Molecular Sequence Accuracy and the Analysis of Protein Coding Regions

DJ States, and D Botstein

PNAS 1991;88;5518-5522
doi:10.1073/pnas.88.13.5518

Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 5518–5522, July 1991
Genetics

Molecular sequence accuracy and the analysis of protein coding regions

(sequence errors/Bayesian alignment/codon usage)

DAVID J. STATES* AND DAVID BOTSTEIN†‡

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; and †Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305

Contributed by David Botstein, February 13, 1991

ABSTRACT **Molecular sequences, like all experimental data, have finite error rates. The impact of errors on the information content of molecular sequence data is dependent on the analytic paradigm used to interpret the data. We studied the impact of nucleic acid sequence errors on the ability to align predicted amino acid sequences with the sequences of related proteins. We found that with a simultaneous translation and alignment algorithm, identification of sequence homologies is resilient to the introduction of random errors. Proteins with >30% sequence identity can be reliably recognized even in the presence of 1% frameshifting (insertion or deletion) error rates and 5% base substitution rates. Incorporation of prior knowledge about the location and characteristics of errors improves tolerance to error of amino acid sequence alignments. Similarly, inclusion of prior knowledge of biased codon utilization by yeast (*Saccharomyces cerevisiae*) allows reliable detection of correct reading frames in yeast sequences even in the presence of 5% substitution and 1% frameshift errors.**

Knowledge of the sequences of residues in proteins and nucleic acids is central to most aspects of modern biology. Sequences are determined experimentally; generally the sequence of nucleotides in DNA is determined and the sequence of amino acids in the encoded protein is deduced from the DNA sequence. At present, comparison among the sequences is most commonly done at the protein (i.e., deduced amino acid sequence) level for the purpose of understanding the relationships among proteins (and their functions) in the same and in different species.

As experimental data, sequences are subject to errors and uncertainties. Errors can arise at many points: in the manipulations required to obtain the DNA clones; during the actual nucleotide sequence determination; in the assembly of stretches of sequence into a continuous whole; and finally in data entry, handling, and storage. Although some level of error in sequences seems inevitable, different sequencing strategies may have different intrinsic error rates and types (1–4). For many reasons, including cost, it therefore seems appropriate to consider the impact errors will have on the uses to which the sequences will be put.

In an analysis of the impact of errors in nucleotide sequences it is vital to consider not only the frequency of errors but also their type (substitution, deletion, or insertion of one or more bases). In particular, derived amino acid sequences are very differently affected by errors of different types. The requirement for a reading frame largely or entirely free of insertions or deletions (i.e., "frameshift" errors) along its entire length places an apparently stringent limit on the number of errors that can be tolerated in the underlying sequence data used to recognize homologies. Yet in the limit it seems clear that sequences identical except for even a large number of frameshift errors should be recognizable as essentially identical if all three reading frames were tested for similarity at all sites: the correct reading frame would rise above statistical significance whenever runs of a few identical amino acid residues were found. These considerations suggest that a suitable set of algorithms and strategies might indeed allow sensitive detection of amino acid similarity even in sequences containing high error rates.

The impact of errors on detection of reading frames and/or sequence similarities might also be modified if one knew in advance that the errors were not uniformly distributed—that some regions of a given sequence were much more likely to contain errors than others. One might then incorporate this knowledge into a Bayesian interpretation strategy. Similarly, other prior knowledge about the sequence, such as known empirical biases in codon usage in the organism, might be incorporated into Bayesian algorithms aimed at detecting reading frames. We therefore focus on the effect of different frequencies and types of error upon the most common use of sequence data: determination of the amino acid sequences of proteins and detection of similarities among derived amino acid sequences by standard sequence similarity searches.

METHODS

To search for coding regions similar to a target peptide sequence in untranslated nucleic acid sequence, a two-stage Bayesian algorithm was used. For error-prone molecular sequence data, this theorem allows us to calculate the probability of the hypothesis that a particular amino acid was coded at a particular site in the sequence in terms of the known genetic code and the uncertainties in the data. In a second step, the probability of a sequence alignment may also be calculated given this set of probabilities for coding for each amino acid at each site in a sequence. With the use of a table of amino acid exchange probabilities (5), Bayes' theorem allows us to calculate the probability of the hypothesis that an amino acid in the target sequence is at that site.

