BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 5 2009, pages 670–671
doi:10.1093/bioinformatics/btp024

Sequence analysis

FrameDP: sensitive peptide detection on noisy matured sequences

Jérôme Gouzy[1], Sébastien Carrere[1] and Thomas Schiex[2,*]

[1]Laboratoire Interactions Plantes Micro-organismes (LIPM) UMR441/2594, INRA/CNRS and [2]Unité de Biométrie et d'Intelligence Artificielle UR 875, INRA, F-31320 Castanet Tolosan, France

Received on October 22, 2008; revised on December 22, 2008; accepted on January 8, 2009

Advance Access publication January 19, 2009

Associate Editor: Alex Bateman

ABSTRACT

Summary: Transcriptome sequencing represents a fundamental source of information for genome-wide studies and transcriptome analysis and will become increasingly important for expression analysis as new sequencing technologies takes over array technology. The identification of the protein-coding region in transcript sequences is a prerequisite for systematic amino acid-level analysis and more specifically for domain identification. In this article, we present FrameDP, a self-training integrative pipeline for predicting CDS in transcripts which can adapt itself to different levels of sequence qualities.

Availability: FrameDP for Linux (web-server and underlying pipeline) is available at {{http://iant.toulouse.inra.fr/FrameDP}} for direct use or a standalone installation.

Contact: thomas.schiex@toulouse.inra.fr

1 INTRODUCTION

The reconstruction of transcripts from fragments of transcript sequences, such as EST (EST clusters, Tentative Consensus) provides a fundamental source of information for genome-wide studies and transcriptome analysis (Journet *et al*., 2002). This source will become widely accessible using new generation sequencing technology. When analyzing such data, the identification of the associated peptide sequence is required for:

- Extensive amino acid-level similarity searches or domain identification for GO-based functional classification.
- The construction of annotated full-length transcripts that can be used as training sets for gene prediction.
- The construction of peptides databases for proteomics analyses.

The prediction of coding regions from eukaryotic matured transcripts is similar to prokaryotic gene prediction, but additional difficulties arise from the fact that (i) EST clusters have heterogeneous sequencing depth which yields consensus cDNA of highly variable robustness; (ii) EST clusters may represent partial cDNAs, possibly missing START/STOP codons; and (iii) they may, in some cases, be derived from different organisms, such as a symbiont or pathogen rather than from the targeted organism. The CDS prediction should therefore be able to deal with 'noisy' sequences, with possible frameshifts, missing signals and being potentially derived from different organisms.

Pure *ab initio* CDS predictors for EST clusters such as ESTscan (Lottaz *et al*., 2003) require training sequences and ignore additional information such as possible protein similarities. Following the increasing trend of information integration in eukaryotic gene prediction, we designed FrameDP, a discriminative integrative CDS predictor for EST clusters. Compared with existing pipelines such as prot4EST (Wasmuth and Blaxter, 2004), FrameDP is a *self-trainable* pipeline and is therefore directly usable on organisms with no curated data. It inherits from FrameD (Schiex *et al*., 2003) the ability to handle noisy sequences and to integrate protein similarities and probabilistic models.

2 INTRINSIC FEATURES OF THE PIPELINE

2.1 FrameD

The core tool for the prediction of coding regions in the pipeline is the FrameD program. FrameD is natively capable of handling sequences formed from all IUPAC-IUB symbols, enabling FrameD to detect degenerated START and STOP codons. To estimate coding/non-coding potential, FrameD uses extended interpolated Markov models (IMMs) that explicitly handle unknown nucleotides 'N'.

Originally based on a weighted graph model, FrameD can be described as a conditional random field (CRF) gene finder. A gene is defined by a CDS composed of one or more regions coding in different frames (according to possible indels), surrounded by non-coding regions. The features used in the CRF model include IMM to estimate the coding/non-coding potential of a region, existence of START and STOP codons, existence of a similarity with known proteins and possible existence of frameshifts. The protein similarity feature favors predictions which are consistent with the observed similarities. CRF scaling parameters for frameshifts and protein similarities, respectively, define frameshift sensitivity (FS) and similarity confidence (SC) parameters. Decoding is performed by a Viterbi-like dynamic programming algorithm. *A posteriori* probabilities are computed using a Forward–Backward-like algorithm (including *a posteriori* probabilities of frameshifts).

2.2 Capturing coding styles with self training

A learning set is automatically extracted from the transcript sequences using regions showing a significant identity over a given

[*]To whom correspondence should be addressed.

length with a reference sequence database [defaults: Swiss-Prot scanned using NCBI-BlastX (Altschul *et al*., 1997) filtered with e = 1e-4, % id. = 40% over 100 amino acids].

Variations in GC content or more generally codon usage are known to significantly influence the predictive quality of statistical Markov models. In order to deal with possibly heterogeneous sets of sequences coming from different organisms, the FrameD pipeline is able to automatically estimate and use different IMMs.

