**SOME PROBLEMS OF GENE RECOGNITION IN PROTIST DNA.**

ASTAKHOVA T.V.1, GELFAND M.S.2, ROYTBERG M.A.1+

Keywords: protist DNA, splicing sites, exons, gene recognition

1 Institute of Mathematical Problems of Biology, Russian Acad. Sci., Pushchino, 142292, Russia

2 Institute of Protein Research, Russian Acad. Sci., Pushchino, 142292, Russia

+ Corresponding author: roytberg@impb.serpukhov.su

One of important steps in many gene recognition algorithms is preliminary filtration of candidate exons. Here we consider specifics of this problem in protist gene identification, in particular, multimodal distribution of splicing site scores and very long exons often occuring in these genomes.

Many gene recognition algorithms are based on exon assembly by dymanic programming [1]. The criteria for choosing the optimal chain of exons can be purely statistical (e.g. GREAT [2]) or based on similarity to some known protein (e.g. Procrustes [3]). The simplest way to generate the set of candidate exons is to consider all sequence fragments between the invariant AG and GT dinucleotides. However, in realistic situations this leads to a very large number of candidate exons that cannot be processed even by the computationally efficient dynamic programming. Thus there arises the problem of filtration of the set of candidate exons. This filtration should decrease the number of candidate exons, but not miss true exons. In addition to making possible processing of siggiciently large genomic DNA sequences, this decreases the combinatorial possibilities for generating high scoring alternatives to true genes in similarity-based analysis (the so called mosaic effect [4]) and thus provides for successful predictions using distant homologs [5].

Protists of the phylum Apicomplexa are an important cause of infectuous diseases both in humans (malaria, toxoplasmosis) and domestic animals. Some of them are objects of major sequencing projects (Plasmodium falciparum, Toxoplasma gondii). However, none of the existing gene recognition programs can process these sequences. We have started to develop algorithms that should fill this gap.

Earlier we have developed an algorithm for statistical filtration of candidate exons in higher eukaryote genes, and implemented it as a pre-processing module of Procrustes [5] and Cassandra [6]. However, it turned out that direct application of this algorithm for gene identification in protist DNA leads to missing an unacceptably large number of true exons.

Statistical analysis of multi-exon Apicomplexa genes (from Plasmodium spp., Theileria parva, Babesia bovis, Eimeria tenella, Toxoplasma gondii) demonstrated that it was caused by two peculiarities. First, the distribution of site scores (defined by some measure of their correspondence to an average frequency profile) is polymodal. Thus setting the usual thresholds for candidate sites leads to loss of true sites and the corresponding exons. Second, a large fraction of genes (approximately 50%) contains exons that are much longer than the remainig exons (up to 8000 nucleotides). Each such huge exon generates a very large number of highly scoring candidate exons thus obscuring some of the remaining short true exons.

To deal with this problem we introduce the notion of "long open reading frame", that is an ORF whose length exceeds some threshold. The exons that are situated within this reading frame are considered separately and thus do not compete with the exons in the remaining parts of the sequence. It turned out that it is sufficient to accept 100 candidate exons for each ORF exceeding 1000 nucleotides irrespective of its length.

This sharply improved the quality of filtration. Testing was performed on the set of 60 genes (total 206 exons) The filtering procedure lost 6 exons in 4 genes, that is comparable to performance of the non-modified filter on human genes [5] (the latter lost 45 protist exons).

This work was supported by grant 97-04-49040 from the Russian Fund of Fundamental Research, grant 14/98 from Russian State "Human Genome" program, and grant from USA Department of Energy.

**References**

1. M.A. Roytberg, T.V. Astakhova, M.S. Gelfand, "Combinatorial approaches to gene recognition". Comput. Chem 21, 229-235 (1997).
2. M.S. Gelfand, L.I. Podolsky, T.V. Astakhova, M.A. Roytberg, "Recognition of genes in human DNA sequences". J. Comput. Biol. 3,223-234 (1996).
3. M.S. Gelfand, A.A. Mironov, P.A. Pevzner, "Gene recognition via spliced sequence alignment". Proc. Natl. Acad. Sci. USA 93, 9061-9066 (1996).
4. S.-H. Sze, P.A. Pevzner, "Las Vegas algorithms for generecognition: suboptimal and error-tolerant spliced alignment". J.Comput. Biol. 4, 297-310 (1997).
5. A.A. Mironov, M.A. Roytberg, P.A. Pevzner, M.S. Gelfand, "Performance guarantee gene predictions via spliced alignment". Genomics 50 (1998).
6. S.-H. Sze, M.A. Roytberg, M.S. Gelfand, A.A. Mironov, T.V. Astakhova, P.A. Pevzner, "Algorithms and software for support of gene identification experiments". Bioinformatics 14, 14-19 (1998).