J Mol Evol (1995) 41:582–586

JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 1995

Sense in Antisense?

D.R. Forsdyke

Department of Biochemistry, Queen's University, Kingston, Ontario, Canada K7L3N6

Received: 18 October 1994 / Accepted: 25 April 1995

Abstract. A correspondence between open reading frames in sense and antisense strands is expected from the hypothesis that the prototypic triplet code was of general form RNY, where R is a purine base, N is any base, and Y is a pyrimidine. A deficit of stop codons in the antisense strand (and thus long open reading frames) is predicted for organisms with high G + C percentages; however, two bacteria (*Azotobacter vinelandii, Rhodobacter capsulatum*) have larger average antisense strand open reading frames than predicted from (G + C)%. The similar codon frequencies found in sense and antisense strands can be attributed to the wide distribution of inverted repeats (stem-loop potential) in natural DNA sequences.

Key words: Open reading frame — Codon usage — GC/AT pressure — Stem-loop

Introduction: Antisense Phenomena

A DNA segment encoding a protein usually has a "sense" strand and a complementary "antisense" strand which acts as a template for RNA polymerase. Conventionally, the sense strand is considered to encode the protein since it has the same sequence as the mRNA. Attention has been drawn to the possibility of long open reading frames in the antisense strand which might encode "antisense" proteins (Meckler 1969; Biro 1981a,b; Blalock and Smith 1984; Merino et al. 1994). Codons for hydrophilic and hydrophobic amino acids on the sense strand may sometimes be complemented, in frame, by codons for hydrophobic and hydrophilic amino acids on the antisense strand. Furthermore, antisense proteins may sometimes interact with high specificity with the corresponding sense proteins (Blalock and Bost 1986; Blalock 1990; Clarke and Blalock 1991). The interactions involve multiple contacts along the lengths of the polypeptide chains (Tropsha et al. 1992).

Yomo and co-workers (1992) interpreted the discovery of long antisense open reading frames in certain plasmid genes as indicating that some "unknown force" is protecting the frames from mutations generating stop codons. Recently, Yomo and Urabe (1994) have generalized these arguments, taking into account the observation that the frequencies of individual codons in sense strands are often similar to their frequencies in the antisense strands (when read in the same phase and with the correct polarity; Alff-Steinberger 1984, 1987; Yomo and Ohno 1989). Others have suggested that the genetic code may initially have evolved to favor the simultaneous emergence of sense and antisense peptides (Alff-Steinberger 1984; Zull and Smith 1990; Konecny et al. 1993).

The physiological relevance of antisense phenomena at the RNA level is well established (e.g., Tomizawa 1984). However, antisense phenomena at the protein level are more controversial (Goldstein and Brutlag 1989; Eberle and Huber 1991; Tropsha et al. 1992; Moser et al. 1993). I here point out that many antisense phenomena at the protein level can be interpreted as incidental by-products of (1) the evolution of interactions between mRNA codons and tRNA anticodons (Eigen and Schuster 1978), (2) the fine-tuning of base composition to avoid *inter*species recombination ("GC/AT pressure"; Filipsky 1990), and (3) the fine-tuning of stem-loop-forming potential to promote *intra*species recombination ("fold pressure"; Forsdyke 1995b–d).

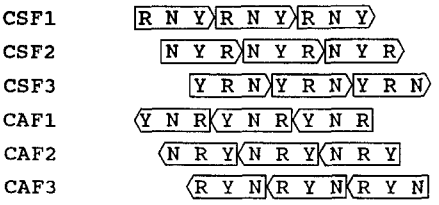

Fig. 1. The proposed prototypic triplet code (*RNY*) predicts a unique relationship between reading frames in sense and antisense strands of DNA. *CSF1, CSF2,* and *CSF3* refer to the three frames in the coding sense strand. *CAF1, CAF2,* and *CAF3* refer to the three corresponding frames in a coding antisense strand. *Boxes with arrows* indicate the polarity of coding triplets (5′ → 3′).

NonStop Frames Use Same Frame as ORFs

A long open reading frame can only exist in the antisense strand if there are no stop codons. Following the convention of Yomo et al. (1992), I here refer to such frames as nonstop frames (NSFs) to distinguish them from the open reading frames (ORFs) of the sense strand.

