Volume 10 Number 17 1982 Nucleic Acids Research

Recognition of protein coding regions in DNA sequences

James W.Fickett

Theoretical Division, Los Alamos National Laboratory, University of California, Los Alamos, NM 87544, USA

Received 6 July 1982; Revised and Accepted 20 July 1982

ABSTRACT

We give a test for protein coding regions which is based on simple and universal differences between protein-coding and noncoding DNA. The test is simple enough to use without a computer and is completely objective. The test has been thoroughly proven on 400,000 bases of sequence data: it misclassifies 5% of the regions tested and gives an answer of "No Opinion" one fifth of the time. We predict some new coding and noncoding regions in published sequences.

INTRODUCTION

There has been for several years now a well known and very general need for a way to distinguish a true protein-coding sequence (PCS) from a merely fortuitous open reading frame (ORF) in known DNA sequences. The need arises mainly when a gene location is only approximately known at the start of sequencing, and the sequence turns out to have more than one candidate ORF. Even when a gene has been located the surrounding sequence may contain other ORF's of unknown character, and a method to distinguish the true PCS's among these yields a powerful tool for the discovery and characterization of new proteins.

We set ourselves the task of finding an objective and self-contained test, (or decision procedure) which when presented with a DNA sequence would classify it as either coding or noncoding (in this paper "coding" will always mean "coding for protein"). Later we decided to allow the test the option of refusing to classify an occasional sequence. To be of practical value such a test should not depend on the subjective evaluation of results by the user, and should have been checked on a large number of sequences so as to be of known reliability. We chose to look for a test depending on the overall statistical properties of the base sequence rather than on specific transcription or translation initiation signals for two reasons. First, initiation signals may be unavailable. This happens frequently when the 5'

end of an interesting ORF is not included in the known sequence. It can also happen that a PCS has no initiation signals at all: cf. for example the lysis gene of phage MS2, which is only translated upon readthrough of the stop codon of the previous gene (1), and the yeast mitochondrial introns which code for protein (reviewed in Ref. 2). Second, the problem of precisely characterizing what is and what is not an initiation signal still looks extrememly difficult. We also chose to find a test which would give a simple coding/noncoding answer for a specific region, rather than trying to map all coding and noncoding regions in a large sequence at once. This makes it easier to do meaningful large-scale reliability testing. Also, though our test is not adapted to finding the exact boundaries of coding regions, it is very well adapted for combination with other relevant algorithms, such as searches for ORF's, ribosome binding sites, intron boundaries, etc.

Four papers have appeared in the last year which describe statistical patterns which are probably characteristic of coding regions in general. All of these patterns have the potential of forming the basis for a useful coding/noncoding test. However we believe that ours is the first paper to give a fully specified and objective test, checked on a large number of sequences. Shulman et al. found (3) patterns in the coding regions of two phage that pointed to the three letter code and to the correct reading frame. However their sample was very small, and they did not investigate the predictive power of their observations. J. C. W. Shepherd, in researching the origin of the genetic code, found (4) periodicities in the autocorrelation functions of single bases and doublets in DNA, and applied this (5) to the problem of discovering the reading frame of a PCS. Though interesting patterns are found, no specific coding/noncoding test is given, and no evidence is presented that noncoding DNA always lacks the patterns supposedly characteristic of coding DNA. Staden and McLachlan have written (6) a computer program for mapping the PCS's in a sequence by measuring the similarity of the codon usage strategy between a known PCS and the ORF under test. The method requires that the PCS used as a standard be closely related (in codon usage patterns) to any PCS discovered. This makes the method highly dependent on the judgement of the user, and may make it inapplicable in some cases.

Another, more popular, vein of research is in trying to characterize the signals for initiation of transcription and translation by which the cell itself recognizes a PCS. For reasons given above we consider this a separate problem, complementary to the one we are considering, and only refer the

reader to the surveys of Gold et al. (7) and Breathnach and Chambon (8), and to the recent computer program of Rodier et al. (9).

CHARACTERISTIC PARAMETERS OF CODING AND NONCODING REGIONS

Many people have noticed patterns, or statistical order, in PCS's, but for the most part it has not been shown that these patterns consistently fail to appear in noncoding DNA. In this section we will give a striking illustration to show that some of the order in PCS's is in fact characteristic of coding regions, and will then define some numerical parameters of sequences whose distributions reveal universal differences between coding and noncoding DNA.

