Automatic simulation of RNA editing in plants for the identification of novel putative Open Reading Frames

Fabio Fassetti[1], Claudia Giallombardo[2] and Ofelia Leone[1] and Luigi Palopoli[1] and Simona E. Rombo[2*] Pierluigi Ruffolo[1] and Adolfo Saiardi[3]

[1]DIMES, University of Calabria, Cosenza, Italy
[2]DMI, University of Palermo, Palermo, Italy
[3]LMCB, MRC, Cell Biology Unit & Department of Developmental Biology, University College, London, UK
*Corresponding Author: simona.rombo@unipa.it

Introduction

In plant mitochondria an essential mechanism for gene expression is RNA editing, often influencing the synthesis of functional proteins. RNA editing alters the linearity of genetic information transfer, introducing differences between RNAs and their coding DNA sequences that hind both experimental and computational research of genes. Thus common software tools for gene search, successfully exploited to find canonic genes, often can fail in discovering genes encrypted in the genome of plants. In this work we propose a novel strategy useful to intercept candidate coding sequences resulting from some possible editing substitutions on the start and stop codons of a given input organism DNA. Our method is based on the simulation of the RNA editing mechanism, in order to generate candidate Open Reading Frame (ORF) sequences that could code for some, yet unknown, proteins. Results obtained on the mtDNA of *Oryza sativa* are promising, since we identified ORF sequences trascripted in *Oriza*, that do not correspond to already known proteins in this organism. Part of the corresponding amino acid sequences present high homologies with proteins already discovered in other organisms, the remaining ones could represent novel proteins not yet discovered in *Oryza*.

Methods

In order to extract the ORF sequences from the genome of a given organism, special nucleotide triplets corresponding to the start and stop of an amino acid sequence have to be intercepted on the DNA sequence. Such triplets are called *start codons* and *stop codons*, respectively. In particular, there exist one start codon, that is atg, and three stop codons, that are tag, tga and taa. Although ORF sequences can be easily searched for in a genomic sequence by exploiting one of the existing software tools, such as for example ORF FINDER [3] and STARORF [4], taking into account the occurrences of such codons, this is not sufficient to intercept possible proteins coming from RNA editing mechanisms. This means that, in plants, several proteins are not found from the ORF sequences returned in output from such existing tools. We propose an automatic simulation of editing mechanisms possibly causing the presence of proteins not imputable to standard ORF sequences. This is rather meaningful in plants, where mtDNA editing mechanisms can often involve nucleotide triplets leading to start and stop codons.

We start from the mtDNA of a specific plant, and suppose that some editing substitutions might have happened causing the generation of some start/stop codons. Among all such possible new codons, only those corresponding to significative potential ORF sequences are taken into account. In particular, only ORF sequences corresponding to amino acid sequences of lenght at least 100 can correspond to potential proteins. Thus, between a start and a stop at least 300 nucleotides have to occur for interesting ORF sequences to be sigled out. Furthermore, the most frequent nucleotide substitution caused by editing is $c \rightarrow u$ at the RNA level, that is, $c \rightarrow t$ if we refer to mtDNA. Therefore we consider only this kind of nucleotide substitution in our analysis.

The input is a nucleotide sequence s_n (e.g., the mtDNA of a plant), that is scanned in all its three possible reading frames (for both the forward and the reverse cases), by considering all the substitu-

tions $c \to t$ that can generate new (*edited*) start/stop codons. Then, the nucleotide subsequences with minimum length 300 between a start and a stop codons are extracted, by taking care that only maximal subsequences are considered. Indeed, if several useful start codons occur before a same stop codon, only the first start codon is considered for the purpose of extracting the corresponding ORF sequence. All the other start codons are traduced as the corresponding amino acid Methionine (M) in the resulting amino acid sequence. This avoids intercepting all the possible subsequences. For what concerns the stop codons, the first one after the chosen start c_{START} is considered, if such a c_{STOP} is an original codon. If c_{STOP} is an edited stop, it is taken into account only if between c_{START} and c_{STOP} there are at least 300 nucleotides. Otherwise it is discarded, and the next c_{STOP} is searched for, by taking care of the same rule. We avoid this way subdividing a potentially significative sequence in several meaningless subsequences, even discarded since not enough long.

Among all the candidate ORF sequences generated as explained above, we consider only those involving some edited (start and/or stop) codons. Let S_{ORF} be the set of such sequences, whose corresponding amino acid sequences are referred to as *candidate protein predictions* in the set P_{ORF}. Sequences in P_{ORF} are compared against known proteins by exploiting available alignment algorithms (e.g., [1]), in order to single out interesting homologies. Some of the sequences in P_{ORF} can be found to be known proteins, in which case we discard them from further analysis. Let $\widehat{P}_{\text{ORF}}$ be the resulting amino acid sequences set, that we can divide in two further subsets $\widehat{P}'_{\text{ORF}}$ and $\widehat{P}''_{\text{ORF}}$. $\widehat{P}'_{\text{ORF}}$ includes amino acid sequences for which significant homologies have been found w.r.t. some proteins belonging to other organisms, while $\widehat{P}''_{\text{ORF}}$ contains the remaining ones. In both cases, a further filtering step is carried out by searching for the presence of possible transcripts by querying the DBEST [2], since this can be considered indicative of gene activity. Eventually, our system returns in output two sets of predicted proteins: P' and P'', respectively containing proteins in $\widehat{P}'_{\text{ORF}}$ and in $\widehat{P}''_{\text{ORF}}$ for which trascripts have been found.

Results

We show in Table 1 the most significant putative ORF sequences resulting from Blastp query for protein sequence similarity search, for *O. sativa*. Due to space contraints, we omit the analogous results obtained by querying the DBEST [2].

START TYPE	STOP TYPE	STRAND TYPE	START CODON	ORF LENGTH	SIMILARITY FOUND ORGANISM
edited	edited	rev	414167, 283246	327	*G. hirsutum*
edited	edited	fwd	372740, 453827	408	*Z. mays subsp. mays*
edited	edited	fwd	283844, 414765	330	*V. faba*
edited	edited	rev	315377	330	*A. duranensis*
edited	main	rev	314493	342	*N. tabacum*
edited	edited	rev	461810, 380723	324	*T. aestivum*
edited	edited	rev	404262	363	*F. rimosivaginus*
edited	main	fwd	316040	309	*C. card. var. scolymus*
edited	main	fwd	142093	312	*S. bicolor*
edited	edited	fwd	364454, 445541	327	*Z. mays subsp. mays*
edited	main	rev	201474	348	*B. napus*
edited	main	fwd	232846	339	*A. thaliana*
edited	main	rev	232370	387	*G. raimondii*

Table 1: Putative ORF sequences predicted for *O. sativa* (Blast bitscore ≥ 80).

Acknowledgements

F. Fassetti, L. Palopoli and S.E. Rombo have been partially supported by the INdAM – GNCS Project 2017 “Efficient Algorithms and Techniques for the organization, management and analysis of biological big data”.

References

[1] S. F. Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[2] M. S Boguski, T. M. Lowe, and C. M. Tolstoshev. dbEST–database for Expressed Sequence Tags. *Nat Genet.*, pages 332–333, 1993.

[3] NCBI. http://www.ncbi.nlm.nih.gov/projects/gorf/. Orf Finder.

[4] MIT. http://web.mit.edu/star/orf/. StarORF, Open Reading Frame Finder Tool.