Short Communication

DicodonUse: The Programme for Dicodon Bias Visualization in Prokaryotes

(codon / gene identification / genomes / open reading frames)

J. PAČES, V. PAČES

Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Prague, Czech Republic and Department of Biochemistry and Microbiology, Institute of Chemical Technology, Prague, Czech Republic

Abstract. The DicodonUse programme is aimed at a fast and simple assessment of genes present in prokaryotic nucleotide sequences. It identifies open reading frames that are not genes, and it distinguishes the genes that inherently belong to the genome in question from the genes that were inserted into the genome in the course of evolution. The programme is based on frequencies of dicodons used by the organism.

Accurate identification of microbial genes is becoming ever more important with the increasing rate of the genome sequencing projects. Several approaches have been adopted to screen long DNA sequences for open reading frames (ORFs). The basic approach is to scan the sequence for stop and initiation codons. This reveals all ORFs but does not show which of them are actual genes. Very short genes are usually missed, and some of the longer ORFs may not be protein or RNA encoding sequences. This is especially true for GC-rich DNA, where stop codons in the non-coding DNA may be relatively rare. In addition, even if an ORF is identified with a gene, the simple scan of stop and initiation codons does not reveal which of the codons thus serves as initiation signal for protein synthesis *in vivo*.

New approaches for gene identification are based on preferences of individual codons in coding sequences. There is a number of programmes for gene prediction based on this approach, e.g. GENEMARK (Borodovsky and McIninch, 1993), GCUA (McInerney, 1998), GENESCAN (Burge and Karlin, 1997), CRITICA (Badger and Olsen, 1999) and GLIMMER (Delcher et al., 1999). These programmes usually give a tabular output of predicted genes without graphical representation.

Results and Discussion

The DicodonUse programme is aimed at a fast and simple assessment of genes present in prokaryotic nucleotide sequences. It can be used for identification of ORFs that with a high probability are genes. In addition, it can distinguish the genes that inherently belong to the genome in question from the genes that were inserted into the genome in the course of evolution by horizontal transfer. Most of these genes are of viral or transposonal origin.

The programmme is based on assessment of frequencies of dicodons (two adjoining codons) used preferentially by organisms. It includes the codon bias of organisms as well as the observation that certain amino acids are only rarely neighbouring in polypeptide chains (Badger and Olsen, 1999). Using dicodons it is possible to distinguish genes from non-gene ORFs on the basis of the encoded protein primary structures. Dicodon frequencies for different genomes can be easily obtained and used for analysis.

The programme calculates logarithms of the probability that the particular dicodon occurs in genes. The sum of these logarithms (usually for the window of 90 nucleotides, but the size of the window can be changed to fit users' needs) then yields six possible probabilities for each reading frame. The programme normalizes them one against the other and draws six graphs together with other relevant information, e.g. locations of start and stop codons.

Figure 1 shows DicodonUse analysis of a short region of *Rhodobacter capsulatus* DNA. The green peaks in the six possible reading frames correspond to genes.

As mentioned above, it is generally difficult to automatically distinguish codons that initiate protein synthesis in individual ORFs from the internal methionine (or valine) codons. Often, ribosome binding sites

Received April 29, 2002. Accepted June 10, 2002.

This work was supported by the Center for Integrated Genomics and by grant MSMT 320050600.

Corresponding author: Jan Pačes, Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Flemingovo 2, 16637 Prague 6, Czech Republic. Fax: +420 224311019; e-mail: paces@img.cas.cz.

Abbreviations: ORF – open reading frame, RBS – ribosome binding site.

