Volume 13 Number 1 1985 Nucleic Acids Research

A method to locate protein coding sequences in DNA of prokaryotic systems

A.S.Kolaskar and B.V.B.Reddy

Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad 500 007, India

Received 24 October 1984; Accepted 26 November 1984

ABSTRACT

cDNA sequence data from E.coli phages, for which complete genome sequences are known, have been analysed, From this analysis thirteen triplets have been identified as markers to distinguish protein-coding frames from fortuitous open reading frames. The region of -18 to + 18 nucleotides around ATG/GTG, has been analysed and used to identify initiator codons from internal ATG/GTG. With the aid of criteria defined above a method has been developed to locate protein coding sequences by a combination of 'gene search by signal' and 'gene search by content' approaches. Application of this method to prokaryotic systems including those which were not part of our data base indicates that it is quite accurate and general in nature.

INTRODUCTION

In prokaryotes the presence of overlapping genes, as well as the fact that the first ATG/GTG triplet from 5'-end of cDNA does not always act as an initiator codon, have attracted attention of several workers to develop methods which could locate protein coding regions in DNA (1-12). Success of these methods is limited mainly because most of these attempts have been aimed at finding signal sequences which can be recognised by the ribosomal machinery, and thus, can be categorised as 'gene search by signal' (6-11). On the other hand, there have been a few attempts where the protein coding cDNA sequences were analysed to find certain features which might be present only in the coding frames but not in noncoding frames (1-5). These features were searched only at the mononucleotide level (1-4), the exception being the recent study by Staden (5,12). In all these studies, cDNA sequence data from prokaryotic as well as from eukaryotic systems have been used to find these characteristic features, even though it is known that the machinery by which proteins are synthesised in prokaryotic systems is quite different from that in eukaryotic systems. Therefore, in our study we have used cDNA sequence data from only prokaryotic systems. Further, in our earlier study we had observed that nonlinear interactions among nucleotides are substantial in magnitude up to triplet level and nonlinearity of these inter-

actions decreases sharply at tetranucleotide level (14). Other studies (13) have also pointed out that DNA sequences are Markov chains of order two. This suggests that the characteristic features of coding sequences and of coding frame may become identifiable only at triplet level or at higher order sequences. Therefore, data at triplet level has been used in our study to develop an algorithm which can locate protein coding regions.

In order to locate initiator codon in coding sequences, 36 nucleotides around every ATG/GTG have been analysed, and, a comparison has been made between the distribution of nucleotides around initiator ATG/GTG and those which are inside the coding sequences. The feature table developed from this analysis has been used to locate the initiator ATG/GTG. We have combined the 'gene search by signal' and 'gene search by content' approach to develop a simple objective method to locate coding sequences with a precise initiator in prokaryotic DNA. The computer program is interactive in nature and is now developed for PDP 11/23 computer but being in BASIC, can be easily used on any system.

METHOD

a) Analysis of Coding Sequence of cDNA

cDNA sequence data of ϕX174, MS2, fd, f1, G4, M13, λ and T7 phages of E.coli have been analysed. Total 174 genes are reported from experimental studies in these systems and they contain altogether 38,441 codons. We have used sequence data from these genes in our analysis. These were the only systems chosen because, it is only for them that the complete genome sequences are known and any bias which might arise due to the usage of certain codons preferentially in specific types of genes, could be avoided.

Potential values are calculated for each of the 64 types of triplets in the three possible frames $X_1X_2X_3$, $X_2X_3X'_1$ and $X_3X'_1X'_2$ where subscript represents the position of nucleotide X in the codon from 5'-end. The superscript is used to show that nucleotides are from adjacent codon. A simple relation is used to calculate the potential values of triplets in each of the three frames.

$$P_1(X_1X_2X_3) = \frac{f(X_1X_2X_3)}{f(X_1X_2) \cdot f(X_3)} \qquad \ldots . 1a.$$

$$P_2(X_2X_3X'_1) = \frac{f(X_2X_3X'_1)}{f(X_2X_3) \cdot f(X_1)} \qquad \ldots 1b.$$

$$P_3(X_3\ X_1'\ X_2') = \frac{f(X_3\ X_1'\ X_2')}{f(X_3\ X_1') \cdot f(X_2)} \quad \ldots 1c.$$

