Journal of Protein Chemistry, Vol. 15, No. 1, 1996

Do "Antisense Proteins" Exist?

Kuo-Chen Chou,[1] Chun-Ting Zhang,[1,2] and David W. Elrod[1]

Received September 22, 1995

A DNA double helix consists of two complementary strands antiparallel with each other. One of them is the sense chain, while the other is an antisense chain which does not directly involve the protein-encoding process. The reason that an antisense chain cannot encode for a protein is generally attributed to the lack of certain preconditions such as a promotor and some necessary sequence segments. Suppose it were provided with all these preconditions, could an antisense chain encode for an "antisense protein"? To answer this question, an analysis has been performed based on the existing database. Nine proteins have been found that have a 100% sequence match with the hypothetical antisense proteins derived from the known *Escherichia coli* antisense chains.

KEY WORDS: DNA; genetic coding; protein sequence match.

The term "antisense protein" is derived from a naturally logical way of thinking about DNA and genetic coding, as illustrated in Fig. 1a. Whether an antisense peptide can fold into a stable antisense protein, or whether the product of such a presumably logical possibility exists in nature, is a long-standing unsolved problem, which has been pondered by biochemists and molecular biologists since the elucidation of the genetic code three decades ago. The main difference between then and now, however, is that we currently have nearly 150,000 coding sequences in the public databases and these can be used directly to approach this problem.

To begin with, let us see if there are any unfavorable constraints imposed on an antisense DNA chain. According to the principle of complementary pairing, if any one of the three codons TTA, CTA, and TCA occurs in a sense-chain sequence, then one of the complementary codons TAA, TAG, and TGA must occur in its antisense chain (Fig. 1b); these codons are nothing but terminators. Usually the codons TTA, CTA, and TCA may occur many times in a coding sequence, implying that there are many terminator codons located in the corresponding antisense chain. A sequence separated by many terminators can of course not be used to encode for a protein. Furthermore, the last codon in the coding sequence of a sense chain is generally not CAT; thus, the first codon in the antisense chain will seldom be the codon ATG, a special initial codon required for a coding sequence. All of these constraints would further reduce the likelihood for an antisense chain to be able to encode for a protein. Nevertheless, it will provide a deeper insight to consider the following problem: If an antisense chain is not divided by many stop codons, and if it is provided with all the working conditions for coding performance, such as a promotor and an initial codon ATG, can it encode for a protein?

In a study using a graphic approach to analyze codon usage in *Escherichia coli* protein coding sequences (Zhang and Chou, 1994), we found that, of the 1562 *E. coli* protein coding sequences (Wada *et al.*, 1991), 110 coding sequences (see Table I for their names in GenBank) do not contain any of the codons TTA, CTA, and TCA. This means that the corresponding 110 antisense sequences will not be separated by stop codons.

[1] Computer-Aided Drug Discovery, Upjohn Laboratories, Pharmacia & Upjohn Inc., Kalamazoo, Michigan 49007-4940.
[2] On sabbatical leave from Department of Physics, Tianjin University, Tianjin, China.