All studies reported here are based on sequence data stored in the Los Alamos Sequence Library, a public databank on the CDC 7600 computers at Los Alamos National Laboratory, currently listing 486,000 bases in 320 sequences. A description of the databank (including references for the sequences) is given in Ref. 10. Each sequence in the library was divided into its coding and noncoding parts, based on the experimental evidence reported by the original authors: sections of sequence for which this information was incomplete were not used. In early experiments we found that sequences under 200 bases (a somewhat arbitrary limit, considered further below) were too small to give reliable results. So for our primary data we took 321 fragments of coding DNA (230877 bases) and 249 fragments of noncoding DNA (158987 bases), each at least 200 bases long. (Thus a coding/noncoding decision made by the test given in this paper is based on the data in the Los Alamos Sequence Library. But we will show that our method is general and can be based on any collection of sequence data.)

Underlying all observations of statistical order in PCS's is the fact that codons are used with unequal frequency (for data and review see the work of Grantham et al. (11-13)). One consequence of this fact, which has been noted several times (3-5,14,15), is that oligonucleotides (and in particular nucleotides) tend to be repeated with a periodicity of three in a PCS. Figure 1 shows the autocorrelation function for thymine in the coding and noncoding parts of the Los Alamos Library (we ignore the distinction between RNA and DNA throughout the paper, so T and U are considered synonymous). The first graph shows that in coding sequences the number of bases separating two T's is much more likely to be 2,5,8,11,... (2+3n) than it is to be 3n or 1+3n. I.e. in coding sequences identical bases are most often found in identical codon positions. The second graph shows that this regularity is absent in noncoding sequences.

We now turn to the definition of eight numerical parameters of DNA sequences which we use to distinguish coding from noncoding regions. The first four parameters, motivated by Figure 1, measure the asymmetry in the distribution of each base among the three codon positions (or the analogous positions in a noncoding sequence). Let

$$(1)\quad \begin{aligned} A_1 &= \text{Number of A's in positions } 1,4,7,10,\ldots \\ A_2 &= \text{Number of A's in positions } 2,5,8,11,\ldots \\ A_3 &= \text{Number of A's in positions } 3,6,9,12,\ldots \end{aligned}$$

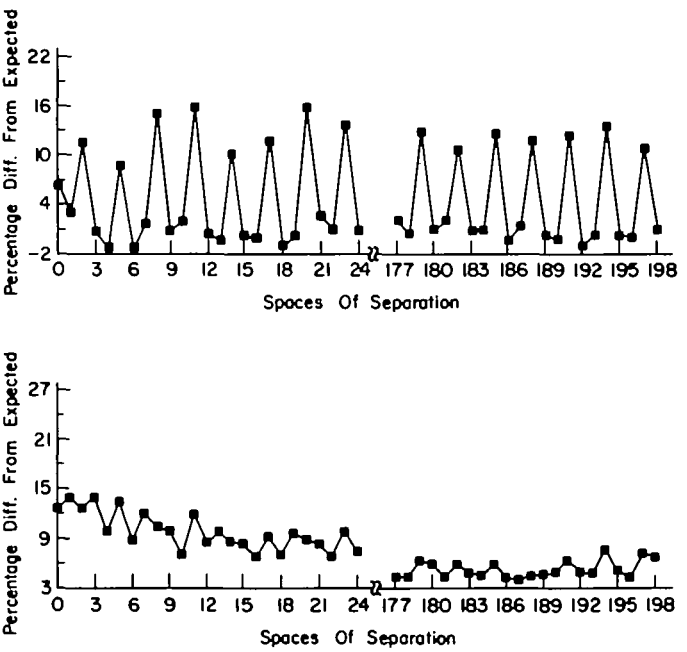

FIGURE 1. Autocorrelation graphs for T (thymine) in the 321 coding and 249 noncoding fragments over 200 bases long in the Los Alamos Sequence Library. Top: For each possible separation k, we counted, in all coding fragments in the Library (a total of 231 kilobases) the number of times two T's appeared with k nucleotides between them, and compared this with the count expected in a model where bases are chosen independently - namely the number of blocks of k+2 nucleotides times the square of the overall T-content of the coding regions. The percent difference is graphed for k running from 0 to 24 and from 147 to 198. Bottom: The same for the noncoding regions (159 kilobases). The wave so conspicuous for the coding regions is absent here. Findings were similar for the other three bases, and for pairs of unlike bases. The high values near the beginning of the noncoding graph are probably due to AT clustering; otherwise the two graphs have about the same average value.

and similarly for C, G and T. Then define

$$(2) \quad \text{A-Position} = \frac{\text{MAX}(A_1, A_2, A_3)}{\text{MIN}(A_1, A_2, A_3)+1}$$

and similarly for C, G and T.

The parameters A-, C-, G- and T-Position measure the degree to which each base is favored in one codon position over another. Note that it is irrelevant which of the three codon positions favors the base; it is only the degree to which the base is favored that is measured - this property gives these four parameters fairly similar distributions in all sequences, regardless of the well known differences in codon usage strategy between organisms.

