BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 3 2010, pages 399–400
doi:10.1093/bioinformatics/btp688

Genome analysis

sORF finder: a program package to identify small open reading frames with high coding potential

Kousuke Hanada[1,2,3,*], Kenji Akiyama[1], Tetsuya Sakurai[1], Tetsuro Toyoda[2], Kazuo Shinozaki[1] and Shin-Han Shiu[3]

[1]Plant Science Center, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, [2]Bioinformatics and Systems Engineering Division, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan and [3]Department of Plant Biology, Michigan State University, East Lansing, MI, 48824, USA

Received on July 13, 2009; revised on November 2, 2009; accepted on December 8, 2009

Advance Access publication December 14, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: sORF finder is a program package for identifying small open reading frames (sORFs) with high-coding potential. This application allows the identification of coding sORFs according to the nucleotide composition bias among coding sequences and the potential functional constraint at the amino acid level through evaluation of synonymous and non-synonymous substitution rates.

Availability: Online tools and source codes are freely available at http://evolver.psc.riken.jp/

Contact: kohanada@psc.riken.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Proteins translated from small open reading frames (sORF) are likely to have important functions in eukaryotes. In yeast, small proteins include mating pheromones, proteins involved in energy metabolism, proteolipids, chaperonins, stress proteins, transporters, transcriptional regulators, nucleases, ribosomal proteins, thioredoxins and metal ion chelators (Kastenmayer *et al*., 2006). In multicellular organisms, many small proteins are known to be associated with developmental processes or hormone function (Butenko *et al*., 2003; Cock *et al*., 2001; Ghaemmaghami *et al*., 2003; Huh *et al*., 2003). However, there are relatively few small proteins in genome annotations because most gene finders tend to miss sORFs (Wang *et al*., 2003).

A number of gene prediction programs distinguish coding sequence (CDS) and non-coding sequence (NCDS) by integrating differences in nucleotide composition, intron splice sites, promoters, translational start/stop sites and polyadenylation signals. However, the integration of these multiple criteria increases the chance that true exons are not predicted as true (high false negative rate) (Claverie, 1997). The issue of false negative prediction is particularly serious for smaller CDSs ($\leq$300 nt; Wang *et al*., 2003).

Because of these limitations, we developed an analysis pipeline for identifying coding sORFs using the hexamer composition bias between CDS and NCDS (Hanada *et al*., 2007). Functional sORF genes experience stronger selective constraints on non-synonymous sites than synonymous ones. Therefore, we applied an additional filter requiring putative sORF genes to have significant selective constraints. Using these criteria, 2376 putative sORF genes (30–100 amino acids) were identified in the intergenic regions of the model plant *Arabidopsis thaliana* (Hanada *et al*., 2007). Despite the importance of sORF genes, currently no dedicated software or web application is available for their identification. Therefore, we developed a new software package and a new web application for identifying sORF genes between 10 and 100 amino acids.

2 METHODS AND IMPLEMENTATION

2.1 Work flow of sORF finder

sORF finder is a program package used to identify sORFs (10–100 amino acids) with high-coding potential in all six frames of a given nucleotide sequence. sORF finder consists of five programs, 'make_model.pl', 'simulate.pl', 'search_sORF.pl', 'collect_homo.pl' and 'examine_SP.pl'. All programs are implemented in Perl. Since each organism has a distinct hexamer composition in CDSs or NCDSs, this method requires as many known CDSs and NCDSs in an organism as possible. The pentamer and hexamer nucleotide composition frequency tables for both CDSs and NCDSs are generated by 'make_model.pl' (details in Section 2.2). In the web application, we have generated nucleotide composition frequency tables for 11 organisms including *Saccharomyces cerevisiae, A. thaliana, Oryza sativa, Caenorhabditis elegans, Drosophila melanogaster, Anopheles gambiae, Ciona intestinalis, Xenopus laevis, Danio rerio, Gallus gallus* and *Homo sapiens*. If a target organism is one of these 11, the web application is useful for identifying coding sORF. However, for other species, all procedures should be conducted on a local computer. After the nucleotide compositions are determining, sORF genes (between 30 and 300 bp) with qualifying coding likelihood is identified with 'simulate.pl' and 'search_sORF.pl' including a Bayes' estimation implementation detailed in Section 2.2. The comparison of synonymous and non-synonymous substitution rates to homologous sequences is performed by 'collect_homo.pl' and 'examine_SP.pl' detailed in Section 2.3.

The run time of each program increases as the nucleotide sequence increases. To estimate the run time of the program package, we applied our methods to a test dataset where the total amount of nucleotide sequences was 4 MB, 4 MB and 3 MB in CDS, NCDS and target sequences, respectively. Using a workstation (Linux RHEL4.1 x84_64 bit) with two Xeno 5160 (Dual core, 3.00 GHz) processors and 32 GB RAM, the total run time was 240 min. Run time details are shown in Supplementary Material 1.

*To whom correspondence should be addressed.