0277-8033/96/0100-0059$09.50/0

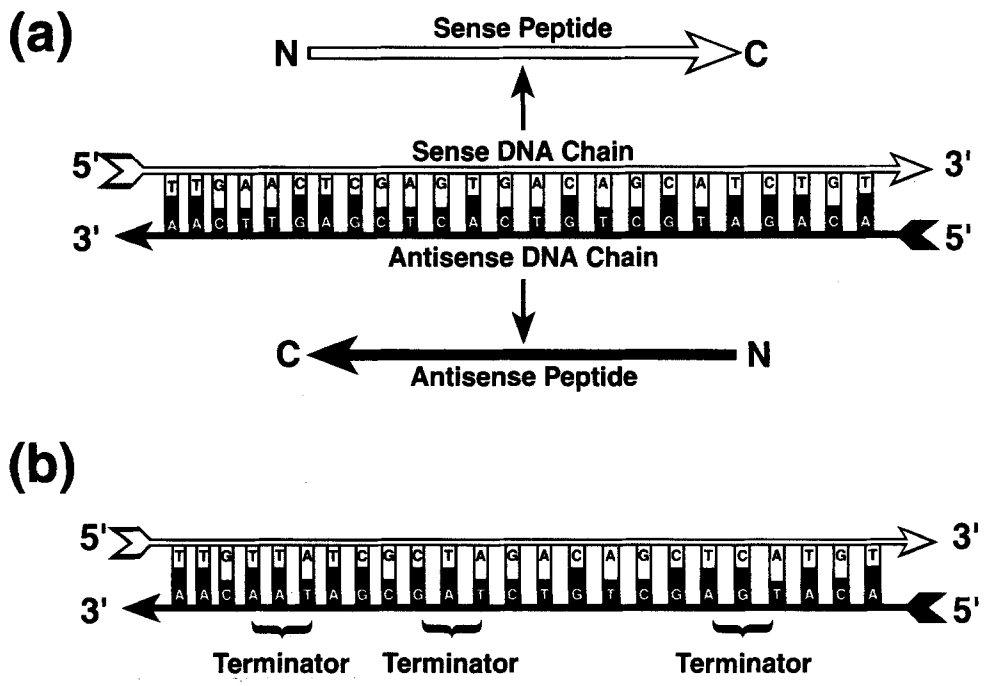

Fig. 1. A schematic illustration to show (a) the relationship between sense and antisense peptides encoded, respectively, by sense and antisense DNA chains, and (b) the occurrence of a terminator in the antisense chain when any one of the three codons TTA, CTA, and TCA occurs in the sense chain. A, C, G, and T represent the four bases adenine, cytosine, guanine, and thymine, respectively.

Table I. The 110 *E. coli* Protein Coding Sequences[a] Whose Antisense Sequences Are Not Separated by Stop Codons

EC2MIN#21	ECADK	ECAPH4	ECOAPPIB	ECOAPTADK#3
ECOAPTADK#6	ECOCSPAA	ECOCYD#2	ECOCYSPTS#2	ECOCYSXE#2
ECODCCDP	ECDCM#2	ECDEOCA1	ECODKSA#2	ECODNAAOP#1
ECFDAPGK#5	ECFDAPGK#6	ECOFIS	ECOFISA	ECOFTSQAB
ECOGLNB	ECOGLTA#4	ECGROES#1	ECGROESL#1	ECGRPE
ECOHATP#2	ECOHATP#8	ECOHIMA#2	ECOHIMB	ECHISOP#1
ECHISG	ECHU2	ECOHYA#4	IS903B#2	ECKCRAB#1
ECKCRAB#2	ECKCRAB#3	ECOLPP	ECLSPDAP#2	ECOXA2#1
ECPAL#1	ECPAP1#3	ECOPHEA	ECOPHEAB#1	ECOPHEAB#2
ECOPHEAC	ECOPHEIS1	ECOPHN#4	ECOPHN#12	ECOPHNAQ#17
ECOPHOS#4	ECOPHOS#5	ECOPHOSYS#1	ECOPHOSYS#3	ECPHOWTU#3
ECPHOWTU#4	ECPRSFMO#2	ECPRSFMO#4	ECOPTSH	ECOPTSHI#1
ECOPTSHI#3	ECR18EEX#2	ECRELB#1	ECORHOA	ECORNPA#1
ECRPA#2	ECSPC#3	ECSPC#5	ECSPC#7	ECSPC#8
ECSPC#9	ECRPOBC#2	ECRPOBC#4	ECORPMBG#1	ECORPMBG#2
ECORPMFA#2	ECRPOS10#2	ECRPOS10#11	ECRPOS10#5	ECRPOS10#7
ECRPOS10#8	ECRPSB#1	ECRPSB#2	ECRPSFRI#1	ECRPSFRI#3
ECRPSFRI#4	ECRPSI#2	ECRPSL	ECORPSRPO#1	ECRRNGTN#1
ECSDHACD#3	ECSOD	ECOSTR1	ECOSYNTGV	ECOTGP#1
ECOTGT#2	ECOTGTUFB	ECOTGY1	ECOTHR#1	ECTHRINF#3
ECTHRINF#4	ECTRMD#4	ECOTRPR#4	ECOTRX	ECOTRXA
ECOUNC#3	ECOUNC#9	ECUNC#11	ECUNC#6	ECVALS

[a] Names used in GenBank (Wada *et al.*, 1991; Benson *et al.*, 1994), where the definition for the symbol # is given.