In our calculations, uncertainties were assigned for each base in the nucleic acid sequence as well as the probability that an insertion or deletion of a single base would occur at that location in the nucleic acid sequence. At the sites of insertion mutations, all four bases were inserted with equal probability. For deletion mutations, no bias was assumed in the identity of the deleted base. The probability of a given codon being accepted at location i in the nucleic acid sequence was the product of the probabilities of the codon bases being present at the required positions:

$$P_{\text{codon}}(i) = P_{\text{b1}}(i)P_{\text{b2}}(i+1)P_{b3}(i+2), \qquad [1]$$

where b1, b2, and b3 are the first, second, and third bases of the codon, respectively.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

‡To whom reprint requests should be addressed.

Separate tables were maintained for the probability of coding for each amino acid with a deletion having occurred within the codon at site i,

$$P_{\text{codon}_{\text{del}}}(i) = \frac{1}{4} P_{\text{del}}(i)P_{\text{b2}}(i)P_{\text{b3}}(i+1) + P_{\text{b1}}(i)\frac{1}{4}P_{\text{del}}(i+1)P_{\text{b3}}(i+1) + P_{\text{b1}}(i)P_{\text{b2}}(i+1)\frac{1}{4}P_{\text{del}}(i+1), \quad [2]$$

and the probability of coding for each amino acid with an insertion mutation having occurred within the codon,

$$P_{\text{codon}_{\text{ins}}}(i) = P_{\text{ins}}(i)P_{\text{b1}}(i+1)P_{\text{b2}}(i+2)P_{\text{b3}}(i+3) + P_{\text{b1}}(i)P_{\text{ins}}(i+1)P_{\text{b2}}(i+2)P_{\text{b3}}(i+3) + P_{\text{b1}}(i)P_{\text{b2}}(i+1)P_{\text{ins}}(i+2)P_{\text{b3}}(i+3). \quad [3]$$

The probability of coding for an amino acid in a given nucleic acid sequence was then calculated as the sum of the probabilities of all the codons that could code for that amino acid:

$$P(\text{aa}|\text{seq}_{\text{DNA}}) = \sum_{\text{codon}\Rightarrow\text{aa}} P_{\text{codon}}(\text{seq}_{\text{DNA}}). \quad [4]$$

In the alignment phase of the algorithm the target protein sequence was matched with the table of coding probabilities to obtain the most likely alignment. The probability, P_a, of an alignment was calculated as the product over all the sites in the protein sequence of the probability, P_c, that a given amino acid was coded at the site and the probability, P_s, that amino acid would substitute for the amino acid present in the protein sequence, based on the PAM250 amino acid similarity matrix (5), although other scoring systems could be used (6).

$$P_a = \prod_{\text{seq}_{\text{prot}}} \prod_{\text{aa}_{\text{coded}}} P_s(\text{aa}_{\text{obs}}|\text{aa}_{\text{coded}})P_c(\text{aa}_{\text{coded}}|\text{seq}_{\text{DNA}}). \quad [5]$$

Logarithmic transformation reduced these products to sitewise separable sums, allowing standard dynamic programming algorithms to be applied. A Smith–Waterman dynamic programming algorithm (7) was used to align the protein coding probability to the target protein sequence. In this algorithm, the alignment is represented as a path through a lattice. As shown in Fig. 1, five possible moves were considered at each site in the dynamic programming lattice: introducing or extending a gap in the protein sequence; introducing or extending a gap in the nucleic acid sequence; matching a codon to an amino acid; matching a codon that assumes a deletion to an amino acid; matching a codon that assumes an insertion to an amino acid.

