PRO-FRAME: SIMILARITY-BASED GENE RECOGNITION IN EUKARYOTIC DNA SEQUENCES WITH ERRORS

Mironov A.A., *Gelfand M.S.

State Scientific Center for Biotechnology NIIGenetika, Moscow, Russia
Anchorgen, Inc., Santa Monica, USA
e-mail: misha@imb.imb.ac.ru

*Corresponding author

Keywords: gene recognition, spliced alignment algorithm, sequencing errors, eukaryotes

Resume

Performance of existing algorithms for similarity-based gene recognition in eukaryotes drops when the genomic DNA has been sequenced with errors. A modification of the spliced alignment algorithm allows for gene recognition in sequences with errors, in particular frameshifts. It tolerates up to 5% of sequencing errors without considerable drop of prediction reliability when a sufficiently close homologous protein is available (normalized similarity score 50% or higher).

Availability:

http://www.anchorgen.com

Analysis of sequence similarity is a powerful tool for gene recognition. It is employed in a number of database search programs, most notably BLASTX (Gish and States, 1993), and programs for exact prediction of exon-intron structure, in particular, Procrustes (Gelfand et al., 1996; Mironov et al., 1998), INFO (Hultner et al., 1994; Laub et al., 1998), GeneWise (Birney et al., 1997). The common idea behind these algorithms is that among numerous possible exon chains, an algorithm chooses the chain having the highest similarity to a related protein (target). This is done by modified dynamic programming treating introns as a special case of gaps (GeneWise) or by spliced alignment (Procrustes).

Testing of similarity-based gene recognition programs demonstrated that given sufficiently close relatives, they produce highly reliable predictions. In particular, the correlation between predicted and real human genes is 96-99% when homologous vertebrate genes are available (Mironov et al., 1998; Laub et al., 1998). However, the quality of gene predictions when the genomic DNA contains sequencing errors is much lower (Burset and Guigo, 1996). One possibility to avoid this problem is to use DNA spliced alignment instead of aligning translated candidate exons with proteins (Sze and Pevzner, 1997). However, it is well known that protein alignments are much more sensitive to distant similarities than nucleotide alignments. Thus it is indicative that there exist numerous protein-DNA alignment algorithms accounting for frameshifts (Posfai and Roberts, 1992; Birney et al., 1996; Guan and Uberbacher, 1996; Zhang et al., 1997; Pearson et al., 1997). However, none of them handles introns.

We have implemented a modified version of the spliced alignment algorithm performing gene recognition in the presence of frameshift errors. The algorithm treats introns as non-penalized gaps that may start only at dinucleotide GT and end at dinucleotide AG. Frameshifts and in-frame stop codons in the genomic sequence are allowed, but heavily penalized. There is an option for acceleration of dynamic programming, using k-tuple alignment technique due to M.Roytberg (Naziava et al., 1995). Since sequencing errors can destroy invariant dinucleotides at splicing sites, the program has a post-processing step. At this step the program identifies local drops of similarity at exon termini, and observing a sharp drop, moves the exon-intron boundary even if there are no suitable dinucleotides.

Results of testing the algorithm on a sample of human genes and related proteins from (Mironov et al., 1998) are given in Fig. 1. The performance at different error levels is estimated using the standard correlation coefficient measure (Burset and Guigo, 1996; Mironov et al., 1998). For comparison we present also the correlation coefficient demonstrated by the original Procrustes algorithm. Since the performance depends on the similarity between the gene and a target, the figure features plots of the correlation coefficient at different similarity levels. The similarity measure is the score of the alignment of the actual and target proteins divided by the halfsum of the scores of (trivial) alignments of the actual protein and the target protein with themselves. Such normalization accounts for varying protein length and amino acid composition. Sequencing errors were modeled as random nucleotide substitutions (80%), insertions (10%) and deletions (10%).

It is noteworthy that in the absence of sequencing errors Pro-Frame performs almost as well as Procrustes when the target protein is close to the analyzed gene, but the performance drops for distant relatives. This agrees with our observations about importance of the statistical filtering procedure implemented in Procrustes.
(Mironov et al., 1998). On the other hand, up to 3% rate of sequencing errors does not considerably influence the reliability of predictions, and further, up to 6% of errors are easily tolerated if the target protein is sufficiently close to the analyzed gene.

The above results demonstrate that Pro-Frame may be a useful tool for analysis of preliminary sequencing data, e.g. phase II output of major sequencing projects.

Acknowledgements

We are grateful to Drs. V. Bafna, J.W. Fickett, P. Pevzner and M.A. Roytberg for useful discussions. This work was partially supported by Anchorgen, Inc. (http://www.anchorgen.com).

References