Evidence for a coding pattern on the non-coding strand of the *E. coli* genome

Steinberger, Cynthia

How to cite

STEINBERGER, Cynthia. Evidence for a coding pattern on the non-coding strand of the *E. coli* genome. In: Nucleic Acids Research, 1984, vol. 12, n° 5, p. 2235–2241. doi: 10.1093/nar/12.5.2235

This publication URL: https://archive-ouverte.unige.ch/unige:128121
Publication DOI: 10.1093/nar/12.5.2235

Volume 12 Number 5 1984 Nucleic Acids Research

Evidence for a coding pattern on the non-coding strand of the *E. coli* genome

C.Alff-Steinberger

Department of Molecular Biology, University of Geneva, 30 Quai Ernest-Ansermet, 1211 Geneva 4, Switzerland

Received 24 October 1983; Accepted 24 January 1984

ABSTRACT

Analysis of codon usage frequency for the combined coding sequences of 52 E. coli genes, taken from the European Molecular Biology Laboratory Nucleotide Sequence Data Library, Release 2, shows that there is a significant positive correlation between the frequency with which a given codon appears on the coding strand and the frequency with which it appears, in phase, on the non-coding strand.

INTRODUCTION

For the purpose of studying the codon distribution in the Escherichia coli genome, a tabulation of codon frequency was made using the sequence data of 52 complete E. coli genes. These compiled data are shown in Table 1. The observation which is the subject of this paper is that there is a significant positive correlation between the frequency with which a given codon appears on the coding strand and the frequency with which the same codon, in phase and with the correct polarity, appears on the non-coding strand.

MATERIAL AND METHODS

Sequence Selection. The sequence data were obtained from the Nucleotide Sequence Data Library, Version 2, of the European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. All complete E. coli coding regions for known proteins were selected, and in the cases where more than one sequence was given for a particular gene, one of the sequences was arbitrarily chosen to be included in the tabulation. Coding sequences were rejected if there were gaps in the sequence, or if the gene product was

not identified, or if the termination codon was missing, or if the length of the sequence given was not a multiple of three. In this way, 32 data library entries containing 52 complete genes were selected. The EMBL identifying codes for the entries, together with the relevant genes in parentheses, are :

EC5388 (TRIMETHOPRIM RESISTANT DIHYDROFOLATE REDUCTASE, PLASMID ASSOCIATED) (1),
ECALAS (ALANYL-TRNA SYNTHETASE) (2),
ECARAC (ARAC) (3), (4),
ECASNA (ASPARAGINE SYNTHETASE) (5),
ECATPX (FIRST FIVE GENES OF ATP SYNTHETASE) (6),
ECATPY (GAMMA AND BETA FROM ATP (UNC) OPERON) (7),
ECCOL1 (COLICIN E1 IMMUNITY) (8),
ECFOLX (FOL GENE FOR DIHYDROFOLATE REDUCTASE) (9),
ECHIS1 (REGULATORY PEPTIDE OF HISG GENE) (10),
ECILVX (ILVL AND ILVG) (11),
ECLACI (LAC REPRESSOR) (12),
ECLACY (LACTOSE PERMEASE) (13),
ECLEXX (LEXA REPRESSOR) (14),
ECLPPX (OUTER MEMBRANE LIPOPROTEIN) (15),
ECNDHX (NADH DEHYDROGENASE) (16),
ECOMPA (OUTER MEMBRANE PROTEIN II) (17),
ECPAP1 (ATPASE GENES PAPD/UNCB, PAPH/UNCE, PAPF/UNCF) (18),
ECPAP2 (ATPASE PAPB, PAPG) (19),
ECPAP3 (ATPASE PAPA, PAPC) (20),
ECPHEA (ATTENUATOR PEPTIDE OF PHENYLALANINE OPERON) (21),
ECPURF (PURF) (22),
ECRECA (RECA) (23), (24),
ECRPOBC (RNA POLYMERASE RPLK, RPLA, RPLJ, RPLL, RPOB, RPOC) (25-29),
ECRPSA (RIBOSOMAL PROTEIN S1) (30),
ECRPSB (RIBOSOMAL PROTEIN S2 AND ELONGATION FACTOR TS) (31),
ECRPSL (RIBOSOMAL PROTEIN RPSJ (S10)) (32),
ECRPST (RIBOSOMAL PROTEIN S20) (33),
ECSTR1 (RIBOSOMAL PROTEIN S12) (34),
ECTHR1 (ATTENUATOR PEPTIDE OF THRA GENE (35),
ECTHRA (THRA GENE OF THREONINE OPERON) (36),
ECTRPR (TRP APOREPRESSOR (37),
ECTRPX (TRYPTOPHAN OPERON GENES TRPE, TRPD, TRPG, TRPB, TRPA) (38).