Note that in the denominator we have used positional dinucleotide and mononucleotide frequencies. Potential values are good indicators of nonlinear interactions between di- and mononucleotides. We have observed that more than 50% types of triplets have potential values outside the range of 0.80 — 1.25 indicating that for these triplets the interactions are highly nonlinear. We have picked up those triplets which have $P_1 \geq 1.5$ or $P_1 \leq 0.64$ but P_2 and P_3 values near unity. There are thirteen triplets of such kind and the potential values for each of them are given in Table I. Though we have analysed only coding sequences, we expect that the patterns in noncoding regions will be similar to those observed in noncoding frames of coding regions. Therefore, we have used these triplets as markers, by assigning weight values, to distinguish coding sequences from noncoding DNA sequences. These weight values are assigned using the following simple rule. In a given DNA sequence:

a) $W(X_1X_2X_3) = 1$ if $P_1(X_1X_2X_3) \geq 1.25$ where $X_1X_2X_3$ = AAA, CTG, CAG, TAT, TAC or TGG and

b) $W(X_1X_2X_3) = 1$ if $P_1(X_1X_2X_3) \leq 0.80$ where $X_1X_2X_3$ = CCC, CGG, CAT, AGG, TCG, GGG or AAT, otherwise $W(X_1X_2X_3) = 0$ is assigned.

b) Signal search analysis for initiator ATG/GTG:

From earlier studies (6, 9, 12), it is clear that the signal for translation is located not only in preinitiator region but also in coding region. Though, preinitiator region up to -75 to -80 nucleotides seems to be important, we found that -18 to +18 region around initiator is sufficient to statistically distinguish initiator triplet from internal AGT/GTG. Therefore, frequencies of occurrence of the four nucleotides T, C, A and G in their corresponding positions from -18 to +18 for 124 initiator ATG/GTG are calculated. Similarly, we have also calculated the frequencies of occurrences of nucleotides around internal ATG/GTG (non-initiator region). The potential values $P(X_i)$ are calculated using the following relation:

$$P(X_i) = \frac{\text{Frequency of } X_i \text{ in the initiator region}}{\text{Frequency of } X_i \text{ in the noninitiator region}} \quad \ldots\ldots 2.$$

TABLE I

Triplets which are used in the prediction algorithm with their potential values in three frames.

Triplet (XXX)	$P_1(X_1X_2X_3)$	$P_2(X_2X_3X'_1)$	$P_3(X_3X'_1X'_2)$
AAA	1.55	0.97	1.12
CTG	1.72	1.18	1.15
CAG	1.77	0.90	0.90
TAT	1.77	1.12	1.03
TAC	1.78	1.08	0.87
TGG	2.26	0.92	1.19
CCC	0.43	0.79	0.81
CGG	0.49	0.91	1.45
CAT	0.49	1.14	1.04
AGG	0.50	0.97	1.24
TCG	0.54	0.88	1.23
GGG	0.56	0.84	0.87
AAT	0.63	1.00	1.07

where X is the nucleotide T, C, A or G in the position 'i', and, 'i' varies from -18 to +18,. i = 0 for all three nucleotides of triplet ATG/GTG. Potential values thus calculated for nucleotides around ATG/GTG, are given in Table II.

The $P(X_i)$ values indicate the preference of X in the ith position around initiator codon to that in the same position around internal ATG/GTG. We have developed a weight matrix using these potential values with the aid of following simple rules:

Rules: $P(X_i)$ values are divided into three regions.

a) $P(X_i) \leq 0.64$, (b) $0.64 < P(X_i) < 1.50$, (c) $P(X_i) \geq 1.5$

Case 1: At least one of the nucleotides in position 'i' has $P(X_i) \leq 0.64$ then

a) $W(X_i) = 0$ if $P(X_i) \leq 0.64$

b) $W(X_i) = 1$ if $0.64 < P(X_i) < 1.5$

c) $W(X_i) = 2$ if $P(X_i) \geq 1.5$

Case 2: None of the nucleotides in position 'i' has $P(X_i) \leq 0.64$

a) $W(X_i) = 0$ if $P(X_i) < 1.5$

b) $W(X_i) = 1$ if $P(X_i) \geq 1.5$

Weight values thus assigned are given in Table II. These weight values are used in our algorithm, discussed below, to locate initiator codons.