The other four parameters we use are just the A-, C-, G- and T-Content of the sequence (i.e. the percentage of the sequence contributed by each of four bases). Note that, as a practical matter, the counts A_1 etc. made in the calculation of the Position parameters yield immediately the Content parameters also.

The relative distribution of these eight parameters between coding and noncoding fragments is shown in Table 1. All eight parameters will be used in a single test in the next section, but note that even in the distribution of individual parameters the differences between coding and noncoding DNA are evident. For example among fragments having a T-Position parameter less than 1.2 (this includes about one fourth of all fragments) there is only a 9% probability of coding function, while among fragments with T-Position parameter over 1.7 (again about one fourth of the total) the probability of coding is over 90%. Table 1 contains all the information about these parameters needed for our decision procedure. The full distributions of the eight parameters, of interest in their own right, are given in Figure 2 and discussed further below.

HOW TO DISTINGUISH CODING FROM NONCODING SEQUENCES

In the last section we gave the distribution of our eight test parameters. Next we will assign weights to each parameter, telling how much attention we should pay to it in making the final coding/noncoding decision. The parameter distribution and weights should need to be recalculated only very occasionally as more sequence data accumulates. Users of the coding/noncoding test will only need to do a very simple calculation detailed below.

From Table 1 it is clear that, for example, the T-Position parameter of a sequence usually tells one a good deal more than its A-Content. To get a

TABLE 1
Characteristic Parameters of Coding and Noncoding Sequences

Position Parameter			Probability of Coding			
0.0	to	1.1	A: .22	C: .23	G: .08	T: .09
1.1		1.2	.20	.30	.08	.09
1.2		1.3	.34	.33	.16	.20
1.3		1.4	.45	.51	.27	.54
1.4		1.5	.68	.48	.48	.44
1.5		1.6	.58	.66	.53	.69
1.6		1.7	.93	.81	.64	.68
1.7		1.8	.84	.70	.74	.91
1.8		1.9	.68	.70	.88	.97
1.9		2.0+	.94	.80	.90	.97
Content Parameter			**Probability of Coding**			
.00	to	.17	A: .21	C: .31	G: .29	T: .58
.17		.19	.81	.39	.33	.51
.19		.21	.65	.44	.41	.69
.21		.23	.67	.43	.41	.56
.23		.25	.49	.59	.73	.75
.25		.27	.62	.59	.64	.55
.27		.29	.55	.64	.64	.40
.29		.31	.44	.51	.47	.39
.31		.33	.49	.64	.54	.24
.33		.99	.28	.82	.40	.28

TABLE 1. The values of the eight parameters, A-, C-, G- and T-Position and A-, C-, G- and T-Content, were calculated for each of the 321 coding and 249 noncoding fragments over 200 bases long in the Los Alamos Sequence Library (see text). The range of each parameter was divided into ten intervals as shown (we use these same intervals for any collection of sequence data). For each interval the percentage of coding and noncoding fragments whose parameter fell therein was recorded. The value "Probability of Coding" shown is the percentage of coding fragments falling in the interval, divided by the percentage of coding plus the percentage of noncoding. This is essentially the fraction of all fragments falling in the interval which are coding, but differs slightly because more coding than noncoding fragments are used.

number telling us how much input each parameter should have in the final decision, we used each parameter alone to predict coding function, as follows: if a sequence fell in an interval where the probability of coding (from Table 1) was greater than one half the sequence was called coding, otherwise not. (I.e. if more coding than noncoding fragments share this parameter value with the fragment in question, we guess it is coding.) The weight for a given parameter is just the percentage of the time that this guess was correct, less 50% (random level). The weights for each of the eight parameters are shown in Table 2. In giving these weights we are not making any important claim about