To achieve this, the learning set identified using BlastX is split equally in subsets based on GC3% (or GC%) and an initial maximum likelihood IMM is built for each subset. Iteratively, each sequence is then reassigned to the model giving it maximum likelihood and new IMMs are estimated based on this new classification. This process is similar to the classification EM (CEM) algorithm of Celeux and Govaert (1992), albeit for class probabilities which are assumed to be identical. Iterations are stopped upon convergence or after a maximum number of iterations are reached.

2.3 Adaptation to sequence heterogeneity

CDS prediction on a given sequence is always done using the Markov model which maximizes its loglikelihood. Because of the variability in quality and origin in the analyzed sequences, FrameD is applied using a set of different parameter combinations.

For SC, two values corresponding to a standard (2) or high (1000) confidence are tried. The high confidence level allows to recover from possible low sensitivity of Markov models when a BlastX match exits.

For each level of SC, in order to deal with different depths of sequencing, three different FS are tried, from the less sensitive (−12) to the most sensitive (−6). These different values have been chosen based on experience and are user configurable.

Each of these combinations yields a corrected sequence together with an associated predicted CDS. Because each EST cluster is a transcribed sequence that likely contains a coding region, predictions with long CDS are preferred. With this aim, predictions are sorted by CDS length in a series of buckets corresponding to increasing CDS lengths from small (typ. 50 codons) to large (typ. 500 codons or more) by fixed steps (of 50 codons). From the longest non-empty bucket, predictions with the lowest SC and then with the lowest FS are preferred, in order to avoid spurious FS predictions.

Note that, since FrameD performs gene prediction on both strands, FrameDP can automatically reverses $3'$–$5'$ oriented EST cluster sequences. It also automatically produces sequences corrected for the detected frameshifts as well as corresponding CDS and amino acid sequences in the standard GFF3 and FASTA formats.

3 VALIDATION AND COMPARISON

The FrameDP pipeline has been used to predict CDS from *Medicago truncatula* EST clusters (Journet *et al*., 2002) and from *Helianthus annuus* EST clusters {{http://www.heliagene.org}}. We evaluated FrameDP on the 87 237 EST clusters of *H.annuus* by performing a global NCBI-BlastX interrogation of the *Arabidopsis thaliana* protein database (TAIR release 8). The initial set of EST clusters showed 19 580 hits with TAIR8 that spanned more than 80% of the *A.thaliana* protein. Following FrameDP frameshift corrections, this number rose to 20 576 (+1096) which shows that the correction method is effective. Thanks to a flexible parallelization script (paraloop), the complete analysis took just 2 days using four CPUs.

Compared with the alternative prot4EST pipeline, FrameDP has strong qualitative advantages. The most important of all is its ability to self-train directly on EST clusters instead of requiring curated cDNA sets to train the underlying ESTScan and DECODER (Fukunishi and Hayashizaki, 2001) software. Thanks to FrameD, FrameDP also directly integrates the similarity information inside the CDS prediction process instead of performing separate predictions. Beyond this, FrameDP can use multiple Markov models and can handle degenerated sequences both for signals (STOP/START codons) and inside Markov models.

4 WEB-SERVER AND STANDALONE PACKAGE

The PERL-CGI server, accessible at {{http://iant.toulouse.inra.fr/FrameDP}}, provides life scientists with a user-friendly interface to the pipeline (limited to batches of 50 sequences). It also provides an automatic protein description based on InterPro domain content. The functional annotation capabilities rely on BioMoby web services and on the REMORA workflow manager (Carrere and Gouzy, 2006).

A package for large-scale local application is provided under the CECILL2 open source licence. It includes FrameD, NCBI-BlastX and paraloop, under their own licenses. The pipeline is controlled by a single program, configurable using one configuration file.

Conflict of Interest: none declared.

REFERENCES

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

Carrere,S. and Gouzy,J. (2006) REMORA: a pilot in the ocean of BioMoby web-services. *Bioinformatics*, **22**, 900–901.

Celeux,G. and Govaert,G. (1992) A classification EM algorithm for clustering and two stochastic versions. *Comput. Stat. Data Anal*., **14**, 315–332.

Fukunishi,Y. and Hayashizaki,Y. (2001) Amino acid translation program for full-length cDNA sequences with frameshift errors. *Physiol. Genomics*, **5**, 81–87.

Journet,E.P. *et al*. (2002) Exploring root symbiotic programs in the model legume *Medicago truncatula* using EST analysis. *Nucleic Acids Res*., **30**, 5579–5592.

Lottaz,C. *et al*. (2003) Modeling sequencing errors by combining hidden Markov models. *Bioinformatics*, **19**(Suppl. 2), ii103–ii112.

Schiex,T. *et al*. (2003) FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *Nucleic Acids Res*., **31**, 3738–3741.

Wasmuth,J.D. and Blaxter,M.L. (2004) prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics*, **5**, 187.