TABLE II

Potential values (PX_i) and assigned weight values ($W(X_i)$) for nucleotides around ATG/GTG triplets. Note all three nucleotides of triplet ATG/GTG have been assigned zero position.

Position (i)	$P(T_i)$	$P(C_i)$	$P(A_i)$	$P(G_i)$	$W(T_i)$	$W(C_i)$	$W(A_i)$	$W(G_i)$
-18	1.55	1.04	0.94	0.62	2	1	1	0
-17	1.27	0.87	0.75	1.28	0	0	0	0
-16	0.62	0.90	2.26	0.85	0	1	2	1
-15	1.82	0.50	1.39	0.64	2	0	1	0
-14	1.22	0.53	1.08	1.34	1	0	1	1
-13	0.40	0.88	2.04	1.77	0	1	2	2
-12	0.47	0.42	1.43	1.43	0	0	1	1
-11	0.12	0.26	0.93	5.24	0	0	1	2
-10	0.10	0.39	1.23	4.11	0	0	1	2
-9	0.48	0.35	1.20	1.50	0	0	1	2
-8	1.01	0.15	1.03	2.46	1	0	1	2
-7	0.56	0.68	1.57	1.60	0	1	2	2
-6	1.15	0.31	1.62	0.98	1	0	2	1
-5	1.14	0.53	1.12	1.14	1	0	1	1
-4	0.76	1.08	1.69	0.59	1	1	2	0
-3	0.65	1.34	1.70	0.52	1	1	2	0
-2	1.09	0.48	1.31	1.35	1	0	1	1
-1	1.08	1.15	0.96	0.73	0	0	0	0
1	1.28	0.33	1.50	1.10	1	0	1	1
2	1.01	0.81	1.00	1.53	0	0	0	1
3	0.81	0.48	1.89	1.25	1	0	2	1
4	0.53	0.86	1.90	0.67	0	1	2	1
5	0.78	0.60	1.58	0.84	1	0	2	1
6	0.68	1.01	1.41	1.33	0	0	0	0
7	1.21	1.27	0.75	0.88	0	0	0	0
8	1.53	0.60	0.80	1.27	2	0	1	1
9	0.78	0.86	1.73	0.74	0	0	1	0
10	0.75	0.65	1.73	1.01	0	0	1	0
11	1.18	0.56	1.76	0.67	1	0	2	1
12	0.84	0.76	1.29	1.73	0	0	0	1
13	0.92	0.55	1.07	1.35	1	0	1	1
14	1.24	0.73	1.09	0.93	0	0	0	0
15	1.01	0.84	1.47	0.61	1	1	1	0
16	1.23	1.00	1.40	0.51	1	1	1	0
17	1.32	1.05	0.70	1.03	0	0	0	0
18	0.91	1.13	1.12	0.87	0	0	0	0

c) **Algorithm to locate protein coding sequences in DNA:**

i) The sequence of given prokaryotic DNA is read from 5'-end and a triplet ATG/GTG is located. Then in the same frame a terminator TAA/TAG/TGA is located. If the DNA fragment from ATG/GTG to its corresponding terminator has more than 60 nucleotides

then the search to find out whether this fragment is coding or not has been continued.

ii) For the above fragment potential values are calculated for the thirteen triplets mentioned in Table I, in the frame of ATG/GTG. Weight values are then assigned using these potential values. If $\Sigma W(X_1X_2X_3) \geq 6$ only then is the fragment termed as potential coding sequence. Otherwise it is rejected as noncoding fragment.

iii) The region around ATG/GTG is now scanned for potential coding fragments and weight values are assigned as per Table II to each nucleotide in this region. These weight values are added and if $I = \sum_{i=-18}^{+18} W(X_i) \geq 26$ then that ATG/GTG is considered as initiator and the fragment is termed as protein coding sequence in the DNA.

iv) Protein coding sequences obtained by the above method were then grouped as per their terminators. If more than one ATG/GTG for same terminator are found as initiators then the one farthest from terminator is assigned as initiators and others as additional potential initiators.