Gaps introduced in alignment were scored independently of single base insertions or deletions considered in codon probabilities, and standard penalties were used for the introduction and extension of gaps in aligned sequences (5). Separate paths and scores were kept for leaving each alignment lattice point along a diagonal, protein gap, or nucleic acid gap. In this way, penalties for introducing gaps into one or the other of the sequences could be separated from the penalty for extending the gap (8, 9). To assess the expected effects of random mutations, multiple trial alignments (typically 100) were performed.

The identification of reading frame by codon utilization was performed by first computing tables of species-specific codon utilization, using all complete coding regions in GenBank Release 62. A similar table of noncoding-region triplet utilization frequencies was compiled from species-specific intervening sequences. For each 90-base segment of genomic sequence, the probability of a reading frame was calculated as the product of the probability that each codon in that reading frame was drawn from the coding pool of codons and the probability that the codon was drawn from the noncoding pool. Probabilities were calculated for all three reading frames and for no reading frame (noncoding) status. A segment was correctly categorized if the true reading frame was the most probable source of the codons in the segment. The confidence of assignment was determined by the probability of the correct assignment relative to the next best alternative.

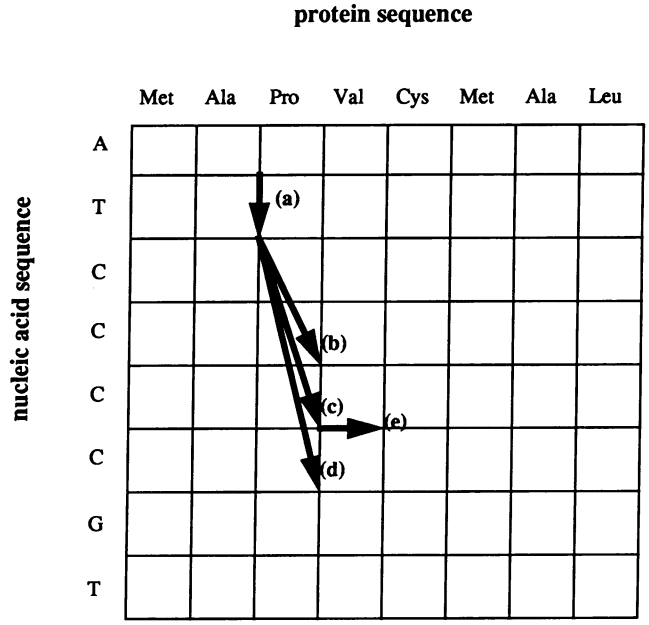

FIG. 1. Dynamic programming alignment of a protein and nucleic acid sequence. The nucleic acid sequence is on the vertical axis and the protein sequence on the horizontal axis of a lattice. An alignment can be represented as a path through this lattice. The five steps considered in the algorithm include (a) introducing a gap in the protein sequence, (b) aligning an amino acid to a codon in which a deletion has occurred, (c) aligning an amino acid to an intact codon, (d) aligning an amino acid to a codon in which an insertion mutation has occurred, and (e) introducing a gap in the nucleic acid sequence. Steps b, c, and d correspond to aligning the proline with 2, 3, or 4 cytosines. A cost or score is associated with each step based on the probabilities of coding for each amino acid in the nucleic acid sequence and the probability that those amino acids would align with the residue in the protein sequence. The optimal alignment is the path through the lattice with the highest score.

Protein sequences for rat trypsin (10), human neutrophil elastase (11), and schistosomal elastase (12) were obtained from the Protein Information Resource database. The mRNA sequence for rat trypsin (10) was obtained from GenBank.

RESULTS

We chose sequence examples from the serine protease family spanning a range of divergence: rat trypsin, human neutrophil elastase, and schistosomal elastase. With fractions of identity generated by the FASTA program (13), rat trypsin and human neutrophil elastase show 33% identity; the schistosomal elastase is more distantly related to the others: 19% and 23% identity to rat trypsin and human elastase, respectively.

Sequence alignment depends on information present in regions of similarity interspersed with regions of greater difference. The dispersed and separable nature of the information content in this alignment suggests that the sequence

alignment process itself may be more resilient to the effects of frameshifting errors than is simple translation. To test this hypothesis, the Bayesian algorithm described above was used to simultaneously translate and align nucleic acid sequences with protein sequences.

