Sense and nonsense DNA

Nucleotide sequences are routinely written in a 5′ to 3′ manner and, like this sentence, should be read from left to right to make sense. During processing of DNA sequence data the polarity can be accidentally reversed – for instance, when the sequence of the complementary strand is read without reversing it. The resulting mirror image does not reflect the properties of the original sequence and could lead to serious misinterpretation. For example, an ATG triplet would appear to be a start codon, but is in fact just a GTA sequence.

This is not merely a hypothetical problem. In 1991, we discovered three such instances by screening databases with the reversed M13mp18 sequence (R. C. L. Olsthoorn, PhD thesis, Leiden University, 1996). A recent survey raised this number to 20, indicating a growing tendency for sequence dyslexia. Searches using the reverse of *E. coli* rDNA and IS sequences yielded another 25 hits. Suspected sequences were mirrored and matched against the GenBank content, revealing additional, non-targeted inversions. Several entries appeared as mosaics composed of multiple 'mirror-image sequences', which made up between 50% and almost 80% of the entire deposited sequence.

Several categories of reversed sequence can be recognized. The first category constitutes those sequences in which the order of nucleotides has been entirely reversed (Fig. 1a). In the original papers, the sequences are presented correctly, although the complementary strands are usually shown.

A second category contains sequences in which mirror images of multiple cloning sites or other vector sequences are present at the 5′ and/or 3′ ends (Fig. 1b). This we conclude from finding restriction sites for enzymes such as *Hin*dIII, *Sph*I, *Pst*I, *Sal*I and *Xba*I in the same order as they appear in the M13mp18 polylinker, but with reversed polarity.

Members of the third category of reversed sequence contain additional parts of their own sequence as mirror images – usually an inversion of a nearby region (Fig. 1c). Interestingly, in none of these cases is the 'mirror image' an exact (100%) reversal of the original sequence. This suggests that, in these cases, the sequences were independently determined, but one of them was read backwards.

A fourth category consists of non-rDNA sequences that are contaminated with rDNA sequence (Fig. 1d). In most cases, the rDNA comes from the same organism (e.g. *Mustc41bb*, which encodes a mouse T-cell receptor protein, contains mouse 18S rDNA). In *Dirdg-2*, a gene encoding an antigen from the worm *Dirofilaria immitis*, the rDNA fragments belong to a possible bacterial endosymbiont.

Finally, members of a fifth category of reversed sequence contain multiple inverted sequences derived from polylinker sites, transposons, IS elements, rDNA or from themselves (Fig. 1e).

In the past, the problem of database corruption with vector DNA sequences has been amply discussed, but apparently without much effect. At present some 150 entries still contain M13-like sequences, even in open reading frames (R. C. L. Olsthoorn and J. van Duin, unpublished). The sequence inversions that we describe here are just another aspect of database contamination. Our current search was based only on sequences from vectors, transposons, IS elements and rDNA. It is possible that we are just signaling the tip of the iceberg. For instance, using the reverse of the rice chloroplast genome as query sequence also yields a number of matches.

Reference

1 Altschul, S. F. *et al.* (1990) *J. Mol. Biol.* 215, 403–410

RENÉ C. L. OLSTHOORN AND JAN VAN DUIN

Institute of Molecular Plant Sciences and Leiden Institute of Chemistry, Gorlaeus Laboratories, Leiden University, PO Box 9502, 2300 RA Leiden, The Netherlands.

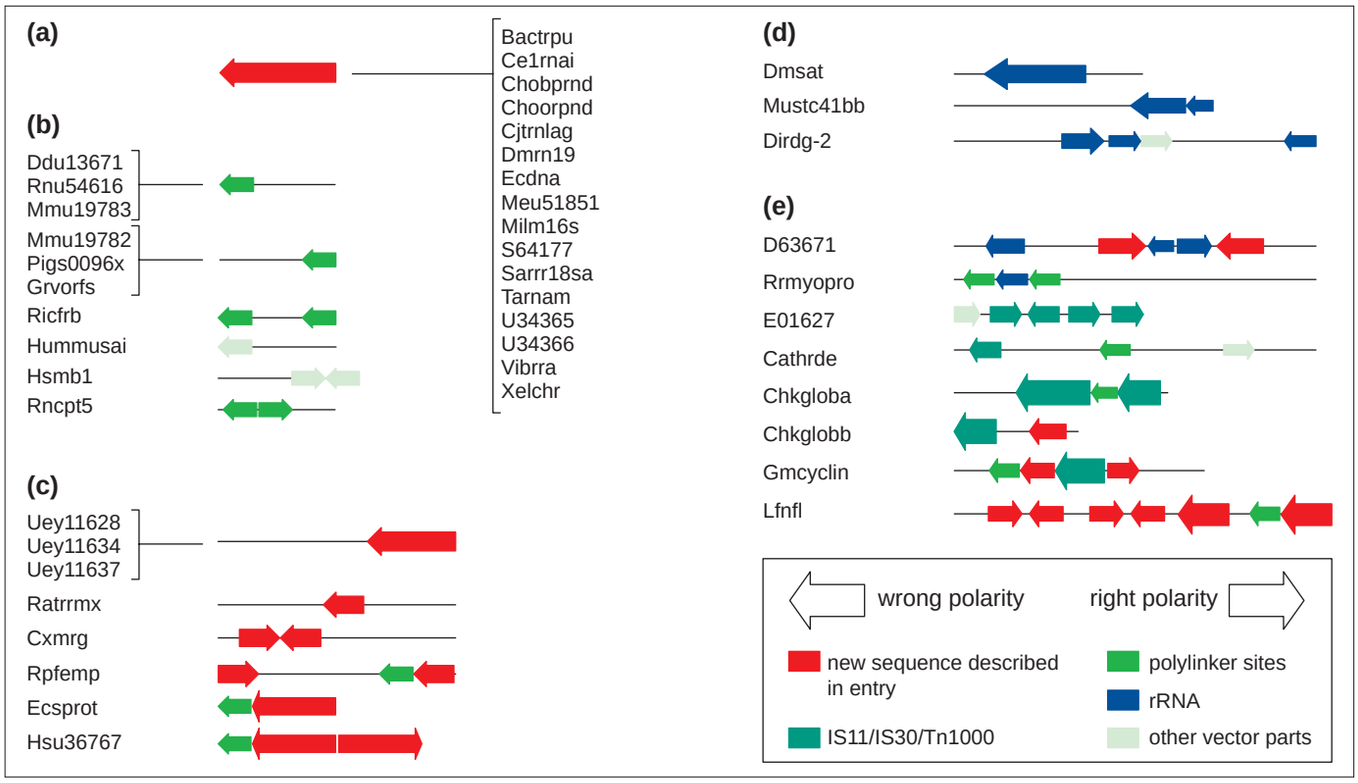

Figure 1

Summary of BLAST[1] searches of the GenBank and EMBL databases through the National Center for Biotechnology Information (NCBI, Bethesda, MD, USA), using the reversed sequences of M13mp18, pUC19 and fragments of the *E. coli* chromosome containing rRNA and tRNA genes.

 0968–0004/98/$19.00 PII: S0968-0004(98)01191-8