Analysis. The tabulation and statistical analysis of the data were done using the CDC Cyber computer of the Cantonal Hospital in Geneva, using programs written by the author. The EMBL Data Library convention is to give the sequence of the non-coding DNA strand, which is homologous to the mRNA transcribed from the coding strand. The frequencies shown in Table 1 are those of the mRNA, from the initiation codon up to and including the termination codon.

Table 1. Codon Distribution from 52 complete E. coli Genes.

UUU	214	PHE	UCU	210	SER	UAU	164	TYR	UGU	62	CYS
UUC	367	PHE	UCC	239	SER	UAC	248	TYR	UGC	83	CYS
UUA	110	LEU	UCA	63	SER	UAA	39	TERM	UGA	9	TERM
UUG	143	LEU	UCG	103	SER	UAG	4	TERM	UGG	113	TRP
CUU	124	LEU	CCU	74	PRO	CAU	121	HIS	CGU	509	ARG
CUC	124	LEU	CCC	45	PRO	CAC	203	HIS	CGC	323	ARG
CUA	30	LEU	CCA	103	PRO	CAA	161	GLUN	CGA	29	ARG
CUG	1114	LEU	CCG	411	PRO	CAG	518	GLUN	CGG	26	ARG
AUU	354	ILEU	ACU	186	THR	AAU	162	ASPN	AGU	53	SER
AUC	604	ILEU	ACC	385	THR	AAC	452	ASPN	AGC	229	SER
AUA	31	ILEU	ACA	52	THR	AAA	672	LYS	AGA	9	ARG
AUG	452	MET	ACG	125	THR	AAG	212	LYS	AGG	6	ARG
GUU	477	VAL	GCU	429	ALA	GAU	407	ASP	GGU	619	GLY
GUC	192	VAL	GCC	358	ALA	GAC	485	ASP	GGC	509	GLY
GUA	292	VAL	GCA	370	ALA	GAA	803	GLU	GGA	61	GLY
GUG	359	VAL	GCG	518	ALA	GAG	302	GLU	GGG	100	GLY

Total number of codons : 16351

RESULTS

The range in the frequency of appearance of individual codons in the distribution shown in Table 1 is wide. It was observed that the codons complementary to those codons appearing frequently also tended to appear frequently, and that the codons complementary to those codons appearing infrequently also tended to be infrequently used. The complementary codon is that which would be obtained if the triplet on the non-coding strand, opposite to the coding strand triplet, were to be transcribed, in the appropriate direction. For example, the codon CAG, for glutamine, is complementary to the codon CUG for leucine, and both are frequently used codons. The same is true for the pair of complementary codons AUC, for isoleucine, and GAU, for aspartic acid; both are frequently used. The codon AGG for arginine is infrequently used and is complementary to the codon CCU for proline, also infrequently used. This relation can be presented graphically for the set of 64 codons. In figure 1, the frequency with which a given codon appears is plotted against the frequency with which the complementary codon appears.

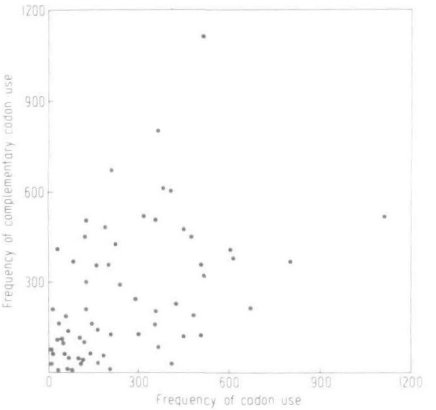

Figure 1. Scatterplot of the frequency of each codon versus the frequency of the complementary codon, using the data from Table 1.