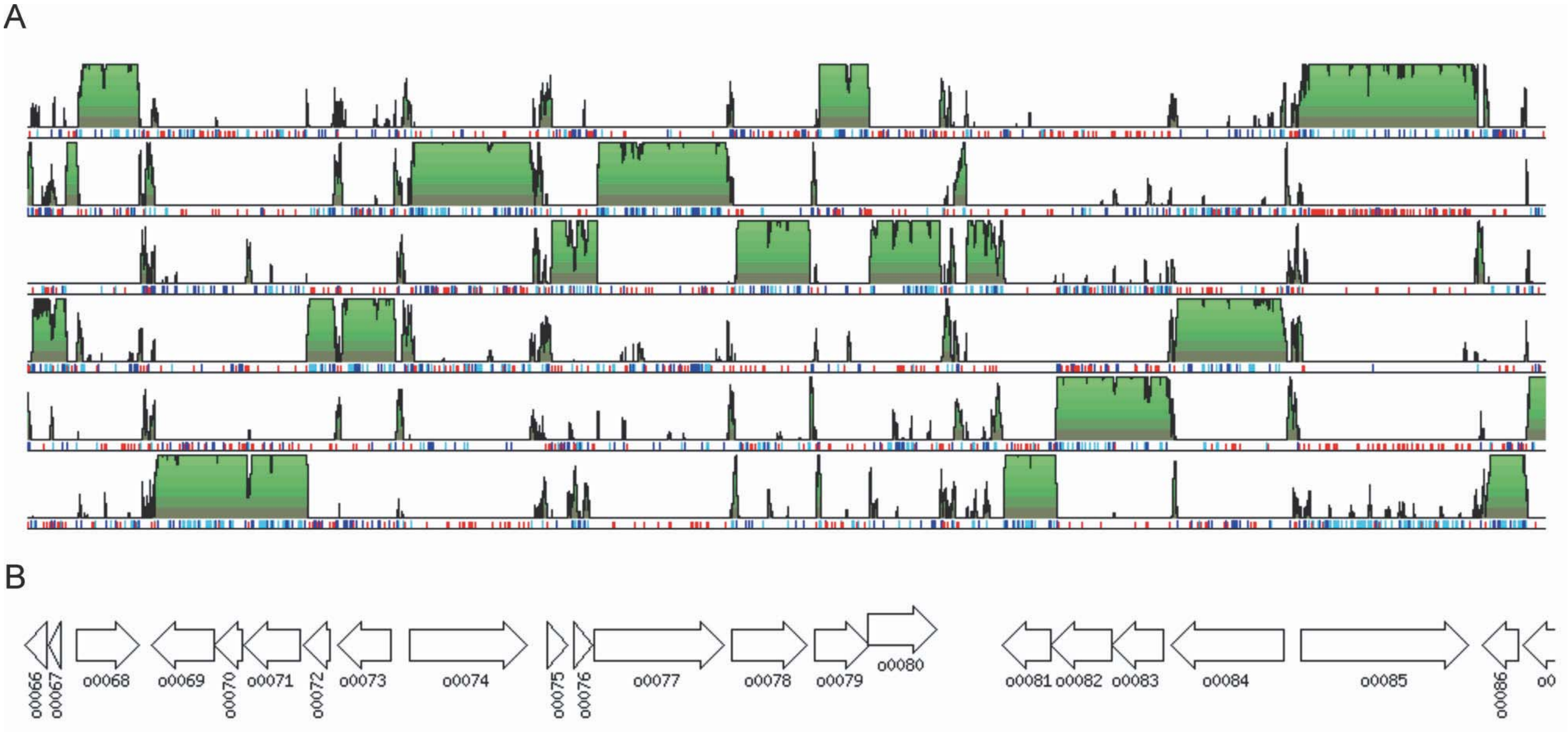

Fig. 1. DicodonUse analysis of a short region of *Rhodobacter capsulatus* DNA

A: Initiation codons are in blue, stop codons are in red. The green peaks are the DNA regions with the high occurrence of dicodons found in protein coding ORFs. B: The arrows indicate orientations of genes identified in the *R. capsulatus* DNA by the DicodonUse programme.