From consideration of the context of interactions between codons and their cognate tRNA anticodons, it has been proposed that the modern genetic code evolved from a prototypic triplet code of general form RNY, where R is a purine base, N is any base, and Y is a pyrimidine base (Eigen and Schuster 1978; Bossi and Roth, 1980). Traces of this code seem to be present in some modern genes and may sometimes be used to predict ORFs (Shepherd 1982). Thus, to simplify, we may write a DNA sequence encoding a protein as a series of RNY codons.

Figure 1 shows this for the three frames in a coding sense strand (CSF1, CSF2, CSF3), and for the three corresponding frames in a coding antisense strand (CAF1, CAF2, CAF3). It is seen that the ORF designated CSF1 is the *only* frame in the sense strand which corresponds to RNY. Similarly, CAF1 is the only frame in the antisense strand which corresponds to RNY. Thus, an antisense NSF should usually be in the same frame as that of the corresponding ORF (Alff-Steinberger 1987). It can also be seen that CSF2 (NYR) corresponds to CAF3, not CAF2. Similarly, CSF3 (YRN) corresponds to CAF2, not CAF3. This was recently confirmed by Yomo and Urabe (1994) in a study of long codon sets generated by combining *E. coli* coding regions into one long sequence. Thus their results are consistent with the proposal of a prototypic RNY code, as well as with the existence of genes in the antisense strand (as they suggest).

Long NSFs in High G + C DNA

If there are no stop codons (TAG, TAA, TGA) in the NSF, then there can be no complementary codons (CTA, TTA, TCA) in the corresponding ORF. CTA and TTA code for leucine. TCA codes for serine. These two amino acids are found frequently in proteins. Thus, except for an unlikely ORF with no leucine or serine, whether there will be a long NSF depends on which synonymous codons for leucine and serine are used in the corresponding ORF. Leucine and serine are two of the three amino acids for which there are six synonymous codons. An important factor determining codon choice is the species-specific pressure on a genome to adopt a particular (G + C)% base composition (''GC/AT pressure''; reviewed by Filipski 1990). AT-rich codons are used infrequently in species with a high (G + C)% (i.e., RNY is usually GNC).

In Fig. 2A the combined average frequencies of the usages of CTA, TTA, and TCA as codons in sets of genes from various species are plotted against the corresponding average percentages of G + C in coding regions calculated from the frequency and GC content of all the codons of each gene set. The latter should provide a measure of GC/AT pressure. The data are taken from the species codon usage profiles of Wada et al. (1990). The right ordinate shows the average length of NSF expected at particular values of the usage of CTA, TTA, and TCA. Thus 1% usage of the codons implies an average NSF of 100 codons. An extreme example is *Rhodobacter capsulatum* (RCA), which uses CTA, TTA, and TCA at a frequency of only 0.05%, implying an average NSF of 2,000 codons in the 21 genes examined by Wada et al. (1990).

As the (G + C)% increases, preferred codons become increasingly GC-rich, and the combined usage of the AT-rich codons CTA, TTA, and TCA decreases. It was found empirically that there is an approximately rectilinear relationship between the logarithm of the sum of the percentages of CTA, TTA, and TCA and the average (G + C)% of the corresponding coding regions. However, at high (G + C) percentages, values for the frequency of CTA + TTA + TCA for two species, *Azotobacter vinelandii* (AVI) and *Rhodobacter capsulatus,* are below the lower limit of the 95% prediction interval (estimated using Minitab statistics software; Ryan and Joiner 1994). TCA participates equally with CTA and TTA in decreasing in frequency disproportionately at high (G + C)% in these species. A species with more GC-rich codons, *Streptomyces* (STM), remains close to the regression line.

The divergence is also apparent in the case of GC-rich plasmid genes (shown as triangles), which encode bacterial nylon-degrading enzymes (Yomo et al. 1992). Except for the gene NA26AHDH, the values for the sum of the percentages of CTA + TTA + TCA as codons in these genes are zero (corresponding to a completely open average NSF). The value for NA26AHDH is close to the regression line and corresponds to an average NSF of 400 codons. However, the difference between NA26AHDH and the other plasmid genes is due to only *one* codon. (TCA is present once in NA26AHDH.) In a series of only four genes this difference is not statisti-