The effects of substitution mutations alone were considered first. The rat trypsin, human neutrophil elastase, and schistosomal elastase protein sequences were each compared with the rat trypsin mRNA sequence in the presence of varying levels of substitution error. One hundred random trials, each with a new set of errors, were run for each sequence pair at each error rate. An equal number of alignments were performed against randomly jumbled sequences having the same base composition as the rat trypsin mRNA sequence. The distributions of alignment scores for each sequence pair were plotted as a histogram (Fig. 2). When the sequences are distantly or closely related (trypsin vs. trypsin or trypsin vs. neutrophil elastase), it is possible to discriminate between the significant and random score distributions even when large numbers of substitution errors are present. For example, the true positive trypsin/neutrophil elastase alignments can be distinguished from random alignments even with substitution errors at the rate of 10% in the nucleic acid sequence. Intuitively, one might think of the errors increasing the divergence from about 150 base changes per hundred amino acids to about 160, causing almost no change in alignment significance. As might be expected, when the sequences have little underlying similarity to begin with (trypsin vs. schistosomal elastase), even low rates (1–2%) of base substitution error essentially prevent discrimination between true positive and random sequence alignments.

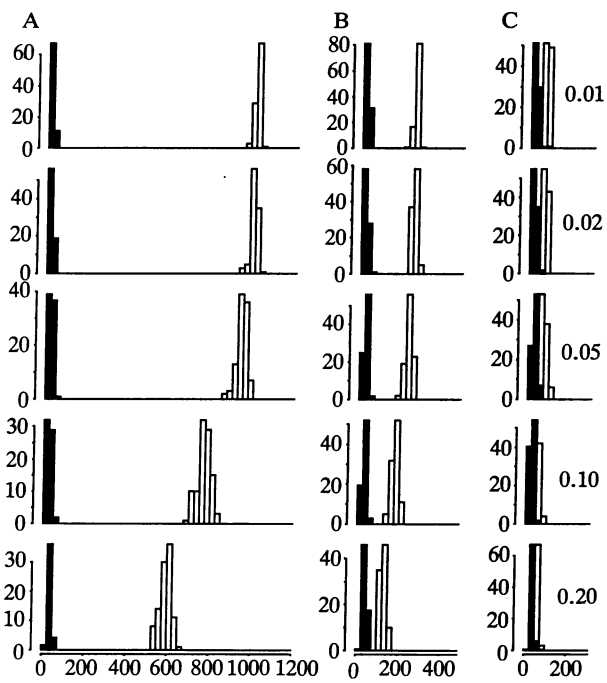

FIG. 2. Effect of substitution errors on sequence similarity alignment. In each panel, alignment scores were determined for 100 trials in which the nucleic acid sequence was randomly mutated at the error rate indicated at right (0.01–0.20), introducing only single base substitution errors. The error rates were incorporated as prior knowledge in the Bayesian algorithm. One hundred additional trial alignments were conducted against random sequence with the same base composition. The distributions of alignment scores for three pairs of sequences are presented: rat trypsin protein and mRNA sequence (*A*); human neutrophil elastase protein and rat trypsin mRNA sequence (*B*); schistosomal elastase protein and rat trypsin mRNA sequence (*C*). Open bars, true positive scores; filled bars, random sequence scores.

Insertion and deletion errors degrade the significance of alignments far more rapidly than substitution errors alone. Fig. 3 demonstrates the effect of uniformly distributed insertion and deletion errors on the alignment of human neutrophil elastase with the rat trypsin gene, performed as above. The results are plotted as histograms showing the score distributions at increasing insertion/deletion error rates. Small numbers of insertion/deletion errors (1%) degrade the score of the alignment against the true sequence, but at these error rates, discrimination between random and true positive alignments is preserved. When the insertion/deletion probability is increased, the significance of the alignments deteriorates rapidly; insertion/deletion rates of 2% completely abolish the ability to recognize true positive alignments between trypsin and neutrophil elastase (Fig. 3).

