimpanzee transcriptome was erroneously mis-classified as coding mRNAs. In other words, pseudogenes share significant similarities with coding genes, and the reason why they are not translated is not yet completely understood.

FUTURE DIRECTION AND ONGOING WORK

Recently, alternative splicing has been proposed as one of the mechanisms driving eukaryote complexity, and invoked as the reason for the difference among the estimated and expected gene

number in complex organisms. Unfortunately, this is not completely true, only about half of the proteins from alternative transcripts seems to be able to fold correctly and/or to perform a function (Tress M et al, 2006). It would seem that regulation of gene expression and non-protein RNAs are the likely explanation of higher organisms complexity. For this reason we are planning to implement a specialized version of DeTECTOR for analysing the alternative splicing products. It will retrieve each (putative) alternative transcript from a given gene, test the presence of regulative elements among splicing variants and assess the feasibility of their protein products providing an estimate of the propensity of a given gene to be a protein coding mRNA.

CONCLUSIONS

Assessing whether an experimentally observed transcript is indeed translated into a functional product is far from being trivial and it is unlikely to be very accurate when simple heuristic rules are used. Automatic learning methods are more promising, however there are problems associated with their development due to the small number of annotated pseudogenes and the lack of clear differences between their features and those of coding transcripts and to the high variability of protein coding transcripts.
As we showed above, combining different approaches from several pipelines provides satisfying results, although the entire process still needs improvement. Some aspects are currently being studied, among which the introduction of additional (albeit less reliable) transcript features in the evaluation pipeline, like the prediction of secondary structure of RNA molecules and the localization of putative protein domains. The latter has been used as a criterion to detect unstable protein structures and thereby to identify pseudogenes (Homma K. et al, 2002). This will clearly require substantial computational resources and, consequently, is inappropriate for large scale analysis..

Acknowledgements.
We want to thank Gianmauro Cuccuru for helping with the web application settings.

REFERENCES

Abdul-Manan N. et al., Williams K.R. 1996. “hnRNP A1 binds promiscuously to oligoribonucleotides: utilization of random and homo-oligonucleotides to discriminate sequence from base-specific binding.” Nucleic Acids Research 1996. ; v.24(20); Oct 154063-70.
Brent M.R. 2005. Genome annotation past, present, and future: how to define an ORF at each locus. Genome Res. 15(12): 1777-86.
Chang Y.F. et al. 2007. The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem. 76: 51-74.
Dykxhoorn D.M. et al. 2003. “Killing the messenger: short RNAs that silence gene expression”. Nat. Rev. Mol. Cell. Biol. 4: 457-467.
Flicek P., et al. 2008. Ensembl 2008. Nucleic Acids Res. 36(Database issue): D707-14.
Gibson G. et al., Muse S. 2004. Introduzione alla genomica. Zanichelli.
Griffiths-Jones S. 2007. Annotating noncoding RNA genes. Annu Rev Genomics Hum Genet. 8: 279-98.

Guigó R., et al. 2006. EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol. 7 Suppl 1: S2.1-31.
Homma K, et al. 2002. A systematic investigation identifies a significant number of probable pseudogenes in the Escherichia coli genome. Gene. 294: 25-33.
Maston G.A. et al. 2006. Transcriptional Regulatory Elements in the Human Genome. Annu. Rev. Genom. Human Genet. 7: 29-59.
Matlin A.J., et al. 2005. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 66(5): 386-98.
McCarthy J.E. 1998. Posttranscriptional Control of Gene Expression in Yeast. Microbiology and Molecular Biology Review. 62(4): 1492–1553.
Mighell, A.J. et al. 2000. Vertebrate pseudogenes. FEBS Lett. 468(2-3): 109-14.
Munroe D. et Jacobson A. 1990. Tales of poly(A): a review. Gene. 91(2): 151-8.
Tress M.L. et al. 2007. The implications of alternative splicing in the ENCODE protein complement. Proc Natl Acad Sci U S A. 104(13): 5495-500.
van Baren, M.J. et al. Brent M.R. 2006. Iterative gene prediction and pseudogene removal improves genome annotation. Genome Res. 16(5): 678-85.
van Helden J et al. 1998. Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies. J. Mol. Biol. 281: 827-842.