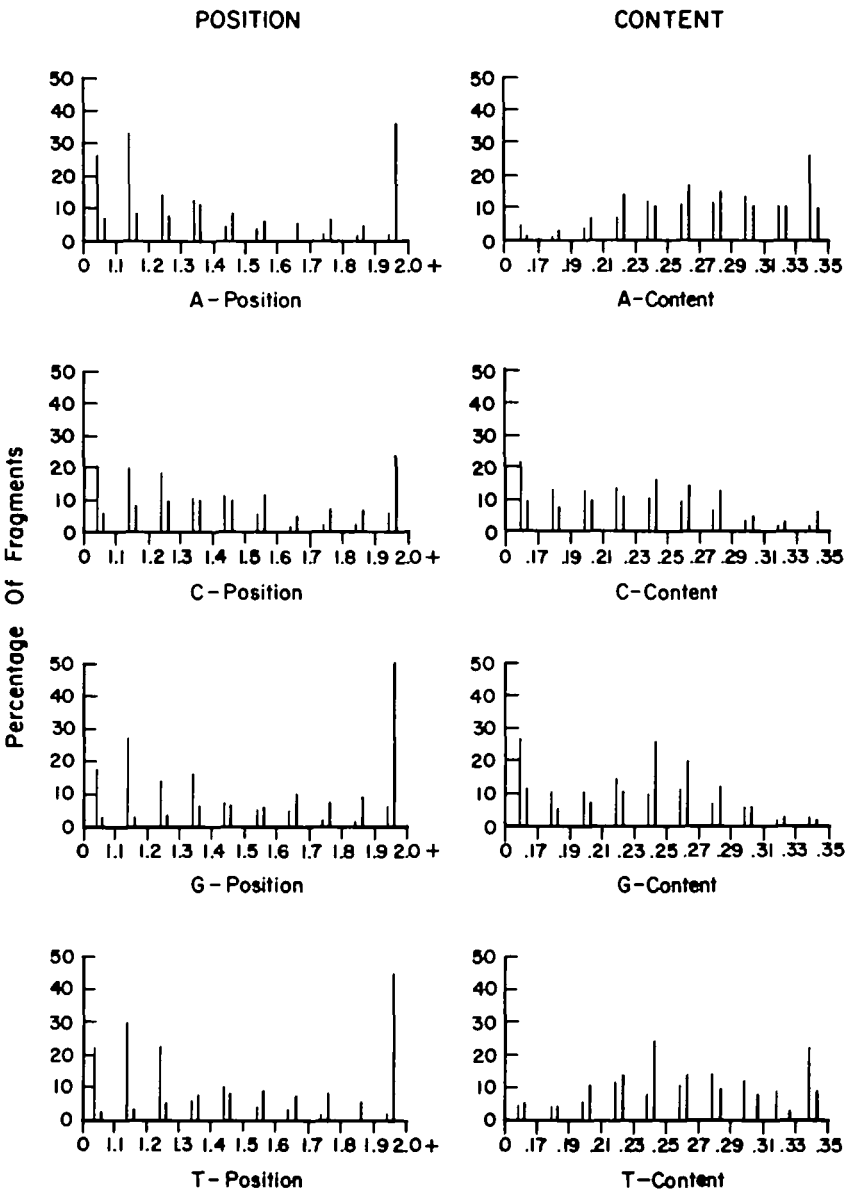

FIGURE 2. The distribution of the Position and Content parameters for coding (heavy bars) and noncoding (light bars) fragments. See the legend of Table 1 for details.

TABLE 2
Weight to be Given to the Individual Parameters

	Position	Content
A	.26	.11
C	.18	.12
G	.31	.15
T	.33	.14

TABLE 2. The weight shown is the percentage of the time (above 50%, the random level) that each parameter alone successfully predicted coding or noncoding function.

these parameters; rather we are just deciding how to use them in our specific decision procedure.

We can now describe TESTCODE, our algorithm for predicting whether a fragment of DNA is coding or not. Given a fragment of DNA, first make the counts A_i, C_i, G_i and T_i, i=1,3 (equation (1)). From these calculate the eight parameters A-, C-, G- and T-Position (equation (2)), and A-, C-, G- and T-Content of the fragment. For each of these parameters look up the "Probability of Coding" value in Table 1; call these probabilities $p_1,\ldots,p_8$. Let the corresponding weights, given in Table 2, be denoted $w_1,\ldots,w_8$. The sum $p_1w_1 + \ldots + p_8w_8$ is the TESTCODE indicator of coding function. Its distribution in the Los Alamos Library, and the predictions corresponding to its different values, are shown in Table 3. (A more familiar way to combine the information from the eight parameters would be to use Bayes' formula. But in using Bayes' formula we assume that the eight parameters are independent, which of course is not the case. So it is not surprising that the method given above worked a little better.)

RELIABILITY OF THE METHOD

From Table 3 it is clear that TESTCODE correctly predicted the function of all but a few of the fragments used in the study. However since we used these same fragments to calculate the parameter distributions which TESTCODE uses, one might object that perhaps the algorithm was just "remembering" special properties of the Los Alamos collection, and would be less reliable for distinguishing coding and noncoding DNA in general. To take care of this objection we divided the Los Alamos Library into two parts, calculated the distribution of our eight parameters on one half, and used this information to predict which fragments in the other half coded for protein. There was only a

TABLE 3
Distribution of the TESTCODE Indicator

TESTCODE Indicator	Probability of Coding	Prediction
0.32 to 0.43	0.00	Noncoding
0.43 0.53	0.04	Noncoding
0.53 0.64	0.07	Noncoding
0.64 0.74	0.29	Noncoding
0.74 0.84	0.40	No Opinion
0.84 0.95	0.77	No Opinion
0.95 1.05	0.92	Coding
1.05 1.16	0.98	Coding
1.16 1.26	1.00	Coding
1.26 1.37	1.00	Coding

TABLE 3. The distribution of the TESTCODE indicator, our predictor of coding function, is shown on all the 321 coding and 249 noncoding fragments used in this study. "Probability of Coding" is calculated just as in Table 1. The last column gives the TESTCODE prediction of function for a fragment whose indicator value falls in the corresponding interval. In calibrating TESTCODE on any set of sequence data there is always a natural cutoff point (in this case .84) above which every interval contains more coding than noncoding fragments, and below which every interval contains more noncoding than coding fragments. We always make the two intervals flanking this cutoff the "No Opinion" range.