From Figure 1, it is clear that although there is a substantial scattering of points, the correlation between the two variables is positive. There are few points in Figure 1 corresponding to frequently used codons whose complements are infrequently used. To quantify this relation, the intraclass correlation coefficient ρ (39), was calculated for the 32 pairs of numbers formed by pairing the frequency of codon use and the frequency of use of the complementary codon. In this calculation, the order of the numbers within a pair is irrelevant. A value of $\rho = .43$ was obtained. If the frequency with which a given codon is used were unrelated to the frequency of the codon complementary to it, an intraclass correlation coefficient of zero would be expected. The standard deviation of the value of ρ obtained was estimated using a bootstrap calculation (40) to be .10; thus the value of $\rho = .43 \pm .10$ is more than two standard deviations from zero and is significantly positive.

DISCUSSIONS

The conclusion reached is that the non-coding strand of the E. coli genome tends to ressemble the coding strand in that the triplets found, in phase with the reading frame, have frequencies related to their frequency of appearance on the coding

strand. This does not imply that the non-coding strand is actually transcribed. If there were reading frames on the non-coding strand, the gene product would not be expected to bear any similarity to the gene product of the coding strand, since, although it is possible for the first and third bases of the codon and its complement to be the same, e.g., CUG is complementary to CAG, the second base, via which the most information about the amino acid is transmitted (41), necessarily changes.

The explanation of the observed correlation is not obvious. Optimization of the codon-anticodon interaction energy (42) may influence codon usage in such a way that energetically favored triplet pairs tend to be frequently used, and energetically disfavored pairs infrequently used. Alternatively, the correlation might be the result of the evolution of codon usage to optimize the efficiency of double stranded coding in the past. Another explanation, suggested by H.M. Krisch, is that the pattern observed may be the result of gene rearrangements involving the frequent incorporation of inverted segments.

Preliminary examination of some coding regions of Homo sapiens also reveal a significant positive correlation between codon appearance on the coding and the non-coding strands, although the codon distribution itself is not consistent with that of E. coli.

ACKNOWLEDGEMENTS

I am grateful to L. Caro, J.E. Dowd, R. Epstein, and H.M. Krisch for discussions and suggestions relating to the biological and statistical aspects of this work, the Cantonal Hospital of Geneva for offering computer facilities, and A. Chappuis and A. Rieben for helping me to use the Cyber. This work was supported by grant No. 3.169-0.81 to L. Caro from the Swiss National Science Foundation.

REFERENCES

1. Zolg, J.W., Haenggi, U.J. (1981) Nucleic Acids Res. 9, 697-710.
2. Putney, S.D., Royal, N.J., Neuman de Vegvar H., Herlihy, W.C.

