Very Fast Identification of tRNA in Genomic DNA

F. Lisacek [1] N. El Mabrouk [2]
lisacek@genetique.uvsq.fr mabrouk@univ-mlv.fr

[1] Laboratoire Genome et Informatique, Universite de Versailles
45 avenue des Etats-Unis, Versailles 78035 cedex France

[2] Institut Gaspard Monge, Universite de Marne-la-Vallee
2 rue de la Butte Verte, Noisy-le-Grand 93166 cedex France

The identification of functional regions in genomic DNA increasingly relies on the coupling of experimental work to computer processing of sequences. Newly sequenced fragments of a genome may be analysed with computer programs and the result of an automated search may guide a new set of experiments. In such a context, this paper focuses on the identification of tRNA sequences.

The role of tRNA in protein synthesis is of key importance. Over the last 20 years, the tRNA molecule has been extensively studied. The corresponding gene is a short sequence which folds in the form of a clover leaf. A large amount of sequences are available and aligned [1]. Conserved regions appear in the alignment which consists in only 76 positions.

There are basically two approaches to the identification of tRNA genes. It is either part of a general purpose method designed for searching and/or folding RNA sequences [2, 3, 4] or a self-contained method tailor made for searching tRNA genes such as [5, 6]. As one can expect, reported results are usually more accurate in the latter case.

Whatever the approach, priority is rarely given to how quickly a search is performed. Nevertheless, within years, complete genomes of various organisms will be available and fast sequence scanning is already becoming a concern.

The first really reliable algorithm, *tRNAscan* [5] is based on the use of "weight" or "consensus" matrices which make the definition of the RNA motifs more flexible and is often part of a general search strategy [7].The algorithm depends on two essential characteristics of the primary and secondary structure of the tRNA gene: (1) the presence of invariant (i.e. universal) and semi-invariant nucleotides located in two highly conserved regions, (2) the clover leaf structure consisting in four arms (paired bases) and three loops (unpaired bases), one of which being of variable size.

Such an approach was pushed further to improve both the flexibility and the speed of the algorithm. To address the question of flexibility, in particular in defining arms, each of the possible ten pairs (regardless of the orientation) is given a weight. These values reflect the stability of a base-pair and its frequency in natural RNA helices. As a result, a selection threshold for a potential arm is not simply a minimal number of Watson-Crick or (G,U) pairs but more accurately a minimal weight of successive pairings. Other changes involving the definition of thresholds and scoring functions were also made.

To address the problem of speed, work on string matching was considered. Searching for invariant or semi-invariant nucleotides is considered as searching for patterns with don't care symbols. The new algorithm is relying on the Shift-Add algorithm [8] defined in the case of search with mismatches. Computing time is also minimized by setting a hierarchy of searching operations. The hierarchy introduced in tRNAscan was slightly modified.

The modified version (FAStRNA) of tRNAscan yields good results. It runs 500 times faster and both rates of false positive (a selected sequence which does not correspond to a known tRNA) and false negative (a non selected sequence corresponding to a known tRNA) are reduced. Scanning is performed in a few seconds for small genomes. The newly sequenced genome of Saccharomyces cerevisiae is scanned both ways in less than 3 minutes and results match annotations found in databanks with three exceptions, two of which being arguably real tRNAs [9].

References

[1] Sprinzl M. et al. *Nucleic Acids Res.* Vol. 24, pp.68-72, 1996.

[2] Gautheret D., Major F., Cedergren R. *Comput. Appl. Biosci.* Vol. 6, pp.325-331, 1996.

[3] Chiu D.K.Y, Kolodziejczak T. *Comput. Appl. Biosci.*Vol. 7, pp.347-352, 1991.

[4] Eddy, S.R, Durbin, R. *Nucleic Acids Res.* Vol. 22, pp.2079-2088, 1994.

[5] Fichant G. A., Burks C. *J. Mol. Biol.* Vol. 220, pp.659-671, 1990.

[6] Pavesi A. et al. *Nucleic Acids Res.* Vol. 22, 1247-1256, 1994.

[7] Dandekar, T., Hentze, M.W. *Trends in Genet.* Vol. 11, 45-50, 1995.

[8] Baeza-Yates R., Gonnet G.H. *Communications of the ACM* Vol. 35, pp.74-82, 1992.

[9] El Mabrouk N., Lisacek F. *J. Mol. Biol.*, in the press, 1996.