5% error rate in these predictions, showing that TESTCODE is almost certainly based on universal differences between coding and noncoding DNA, independent of the Los Alamos collection.

In more detail our procedure was as follows: We numbered the coding fragments from 1 to 321 and the noncoding from 1 to 249. We then calculated the relative distribution of our eight parameters, as in Table 1, and the weights to use with them, as in Table 2, but using only the odd-numbered fragments as our data set. We then used the resulting parameter distributions to calculate a TESTCODE indicator for each of the even-numbered fragments. The range of the indicator was divided into 10 equal intervals, as in Table 3. Any fragment whose indicator fell in the top four intervals was judged coding, any in the bottom four noncoding, and in the middle two intervals no answer was given. The TESTCODE prediction was "No Opinion" on 18% of the fragments. 6% of the coding segments were judged incorrectly as "Noncoding", and 3% of the noncoding segments were judged incorrectly as "Coding". The actual distribution of the TESTCODE indicator is given in Figure 3.

In the future, when a larger sample of sequences is available, it may be

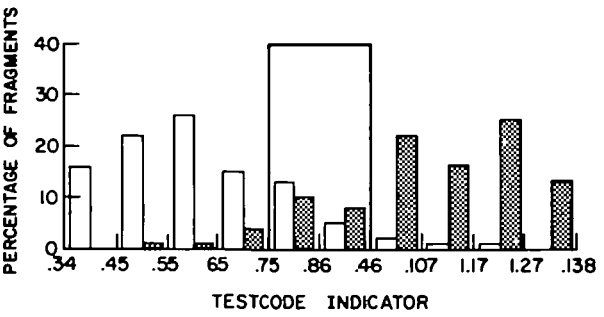

FIGURE 3. Results of the reliability test for TESTCODE. After the TESTCODE indicator was calculated for all even fragments the range of the indicator was divided into ten equal intervals, whose endpoints are marked on the abscissa. The percentage of coding (shaded bars) and noncoding (open bars) fragments whose TESTCODE indicator fell in each interval is graphed. The "No Opinion" range is boxed.

worthwhile to use separate data sets when using TESTCODE on fragments from different taxonomic classes. For example when we ran the kind of reliability test just described using only vertebrate nuclear sequences, we found that TESTCODE returned "No Opinion" on only 12% of the fragments used, and only misclassified 3%. For the vertebrate study we used 82 coding and 102 noncoding fragments; other taxonomic groups are still rather small for this kind of reliability test.

Throughout this study we have restricted attention to fragments over 200 bases long. It turns out that in fact TESTCODE's reliability is unacceptable on shorter fragments. When we used TESTCODE (just as specified in the preceeding section) to predict the function of the 57 noncoding and 159 coding fragments in the library between 100 and 199 bases in length, the predictions were incorrect 13% of the time, and the "No Opinion" rate was 29%. 200 bases seems to be a reasonable minimum, for when predictions were made in the length ranges 200-299, 300-399, 400-499, 500-599, 300+ and 600+ the error rate was always close to 5%. The chief effect of the length, above 200 bases, seems to be on the "No Opinion" rate, which is 24% for fragments of 200-299 bases, but under 15% for longer fragments.

PREDICTION OF CODING AND NONCODING REGIONS IN PUBLISHED SEQUENCES

We have scanned the Los Alamos Sequence Library for ORF's not associated with a known protein, and have rated them all with TESTCODE. In this section we give a few of our more interesting findings. Our predictions are summarized in Table 4; further comments on some are given below. A general

TABLE 4
Predicted Coding and Noncoding ORF's

Organism	Reported Sequence	Ref.	Open Frame	Prediction
Adenovirus7	Transforming Region	(17)	402 to 166*+	Coding (.92)
A. nidulans	Cytochrome B	(18)	507 to 713+#	Coding (.98)
E. coli	Insertion Element I	(19,20)	250 to 753#	Coding (.98)
			56 to 331#	No Opinion(.77)
E. coli	Origin of Replication	(21-24)	734 to 291*#	Coding (.92)
			1282 to 824*	Coding (1.0)
E. coli	Ribosomal Operon B	(25-28)	275 to 1144#	Coding (.98)
			2699 to 2959#	No Opinion(.77)
			6916 to 7506#	Coding (1.0)
Human	δ-hemoglobin	(29)	1493 to 1810	Coding (.98)
Yeast	18S rRNA	(30)	1349 to 1149*	Coding (.98)
Yeast	2μ plasmid	(31,32)	5570 to 523#	Noncoding (.29)
			2008 to 887*#	No Opinion(.77)
			5198 to 4308*#	Coding (.98)
			2271 to 2816#	No Opinion(.40)
			6258 to 5905*#	Noncoding (.04)