Biemann, K., Schimmel, P. (1981) Science 213, 1497-1501.
3. Miyada, C.G., Horwitz, A.H., Cass, L.G., Timko, J., Wilcox, G. (1980) Nucleic Acids Res. 8, 5267-5274.
4. Wallace, R.G., Lee, N., Fowler, A.V. (1980) Gene 12, 179-190.
5. Nakamura, M., Yamada, M., Hirota, Y., Sugimoto, K., Oka, A., Takanami, M. (1981) Nucleic Acids Res. 9, 4669-4675.
6. Gay, N.J., Walker, J.E. (1981) Nucleic Acids Res. 9, 3919-3926.
7. Saraste, M., Gay, N.J., Eberle, A., Runswick, M.J., Walker, J.E. (1981) Nucleic Acids Res. 9, 5287-5296.
8. Oka, A., Nomura, N., Morita, M., Sugisaki, H. Sugimoto, K., Takanami, M. (1979) Molec. Gen. Genet. 172, 151-159.
9. Smith, D.R., Calvo, J.M. (1980) Nucleic Acids Res. 8, 2255-2274.
10. Verde, P., Frunzio, R., Di Nocera, P.P., Blasi, F., Bruni, C.B. (1981) Nucleic Acids Res. 9, 2075-2086.
11. Lawther, R.P., Calhoun, D.H., Adams, C.W., Hauser, C.A., Gray, J., Hatfield, J.W. (1981) Proc. Natl. Acad. Sci. 78, 922-925.
12. Farabaugh, P.J., (1978) Nature 274, 765-769.
13. Buechel, D.E., Gronenborn, B., Mueller-Hill, B. (1980) Nature 283, 541-545.
14. Markhan, B.E., Little, J.W., Mount, D.W. (1981) Nucleic Acids Res. 9, 4149-4161.
15. Nakamura, K., Inouye, M. (1979) Cell 18, 1109-1117.
16. Young, J.G., Rogers, B.L., Campbell, H.D., Jaworowski, A., Shaw, D.C. (1981) Eur. J. Biochem. 116, 165-170.
17. Beck, E., Bremer, E. (1980) Nucleic Acids Res. 8, 3011-3027.
18. Kanazawa, H., Mabuchi, K., Kayano, T., Noumi, T., Sekiya, T., Futai, M. (1981) Biochem. Biophys. Res. Comm. 103, 613-620.
19. Kanazawa, H., Kayano, T., Kiyasu, T., Futai, M. (1982) Biochem. Biophys. Res. Comm. 105, 1257-1264.
20. Kanazawa, H., Kayano, T., Mabuchi, K., Futai, M. (1981) Biochem. Biophys. Res. Comm. 103, 604-612.
21. Zurawski, G., Brown, K., Killingly, D., Yanofsky, C. (1978) Proc. Natl. Acad. Sci. 75, 4271-4275.
22. Tso, J.Y., Zalkin, H., Van Cleemput, M., Yanofsky, C., Smith, J.M. (1982) J. Biol. Chem. 257, 3525-3531.
23. Sancar, A., Stachelek, C., Konigsberg, W., Rupp, W.D. (1980) Proc. Natl. Acad. Sci. 77, 2611-2615.
24. Horri, T., Ogawa, T., Ogawa, H. (1980) Proc. Natl. Acad. Sci. 77, 313-317.
25. Post, L.E., Strycharz, G.D., Nomuda, M., Lewis, H., Dennis, P.P. (1979) Proc. Natl. Acad. Sci. 76, 1697-1701.
26. Ovchinnikov, Y.A., Monastyrskaya, G.S., Gubanov, V.V., Guryev, S.A., Chertov, O.Y., Modyanov, N.N., Grinkevich, V.A. Makarova, I.A., Marchenko, T.V., Polovnikova, I.N., Lipkin, V.M., Sverdlov, E.D. (1980), Dokl. Akad. Nauk, SSSR 253, 994-998.
27. Delcuve, G., Downing, W., Lewis, H., Dennis, P.P. (1980), Gene 11, 367-373.
28. Ovchinnikov, Y.A., Monastyrskaya, G.S., Gubanov, V.V., Guryev, S.O., Modyanov, N.N., Grinkevich, V.A., Makarova, I.A., Marchenko, T.V., Lipkin, V.M., Sverdlov, E.D.

(1981) Eur. J. Biochem. 116, 621-629.
29. Ovchinnikov, Y.A., Monastyrskaya, G.S., Gubanov, V.V., Guryev, S.O., Salomatina, I.S., Shuvaeva, T.M., Lipkin, V.M., Sverdlov, E.D. (1981) Dokl. Akad. Nauk., SSSR 261, 763-768.
30. Schnier, J., Kimura, M., Foulaki, K., Subramanian, A.R., Isono, K., Wittmann-Liebold, B. (1982) Proc. Natl. Acad. Sci. 79, 1008-1011.
31. An, G., Bendiak, D.S., Mamelak, L.A., Friesen, J.D. (1981) Nucleic Acids Res. 9, 4163-4172.
32. Olins, P.O., Nomura, M. (1981) Cell 26, 205-211.
33. Mackie, G.A. (1981) J. Biol. Chem. 256, 8177-8122.
34. Post, L.E., Nomura, M. (1980) J. Biol. Chem. 255, 4660-4666.
35. Gardner, J.F. (1979) Proc. Natl. Acad. Sci. 76, 1706-1710.
36. Katinka, M., Cossart, P., Subilli, L., Saint-Girons, I. Chalvignac, M.A., Le Bras, G., Cohen, G.N., Yaniv, M. (1980) Proc. Natl. Acad. Sci. 77, 5730-5733.
37. Gunsalus, R.P., Yanofsky, C. (1980) Proc. Natl. Acad. Sci. 77, 7117-7121.
38. Yanofsky, H., Platt, T., Crawford, I.P., Nichols, B.P., Christie, G.E., Horowitz, H., Van Cleemput, M., Wu, A.M. (1981) Nucleic Acids Res. 9, 6647-6668.
39. Kempthorne, O. and Folks, L. (1971) Probability, Statistics and Data Analysis, 1st edn., p. 454, p. 467, The Iowa State University Press, Ames, Iowa.
40. Efron, B. (1979) SIAM Review 21, 460-480.
41. Alff-Steinberger, C. (1969) Proc. Natl. Acad. Sci. 64, 584-591.
42. Grosjean, H. and Fiers, W. (1982) Gene 18, 199-209.