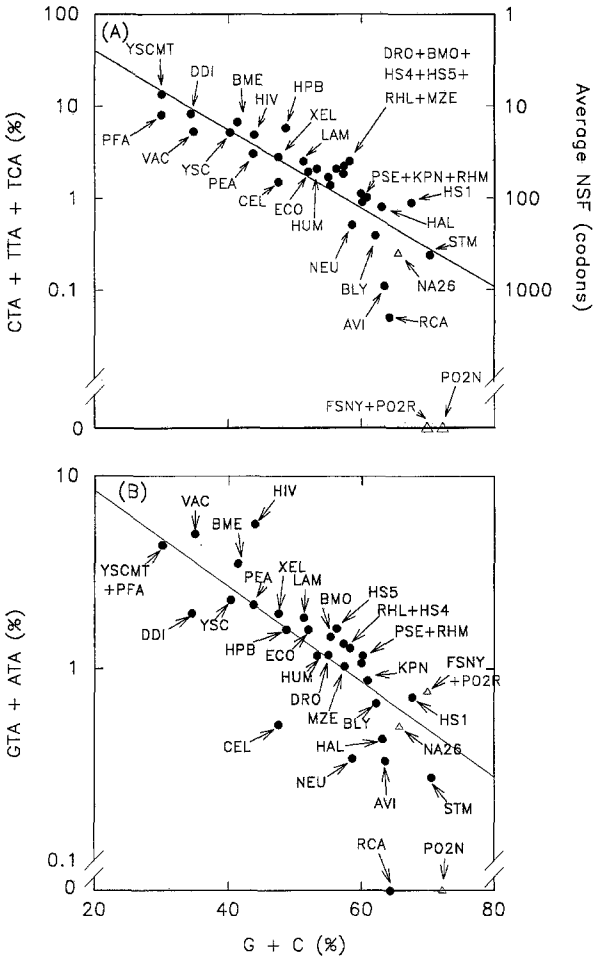

Fig. 2. Potential length of open reading frames in the antisense strand (NSF) increases as (G + C)% of total codons increases. Codon usage tables for various species were used to calculate the sum of the frequencies of **(A)** the codons CTA, TTA, and TCA, which are the complements of stop codons (TAG, TAA, TGA), and **(B)** the codons GTA and ATA, which are members of the set of codons of general form NTA. From knowledge of the frequency and GC content of individual codons, the overall GC content of the coding regions used to construct the tables was also calculated (abscissa). Data-points for individual species are shown as *filled circles.* Data-points for plasmid genes encoding enzymes which degrade nylon oligomers are shown as *triangles.* Species codes are taken from Wada et al. (1990): *AVI, Azotobacter vinelandii: BLY,* barley; *BME, Bacillus megaterium; BMO, Bombyx mori; CEL, Caenorhabditis elegans; DDI, Dictyostelium discoideum; DRO, Drosophila melanogaster; ECO, Escherichia coli; HAL, Halobacterium halobium; HIV,* human immunodeficiency virus type 1; *HPB,* hepatitis B virus; *HS1, Herpes simplex* virus type 1; *HS4,* Epstein-Barr virus; *HS5,* human cytomegalovirus; *HUM, Homo sapiens; KPN, Klebsiella pneumonii; LAM,* bacteriophage lambda; *MZE,* maize; *NEU, Neurospora crassa; PEA,* pea; *PFA, Plasmodium falciparum; PSE, Pseudomonas; RCA, Rhizobium capsulatus; RHL, Rhizobium leguminosarum; RHM, Rhizobium meliloti; STM, Streptomyces; VAC,* vaccinia virus; *XEL, Xenopus laevis; YSC, Saccharomyces cerevisiae; YSCMT,* mitochondria of the latter. The GenBank designations of nylon oligomer-degrading plasmids are: *FSNY,* FSNYLBB; *NA26,* NA26AHDH; *PO2N,* PO2NYLC; *PO2R,* PO2RSB. Correlation coefficients (r) for the least square regression lines are **(A)** 0.84, and **(B)** 0.82.

cally significant, and so does not support the suggestion that there is some special mechanism protecting the plasmid NSFs from accumulating mutations that generate stop codons (Yomo et al. 1992; Yomo and Urabe 1994). Such a special mechanism might be expected to affect only CTA, TTA, and TCA. However, in the plasmids there are zero frequencies not only of these codons, but also of seven of the eight codons of general formula W_3 (e.g., TTT, ATA, etc.), and of about half the 24 codons of general formula W_2S (e.g., AGA, TGT, etc.). It seems likely that, as suggested by Ikehara and Okazawa (1993), the plasmid genes are merely responding to GC/AT pressure, and not to some "unknown force."

An "Unknown Force" in *Rhodobacter* and *Azotobacter?*

Assuming that the observed regression line (Fig. 2A) is a close representation of the relationship between GC/AT pressure and the frequencies of CTA + TTA + TCA, species whose frequencies of these codons fall close to the regression line would appear to be responding simply to GC/AT pressure. The species whose frequencies of the codons diverge significantly below the line at high GC/AT pressures would lend some credence to the "unknown force" postulate. The average NSF for *all* the 37 genes of *Azotobacter vinelandii,* and for *all* the 21 genes of *Rhodobacter capsulatus* (that were studied by Wada et al. 1990), is greater than predicted by the regression line. The "unknown force" would seem to have acted on all the genes of these species. It would seem unlikely that all 58 genes would have sufficient flexibility in the coding of the sense-strand product to permit the evolution of a useful antisense-strand product. Thus, either pressures to form each antisense protein have not been causing these genes to lose CTA, TTA, and TCA or the only way pressure to form antisense protein can work is by committing numerous "innocent bystander" genes to the same strategy; that is, the evolution of one particular antisense product was so advantageous that all the genes in the organism were required to respond to some "unknown force" in order to acquire the potential to generate an antisense product. Alternatively, the organisms could be responding to some intrinsic species-specific "unknown force" the blind consequences of which would be NSFs longer than predicted from (G + C)%.

Role for TpA Pressure in Rhodobacter?

In *Rhodobacter capsulatus* two codons of general form NTA (CTA and TTA) have zero frequency, and TCA has a frequency of 0.05%. The latter value is the lowest of all the species tabulated by Wada et al. (1990), except for *Azotobacter vinelandii* (0.01%). The low values are not specific to these three codons. Figure 2B shows plots of

the combined frequencies of two other codons of general form NTA (GTA, encoding valine; ATA, encoding isoleucine). As in Fig. 2A, most of the data-points are close to the regression line. The point corresponding to *Rhodobacter capsulatus* diverges below the regression line. Indeed, in this species the usage of GTA + ATA is zero, which is beyond the 95% prediction interval (Ryan and Joiner 1994). Thus, it is possible that the divergence may be ascribed, in part, to "TpA pressure," which is a "known" but not well understood pressure (Nussinov 1984; Alff-Steinberger 1987; Barrai et al. 1991). TpA pressure affects all codons of general forms NTA and TAN (Forsdyke 1995d).

The point corresponding to *Azotobacter vinelandii* is closer to the regression line and is within the 95% prediction interval. In this species the average usages of GTA and ATA are 0.27% and 0.06%, respectively. With respect to these codons this organism may simply be responding to GC/AT pressure. Thus TpA pressure does not appear to explain the low levels of CTA and TTA in *Azotobacter vinelandii.* Here the "unknown force" postulate may be valid.

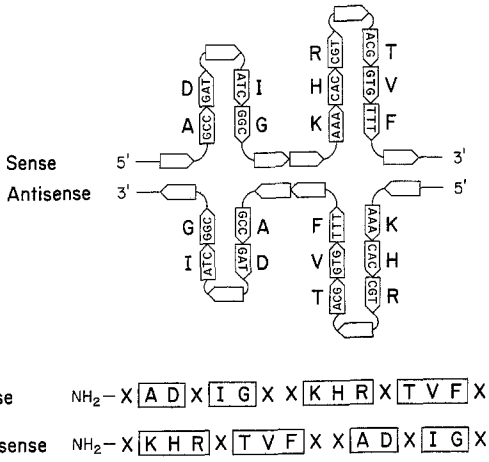

Fig. 3. If there are palindromes in exons then there is a potential for sense and antisense proteins to have motifs in common. **Upper:** Codons in duplex DNA, which has adopted cruciform configurations, are shown as *arrowheaded boxes,* as in Fig. 1. Codons which complement each other to form the stems of stem-loops are shown *within boxes.* The amino acids corresponding to each codon are shown in *bold lettering.* **Lower:** Amino acid sequence of the sense and antisense proteins derived from the above DNA sequence. Amino acids in boxes represent motifs present in both sequences.

Similar Codon Frequencies in Sense and Antisense Strands

Sense and antisense strands have similar codon frequencies in various organisms (Alff-Steinberger 1984, 1987). Furthermore, trinucleotides with the potential to encode hydrophilic amino acids (e.g., GAA; glutamate), are present at approximately the same frequencies as their complements, which often have the potential to encode hydrophobic amino acids (e.g., UUC; phenylalanine). However, studies of the frequencies of dinucleotides (Nussinov 1984), trinucleotides (Yomo and Ohno 1989; Ohno and Yomo 1991), and higher oligonucleotides (Pradhu 1993) show this to be a general characteristic of DNA sequences, *both* coding and in noncoding. One explanation for this, proposed by Nussinov (1982), was that DNA might contain numerous inverted repeats. These would confer on DNA the ability to form stem-loop, cruciform, structures. Indeed, evidence obtained by comparing the optimum folded structures of "windows" in natural sequences with the optimum folded structures of the same windows in which the order of bases have been randomized suggests that all DNA sequences have been under an evolutionary selection pressure to maximize the ability to form local stem-loop structures ("fold pressure"). This affects both protein-coding and noncoding regions (introns and intergenic regions; Forsdyke 1995b,c).

The automatic consequence of this for protein sequences has been pointed out by Ohno and Yomo (1991) and is illustrated in Fig. 3. Some motifs present in sense-strand-derived proteins will also be present on the antisense-strand-derived proteins. Thus, the properties of the latter may include some of those of the former. Since under appropriate conditions proteins can be induced to self-aggregate (Lauffer 1975; Forsdyke 1994, 1995a), there is the possibility of reaction of sense-strand-derived proteins with the corresponding "near-self" proteins encoded by antisense strands. Should this or any other property of the antisense protein prove advantageous, further evolutionary selection would be possible, which might result in further modification both of sense and antisense-derived proteins. However, for this opportunity to have arisen, two prior evolutionary forces are likely to have acted. First, GC-pressure should have modified the genome such that long NSFs were feasible. Second, fold pressure should have attained equilibrium with other selective forces acting on the protein. As discussed elsewhere (Forsdyke 1995b), a section of DNA encoding a protein can adapt to fold pressure either by (1) adopting an appropriate synonymous codon or by (2) allowing a conservative amino acid change, or by (3) encoding the protein in discrete units (exons) interrupted by sequences where stem-loop potential is less restrained (introns).

Acknowledgments. I thank D. Bray of Queen's University StatLab for advice, and I. Chaiken for assistance in tracing the early antisense literature. The work was supported by a grant from the Medical Research Council of Canada.

References

Alff-Steinberger C (1984) Evidence for a coding pattern on the non-coding strand of the *E. coli* genome. Nucleic Acids Res 12:2235–2241

Alff-Steinberger C (1987) Codon usage in *Homo sapiens:* evidence for a coding pattern on the non-coding strand and evolutionary implications of dinucleotide discrimination. J Theor Biol 124:89–95

Barrai I, Scapoli C, Gambari R, Brugnoli F (1991) Frequencies of codons in histones, tubulins and fibrinogen: bias due to interference between transcriptional signals and protein function. J Theor Biol 152:405–426

Biro J (1981a) The complementary coding of some proteins as the possible source of specificity in protein-protein interactions. Med Hypothesis 7:981–993

Biro J (1981b) Models of gene expression based on the sequential complementary coding of some pituitary proteins. Med Hypothesis 7:995–1007

Blalock JE (1990) Complementarity of peptides specified by "sense" and "antisense" strands of DNA. Trends Biotechnol 8:140–144

Blalock JE, Bost KL (1986) The binding of peptides that are specified by complementary RNAs. Biochem J 234:679–683

Blalock JE, Smith EM (1984) Hydropathic anti-complementarity of amino acids based on the genetic code. Biochem Biophys Res Commun 121:203–207

Bossi L, Roth JR (1980) The influence of codon context on genetic code translation. Nature 286:123–127

Clarke BL, Blalock JE (1991) Characteristics of peptides specified by antisense nucleic acids. In: Mol JNM, Van der Krol A (eds) Antisense nucleic acids and proteins. Marcel Dekker, Basel, pp 169–185

Eberle AN, Huber M (1991) Antisense peptides of ACTH and MSH: tools for receptor isolation? In: Mol JNM, Van der Krol A (eds) Antisense nucleic acids and proteins. Marcel Dekker, Basel, pp 187–203

Eigen M, Schuster P (1978) The hypercycle. A principle of natural organization. Naturwissenschaft 65:341–369

Filipski J (1990) Evolution of DNA sequence. Contributions of mutational bias and selection to the origin of chromosomal compartments. Adv Mutagen Res 2:1–54

Forsdyke DR (1994) Relationship of X chromosome dosage compensation to intracellular self/not-self discrimination: a resolution of Muller's paradox? J Theor Biol 167:7–12

Forsdyke DR (1995a) Entropy-driven protein self-aggregation as the basis for self/not-self discrimination in the crowded cytosol. J Biol Sys 3:273–287

Forsdyke DR (1995b) A stem-loop "kissing" model for the initiation of recombination and the origin of introns. Mol Biol Evol (in press)

Forsdyke DR (1995c) Different biological species "broadcast" their DNAs at different (G + C)% "wavelengths." Proc Can Fed Biol Socs 38:107

Forsdyke DR (1995d) Relative roles of primary sequence and (G + C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. J Mol Evol 41:573–581

Goldstein A, Brutlag DL (1989) Is there a relationship between DNA sequences encoding peptide ligands and their receptors? Proc Natl Acad Sci USA 86:42–45

Ikehara K, Okazawa E (1993) Unusually biased nucleotide sequences in sense strands of *Flavobacterium* sp. genes produce nonstop frames on the corresponding antisense strands. Nucleic Acids Res 21:2193–2199

Konecny J, Eckert M, Schoniger M, Hofacker GL (1993) Neutral adaptation of the genetic code to double-strand coding. J Mol Evol 36:407–416

Lauffer MA (1975) Entropy-driven processes in biology. Springer-Verlag, New York

Meckler LB (1969) Specific selective interactions between amino acid residues of peptide chains. Biofizika 14:581–584

Merino E, Balbas P, Puente JL, Bolivar F (1994) Antisense overlapping open reading frames in genes for bacteria to humans. Nucleic Acids Res 22:1903–1908

Moser M, Oesch B, Bueler H (1993) An antiprion protein? Nature 362:213–214

Nussinov R (1982) Some indications for inverse DNA duplication. J Theor Biol 95:783–793

Nussinov R (1984) Doublet frequencies in evolutionarily distinct groups. Nucleic Acids Res 12:1749–1763

Ohno S, Yomo T (1991) The grammatical rule for all DNA: junk and coding sequences. Electrophoresis 12:103–108

Pradhu VV (1993) Symmetry observations in long nucleotide sequences. Nucleic Acids Res 21:2797–2800

Ryan BF, Joiner BL (1994) Minitab handbook, 3rd ed. Wadsworth Publishing, Belmont, CA, pp 287–288

Shepherd JCW (1982) From primeval message to present day gene. Cold Spring Harb Symp Quant Biol 47:1099–1108

Tomizawa J (1984) Control of ColE1 plasmid replication: the process of binding of RNA I to the primer transcript. Cell 38:861–870

Tropsha A, Kizer JS, Chaiken IM (1992) Making sense of antisense: a review of experimental data and developing ideas on sense-antisense peptide recognition. J Mol Recognit 5:43–54

Wada K, Aota S, Tsuchiya R, Ishibashi F, Gojobori T, Ikemura T (1990) Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res 18:2367–2403

Yomo T, Ohno S (1989) Concordant evolution of coding and non-coding regions of DNA made possible by the universal rule of TA/CG deficiency-TG/CT excess. Proc Natl Acad Sci USA 86: 8452–8456

Yomo T, Urabe I, Okada H (1992) No stop codons in the antisense strands of genes for nylon oligomer degradation. Proc Natl Acad Sci USA 89:3780–3784

Yomo T, Urabe I (1994) A frame-specific symmetry of complementary strand of DNA suggests the existence of genes on the antisense strand. J Mol Evol 38:113–120.

Zull JE, Smith SK (1990) Is genetic code redundancy related to retention of structural information in both DNA strands? Trends Biochem Sci 15:257–261