TABLE 4. As far as we know none of the ORF's listed here has been shown to be coding or noncoding. Numbering of the sequence is as in the (first of the) reference(s) cited. The ranking by TESTCODE (from Table 3) is given in the last column.

*Complemetary strand from that given in reference.
+No start codon.
#Possible coding function suggested in reference.

experimental method for identifying the protein product of any ORF, if it exists, has been given (16), so these predictions provide a way to assess the usefulness of TESTCODE as an exploratory tool.

The gene products of the Adenovirus transforming region are of great interest, yet we have seen no mention of the Adenovirus ORF listed in Table 4. Although it has no start codon, it might be spliced with other ORF's upstream on the same strand.

It has been shown that the box3 intron of yeast cytochrome b codes for a protein maturase, and other yeast mitochondrial introns are suspected of coding (reviewed in Ref. 2). Waring et al. (18) have shown that the

situation in Aspergillus nidulans is similar to that in yeast; the single intron in the cytochrome b gene of A. nidulans has a long ORF which continues in phase with the previous exon. Since the probability, according to TESTCODE, that this ORF codes is .98, it looks very likely that coding introns will be found in organisms other than yeast.

There is considerable interest in protein products which may be coded by movable DNA elements and which may help to insert and excise them. TESTCODE ranks very highly one long ORF in Insertion Element I of E. coli. Ohtsubo et al. (20) have sequenced an analogue of this insertion element in Shigella dysentariae and have shown that in this ORF (and another which is ranked ambiguously by TESTCODE) many more of the differences from Insertion Element i occur in third codon position than in first or second - a strong indication that both ORF's code.

The first ORF listed for the E. coli replication origin has been noted before, and in fact evidence that supports its probable coding function is given in Ref. 23. The second ORF listed, however, seems to have escaped attention.

The 3' flanking regions of many vertebrate genes have short ORF's, partly overlapping the gene, which rank highly. We include one fairly long one associated with Human δ-hemoglobin, which is clearly separate from the main gene. (25% of the designated ORF overlaps the hemoglobin gene. The remaining 75% of the ORF was tested separately and found to have a .92 probability of coding.)

The possible PCS listed for Yeast 18S rRNA is particularly interesting because no PCS is known to overlap a ribosomal RNA gene. Many ribosomal RNA genes in the Los Alamos Library contain long ORF's; the second ORF listed from the E. coli RRNB operon is another.

We have examined all the ORF's of an important cloning vector, the yeast 2 micron plasmid, and offer our opinion on its overall coding capacity.

WHY TESTCODE WORKS

In this section we show that TESTCODE's success can be understood in terms of two simple facts: 1) Any kind of consistent non-random codon use results in uniformly high Position parameters, and 2) Coding sequences have higher GC-content, on average, than noncoding sequences. We begin by explaining more fully the connection between codon usage and our Position parameters.

Suppose we had an organism in which A was suppressed in third codon position, but shared first and second codon position equally with the other

three bases. Thus the probability that the first base of a codon was A would be .25, and likewwise the second, but the probability that the third base was A might be only .15. Then in a PCS of length N we would have, approximately, A_1=.25N, A_2=.25N and A_3=.15N, so that the expected value of A-Position would be about .25/.15, or 1.7. Now note that if we had another organism in which third position A was favored instead of suppressed, so that the probabilites of finding an A in each of the three positions was, say .22, .25 and .35, respectively, the expected value of the A-Position parameter would be .35/.22=1.6, a similar value. Thus it turns out that all the very different coding strategies used by different creatures lead to the same result - Position parameter values mostly in the range 1.5 to 4.0 (whereas noncoding fragments have Position values, generally, in the range 1.0 to 1.5). As we mentioned earlier, this is what makes our one calculation applicable to all different kinds of sequences.

To take an actual example, the probabilities of finding an A in each of the three codon positions in vertebrates are .27, .31, and .15. We would predict from this an average A-Position parameter of .31/.15=2.1, while the actual average is 3.2. The true average is higher because the PCS's exhibit stronger codon usage preferences individually than one sees in the overall average. In the same way the predicted average C-, G-, and T-Parameter values are 1.6, 1.6 and 1.5 respectively, while the actual averages are 1.8, 1.9 and 1.9.