2.2 Bayes' estimation of coding likelihood

For a given sequence segment F, the posterior probability that F appears in the coding regions of a genome (P(CDS|F)) can be defined using the Bayes' theorem as follows:

$$P(CDS|F) = \frac{P(F|CDS)P(CDS)}{P(F|CDS)P(CDS)+P(F|NCDS)P(NCDS)}$$

Here, P(F|CDS) and P(F|NCDS) are the probabilities that F is derived from the CDS and NCDS training sequence, respectively. For identifying sORFs in intergenic regions, P(CDS) and P(NCDS) of prior probabilities should be set to be the proportion of total base pairs in CDS and NCDS, respectively. However, P(CDS) and P(NCDS) is unknown for identifying sORFs in mRNA sequences. Therefore, the default values of P(CDS) and P(NCDS) are set to be 0.5 and 0.5, respectively. P(F|CDS) and P(F|NCDS) are calculated from the Markov chain models of CDS and NCDS in an organism as follows:

$$P(F|S) = P_S(f_1f_2f_3f_4f_5)\prod_{i=1}^{i=L-5} P_S(F_{i+5}|f_if_{i+1}f_{i+2}f_{i+3}f_{i+4}), \quad S \in \{CDS, NCDS\}$$

Here, f_i is the *i*-th nucleotide in a given F sequence. *L* is the total length of the F sequence. The initiation probabilities $P(f_1,f_2,f_3,f_4,f_5)$ of P(F|S) are identical to the frequencies of a pentanucleotide in CDS and NCDS, and the transition probabilities are identical to conditional probabilities that a base appears at the next position in a given hexanucleotide. Detailed computation of each step is shown in Supplementary Material 2. The initiation and transitional probabilities of either P(F|CDS) or P(F|NCDS) in a target organism were calculated by the table generated by 'make_model.pl' in our program package.

To elucidate whether a target sequence is coding or not, P(CDS|F) was calculated on consecutive windows of 30 bp with a step size of 3 bp. The CI (coding index) for a given sequence is the summed posterior probabilities of all windows within a sequence. The CI is calculated by generating random CDS and NCDS training sequences from an organism. Since the CI is influenced by sequence length, 10 000 random sequences for each sequence length are generated and the CI is calculated for all random sequences. When the CI of a sequence is higher than the bottom 1% CI in random CDS and the top 1% CI in random NCDS, the target sequence is defined to have a significant coding potential. To present a numerical estimate of coding potential, a 'coding score' is defined as the percentile of CI in 10 000 random CDS sequences below the CI of the target sequence. CI of the random CDS and NCDS can be generated by 'simulate.pl' in our program package. The output files of 'make_model.pl' and 'simulate.pl' can be processed by 'search_sORF.pl' to identify putative coding sORF (10–100 amino acid).

2.3 Estimation of purifying selection

To assess the degree of functional constraints for sORFs with high-coding potential, sequences that are likely homologous to coding sORFs can be identified first by BLAST search (Altschul *et al*., 1997). The total number of synonymous and non-synonymous substitutions in the homologous sequences can be estimated by the modified Nei–Gojobori method (Zhang *et al*., 1998) because, in most CDSs, synonymous substitutions (nucleotide substitution without amino acid changes) occurs more frequently than non-synonymous substitutions (nucleotide substitution with amino acid changes) (Zhang *et al*., 1998). The null hypothesis is that the expected probability of non-synonymous substitutions is the same as that of synonymous substitutions. Then, whether synonymous substitutions occur more frequently than non-synonymous substitutions can be evaluated by the chi-square test. These procedures are applied by 'collect_homo.pl' and 'examine_SP.pl'.

3 PERFORMANCE

To evaluate the identification of novel small protein genes with respect to coding potential, our method was applied to two small protein gene datasets in *S.cerevisiae* and *A.thaliana* by 'search_sORF.pl'. These small protein genes were not annotated in the original genome release, but were identified in intergenic regions by similarity searches and/or functional studies (Butenko *et al*., 2003; Cock *et al*., 2001; Ghaemmaghami *et al*., 2003; Huh *et al*., 2003). There are a total of 123 small protein genes (<300 bp) in *S.cerevisiae* and *A.thaliana* (Supplementary Material 3). Out of 123, 113 were identified as coding sORF by our method, indicating that the false negative rate of our method is ∼9%. To furthermore evaluate the performance, we estimated false positive rates based on the intron annotation from the 11 species mentioned earlier. We identified sORF (10–100 amino acids) in intron sequences of 11 organisms and examined the coding potentials. Out of 46 457 958 sORFs in introns, 2 078 620 were recognized as having a high-coding potential (above-threshold CI values) by 'search_sORF.pl', indicating that the false positive rate is only 4%. However, even if the false positive rate is low, this method falsely identified many sORF with high-coding potentials. Therefore, we recommend examining the degree of functional constrains by 'collect_homo.pl' and 'examine_SP.pl'. Although the test of functional constrains may miss real protein genes (high false negative rate), the test produces more reliable sORFs (low false positive rate). Furthermore, it is important to combine experimental data with our method to identify coding regions translated into novel small proteins.

ACKNOWLEDGEMENTS

We thank Drs Kei Iida and Shuji Kawaguchi for helpful discussions.

Funding: RIKEN Plant Science Center, a program for promotion of Basic Research Activities for Innovative Biosciences (PROBRAIN) (to K.H.); National Science Foundation grant (MCB-0749634 to S.-H.S.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generate of protein database search programs. *Nucleic Acid Res.*, **25**, 3389–3402.

Butenko,M.A. *et al*. (2003) Inflorescence deficient in abscission controls floral organ abscission in Arabidopsis and identifies a novel family of putative ligands in plants. *Plant Cell*, **15**, 2296–2307.

Claverie,J.M. (1997) Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.*, **6**, 1735–1744.

Cock,J.M. *et al*. (2001) A large family of genes that share homology with CLAVATA3. *Plant Physiol.*, **126**, 939–942.

Ghaemmaghami,S. *et al*. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.

Hanada,K. *et al*. (2007) A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome Res.*, **17**, 632–640.

Huh,W.K. *et al*. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.

Kastenmayer,J.P. *et al*. (2006) Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome Res.*, **16**, 365–373.

Wang,J. *et al*. (2003) Vertebrate gene predictions and the problem of large genes. *Nat. Rev. Genet.*, **4**, 741–749.

Zhang,J. *et al*. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl Acad. Sci. USA*, **95**, 3708–3713.