Table II. The Nine Sequence Matches with 100% Identify Found Between the Antisense Proteins Derived from Table I and the Proteins from the PIR Database by Using the BLAST Program (Altschul *et al.*, 1990)

Antisense chains				Real proteins		
GenBank[a]				PIR[b]		
Accession number[c]	Locus[d]	Begin	End	Accession number	Begin	End
$\overline{\text{X13330}}$	$\overline{\text{ECDCM\#2}}$	808	1314	JS0263	171	339
$\overline{\text{M34333}}$	$\overline{\text{ECOCYSXE\#2}}$	527	919	XYECSA	103	233
$\overline{\text{X04830}}$	$\overline{\text{ECPRSFMO\#2}}$	95	379	S07319	13	107
$\overline{\text{X04830}}$	$\overline{\text{ECPRSFMO\#4}}$	1298	1636	S10916	241	353
				JH0126	241	353
$\overline{\text{X06046}}$	$\overline{\text{ECOXA2\#1}}$	361	765	C26839	180	314
				S04809	180	314
				A42646	180	314
				S32184	180	314

[a] GenBank database (Wada *et al.*, 1991; Benson *et al.*, 1994).
[b] The PIR-international protein sequence database (George *et al.*, 1994).
[c] An accession number with an overbar represents the corresponding antisense chain.
[d] A locus with an overbar represents the corresponding antisense chain.

Based on the 110 coding sequences in Table I, the GCG REVERSE program (Devereux, 1994) was used to generate the corresponding antisense sequences, followed by using the GCG TRANSLATE program (Devereux, 1994) to generate the corresponding antisense protein sequences.

Finally, the BLAST program (Altschul *et al.*, 1990) was used to search the 75,511 protein sequences in PIR database (George *et al.*, 1994) for homologs with the 110 antisense protein sequences. It has been found that there are 21 or 24 sequence matches with identity or with homology greater than 70%, respectively. Of these, 9 sequence matches have 100% identity, as shown in Table II. This indicates that some antisense proteins may really exist in nature.

The nine proteins which have the identical peptide sequences as those coded by the antisense DNA chains are (according to the order in Table II) a site-specific DNA-methyltransferase (cytosine-specific), serine-O-acetyltransferase, hypothetical protein, mobA protein, mobA protein precursor, ORF3 protein, integrase-like protein E2, integrase-*Pseudomonas aeruginosa,* and integrase-*Klebsiella pneumoniae.* Interestingly, of the above nine proteins, the first seven are from the *Escherichia coli* organism, and the last two from other bacterial organisms.

It is anticipated that this finding of antisense proteins will stimulate the investigation into their functions, as well as have a great impact for understanding the antisense chain at a deeper level.

A codon usage analysis for the antisense chains by means of the graphical approach (Zhang and Zhang, 1991; Chou and Zhang, 1992) developed recently is under way.

ACKNOWLEDGMENTS

We would like to thank Prof. T. Ikemura and Dr. Y. A. Nakamura for supplying the relevant data for codon usages.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool, *J. Mol. Biol.* **215,** 403–410.

Bairoch, A., and Boeckmann, B. (1994). The SWISS-PROT protein sequence data bank, *Nucleic Acids Res.* **22,** 3575–3580.

Benson, D. A., Boguski, M., Lipman, D. J., and Ostell, J. (1994). GenBank, *Nucleic Acids Res.* **22,** 3441–3444.

Chou, K. C., and Zhang, C. T. (1992). Diagrammatization of codon usage in 339 human immunodeficiency virus proteins and its biological implications, *AIDS Research and Human Retroviruses* **8,** 1967–1976.

Devereux, J. (1994). The Wisconsin Sequence Analysis Package, Version 8.0, Genetics Computer Group, Madison, Wisconsin.

George, D. G., Baker, W. C., Mewes, H. W., Pfeiffer, F., and Tsugita, A. (1994). The PIR-international protein sequence database, *Nucleic Acids Res.* **22,** 3569–3573.

Wada K., Wada Y., Doi, H., Ishibashi, F., Gojobori, T., and Ikemura, T. (1991). Codon usage tabulated from the GenBank genetic sequence data, *Nucleic Acids Res.* **19**(Suppl), r1981–r1986.

Zhang, C. T., and Chou, K. C. (1994). A graphic approach to analyzing codon usage in 1562 *E. coli* protein coding sequences, *J. Mol. Biol.* **238,** 1–8.

Zhang, C. T., and Zhang, R. (1991). Analysis of distribution of bases in the coding sequences by a diagrammatic technique, *Nucleic Acid Res.* **19,** 6313–6317.