Pro-Frame: similarity-based gene recognition in eukaryotic DNA sequences with errors

Andrey A. Mironov, Pavel S. Novichkov and Mikhail S. Gelfand

State Scientific Center for Biotechnology NIIGenetika, Moscow, 113545, Russia

Received on September 22, 1999; revised on June 5, 2000; accepted on July 27, 2000

ABSTRACT

Summary: 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 evolutionary distance similarity score 50% or higher).

Availability: The program is free for academic users and available upon request at http://www.anchorgen.com

Contact: mgelfand@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 and Smith, 1998), GeneWise (Birney and Durbin, 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 the 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 and Smith, 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 the 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 the dynamic programming, using the k-tuple alignment technique due to M.Roytberg (Nazipova et al., 1995). Since sequencing errors can destroy the invariant dinucleotides at splicing sites, the program has a post-processing step. At this step the program identifies runs of deletions at exon termini, and moves the exon–intron boundary even if there are no suitable dinucleotides. More exactly, observing more than 50% deleted positions in the region (−30, +30) around the exon junction, the program searches for the optimal position of the donor and acceptor splicing sites allowing for a single deviation from the invariant dinucleotide at each site. The program outputs the exon positions before and after the correction and the alignment of the predicted exons and the target protein.

Results of testing the algorithm on a sample of human genes and related proteins from (Mironov et al., 1998) are given in Figure 1. This sample consists of 256 genes. The average length of genomic sequences is approximately 8100 nucleotides, with the longest sequence exceeding 180,000 nucleotides. The number of exons ranges from 1 through 54, the average number of exons per gene is 5.5. The average length of exons in multi-exon genes is 140 nucleotides. The total number of protein targets is 731, their average length is 575 amino acids. Five independent rounds of mutations were performed for each sequence. The 3655 predictions (five times 731 comparisons) were done in about 5 h on a PC with Pentium II 400 MHz processor under Windows NT.

The performance at different error levels is estimated using the standard correlation coefficient measure (Burset
Fig. 1. Testing of Pro-Frame on a sample of human genes. Horizontal axis: similarity between actual genes and related proteins (in %, see the text for definition). Vertical axis: correlation coefficient (in %). Each curve corresponds to a specific level of sequencing errors (the percent of erroneous positions is given in the legend on the right). ‘Pro’ corresponds to the original Procrustes algorithm.

and Guigo, 1996; Mironov et al., 1998). For comparison we also present the correlation coefficient demonstrated by the original Procrustes algorithm and results of gapped BLASTX (Altschul et al., 1997) using the target protein as the query sequence (Figure 2). Since the performance depends on the similarity between the gene and a target, the figures feature 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 half-sum 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%); the latter two types of errors had length one through three with equal probabilities. For instance, at the error rate 15% this means that nucleotides at 12% of all positions have been changed, 3% of nucleotides were deleted (0.5% of single nucleotide deletions, 0.5% of dinucleotide deletions, and 0.5% of trinucleotide deletions), and 1.5% positions contained insertions (with 0.5% being single nucleotides, 0.5%, dinucleotides, and 0.5%, trinucleotides). Errors at GT-AG invariant intron termini were not treated as a special case, however, the fraction of mutated termini can be easily estimated given the overall error rate.

At all similarity and error levels Pro-Frame provides better recognition than straightforward BLASTX (cf. Figures 1 and 2). 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 I or II output of major sequencing projects, draft human genome sequences, etc. The initial identification of target proteins should be done by BLASTX, and then Pro-Frame can be used to exactly map the exon boundaries and to find relatively short exons that could be missed by the straightforward similarity analyses. Since Pro-Frame does not rely on the statistical properties of the analyzed genome (the only requirement is the GT–AG rule for the intron termini), the program can be used for gene recognition in invertebrate, plant, fungal, and even prokaryotic sequences.

ACKNOWLEDGEMENTS

We are grateful to Drs V.Bafna, J.Fickett, P.Pevzner and M.Roytberg for useful discussions. This work was partially supported by grants from the Russian State Scientific Program ‘Human Genome’, the Russian Fund for Basic Research (99-04-48347 and 00-15-99362), the Merck Genome Research Institute (244), and INTAS (99-1476).

REFERENCES


Birney,E., Thompson,J.D. and Gibson,T.J. (1996) PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation


