CRITICA: Coding Region Identification Tool Invoking Comparative Analysis

Jonathan H. Badger and Gary J. Olsen

Department of Microbiology, University of Illinois

Gene recognition is essential to understanding existing and future DNA sequence data. CRITICA (Coding Region Identification Tool Invoking Comparative Analysis) is a suite of programs for identifying likely protein-coding sequences in DNA by combining comparative analysis of DNA sequences with more common noncomparative methods. In the comparative component of the analysis, regions of DNA are aligned with related sequences from the DNA databases; if the translation of the aligned sequences has greater amino acid identity than expected for the observed percentage nucleotide identity, this is interpreted as evidence for coding. CRITICA also incorporates noncomparative information derived from the relative frequencies of hexanucleotides in coding frames versus other contexts (i.e., dicodon bias). The dicodon usage information is derived by iterative analysis of the data, such that CRITICA is not dependent on the existence or accuracy of coding sequence annotations in the databases. This independence makes the method particularly well suited for the analysis of novel genomes. CRITICA was tested by analyzing the available *Salmonella typhimurium* DNA sequences. Its predictions were compared with the DNA sequence annotations and with the predictions of GenMark. CRITICA proved to be more accurate than GenMark, and moreover, many of its predictions that would seem to be errors instead reflect problems in the sequence databases. The source code of CRITICA is freely available by anonymous FTP (rdp.life.uiuc.edu in /pub/critica) and on the World Wide Web (http://rdpwww.life.uiuc.edu).

Introduction

The recent publication of complete genome sequences for several organisms (e.g., Fleischmann et al. 1995; Fraser et al. 1995; Bult et al. 1996; Himmelreich et al. 1996; Kaneko et al. 1996; Blattner et al. 1997) raises the question of whether the tools for sequence analysis are keeping pace with the data. One deceptively simple problem is that of how to identify the protein-coding regions of DNA sequences even in the absence of introns. Doing so requires accurate recognition of those genomic sequences most consistent with protein coding and the choice of the appropriate translation start and end points.

Numerous approaches to identifying coding sequences have been proposed (for a review of early work, see Fickett and Tung 1992; more recent work includes Borodovsky and McIninch 1993; Gish and States 1993; States and Gish 1994; Snyder and Stormo 1995; Gelfand, Mironov, and Pevzner 1996; Uberbacher, Xu, and Mural 1996; Burge and Karlin 1997; Salzberg et al. 1998). On the simplest level, DNA sequences are often analyzed by looking for open reading frames (ORFs), which are a series of coding triplets uninterrupted by a terminator codon. Typically, an ORF capable of producing a peptide of at least 60–75 amino acids is retained. Although this approach has the advantage of making few assumptions about the nature of coding DNA, it misses genes encoding proteins shorter than the 60–75-amino-acid threshold; yet, even when the threshold length is set this high, the analysis produces a significant number of false-positive ORFs that occur by chance alone (particularly in G+C-rich DNA). Another common problem is ambiguity as to which triplet is the actual initiator codon.

More sensitive approaches to protein prediction exploit the fact that an absence of terminators is not the only nonrandom property of coding sequences. In particular, the use of synonymous codons is generally biased (Staden and McLachlan 1982), and even more so is the use of dicodons, hexamer DNA sequences defining adjacent codons (Claverie and Bougueleret 1986). Analyses based on dicodon usage and a related measure based on a fifth-order Markov model of sequences (Borodovsky and McIninch 1993) are among the most powerful current methods for defining the coding regions of a new DNA sequence.

Other approaches to identifying coding frames in a DNA sequence are based on comparative analysis. If a nucleotide sequence can be translated to yield a product with significant similarity to a known protein, then that DNA is reasonably assumed to be protein-coding in the chosen frame (Gish and States 1993). Analysis using the BLASTX program (Gish and States 1993) depends on a database of previously defined proteins and therefore cannot find genes that encode new types of proteins. In contrast, the TBLASTX program (W. Gish, unpublished), which relies only on a DNA sequence database translated in all six reading frames, identifies DNA sequences that would make similar proteins but does not directly distinguish between those that are protein-coding and those that are merely similar DNA sequences. Recently, an algorithm similar in spirit to BLASTX was developed for the analysis of intron-containing sequences (Gelfand, Mironov, and Pevzner 1996). First, all possible exons in the sequence being analyzed are compiled, then the hypothetical proteins resulting from the various possible splicings are compared against a database of known proteins. The splicing that yields a protein most similar to a known protein in the database is

Key words: coding sequence prediction, sequence analysis, dicodon bias, genomics, *Salmonella typhimurium.*

Address for correspondence and reprints: Gary J. Olsen, Department of Microbiology, University of Illinois, B103 Chemical and Life Sciences Laboratory, 601 South Goodwin Avenue, Urbana, Illinois 61801. E-mail: gary@phylo.life.uiuc.edu.

Mol. Biol. Evol. 16(4):512–524. 1999

assumed to be the correct one. Again, this method depends on the presence and accuracy of protein homologs in the databases.

We have developed a new method for identifying coding DNA. Its novelty lies in comparing query DNA sequences to related DNA sequences from other species to find those regions of DNA in which the encoded amino acids display more sequence identity than would be expected from the observed amount of DNA sequence divergence. Such excess identity provides evidence of amino acid conservation and, hence, translation. The method does not rely on the annotation of any of the sequences in the DNA data banks; hence, it is particularly well-suited for the analysis of novel genes and genomes. Because it incorporates comparative analysis, the method's value and accuracy increase with increasing DNA data.

In this paper, we describe CRITICA (Coding Region Identification Tool Invoking Comparative Analysis), a set of programs that implement this approach. We evaluated it by analyzing the DNA sequence data available from *Salmonella typhimurium*, comparing CRITICA's predictions with the corresponding database annotations and with the coding regions suggested by similarity searches using BLASTP. We also compare CRITICA's performance to that of GenMark (also called GeneMark; Borodovsky and McIninch 1993), perhaps the most widely used and accurate alternative currently available.

The CRITICA Algorithm

Rationale

The problem to be solved is one of finding regions in a DNA sequence with high ''evidence of coding.'' CRITICA uses four steps to analyze a given DNA sequence (the query): (1) Give each trinucleotide (triplet) in the DNA a numerical score based on how much more it resembles a codon in a coding sequence than it resembles a triplet in a noncoding region. This score is (usually) the sum of two components: a comparative score based on the relative identities of the nucleotides and the corresponding potential amino acids, and a noncomparative score based on dicodon bias in coding frames. (2) Identify regions of sequence that have higher-than-random scores for coding. (3) Extend the candidate coding region to a terminator codon or to the end of the query sequence. (4) Examine the effect of choosing each of the available initiator codons by incorporating an initiator codon preference score and a score for any potential Shine-Dalgarno sequence (ribosome-binding site). If the resulting overall evidence of coding is sufficiently high, the DNA sequence is predicted to be coding. Each of these steps is considered in more detail below.

Assessing Comparative Evidence for Coding

To understand comparative evidence of coding, consider DNA sequences A and B, stemming from a common ancestral sequence. Subsequent to their separation, the A and B lineages have independently accumulated nucleotide changes. If these sequences do not code for protein, then the sites of nucleotide substitutions should be distributed randomly with respect to coding potential; that is, there will be no special conservation of conceptual translation products. If, on the other hand, the sequences encode a protein, excess identity of the amino acids may be observed, which can then be taken as evidence of translation. To carry out the analysis, DNA sequences that are similar enough to the query sequence to be probable homologs are first found and aligned. The aligned triplets in the sequences are then analyzed in terms of percentage identity of nucleotides versus percentage identity of coded amino acids.

CRITICA uses the BLASTN program (Altschul et al. 1990) to locate sequences in a DNA database that are sufficiently similar to the query as to be likely homologous. The BLASTN search parameters E and E2 (expected number of randomly matching sequence segments) are typically set to 10^{-4}. Following removal of matches to the query organism, the remaining local alignments (the high-scoring segment pairs [HSPs]) produced by BLASTN are used directly. The fact that these alignments do not include alignment gaps is an advantage, because the comparative detection of coding (below) assumes a consistent relative reading frame in the aligned DNAs. A typical BLASTN alignment of related DNA sequences from *S. typhimurium* and *Pasteurella haemolytica* is shown in figure 1*A*.

Given one of the alignments from the preceding step (fig. 1*A*), we test the hypothesis that the locations of sequence changes are not related to coding potential. The analysis is carried out in each of the six possible translation frames (three forward and three reverse), although only one is shown. First, the aligned sequences are broken into triplets (fig. 1*B*), and the differences in the DNA sequences of the aligned triplets are counted (fig. 1*C*). For each triplet, the encoded amino acid is determined (fig. 1*D*), and the locations of differences are noted (fig. 1*E*). From these observations, a score that summarizes the contribution to coding evidence is assigned to each triplet (fig. 1*F*).

In general, identical aligned triplets are assigned a score of zero, since they will always encode identical amino acids and, hence, can carry no comparative information about coding. A positive coding score is given to aligned triplets that are synonymous, such as TTC and TTT (phenylalanine) or CTA and TTG (leucine). To compensate for positive scores arising from the random occurrence of synonymous triplets, a negative score must be assigned to aligned triplets encoding different amino acids, that is, nonsynonymous codons such as CAG (glutamine) and CAT (histidine). The random probabilities of coding the same amino acid, averaged over all pairs of the 61 coding triplets weighted equally, for triplets that differ by zero, one, two, or three nucleotides are listed in table 1. This equal weighting is an appropriate model for the *S. typhimurium* DNA analyzed in this paper (which has a G+C content very close to 50%), but might be productively modified for organisms with highly biased G+C contents. The random chance of coding the same amino acid is much lower when more nucleotides differ. Therefore, synonymous triplets

A	aligned DNA:	*S. typ.*	...TTTCGCCAATTGATTCAGGTA...					
		P. hae.	...TTCAAACAACTAGTCCATTTA...					
B	aligned triplets:	*S. typ.*	TTT	CGC	CAA	TTG	ATT	CAG	GTA
		P. hae.	TTC	AAA	CAA	CTA	GTC	CAT	TTA
C	differences per triplet:		1	3	0	2	2	1	1
D	encoded amino acids:	*S. typ.*	F	R	Q	L	I	Q	V
		P. hae.	F	K	Q	L	V	H	L
E	amino acid comparison:		same	diff.	same	same	diff.	diff.	diff.
F	comparative evidence score:		52	–14	0	164	–16	–36	–36
G	dicodon frequency score:		–21	-6	–79	33	37	33	8
H	combined evidence score:		31	–20	–79	197	21	–3	–28
I	comparative running total:		52	38	38	202	186	150	114
J	combined running total:		31	11	0	197	218	215	187

FIG. 1.—Elements of the analysis of evidence for coding performed by CRITICA. *A,* BLASTN is used to align an *Salmonella typhimurium* DNA sequence (part of the coding region for catabolite activator protein [CAP]) with a related sequence from *Pasteurella haemolytica. B,* The aligned sequences and their complements are broken into triplets representing the six possible coding frames, only one of which is shown. *C,* The number of nucleotide differences per aligned triplet is determined. *D,* The triplets from each sequence are translated. *E,* The locations of amino acid differences in the conceptual translation products are noted. *F,* A comparative-evidence-of-coding score is assigned to each triplet based on the number of nucleotide differences and the conservation or nonconservation of the encoded amino acid. The scores shown are for 32 informative triplets (see table 2 and the text). *G,* A dicodon frequency (noncomparative) score is assigned to each triplet based on the relative frequency with which the given triplet follows its preceding triplet in coding frames versus noncoding contexts (eq. 1). To get the first score on the left, it is necessary to know that the preceding triplet is AAA. *H,* The comparative and noncomparative scores are added for each triplet. *I,* A running total of the comparative scores is taken, or (*J*) a running total of the combined scores is taken. The total is not allowed to a go below zero.

differing at two or three positions are more informative and will receive more emphasis in the scoring than synonymous triplets differing by one nucleotide.

Altschul (1993) pointed out that for analyses of this type, a log-odds score has favorable properties. This was also recognized by Snyder and Stormo (1995) in their coding region identification program. In the current context, we define the coding evidence due to two aligned triplets as the logarithm of the probability of finding this combination of triplets in a coding frame divided by the probability of finding these triplets aligned in a noncoding frame. Unfortunately, it is not possible to directly compute these values, since they depend on several factors, including the degree of protein sequence divergence in each real coding frame. In principle, an empirical compilation of these frequencies would be possible, but there is no reason to believe that these could be generalized, because different organisms have different codon usage, and different genes have different extents of divergence. Given these limitations on finding an ''optimal'' solution, we created several scoring matrices (table 2) that differ in the assumed amount of sequence divergence (Altschul 1993). In essence, each matrix was heuristically constructed to detect coding within a region containing a specific number of informative triplets (8, 16, 32, 64, or 128; see appendix). Thus, the matrix called ''8'' is optimized for a few informative triplets with very high amino acid conservation, while the matrix called ''128'' is optimized for many informative triplets with only a small excess of amino acid conservation. For each query sequence, CRITICA performs the comparative scoring with each of the five matrices, keeping alignments that are significant for any of these matrices. The example in figure 1*F* uses the matrix for 32 informative triplets.

An additional consideration was how to combine the comparative evidence scores when BLASTN alignments (HSPs) from multiple database sequences cover the same region of the query. In a Bayesian approach, independent probabilities are multiplied, or, equivalently, their logarithms are added. But in the present case,

Table 1
Probabilities of Trinucleotide Sequences with a Given Number of Differences Coding the Same or a Different Amino Acid

Number of Nucleotide Differences	Probability of Coding the Same Amino Acid	Probability of Coding Different Amino Acids
0	1.000	—
1	0.255	0.745
2	0.018	0.982
3	0.008	0.992

NOTE.—It is assumed that all nucleotides are equally likely and that all nucleotide differences are equally likely. Terminator codons are excluded.

Table 2
Comparative Scores for Aligned Triplets

Optimal Number of Triplets[a]	Score for Aligned Triplets with 0, 1, 2, or 3 Nucleotide Differences						
		1		2		3	
	0	Same aa[b]	Different aa[c]	Same aa	Different aa	Same aa	Different aa
8	0	76	−99	218	−48	243	−28
16	0	65	−59	197	−31	230	−22
32	0	52	−36	164	−16	207	−14
64	0	41	−23	141	−10	171	−7
128	0	32	−16	108	−5	135	−4

[a] The scores in each row are chosen to assess coding in a region containing the given number of aligned triplets with one or more nucleotide differences.
[b] Score assigned if the aligned triplets encode the same amino acid.
[c] Score assigned if the aligned triplets encode different amino acids.

the sequences found by BLASTN are generally related to one another, since they are each related to the query. Therefore, the nucleotide differences might not be independent, and their scores cannot always be added. In CRITICA, each HSP from BLASTN is scored separately, and then the nonzero comparative scores for a given triplet in the query are averaged. This simple compromise minimizes the required bookkeeping; however, if the independence of the nucleotide differences within a triplet for the various HSPs were assured (e.g., changes to different nucleotide identities or changes at different codon positions), adding these scores, rather than averaging them, would be more sensitive.

Assessing Noncomparative Evidence for Coding

Because there is additional useful information in codon usage patterns (and we wish to analyze sequences that lack identifiable homologs and, hence, comparative information), we also incorporate noncomparative information into our analysis through a version of the dicodon method (Claverie and Bougueleret 1986). If the triplets are numbered along the sequence, then triplet i is assigned an integer-valued score (S_{dicodon}) that is a function of the sequence of triplet i (t_i) and the sequence of the preceding triplet (t_{i-1}):

$$S_{\text{dicodon}}(t_i, t_{i-1}) = \text{nint}\left(\alpha \frac{1}{\lambda_C} \ln\left(\frac{f_{\text{coding}}(t_i \mid t_{i-1})}{f_{\text{noncoding}}(t_i \mid t_{i-1})}\right)\right), \quad (1)$$

where $f_{\text{coding}}(t_i \mid t_{i-1})$ and $f_{\text{noncoding}}(t_i \mid t_{i-1})$ are the frequencies of triplets of sequence t_i in coding and noncoding contexts, respectively, given that the preceding triplet is of sequence t_{i-1}. λ_C is used to scale the log-odds scores from equation (1) so that they can be combined with the comparative scores from table 2. It is the λ parameter from Karlin and Altschul (1990; see below for details) for an analysis based solely on comparative scores. Its value is calculated for the given genome and comparative scoring matrix (as defined by a row in table 2), but is generally close to 0.015. The parameter α is an empirically evaluated factor slightly smaller than one (see *Results*) that helps compensate for the fact that the dicodon scores assume that the sequence of triplets as the outcome of a first-order Markov process, whereas Karlin-Altschul statistics assume a series of independent scores. Scores are rounded using nint, the ''nearest integer'' function. Figure 1*G* shows the dicodon scores for the example query sequence. When used, the noncomparative score for each triplet is added to the corresponding comparative scores (fig. 1*H*).

Since our goal is to analyze novel sequences, we have chosen to work without using any of the available annotations. We use an iterative approach. Initially, only reading frames with significant comparative evidence are explicitly called coding, leaving much of the sequence data unclassified—a mixture of coding and noncoding DNAs. In the first cycle of dicodon analysis, CRITICA uses the observed dicodon frequencies in the regions explicitly called coding and a user-supplied, a priori estimate of the fraction of the DNA that is coding to estimate the number of occurrences of each dicodon in all coding regions (G. D. Pusch, personal communication). The difference between the total occurrences of a dicodon (hexanucleotide sequence) in the DNA data and the estimated number of occurrences in all coding regions provides the number (and, hence, frequency) in noncoding regions. In all subsequent iterations, all sequences are explicitly classified as coding or noncoding by CRITICA's calls in the previous iteration, so there is no further use of the user's a priori estimate of the fraction of the DNA that is coding.

Finding Regions with Statistically Significant Evidence of Coding

Given the evidence of coding for each triplet, we now seek regions of sequence sufficiently high in coding support to declare the behavior nonrandom and, thus, the sequence probably coding. This problem is related to other well-studied problems to which the method of maximal segment analysis has been applied (e.g., Karlin and Altschul 1990, 1993). To start, we take a running total of the evidence of coding (fig. 1*I* and *J*), not allowing the total to go below zero. A high-scoring segment (HSS) in this nonnegative running total, being a region enriched in coding evidence, starts with a step up from zero and ends with the maximum value reached before either (1) the running total declines back to zero, (2) the query sequence contains a triplet that would be a stop codon, or (3) the end of the query sequence is reached. Functionally, this results in the identification of all HSSs defined as arbitrary contiguous segments in the query sequence that do not contain any stop codons and that have been extended through either end point as far as possible by adding on triplets of positive or zero score. An HSS can contain triplets of negative score if this permits adding an equal or greater amount of positive scores. Each end of an HSS is bounded by a codon pair of negative score, the beginning or end of the query sequence, or a stop codon.

Karlin and Altschul (1990) provide formulas for assessing the statistical significance of an HSS under the assumption that the scores at each site (in this case, each triplet) are assigned independently from a fixed probability distribution. One can then compute two parameters (K and λ) that define the approximate distribution of the largest score among all of the HSSs contained in a sequence of length N. The theoretical distribution of the largest score is used to assign P values to individual HSSs. The probability that one or more intervals will have a score of S or greater is approximately

$$P(S) = 1 - e^{-KNe^{-\lambda S}}. \quad (2)$$

For the comparative component of an analysis, calculating K and λ requires the random probabilities and scores for each of the seven outcomes in table 1. The probabilities of aligned triplets with zero, one, two, and three nucleotide differences are estimated from their observed frequencies in the BLASTN HSPs. Because the BLASTN alignments for a short individual query sequence sometimes have few sites with changes, we dampen the sampling variations by adding a fixed number of triplets with the average balance of nucleotide

differences (36, 9, 4, and 1 for 0, 1, 2, and 3 differences, respectively) observed in our preliminary analyses of bacterial DNA sequences. For a given number of differences, the random probabilities of encoding the same or a different amino acid are taken from table 1, and the associated score is taken from the appropriate row of table 2. For example, for the query sequence in figure 1, the probability that the aligned triplets differ by one nucleotide was estimated to be 0.18. For sites with exactly one nucleotide difference (as in the first pair of triplets), the random probability of the triplets being synonymous is 0.255 (table 1). Thus, at random, about 0.0459 (0.18 $\times$ 0.255) of all triplets would be aligned with a synonymous triplet that differs by one nucleotide. If the analysis were for the 32-informative-triplet matrix, then the associated score (from table 2) would be 52. The value of λ computed for a comparative analysis encompassing all of the sequences to be analyzed from a given genome is called λ_C and is subsequently used to scale the dicodon scores (above) and the initiator codon and Shine-Dalgarno sequence scores (below).

For the noncomparative (dicodon) component of an analysis, there are 4,096 (64 $\times$ 64) combinations of adjacent triplets. The probability of a specific dicodon is estimated from the empirical frequencies of the sequence in noncoding contexts (which is possible only after the first round of comparative analysis). For example, the frequencies of TTT preceded by AAA in *S. typhimurium* DNA are 0.000281 in coding sequences and 0.000417 in noncoding sequences. The corresponding score from equation (1) is -21 (if $\lambda_C = 0.015$ and $\alpha = 0.8$).

When simultaneously analyzing both comparative and dicodon information, calculating K and λ requires the random probabilities and scores for all 28,672 (7 $\times$ 4,096) combinations of possible comparative and dicodon outcomes. For the above examples, the combined event (a TTT triplet aligned with a synonymous triplet that differs by one nucleotide and preceded by a AAA triplet) has a random probability of 1.91×10^{-5} (0.0459 $\times$ 0.000417), and its score is 31 (52 + -21).

Whether for comparative analysis or for combined comparative and dicodon analysis, CRITICA computes the values of K and λ for each combination of query sequence and scoring matrix, allowing for different amounts of comparative data. However, to provide more uniform behavior for query sequences of different lengths, the assumed number of events, N, is held at a constant value of 2,000, the approximate number of triplets analyzed per gene. For each region of coding evidence, if the P value for the score (eq. 2) is less than a predefined threshold for any of the five matrices, then the region is considered potentially significant and is kept for further processing. Otherwise the region is discarded.

Adjusting the Ends of Potential Coding Regions

A region of coding evidence ends with the last positive evidence for coding. However, in the absence of introns, real coding regions end at stop codons. Therefore, we extend the 3′ (C-terminal) end of each potential coding region to a stop codon or the last complete codon in the query sequence, whichever comes first, and adjust the score of the region to include these triplets.

Similarly, a region of coding evidence starts with the first positive evidence for coding, while real coding regions start with initiator codons. This situation is more complicated because, unlike stop codons, initiator codons also serve a function within a coding sequence. Therefore, the problem is deciding which potential initiator codon to use or whether to extend the region upstream to the first complete codon of the query. If a region is extended upstream, additional triplets and associated scores are added to the coding region; if a downstream initiator is chosen, triplets and associated scores are removed. The score is also adjusted for the sequence of the initiator triplet (t = ATG, GTG, or TTG). The log-odds score for an initiator triplet of identity t is

$$S_{\text{initiator}}(t) = \text{nint}\left(\frac{1}{\lambda_C}\ln\left(\frac{f_{\text{initiator}}(t)}{f(t)}\right)\right), \qquad (3)$$

where $f_{\text{initiator}}(t)$ is the frequency of triplet t among all initiators, and $f(t)$ is the fraction of triplet t among all ATG, GTG, and TTG triplets in the sequences being analyzed.

The score for a potential start site is also adjusted to reflect the quality of a Shine-Dalgarno sequence (ribosome binding site; Shine and Dalgarno 1974) when such a sequence is present. For this purpose, a Shine-Dalgarno sequence is defined as 4 or more contiguous nucleotides starting within 16 nucleotides upstream of the initiator that match a subsequence of the consensus sequence RGGRGGTGAT (where R = A or G; Shine and Dalgarno 1974). The score assigned to a Shine-Dalgarno sequence s is

$$S_{\text{sd}}(s) = \text{nint}\left(\frac{1}{\lambda_C}\ln\left(\frac{f_{\text{sd}}(s)}{f(s)}\right)\right), \qquad (4)$$

where $f_{\text{sd}}(s)$ is the frequency of s being the longest match to the Shine-Dalgarno consensus in sequences adjacent to each highest-scoring translation start site and $f(s)$ is the frequency of s being the longest match to the consensus when analyzing other plausible, but lower scoring, start sites. In this context, the score of a translation start site refers to the score of the region after adjusting the start point to the given initiator codon but without the adjusting for the Shine-Dalgaro score. To limit the region evaluated for Shine-Dalgarno sequences to the most plausible locations, we consider only start positions that would yield a coding score with a P value within two orders of magnitude of the P value associated with the highest-scoring start position (before considering the Shine-Dalgarno sequence). The lack of a ribosomal binding site is treated in an analogous manner; the score is based on the frequency of no ribosome-binding site occurring at high-scoring starts divided by the frequency of this condition at lower-scoring starts.

Thus, for each plausible start point, the score for the region is adjusted for the change in the start point, for the identity of the initiator, and for the quality of the

best Shine-Dalgarno sequence. The start point resulting in the highest score is usually chosen, although we can retain all potential starts whose scores fall within a defined interval of the best. The *P* value of the resulting score is computed according to equation (2), and the region is retained only if the resulting value is more significant than a defined threshold (usually a random probability of 10^{-4}).

There is one more essential element of CRITICA's algorithm. Much of the comparative score of coding sequences is contributed by silent changes in the third position of the codon (e.g., fig. 1). Because the spacing of third-position changes is uniform, there is also a corresponding frame on the complementary strand that has an excess of third-position changes, even though this latter frame does not code; these must be excluded. CRITICA deals with this in a simple manner: for each triplet predicted to be coding, we locate the triplet on the opposite strand that shares the same third position and set its comparative score to zero. This eliminates a known source of bias, without forbidding the prediction of overlapping reading frames. For this simple treatment to work, it is necessary that CRITICA commit to coding-region predictions in the order of most support to least support.

Implementation

CRITICA was implemented as a series of ANSI-C programs and Perl 5.0 scripts. The code has been run on a variety of Sun SPARCstations running SunOS 5.4, a Silicon Graphics workstation running IRIX 5.3, and an IBM-PC-compatible (586) system running Linux 2.0. It should be portable to any system running UNIX or a UNIX-like operating system. BLASTN 1.4.7MP (Altschul et al. 1990) was used to obtain presumptive DNA homologs. BLASTP 1.4.8MP (Altschul et al. 1990) was used to find proteins similar to potential gene products. GenMark (Borodovsky and McIninch 1993) and the *S. typhimurium* fifth-order matrix were kindly supplied by M. Borodovsky.

Results

Defining the Test Data

To test the algorithm, we used CRITICA to predict protein-coding regions in *S. typhimurium* DNA. This organism was chosen because there are many sequences available (GenBank 100 has 523 *S. typhimurium* sequences, totaling 946,808 nt), and most of these data have at least one likely homolog elsewhere in GenBank.

For each *S. typhimurium* sequence, we used the BLASTN program (Altschul et al. 1990) to retrieve a set of similar sequences from GenBank. We accepted sequence matches for which each region of similarity (HSP) had a random expectation (E = E2) of 10^{-4} or less; therefore, we expect few false positives in the search of the 523 *S. typhimurium* sequences. Tenfold variations in this threshold had little overall effect on our results (results not shown).

In these studies, we discarded matches to *S. typhimurium* (the query organism), since self-similarity is uninformative to CRITICA. However, this approach to the problem is less than optimal in that it also discards potentially useful information from similarities to other members of a gene family within the query organism (paralogs). After removing the matches to *S. typhimurium*, one or more BLASTN matches covered 689,278 of the 946,808 query nucleotides.

Using Comparative Data to Estimate Dicodon and Initiator Codon Usage Frequencies

One of the decisions made in designing and implementing CRITICA was to ignore sequence annotations; the inference of dicodon usage in coding regions is based on CRITICA's coding predictions alone. At the beginning of CRITICA's analysis, there are no dicodon or initiator codon scores, but the score-based prediction method used allows inference based solely on comparative evidence. For each predicted coding region in the first cycle of the analysis, a start point (initiator codon or first complete codon of the query) is chosen to maximize the length of the frame without lowering its score. As described above, in the first iteration of a CRITICA analysis, the estimation of dicodon frequencies in coding and noncoding regions involves an extrapolation from the coding regions explicitly identified by comparative data. In the present work, we assumed that 80% of the *S. typhimurium* DNA codes in one of the six reading frames—a conservative estimate for the coding regions in well-characterized prokaryotic genomes. Initial estimates of relative initiator codon usage frequencies were directly taken from the regions explicitly called coding in the comparative analysis. From these data, the corresponding scores were derived for the following studies.

One problem mentioned earlier is that the dicodon scores, which depend on the previous codon, are not independent; therefore, using Karlin-Altschul statistics may not yield a reliable estimate of the significance of a coding region. Karlin and Dembo (1992) provide a method for computing the significance of high-scoring segments of Markov-dependent scores, but it is not computationally feasible in our case. Instead, we explored the issue empirically. We created a simulated sequence of 10^8 codons using the Markov dicodon frequencies in *S. typhimurium* regions considered to be noncoding by CRITICA. These codons were assigned CRITICA-generated dicodon scores, and the observed frequencies of high-scoring regions were compared with those predicted by Karlin-Altschul statistics (table 3). As expected, the nonindependence of the dicodon scores causes overestimation of the significance of any given score. The last column of the table shows that multiplying the log-odds dicodon scores by 0.8 results in a conservative estimate of the random expectation. We accomplish this in CRITICA by setting α to 0.8 in equation (1).

Evaluating CRITICA

To assess the accuracy of CRITICA's predictions, a method of measurement had to be chosen. Earlier evaluations of coding region identification methods (Fickett and Tung 1992; Borodovsky and McIninch 1993) tested

Table 3
Comparison of Predicted Versus Observed Occurrences of High-Scoring Regions Over a Given Score in a Markov Chain of 10^8 Dicodon Scores

Score	*P* value	Predicted ($\alpha = 1$)	Observed	Predicted ($\alpha = 0.8$)
500...	1.26×10^{-1}	6,284	18,278	23,145
600...	2.86×10^{-2}	1,430	5,252	8,340
700...	6.25×10^{-3}	312	1,511	2,608
800...	1.36×10^{-3}	67	414	780
900...	2.93×10^{-4}	14	107	230
1,000...	6.33×10^{-5}	3	32	67
1,100...	1.37×10^{-5}	0	14	20
1,200...	2.96×10^{-6}	0	3	6

NOTE.—The value of *K* for this dicodon table was 0.1425, the value of λ was 0.01532, and the value of *N* used for the calculated *P* value was 2,000.

algorithms for their ability to correctly identify segments that were entirely coding or noncoding DNA. Because experimentally derived data are not so neatly divided, we chose to view the sequences in terms of all triplets in the DNA sequence and its complement, so the number of triplets evaluated is about twice the total sequence length. The coding predictions and the coding annotations (presumed coding regions) are mapped onto the triplets. Five outcomes are distinguished for each triplet: (1) noncoding in both the prediction and the annotation, (2) coding in both the prediction and the annotation, (3) coding in the prediction and noncoding in the annotation (false-positive), (4) noncoding in the prediction and coding in the annotation (false-negative), and (5) coding in the prediction and coding in a different frame in the annotation (wrong frame). This last category was distinguished primarily for future exploration of frameshift detection. To evaluate the reliability of CRITICA, we accepted each predicted coding region in *S. typhimurium* DNA that had a combined comparative evidence and dicodon score with less than a 10^{-4} probability of occurring by chance. The results are reported in table 4.

We performed three different analyses using GenMark (Borodovsky and McIninch 1993). The first GenMark analysis of the data was performed on the WebGeneMark server (http://genemark.biology.gatech.edu/GeneMark/webgenemark.html) on July 1 and 2, 1997. The *S. typhimurium* matrix was selected, and the other parameters were left at their default values (window size = 96, step size = 12, and threshold = 0.5). The second analysis was performed using a local copy of GenMark and a coding matrix created in 1994 (presumably from GenBank annotation, although the details are unavailable). The third analysis was performed using the local copy of GenMark with a matrix created by the recently released (late 1997) utility ''makemat'' (W. Hayes and J. McIninch, unpublished), which can generate a GenMark matrix from sequence data alone, assuming that ORFs over a certain length (default 700 bases) represent true coding regions. In our initial assessment, we compared the CRITICA and GenMark predictions with the ''coding sequence'' (CDS) annotations in GenBank 100. In this test, WebGeneMark did not perform as well as CRITICA, with a total error rate of about 2.6% to CRITICA's 2.2% (table 4). Unexpectedly, in this initial test, the local copy of GenMark performed better using both the 1994 matrix and the new matrix based on the sequence data alone (2.4% total error in both cases).

Table 4
Evaluation of CRITICA and GenMark by Comparison with Presumed Coding Regions Defined by Alternative ''Authorities''

''Authority'' of Presumed Coding Regions for Evaluating Accuracy	Prediction Method	False Positives	False Negatives	Different Frame	Total Error
GenBank annotation[a]	CRITICA	0.0126	0.0088	0.0005	0.0219
	GenMark[d]	0.0127	0.0129	0.0005	0.0260
	GenMark[e]	0.0103	0.0130	0.0005	0.0238
	GenMark[f]	0.0166	0.0065	0.0005	0.0235
GenBank + BLASTP[b]	CRITICA	0.0026	0.0125	0.0032	0.0183
	GenMark[d]	0.0041	0.0179	0.0032	0.0251
	GenMark[e]	0.0032	0.0196	0.0029	0.0257
	GenMark[f]	0.0063	0.0096	0.0034	0.0193
GenBank + consistent BLASTP[c]	CRITICA	0.0028	0.0116	0.0005	0.0149
	GenMark[d]	0.0044	0.0170	0.0006	0.0220
	GenMark[e]	0.0034	0.0184	0.0006	0.0223
	GenMark[f]	0.0068	0.0090	0.0007	0.0164

NOTE.—The False Positives, False Negatives, Different Frame, and Total Error values are the fractions of the nucleotide triplets (in all six frames) for which the coding predictions of CRITICA (or GenMark) disagree with those defined by the ''authority'' in the first column of the table. These analyses cover 1,893,616 overlapping triplets.

[a] Protein-coding sequences defined in GenBank 100 annotations.

[b] Protein-coding sequences defined in GenBank 100 annotations, plus open reading frames of at least 30 amino acids that have a BLASTP match to a sequence in the NCBI nonredundant protein database. See text for additional details.

[c] The same as GenBank + BLASTP, but excluding sequences with reading frames identified by BLASTP that overlap annotated reading frames by 10 or more codons (in a different frame).

[d] Results from using WebGeneMark on July 1 and 2, 1997.

[e] Results from using a local copy of GenMark and a *Salmonella typhimurium* matrix file dated September 27, 1994.

[f] Results from using a local copy of GenMark and a matrix file generated by ''makemat'' on the *S. typhimurium* sequence data (assuming ORFs over 700 nt are true coding regions).

Table 5
Regions in *Salmonella typhimurium* Sequences Predicted to Be Coding by One or More Methods but Not Annotated as Such in GenBank Release 100

Combination of Methods Predicting the Coding Region[a]	Predicted Regions Agreed on by the Combination of Methods	
	Including All BLASTP[b]	Consistent BLASTP[c]
CRITICA + GenMark[d] + BLASTP	90	85
CRITICA + BLASTP	42	41
CRITICA + GenMark	36	41
GenMark + BLASTP	18	13
CRITICA only	29	30
GenMark only	24	29
BLASTP only	106	50
CRITICA total	197	
GenMark total	168	
BLASTP total	256	189
Grand total	345	289

[a] The methods are said to agree if they predicted coding sequences that end with the same terminator; selection of the same start codon is not necessary.

[b] Open reading frames (ORFs) of at least 30 amino acids with similarity to an entry in the NCBI nonredundant protein database.

[c] ORFs with similarity to an entry in the nonredundant protein database and that do not overlap any annotated *S. typhimurium* coding sequence in another frame by 10 or more amino acids.

[d] GenMark in this table refers to the analyses performed with WebGeneMark on July 1 and 2, 1997.

About 1.2% of the DNA triplets seemed to be erroneously called coding by both CRITICA and GenMark. Given the relatively conservative threshold used in the CRITICA analysis, this seemed unreasonably high, and even increasing the stringency of the prediction threshold by orders of magnitude did not change the predictions much (see *Discussion*). BLASTP (Altschul et al. 1990) searches using the ''seg'' filter on the query sequence (Wootton and Federhen 1993) often revealed sequences in the NCBI nonredundant protein database that were similar to translations of CRITICA's and GenMark's false positives, suggesting that there are numerous omissions in the GenBank CDS annotations. Applying this strategy to all *S. typhimurium* ORFs of 30 or more amino acids suggested as many as 256 unannotated coding regions (table 5). Although many of these regions appear to be incomplete protein-coding sequences that run off an end of the sequenced DNA fragment, some would define a complete protein-coding sequence. Of the 256 unannotated coding regions suggested by the BLASTP searches, CRITICA identified 132 (52%), and GenMark identified 108 (42%) as coding (table 5).

The large number of sequence positions that are not annotated as coding but for which BLASTP suggests otherwise indicates that the quality of the annotations might be the limiting factor in assessing the performance of coding prediction methods. To partially relieve this problem, we added the 256 regions suggested by BLASTP matches to the list of presumed coding regions. In each case, the starts of these latter regions were adjusted without including an in-frame terminator codon to (1) the closest upstream initiator, (2) the first complete codon in the sequence, or (3) the closest downstream initiator, in that order of preference. Similarly, the ends of these regions were extended to a terminator codon or the last complete codon in the DNA sequence. Adding these regions to the list of those presumed to be coding dramatically reduced the number of false positives for CRITICA and GenMark (table 4). However, the numbers of false negatives and different-frame triplets were substantially increased by this change.

Further examinations of the data revealed that of the 256 regions added, 67 overlap (by at least 10 amino acids) an annotated *S. typhimurium* coding sequence in a different frame or another BLASTP match of higher significance. Although overlapping coding sequences are known, this seemed too common. Furthermore, most of these 67 BLASTP matches were to database proteins identified only as the products of an ''open reading frame.'' It seems that many of these 67 overlapping ORFs were incorrectly identified as coding by the BLASTP analysis. Subsequently, we treated the 67 regions that overlap an annotated CDS as being ''inconsistent'' with the explicit GenBank annotations for *S. typhimurium*. When these regions are excluded from consideration, there is an increase in the fraction of the BLASTP-based reading frames that are also predicted by CRITICA (126 of 189, or 67%) and GenMark (98 of 189, or 52%) (table 5). Removing these overlapping ORFs from the list of presumed coding regions reduced the false negatives and different-frame errors of both CRITICA and GenMark (table 4).

The performance of CRITICA slightly improved when the dicodon usage and initiator frequency tables were iteratively refined. The tables used above were based on coding frames predicted from an extrapolation of the comparative data alone (the first iteration), so the dicodon frequencies for coding frames and for noncoding sequences were somewhat crude. The second iteration of the dicodon and initiator codon tables provides a small improvement over the first, and the third iteration shows no change (table 6). A similar tendency for the results to improve and then stabilize was observed in analyses of other genomic DNAs as well (data not shown). In all cases that we have tested, it seems that after one or two iterations, the limitations of our model and test data have been reached. The recent release of ''makemat'' to the public has allowed us to iterate GenMark analyses as well (table 6). As in CRITICA, the second iteration provides a small improvement, and further iterations do not change the results significantly. Unexpectedly, even the first iteration using a GenMark matrix generated on sequence data alone yielded significantly better results than those of WebGeneMark, which presumably used a matrix based on coding annotation.

Discussion

CRITICA provides a novel method for the identification of protein-coding sequences in genomic DNAs. The performance of the method appears to be better than

Table 6
Increase in the Accuracy of CRITICA and GenMark Through Successive Approximations of Coding Properties

Prediction Method	Iteration	False Positives	False Negatives	Different Frame	Total Error	Improvement
CRITICA . . .	1	0.0028	0.0116	0.0005	0.0149	—[a]
	2	0.0029	0.0112	0.0005	0.0146	0.0003
	3	0.0029	0.0112	0.0005	0.0146	0.0000
GenMark[b] . .	—	0.0044	0.0170	0.0006	0.0220	—[a]
GenMark[c]. . .	1	0.0068	0.0090	0.0007	0.0164	—[a]
	2	0.0058	0.0096	0.0006	0.0159	0.0005
	3	0.0057	0.0097	0.0006	0.0159	0.0000

NOTE.—CRITICA iteration 1 is the same as that in table 4 and uses the dicodon and initiator tables derived from the coding regions defined from comparative analysis only. The GenMark iterations were performed by supplying the coding regions found in the previous iteration to ''makemat'' and generating a new GenMark matrix. The accuracy is evaluated relative to regions annotated as coding in GenBank 100, plus the regions that are supported by BLASTP matches and do not substantially overlap in a different frame an annotated CDS in GenBank. These analyses cover 1,893,616 triplets.

[a] Not applicable.

[b] WebGeneMark.

[c] Local GenMark using ''makemat''-generated matrix.

that of GenMark, particularly when compared with WebGeneMark, the version of the software most accessible to the public. To date, we have used CRITICA in our analyses of three complete genomes (Bult et al. 1996; Klenk et al. 1997; Deckert et al. 1998).

A positive feature of CRITICA is that it does not depend on the existence or accuracy of annotations in the databases. During the development and testing of the program, we repeatedly encountered sequence database problems. Coding regions that are not annotated caused us to observe an artificially high frequency of false-positive predictions. That many of these regions are likely to be coding was documented by identifying sequences in the protein databases that were similar to translations of unannotated ORFs in the *S. typhimurium* DNA data (table 5); of 197 unannotated coding regions predicted by CRITICA, 132 had BLASTP hits in the protein data banks (of which 6 were subsequently treated as ''inconsistent'' and discarded). Since not all *S. typhimurium* proteins have homologs in the protein sequence databases, it seems likely that the remaining 65 false-positive predictions of CRITICA include additional ORFs that are actually coding. In this regard, we note that 36 of these 65 ''false-positive'' reading frames predicted by CRITICA were also called by GenMark (table 5).

Analyses of the 256 unannotated *S. typhimurium* ORFs ($\geq$30 amino acids) that have BLASTP matches to database proteins revealed another problem: the data banks include many proteins defined by GenBank CDS annotations solely because there is a region of DNA with no terminator codons. In an attempt to remove some of this noise, we dubbed 67 of the 256 ORFs ''inconsistent'' and discarded them because they overlap (in a different frame) by $\geq$10 amino acids an annotated CDS in *S. typhimurium* or a BLASTP match with a higher significance, admittedly a very superficial treatment. There are certainly additional false proteins among the 256 (our strategy for finding them was far from comprehensive), and of the 67 that we called ''inconsistent,'' careful examinations of the data suggest that 10 or more of them are apt to be real protein-coding regions. However, to minimize the introduction of bias in our evaluations of CRITICA, we chose to apply well-defined (if somewhat arbitrary and simplistic) criteria in defining our list of ''presumed coding sequences''—our standard for measuring accuracy.

The issues raised by the completeness and accuracy of database annotations are deep and pervasive. Bork and Bairoch (1996) emphasized that once a sequence with erroneous annotation is introduced to a public database, sequences similar to it will often be assigned corresponding erroneous properties when they are submitted to the databases; thus, errors will propagate from their original source. Even when the original annotation is subsequently corrected, the secondary errors built upon it generally remain. Given this, we stress that although our analyses of CRITICA's accuracy are dependent on database annotations, the method itself only examines the nucleotides. Thus, CRITICA minimizes the propagation and perpetuation of annotation errors. Further, the lack of dependence on preexisting annotations makes CRITICA particularly suitable for genome analysis in phylogenetic groups in which little is known about the organisms and their genes. Finally, the improvement in performance of GenMark when a matrix based only on the sequence data is used (table 6) implies that current annotations may actually be a hindrance to coding prediction schemes.

Returning to the problems involving the 256 additional *S. typhimurium* coding regions suggested by our BLASTP analysis, our conclusion that there are about 67 erroneous predictions among them might seem serious. However, this is an artifact of not including the much larger number of coding regions that were suggested by BLASTP and that are also present in the annotations of the *S. typhimurium* DNA; hence, the absolute frequency of false-positive errors when BLASTP or BLASTX is used to identify protein-coding sequences (Gish and States 1993) is low. This potential problem can be further reduced by incorporating codon bias information into the evaluation of protein database hits (States and Gish 1994). A far more serious concern about relying on BLASTX for coding region identification is that truly novel genes (representing new families) cannot be found by searching existing protein databases.

The design and implementation of CRITICA required several choices that merit additional comment. First, in the comparative component of the analysis, when there were multiple comparative scores of a single triplet (due to BLASTN HSPs with different database sequences), the nonzero scores were averaged. This was done to avoid multiple counting of what might be a single evolutionary change to a residue that was inherited by several database sequences. In principle, if it is known that the nucleotide differences contributing to a comparative score arose from different evolutionary events, then the scores could be added. Doing this would

Table 7
Comparison of Different Thresholds for Identifying Coding Regions in *Salmonella typhimurium* using CRITICA

Threshold	False Positives	False Negatives	Different Frame	Total Error
1×10^{-3}...	0.0033	0.0110	0.0005	0.0148
1×10^{-4}...	0.0029	0.0112	0.0005	0.0146
1×10^{-5}...	0.0022	0.0128	0.0005	0.0155
1×10^{-6}...	0.0020	0.0135	0.0005	0.0160

NOTE.—These analyses cover 1,893,616 overlapping triplets. The data reported come from the third iteration of the CRITICA run and are based on the comparison of CRITICA's analyses with the regions annotated as coding in GenBank 100 plus the regions believed to be coding via consistent BLASTP analyses.

make CRITICA more sensitive. Even without making any assumptions about the histories of the sequences, there are a variety of circumstances under which this could be done (e.g., changes to different residues or changes at different codon positions); needless to say, adding this ability will require significant additional bookkeeping in CRITICA.

Another choice made in the current implementation of CRITICA is the nature of the dicodon table used. Equation (1) makes the scoring sensitive to the encoded amino acid sequence, not just to the codon preferences. This has the effect of making the matrix more sensitive to proteins of ''canonical'' composition and amino acid nearest neighbors, but less sensitive to proteins of unusual sequence. This scoring choice could easily be changed without altering the central components of CRITICA. However, any coding sequence prediction algorithm that examines codon or dicodon usage will be potentially misled when a gene with unusual codon bias is encountered (e.g., genes acquired by recent lateral transfer or extremely high or low levels of expression). For *Escherichia coli,* GenMark matrices optimized for different amounts of codon bias have been created (Borodovsky et al. 1995). These matrices were based on an extensive classification of known *E. coli* coding regions into three bias categories (Médigue et al. 1991). While a similar classification of *S. typhimurium* coding regions is possible, to the best of our knowledge, it has not been attempted; for organisms with few known coding regions, it may in fact be impractical or impossible. In these situations, CRITICA's use of comparative information, in addition to dicodon usage, is a distinct advantage.

In the description of the CRITICA algorithm, several thresholds for scores were mentioned. Generally, these thresholds have been set at levels such that anything that might ultimately be called a coding sequence is analyzed all the way through to the final significance test. In the work described above, we set this final threshold to accept scores with a P value of 10^{-4} or less. This value seems to be close to the optimum for the *S. typhimurium* data, but substantial changes in the value have little effect on the overall error rate (table 7). That is, there seem to be relatively few marginal cases. This has the fortunate consequence that there is no need to readjust the threshold for new organisms, which is important in the analysis of new genomes, for which there are no annotations with which to assess accuracy and find an optimal value. In more general terms, the use of a Karlin-Altschul P value is a convenient heuristic for defining a cutoff score; CRITICA does not depend on the value being a literal probability estimate. In this vein, we note that the score adjustments for moving the start and end points are not covered by Karlin-Altschul statistics. However, when we consider cases with a low P value ($P \ll 1$), the use of log-frequency ratios for these score adjustments is mathematically equivalent to calculating a Bayesian posterior probability ratio of noncoding versus coding, with the Karlin-Altschul P value as the prior probability and the score adjustments being the conditional probabilities of observing the new data given the alternative hypotheses (noncoding or coding). Regardless of this qualitative argument, our modifications to the HSS scores are not covered by the maximal-segment analysis model, and although we refer to the P value of a score (to avoid introducing an almost certainly more confusing term), these cannot be interpreted as literal probabilities of a sequence region being noncoding.

When analyzing highly novel genomic DNAs, one of the limiting factors is the ability to find homologs of the query sequence. We explored two strategies to improve this situation. First, the comparative analysis component of CRITICA currently uses BLASTN to find presumptive homologs. In principle, the same task could be performed by TBLASTX with added sensitivity for finding homologs in protein-coding regions. Because evaluation is based on the relationship between nucleotide divergence and amino acid divergence, this should increase the available signal (by bringing in more distantly related homologs) without increasing the false positives. Although we have carried out preliminary tests of this strategy, the database search time was prohibitive for routine use. The second approach that we explored is to take advantage of the fact that comparisons within a genome often reveal paralogous genes that contribute to the comparative analysis. The structure of CRITICA allows comparative analysis data from any number of sequence similarity searches to be combined. Thus, the comparative component of the analysis of a novel genome can include searches for related sequences in the public DNA databases, in the genome itself, and in any other locally available DNA data.

In its present form, most of CRITICA's errors are due to entirely missing some coding sequences. Specifically, of the total error rate of 0.0146 per triplet evaluated (table 6), 0.0088 is due to missing coding regions (74 regions averaging 78 amino acids), 0.0028 is due to asserting the existence of unannotated coding regions (215 regions averaging 72 amino acids), 0.0024 is due to late start site calls, 0.0001 is due to early start site calls, and 0.0005 is due to errors that we classified as different-frame. The systematic tendency to start coding sequences later than the annotations suggest may be due in part to the tendency of codon usage to differ in the early parts of genes (Bulmer 1988), although this re-

quires further investigation. When GenMark was iteratively trained with makemat, the increase in false positives relative to CRITICA was almost equally distributed among prediction of additional coding regions (117 regions averaging 65 amino acids) and inclusion of extra upstream sequences. There was also a small increase in false negatives due to missing entire coding sequences (281 regions averaging 62 amino acids). WebGeneMark was worse than CRITICA in all components of the error, but most of the increase was due to completely missing coding regions (281 regions averaging 91 amino acids).

There are several features of CRITICA that would be particularly fruitful for additional development. One is incorporation of an improved model for the Shine-Dalgarno sequence, which is currently very relaxed. Preliminary explorations in this area have shown that many seemingly reasonable combinations of sequence and placement of the Shine-Dalgarno sequence are rarely, if ever, used (unpublished data). A second area for possible improvement is the introduction of more sophisticated scoring of comparative data. In particular, it might be preferable to assign positive comparative scores to conservative amino acid changes. For the moment, this has not been done because the additional information would be most evident when analyzing distantly related sequences, yet these are not always found by BLASTN. Thus, a change in the scoring model is best incorporated in conjunction with the use of a more sensitive search for homologous sequences. Another productive change would be the creation of scoring matrices that are not based on the assumption of equal frequency of codons in noncoding data (in table 1). This would allow more sensitive coding region identification in sequences with a biased G+C content. The treatment of the ends of regions with coding evidence could be changed to introduce two additional features. An alternative to adjusting regions of coding evidence to coincide with initiator and terminator codons would be to consider possible frameshifts as well. In a similar manner, one could permit intron sequences within the coding frame; detecting intron-exon boundaries is an area in which comparative analysis could be more fully exploited. This last feature would be of particularly broad interest and would share some features with programs including GRAIL (Uberbacher, Xu, and Mural 1996), GeneParser (Snyder and Stormo 1995), and the Spliced Alignment algorithm (Gelfand, Mironov, and Pevzner 1996). The modular, score-based approach to the design of CRITICA will facilitate the introduction of these and other features.

Acknowledgments

We are grateful to C. R. Woese, R. Overbeek, G. D. Pusch, C. I. Reich, and L. McNeil for encouragement, useful discussions, and comments on the manuscript. We thank M. Borodovsky for providing copies of the GenMark program and the *S. typhimurium* analysis matrix. J.H.B. received training grant support from the National Institutes of Health (GM07283) and the Department of Education (DE-P200A-40706). Early parts of this work were supported by a Presidential Young Investigator Award (NSF DIR 89-57026) to G.J.O.

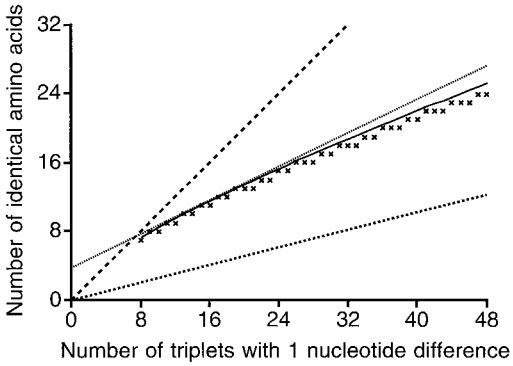

FIG. 2.—Constructions used in the derivation of comparative scores used by CRITICA. The diagram illustrates the analysis of aligned triplets differing by one nucleotide. The elements in the graph are (- - - - -) the line of 100% amino acid identity; (- - - - - - - - -) the line of random identity for triplets differing by one nucleotide (from table 1); (*) the largest number of identical amino acids that is not significantly greater than random at a threshold of $P < 0.0001$, calculated using the exact binomial; (——) an approximation of this significance threshold in the vicinity of 16 triplets based on the normal approximation of the binomial distribution; and (·····) the tangent to (the straight line approximation of) the continuous curve at the point of 16 triplets. Further details and explanations of these lines are in the text.

APPENDIX

Derivation of the Comparative Scores

The comparative scores in table 2 are based on a simple procedure to assign scores for analyzing a given number of aligned triplets, each of which differ by a given number of nucleotides. For concreteness, most of the procedure will be described in the context of inferring the scores for 16 triplets, each of which differs by one nucleotide. In doing so, we will ignore all other types of triplets (those differing by zero, two, or three nucleotides). Similar analyses are performed for triplets differing by two nucleotides and for triplets differing by three nucleotides. Also, the multiple hypothesis testing that is implicit in using maximal-segment analysis is ignored. These simplifications are permissible, since we are only using this procedure to arrive at a useful set of scores; the evaluation of data is based on equation (2).

The first steps in the derivation of the scores are discussed in the context of figure 2. The plot shows the number of identical amino acids (m) among n aligned triplets, each with exactly one nucleotide difference. The accessible portion of the graph is bounded above by $m = n$, the line of complete amino acid identity (- - - - -). The lower line (- - - - - - - - -) is defined by $m = np$, where p is the random frequency of amino acid identity for the given number of nucleotide differences per triplet. From table 1, we have $p = 0.255$ for triplets differing by one nucleotide.

Using maximal-segment analysis to find high-scoring segments and screening the segment scores against a threshold is equivalent to finding all segments that fall above a straight line in figure 2; if we can define the desired line, then we can derive appropriate scores. The

✖ symbols in figure 2 mark the largest values of m that are not significantly greater than random (at $P > 0.0001$), as calculated using the one-tail binomial distribution for the given values of n and p. Thus, for a given n, any greater value of m would be considered significant. The continuous curve (——) approximates this significance threshold in the vicinity of 16 triplets. This line is defined by the number of identical amino acids giving 4.3 standard deviations greater than random identity according to the normal approximation of the binomial distribution function. More precisely, the equation of the line is

$$m = np + Z\sigma,$$

where $\sigma = \sqrt{np(1 - p)}$ and $Z = 4.3$. Z is an adjustable parameter whose value was chosen to position the curve just above the ✖ symbols in the vicinity of $n = 16$. Thus, for a given value of n close to 16, any value of m above this curve is significant, and there are few, if any, significant values of m that are not above the curve. Because maximal-egment analysis efficiently finds combinations of n and m above a straight line, we use the tangent at $n = 16$ triplets (·····) as the best straight-line approximation of the curve. In general, the equation of the tangent line at $n = n_0$ to the curve for Z standard deviations above random is

$$m = an + b,$$

where

$$a = p + \frac{Z}{2}\sqrt{\frac{pq}{n_0}}$$

$$b = \frac{Z}{2}\sqrt{n_0 pq}$$

All points above this line are significant, and in the vicinity of $n = 16$, few, if any, significant combinations of n and m are missed. If M is the score assigned to triplets encoding identical amino acids, then a scoring scheme based on this tangent line must have a score for differing amino acids (N) given by

$$N = \frac{a}{a - 1}M.$$

Because $0 < a < 1$, it follows that $a - 1 < 0$, and N is negative.

All that remains is to choose the magnitude of M (or N). First, we choose an arbitrary, relatively large value of M, compute the corresponding value of N, and then calculate the Karlin and Altschul (1990) parameter λ for a scoring scheme of only two possible outcomes: score M with probability p, and score N with probability $1 - p$. Up to this point, triplets with one, two, and three nucleotide differences (for a given n_0) have been analyzed independently; there are three values of M and N, and the resulting three values of λ. Because λ is inversely proportional to the magnitudes of M and N, it is straightforward to adjust the magnitudes of the scores to give approximately equal values of λ, while maintaining sufficiently large scores to avoid large round-off errors when converting M and N to integers. We found it useful to adjust the scores so that each $\lambda \approx 0.015$. The complete process was performed for $n_0 = 8$, 16, 32, 64, and 128, and the resulting values of M and N are entered in table 2. All subsequent evaluations of high-scoring segments use only the values of the scores; they are independent of the calculations of λ and any simplifying assumptions made in this appendix.

LITERATURE CITED

ALTSCHUL, S. F. 1993. A protein alignment scoring system sensitive at all evolutionary distances. J. Mol. Evol. **36**:290–300.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

BLATTNER, F. R., G. PLUNKETT, C. A. BLOCH III, et al. (14 co-authors). 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277**:1453–1462.

BULMER, M. 1988. Codon usage and intragenic position. J. Theor. Biol. **133**:67–71.

BORK, P., and A. BAIROCH. 1996. Go hunting in sequence databases but watch out for the traps. Trends Genet. **12**:425–427.

BORODOVSKY, M., and J. MCININCH. 1993. GenMark: parallel gene recognition for both DNA strands. Comput. Chem. **17**: 123–133.

BORODOVSKY, M., J. MCININCH, E. V. KOONIN, K. E. RUDD, C. MÉDIGUE, and A. DANCHIN. 1995. Detection of new genes in a bacterial genome using Markov models for three gene classes. Nucleic Acids Res. **23**:3554–3562.

BULT, C. J., O. WHITE, G. J. OLSEN et al. (37 co-authors). 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science **273**:1058–1073.

BURGE C., and S. KARLIN. 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268**:78–94.

CLAVERIE, J. M., and L. BOUGUELERET. 1986. Heuristic informational analysis of sequences. Nucleic Acids Res. **14**:179–196.

DECKERT G., P. V. WARREN, T. GAASTERLAND et al. (12 co-authors). 1998. The complete genome of the hyper-thermophilic bacterium *Aquifex aeolicus*. Nature **392**:353–358.

FICKETT, J. W., and C. TUNG. 1992. Assessment of protein coding measures. Nucleic Acids Res. **20**:6441–6450.

FLEISCHMANN, R. D., M. D. ADAMS, O. WHITE et al. (37 co-authors). 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269**:496–512.

FRASER, C. M., J. D. GOCAYNE, O. WHITE et al. (26 co-authors). 1995. The minimal gene complement of *Mycoplasma genitalium*. Science **270**:397–403.

GELFAND, M. S., A. A. MIRONOV, and P. A. PEVZNER. 1996. Gene recognition via spliced sequence alignment. Proc. Natl. Acad. Sci. USA **93**:9061–9066.

GISH, W., and D. J. STATES. 1993. Identification of protein coding regions by database similarity search. Nat. Genet. **3**: 266–272.

HIMMELREICH, R., H. HILBERT, H. PLAGENS, E. PIRKL, B. C. LI, and R. HERRMAN. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. **24**:4420–4449.

KANEKO, T., S. SATO, H. KOTANI et al. (21 co-authors). 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. **3**:109–136.

KARLIN, S., and S. F. ALTSCHUL. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA **87**:2264–2268.

———. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. Proc. Natl. Acad. Sci. USA **90**:5873–5877.

KARLIN, S., and A. DEMBO. 1992. Limit distributions of maximal segmental score among Markov-dependent partial sums. Adv. Appl. Prob. **24**:113–140.

KLENK, H. P., R. A. CLAYTON, J. F. TOMB et al. (48 co-authors). 1997. The complete genome sequence of the hyperthermophilic, sulfate-reducing archeon *Archaeoglobus fulgidus.* Nature **390**:364–370.

MÉDIGUE, C., T. ROUXEL, P. VIGIER, A. HÉNAUT, and A. DANCHIN. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. J. Mol. Biol. **222**:851–856.

SALZBERG, S. L., A. L. DELCHER, S. KASIF, and O. WHITE. 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. **26**:544–548.

SHINE, J., and L. DALGARNO. 1974. The 3′ terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc. Natl. Acad. Sci. USA **71**:1342–1346.

SNYDER, E. E., and G. D. STORMO. 1995. Identification of protein coding regions in genomic DNA. J. Mol. Biol. **248**:1–18.

STADEN, R., and A. D. MCLACHLAN. 1982. Codon preference and its use in identifying protein coding regions in long DNA sequences. Nucleic Acids Res. **10**:141–156.

STATES, D. J., and W. GISH. 1994. Combined use of sequence similarity and codon bias for coding region identification. J. Comput. Biol. **1**:39–50.

UBERBACHER, E. C., Y. XU, and R. J. MURAL. 1996. Discovering and understanding genes in human DNA sequence using GRAIL. Methods Enzymol. **266**:259–281.

WOOTTON, J. C., and S. FEDERHEN. 1993. Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. **17**:149–163.

STANLEY A. SAWYER, reviewing editor

Accepted January 7, 1999