RESULTS AND DISCUSSION

The algorithm discussed above has been applied to DNA sequences of G4, ϕX174, fd, f1, MS2, M13 phages and results are presented in Table III. It can be seen from Table III that we are able to locate all coding sequences except one each from G4 and ϕX174. Predicted protein coding DNA sequences which agree with experimental data are underlined in Table III. Thus, the success rate of locating protein coding sequences seems to be around 96% in the data base analysed. Secondly, we have also predicted few additional coding sequences which have not been noticed yet by experimental approach; confirmation of their presence requires further study. However, the number of such predicted new sequences is not large, and thus suggests that they might be real coding sequences. It may be noted that additional genes predicted by our algorithm are mostly in the range of 61-100 nucleotides in length and thus code for polypeptides of length 22-30 amino acids. We have also mentioned in parenthesis the potential initiators. For example in phage G4 the sequence from 59-1720 is a protein coding sequence. Our method predicts that the translation can be initiated from nucleotides 59, 698, 1337 and 1412. From the experimental studies it is known that ATG at 698 is also an initiator in addition to one at 59. Thus, these results point out that the method developed above is quite accurate and has the capability of correct prediction.

The systems on which we checked the validity of our method and the results

TABLE III

Protein coding sequences predicted using the algorithm. Sequences underlined are experimentally observed and are reported earlier. In parenthasis additional potential initiators are also given. Note that except two, all experimentally observed sequences are predicted.

Systems	Predicted sequences and the additional initiators given in the bracket	Observed coding sequences missed in the prediction
G4 Phage	59-1720 (698, 1337, 1397, 1412), 670-759, 1638-1805 (1713) 1720-1971 (1885), 1976-2431, 2154-2441, 2477-2551, 2600-3880, 4020-4550 4615-5574, (5017, 5050), 5486-5551 (5589)	1276-1635
oX174 Phage	51-218, 133-390, 390-845, 848-961, 1001-2281, 1266-1385 (1272) 1449-1550 (1464), 1653-1772, 2395-2919, 2931-3914 (3399), 3073-3681 (3076, 3247, 3283, 3439, 3508, 3517), 5169-5381 (5268), 4429-4497, 4621-4854 (4642), 4966-5061, 5075-5383, 4884-133	568-840
fd Phage	226-1455 (1123), 461-553, 728-805, 1133-1315 (1265), 1470-1730 1735-1833, 1833-1928, 1928-2146, 2206-3477, 3086-3368 (3200) 3483-3818, 3824-4861, 4122-4238, 4848-6125, 5947-6036, 5679-0003	
f1 Phage	101-178, 496-828, 506-619, 843-1103, 1108-1206, 1206-1301, 1301-1519, 1579-2850, 2459-2641 (2573), 2856-3191, 3196-4239 4220-5497, 4885-4979, 5390-5408, 6006-0828	
MS2 Phage	130-1308 (748, 913, 988), 1218-1304, 1335-1724 (1536), 1761-1902	
M13 Phage	101-178, 496-828, 506-619, 843-1103, 1108-1206, 1206-1301 1301-1519, 1579-2850, 2459-2641 (2573), 2856-3191, 3196-4239 4220-5497, 4887-4979, 5319-5408, 6006-0828	

TABLE IV

Predicted proteins coding sequences in DNA of Prokaryotic Systems which were not part of data base used in the development of algorithm. All experimentally observed protein coding sequences are properly located, and are underlined. Additional initiators are given in parenthesis.