These results should be compared with the degradation of the ability to maintain reading frame in the presence frameshifting errors. Based on Poisson statistics, and in the limit of low error rates, the probability of preserving an open reading frame is the product of the probabilities of not having a frameshift mutant at each site along the sequence. For a protein of 330 amino acids, an insertion or deletion error rate of 0.001 would result in a 0.63 chance of at least one insertion occurring and disrupting the translation. As we have seen in Fig. 3, in cases of ≥33% underlying amino acid sequence identity, alignment can still be successfully carried out in the presence of 10-fold higher rates of frameshift error.

If information is available on the location of sites that are more or less likely to undergo frameshift errors, a Bayesian alignment algorithm can be devised to use this knowledge, as shown in Fig. 4. The distribution of alignment scores between the neutrophil elastase amino acid sequence and the rat trypsin mRNA sequence with no prior knowledge of error location is compared with the alignment score distributions when insertion and deletion errors are known to be confined to specific regions of sequence. Without prior knowledge of error location, discrimination between the true alignments and the random sequence alignments is not possible at the 2% frameshift error rate. When the same number of frameshifting errors are known to be confined to two-thirds of the sequence

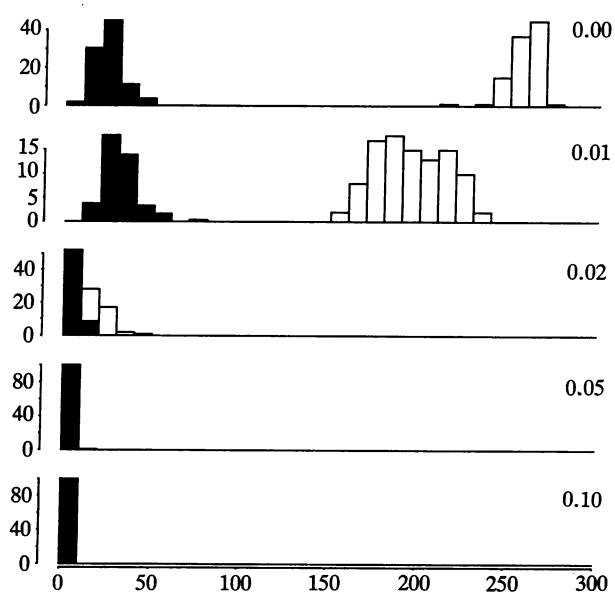

FIG. 3. Effect of introducing insertion/deletion errors on the distribution of alignment scores for the human neutrophil elastase protein sequence aligned with the rat trypsin mRNA sequence. Method was as for Fig. 2, with the frequencies of insertion and deletion error introduction varied from 0 to 10 per 100 bases, as indicated at right. Substitution errors were introduced at a fixed rate of 1 per 100 bases in each case.

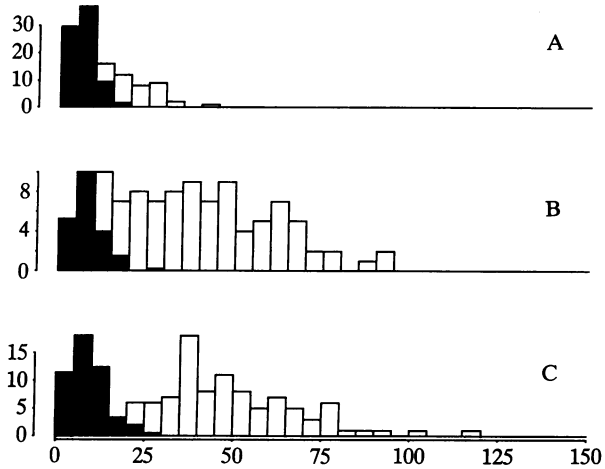

FIG. 4. Effect of incorporating limited prior knowledge about the location of insertion/deletion errors on alignment score. Method was as for Figs. 2 and 3, with insertion and deletion errors introduced at a rate of 2 per 100 bases and substitution errors introduced at a rate of 1 per 100 bases. (*A*) Distribution of scores when insertion/deletion errors are randomly distributed throughout the sequence and there is no prior knowledge of their location. (*B*) Distribution of scores as above but with prior knowledge that 40% of the sequence is insertion/deletion error-free. (*C*) Distribution of scores with prior knowledge that 64% of the sequence is free of insertion/deletion errors.