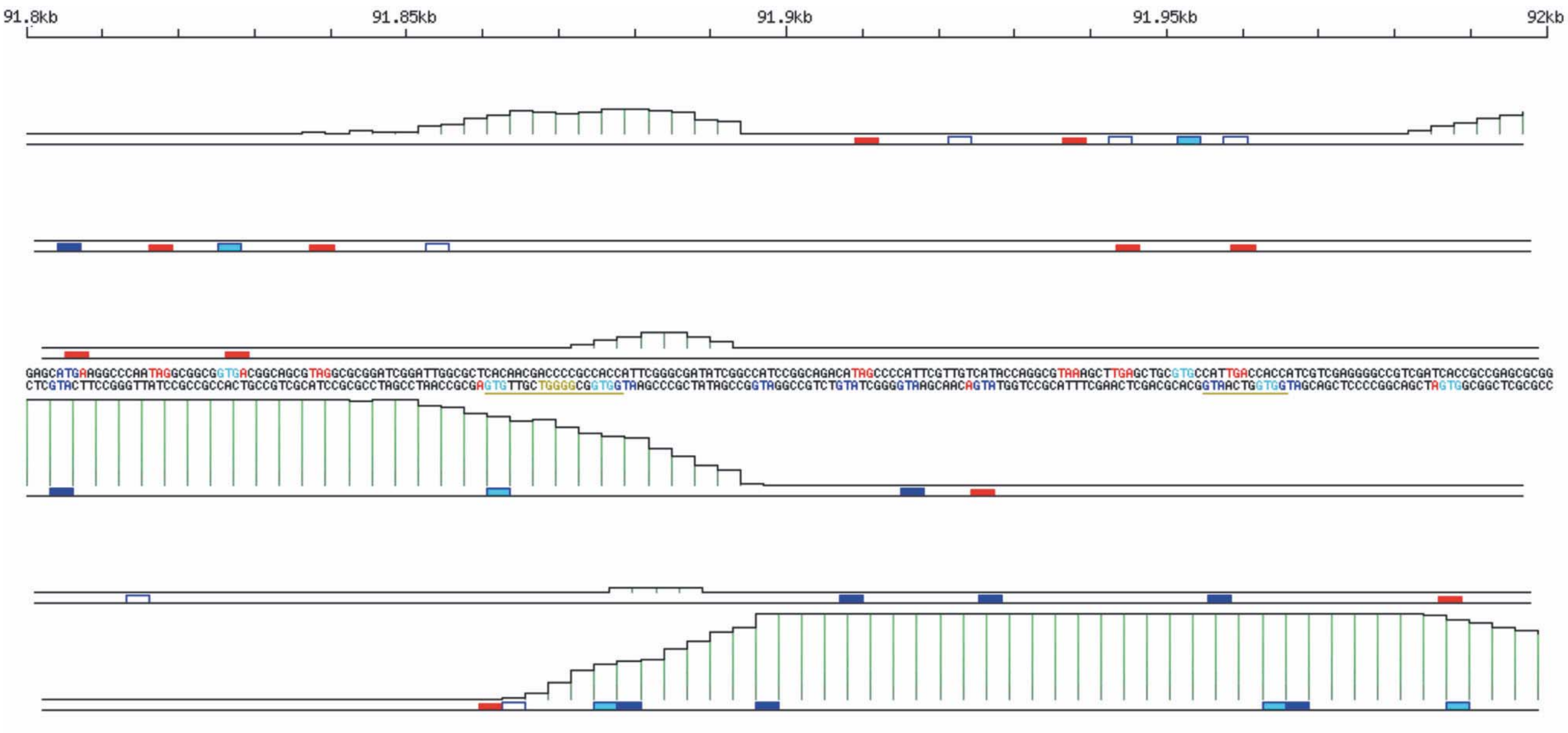

Fig. 2. Example of correct initiation codon identification by the DicodonUse programme. The Figure shows the probability of dicodon occurrence in the nucleotide sequence. The analysed sequence is in the middle of the Figure. The red boxes (and red letters in the sequence) are for stop codons, the dark blue boxes (and letters) are for ATG triplets, the light blue boxes (and letters) are for GTG and TTG triplets. The three reading frames oriented from left to right are in the upper part of the Figure, the three reading frames oriented from right to left are in the lower part of the Figure. One gene has its stop codon marked by the arrow; the neighbouring gene has its initiation codon marked by the asterisk. The potential RBSs in the nucleotide sequence are yellow and underlined.

(RBSs) are looked for that precede initiation codons in a proper distance. However, RBS may not be easily recognizable. Figure 2 shows an example of DicodonUse assessment of correct initiation codons. For the RBS consensus sequence, RGGRGGTGATN$_{(4\text{-}12)}$DTG was used (Badger and Olsen, 1999; Delcher et. al., 1999). Note that two potential RBSs were marked in the nucleotide sequence by the programme. However, the correct RBS can be easily identified in the context of dicodons found in the nucleotide sequence.

The DicodonUse programme package is available at http://genomat.img.cas.cz.

References

Badger, J. H., Olsen, G. J. (1999) CRITICA: Coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.* **16,** 512-524.

Borodovsky, M., McIninch, J. (1993) GeneMark: parallel gene recognition for both DNA strands. *Computers & Chemistry* **17,** 123-133.

Burge, C., Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78-94.

Delcher, A. L., Harmon, D., Kasif, S., White, O., Salzberg, S. L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27,** 4636-4641.

McInerney, J. O. (1998) GCUA (General Codon Usage Analysis). *Bioinformatics* **14,** 372-373.