As one can see from Table 2, TESTCODE's decision is based mainly on the Position parameters. However the base content of the sequence shows some clear trends and does contribute a few percent to the reliability. The most noticeable trend in the base content data is that the GC-content of coding sequences tends to be higher than that of noncoding sequences.

To test whether these statistical trends really account for TESTCODE's performance, we generated artificial random "coding" and "noncoding" sequences and rated them with TESTCODE. For our synthetic "coding" sequences we generated succesive codons independently and at random, with the same frequencies as genuine vertebrate sequences. (The Library as a whole does not show strong codon preference rules, so we needed to limit ourselves to a more internally consistent set of data. There is no reason to think that the choice of vertebrate instead of, say, E. coli sequences is significant.) For our "noncoding" sequences we generated successive bases independently and at random, with frequency .27 for A and T, and .23 for G and C (again the frequencies of vertebrate sequences). We generated 100 coding and 100

noncoding random sequences, each 600 bases long (the average length of the real coding and noncoding fragments used). TESTCODE, using the data from real sequences listed in Tables 1-3, classified only 2% of the random sequences incorrectly, and gave an answer of "No Opinion" on only 17%.

SUMMARY AND DISCUSSION

We have used certain universal differences between protein-coding and noncoding regions to produce a simple algorithm TESTCODE which distinguishes coding from noncoding DNA with high reliability. When TESTCODE was calibrated on one half of the Los Alamos Sequence Library and then used to predict the coding or noncoding regions in the other half it gave an answer of "No Opinion" on 18% of the regions tested, and had an overall error rate of only 5%. We have used TESTCODE to predict a number of new coding and noncoding regions in published sequences.

A method for distinguishing coding from noncoding DNA has a large number of potential uses. First, after a fragment of DNA known to contain the gene for a certain protein has been isolated and sequenced, it often turns out to contain several ORF's from among which one must choose the correct one. A recent example is the search for the E. coli trpR gene by Singleton et al. (33). The authors considered three possible ORF's and discovered the correct one by mutation analysis. TESTCODE rates only the correct one as coding. Thus TESTCODE (or a related algorithm) may be able to reduce the experimental work in such cases to a single confirmatory experiment. Second, when newly sequenced DNA is found to contain an ORF of unknown function, TESTCODE may be used to decide whether it is likely to code for a new protein. This could be a powerful technique for discovering new proteins. One can even imagine the day when semi-automated sequencing of entire genomes followed by computer analysis of the results could fully catalogue the proteins of an organism. A third use for TESTCODE is in checking the accuracy of the data in computer-based sequence libraries. We discovered several errors in the Los Alamos Library with the help of TESTCODE.

We think that TESTCODE will prove to be useful both to experimentalists in their initial analysis of sequence data and to theoreticians as they learn about the differences between coding and noncoding DNA. However we do not claim to have discovered the ultimate coding/noncoding test. Indeed, the main value of this paper as we see it is that it presents one method for recognizing coding sequences which is spelled out in complete detail and has been tried out on a large collection of sequence data. Thus other people can

easily use TESTCODE and know how to interpret the results. We will gladly make available our programs and data to anyone wishing to more fully develop and test other methods. (They are available on-line to users of the Los Alamos Library. Others may request a tape by mail.)

Research on TESTCODE-like algorithms is complementary to several other lines of research. For example on the one hand TESTCODE only has a resolution of 200 bases and can not pinpoint the exact boundaries of a PCS, while on the other hand methods for recognizing signals for the initiation of transcription, initiation of translation, and intron splicing are poorly developed and require additional confirmation; thus these two methods can profitably be combined. Also, since TESTCODE is completely insensitive to phase, it can only be used to tell when a region is coding, and not what the coding frame is. This limitation can usually be overcome by combining TESTCODE with a search for ORF's, but when two ORF's overlap in different phases, another method is needed to decide which is the correct one. This can very likely be done using published methods mentioned in the introduction (3-6). Users of TESTCODE should be aware of one other point: we have not checked TESTCODE on regions of mixed coding/noncoding character. Thus it would be best to apply TESTCODE to regions that will be either fully coding or fully noncoding, for example ORF's starting at the last probable fMET codon.

There is some interesting regularity in the errors that TESTCODE makes. In coding sequences which are incorrectly classified as noncoding it often seems that some use is being made of the DNA which causes the usual codon preference rules to be overridden. For example one of two overlapped viral genes is sometimes classified as noncoding. Also, variable regions of immunoglobulin genes often are rated noncoding, presumably because the mechanism which generates diversity of these regions is stronger than whatever force encourages consistent codon preference. A very interesting example pertains to the yeast mating type loci. The four presumptive PCS's there are rated noncoding - possibly this means that some other pattern is present in this region of the DNA which is necessary to enable transposition.