Sl. No.	Name of the system	EMBL Sequence data code	Predicted coding frames, additional initiators are given in brackets
1.	E.coli genes	ECATPXB	27-143 (36), 155-967 (431, 497, 656) 1077-1253, 1405-1785, 1803-2333 (1932, 2244)
		ECPAP3	1172-1279, 1307-1444, 1592-2452
2.	Cynobacteria Anebaena	A7NIFH	195-1091
3.	Bacillus licheniformis	BLPENC	212-1186, (266), 678-749, 831-926
4.	Salmonella typhimurium	STTRPB	1-1193*(49,301),773-844
		STTRPA	72-875 (387), 634-855
5.	Klebsiella aerogenes	KATRPA	60-866, 610-711

*Preinitiator region at position 1 is not available.

discussed above are part of the data base which have been used to derive the rules underlying this method. In order to establish the generality of our method for prokaryotic systems, we have also applied our method to locate coding sequences in DNA which are not part of our data base. DNA sequences from E.coli, Cynobacteria Anabaena, Bacillus licheniformis, Salmonella typhimurium and Klebsiella aerogenes. Our results given in Table IV prove that the method developed can be used for any prokaryotic system. We feel, one might require little modification in our index values when one applies the method to eukaryotic systems.

These studies bring out (i) the 'initiator region', -18 to +18 nucleotides around ATG/GTG, is important as a whole and gives a specific structure to mRNA which is recognised by the machinery which initiates translation and not small sequences such as Shine and Dalgarno or TATA regions. (ii) There are certain common features for all prokaryotic protein coding DNA sequences which are recognised by transcriptional and translational machinery. These common features can be picked up at trinucleotide level. These triplet patterns may give rise to certain specific tertiary structures to DNA. Recent kinetic studies on oligomers of CG indicate that DNA molecule has specific tertiary structure based on its sequence (15 , 16)

We feel that this approach can be used to suggest the minimal changes in DNA sequences, which when affected, can either shut off the gene specifically at transcriptional or translational level. Similar logic can be used to suggest the changes which can convert a noncoding region of DNA into protein coding sequence.

CONCLUSIONS

A simple method is developed by combining the 'gene search by signal' and 'gene search by content' approaches, which is quite accurate and has capability to predict coding sequences which are yet to be discovered. The software is developed in BASIC and uses the EMBL nucleic acid sequence data library. The approach being simple and objective, search of any DNA sequence can be carried out quite easily.

ACKNOWLEDGEMENT

One of us, B V B Reddy, acknowledges the financial assistance from CSIR (India) in the form of JRF. We also acknowledge the EMBL for providing the data on Magnetic Tape free of cost. We acknowledge useful discussions and interest of Dr P M Bhargava.

REFERENCES

1. Shepherd, J,C.W. (1981) Proc. Natl. Acad. Sci. (USA) 78; 1596-1600.
2. Staden, R. and Mc Lachlan, A.D. (1982) Nucl. Acids Res. 10; 141-156.
3. Fickett, J.W. (1982) Nucl. Acids Res. 10; 5303-5318.
4. Griboskov, M., Devereux, J. and Burgess, R.R. (1984) Nucl. Acids Res. 12; 539-549.
5. Staden, R. (1984) Nucl. Acids Res. 12; 505-519.
6. Shine, J. and Dalgarno, L. (1974) Proc. Natl. Acad. Sci. (USA) 71; 1342-1346.
7. Steitz, J.A. (1979) Ribosomes. ed. G. Chambliss, G.R. Craven, J. Devies, K. Devies, L. Kahan, M. Nomura, pp. 479-495. Univ. Park Press.
8. Alkins, J.F. (1979) Nucl. Acids Res. 7; 1035-1041.
9. Scherer, G.F.E., Walkinshaw, M.D., Arnott, S and Morre, D.J. (1980) Nucl. Acids Res. 8; 3895-3907.
10. Gold. L., Pribnow, D., Schneider, T., Scinedling, S., Swebilius, S. and Stormo, G. (1981) Ann. Rev. Microbiol. 35; 365-403.
11. Stormo, G.D., Schneider, T.D., Gold, L. and Ehrenfeucht, A. (1982) Nucl. Acids Res. 10; 2971-2996.
12. Staden, R. (1984) Nucl. Acids Res. 12; 521-538.
13. Almagor, H.A. (1984) J. Theor. Biol. 104; 633-645.
14. Kolaskar, A.S. and Reddy, B.V.B. (1984) communicated to J. Bioscience.
15. Drew, H.R. (1984) J. Mol. Biol. 177; 535-557.
16. Fuchs, R.P.P. (1984) J. Mol. Biol. 177; 173-180.