ACKNOWLEDGEMENTS

We gratefully acknowledge the helpful criticism and encouragement of W. Beyer, M. Dembo, W. Goad, B. Goldstein, B. Nelson and the referees. This work was performed under the auspices of the U. S. Department of Energy.

REFERENCES

1. Kastelein, R.A., Remaut, E., Fiers, W. and van Duin, J. (1982) Nature 295, 35-41
2. Borst, P. and Grivell, L.A. (1981) Nature 289, 439-440
3. Shulman, M.J., Steinberg, C.M. and Westmoreland, N. (1981) J. Theor. Biol. 88, 409-420
4. Shepherd, J.C.W. (1981) J. Mol. Evol. 17, 94-102
5. Shepherd, J.C.W. (1981) Proc. Nat. Acad. Sci. USA 78, 1596-1600
6. Staden, R. and McLachlan, A.D. (1982) Nucleic Acids Res. 10, 141-156
7. Gold, L, Pribnow, D., Schneider, T, Shinedling, S, Singer, B.S., and Stormo, G. (1981) Ann. Rev. Microbiol. 35, 365-403
8. Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349-383
9. Rodier, F., Gabarro-Arpa, J., Ehrlich, R. and Reiss, C (1982) Nucleic Acids Res. 10, 391-402
10. Fickett, J.W., Goad, W.B. and Kanehisa, M. (1982) Los Alamos National Laboratory Report LA-9724-MS
11. Grantham, R., Gautier, C. and Gouy, M. (1980) Nucleic Acids Res. 8, 1893-1912
12. Grantham, R., Gautier, C, Gouy, M, Jacobzone, M and Mercier, R. (1981) Nucleic Acids Res. 9, r43-r74
13. Grantham, R. (1980) Trends Biochem. Sci. 5, 327-331
14. Trifonov, E.N. and Sussman, J.L. (1980) Proc. Nat. Acad. Sci. USA 77, 3816-3820
15. Eigen, M. and Winkler-Oswatitsch, R. (1981) Naturwissenschaften 68, 282-292
16. Sutcliffe, J.G., Shinnick, T.M., Green, N., Liu, F.-T., Niman, H.L., and Lerner, R.A. (1980) Nature 287, 801-805
17. Dijkema, R., Dekker, B.M.M. and Van Ormondt, H. (1980) Gene 9, 141-156
18. Waring, R.B., Davies, R.W., Lee, S., Grisi, E., Berks, M.M., and Scazzocchio, C. (1981) Cell 27, 4-11
19. Ohtsubo, H. and Ohtsubo, E. (1978) Proc. Nat. Acac. Sci. USA 75, 615-619
20. Ohtsubo, H., Nyman, K., Doroszkiewicz, W. and Ohtsubo, E. (1981) Nature 292, 640-643
21. Sugimoto, K., Oka, A., Sugisaki, H., Takanami, M., Nishimura, A., Yasuda, Y. and Hirota, Y. (1979) Proc. Nat. Acad. Sci. USA 76, 575-579
22. Meijer, M., Beck, E., Hansen, F.G., Bergmans, H.E.N., Messer, W., von Meyenburg, K. and Schaller, H. (1979) Proc. Nat. Acad. Sci. USA 76, 580-584
23. Lother, H. and Messer, W. (1981) Nature 294, 376-378
24. Nakamura, M., Yamada, M., Hirota, Y., Sugimoto, K., Oka, A. and Takanami, M. (1981) Nucleic Acids Res. 9, 4669-4676
25. Brosius, J., Dull, T.J., Sleeter, D.D. and Noller, H.F. (1981) J. Mol. Biol. 148, 107-127
26. Csordas-Toth, E., Boros, I. and Venetianer, P. (1979) Nucleic Acids Res. 7, 2189-2197
27. Brosius, J., Palmer, M.L., Kennedy, P.J. and Noller, H.F. (1978) Proc. Nat. Acad. Sci. USA 75, 4801-4805
28. Brosius, J., Dull, T.J. and Noller, H.F. (1980) Proc. Nat. Acad. Sci. USA 77, 201-204
29. Spritz, R.A., Deriel, J.K., Forget, B.G. and Weissman, S.M. (1980) Cell 21, 639-646
30. Rubtsov, P.M., Musakhanov, M.M., Zakharyev, V.M., Krayev, A.S., Skryabin, K.G. and Bayev, A.A. (1980) Nucleic Acids Res. 8, 5779-5794
31. Hartley, J.L. and Donelson, J.E. (1980) Nature 286, 860-864
32. Hindley, J. and Phear, G.A. (1979) Nucleic Acids Res. 7, 361-375
33. Singleton, C.K., Roeder, W.D., Bogosian, G., Somerville, R.L. and Weith, H.L. (1980) Nucleic Acids Res. 8, 1551-1560