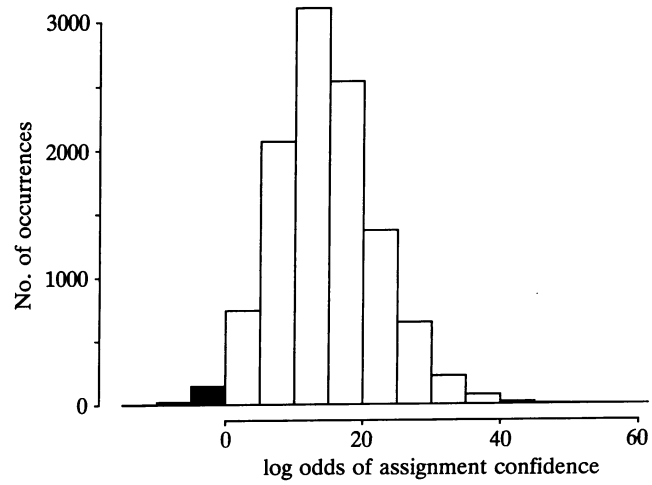

FIG. 5. Distribution of the logarithm of odds scores for correctly assigning the coding status and reading frame of sequence segments based on codon usage. All complete yeast coding sequences were taken from GenBank release 63 and divided into segments 90 bases in length. For each reading frame in each segment, the probability of drawing the observed set of codons from coding vs. noncoding pools of sequence were calculated and converted to logarithmic units. The segment was assigned to the category (coding vs. noncoding, and if it was coding, to a particular reading frame) most likely to have been the source of the 30 observed base triplets. The ratio of codon probability based on the correct reading frame relative to the probability of the best alternative is the confidence with which a segment may be assigned. The distribution of assignment confidence for all 90-base segments derived from yeast coding sequences in the GenBank database is shown. The abscissa is the natural logarithmic units of the reading-frame assignment confidence. The black area indicates segments not assigned to correct reading frame.

with one-third insertion- and deletion-free, some degree of discrimination is possible. When two-thirds of the sequence is known *a priori* to be insertion- and deletion-free, the alignment score distributions have little overlap and reliable discrimination between the true positive and random alignments becomes possible. It is significant that the scores for similarity to the random-sequence remain the same; it is the scores for the true positives that improve with application of the prior knowledge of the location of the frameshift errors.

The identification of coding regions by similarity to homologs depends on the presence of previously sequenced homologs in the database, but coding regions do have a number of statistical features which suggest that it may be possible to recognize them in the absence of a homolog query (14). To examine the reliability with which coding regions can be recognized on the basis of codon utilization statistics, four separate databases containing all yeast or human coding or intervening sequences were prepared from GenBank Release 63. Sequences in each of these databases were divided into 90-base segments and the codon utilization frequencies in these segments were compared to reference in-frame codon frequencies for these species and to average noncoding region base-triplet frequencies. The probability of a segment being a coding region in frame was calculated as the product of the probability of drawing each of its 30 triplets from the codon pool. The probability of a segment being a noncoding region was calculated as the probability of drawing each of its base triplets from the noncoding pool. Fig. 5 shows a histogram of the log-probability scores for assigning yeast coding regions correctly as coding with a particular reading frame. The vast majority of segments are assigned correctly; the mean probability indicates a high degree of confidence in the assignments.

The results for coding-region assignment by this method depend on the strength of the codon usage bias, which varies from species to species. Some species, such as *Saccharomyces cerevisiae* and *Escherichia coli*, have strong biases and many very rarely used codons (15, 16), while mammalian species typically have comparatively weak biases (17). Some results for reading-frame assignment based on codon usage statistics in the presence of sequence errors are shown in Table 1. For yeast, with a strong codon bias, most segments can be correctly assigned both with respect to coding status and reading frame even in the presence of a high substitution rate (5%). For human sequences, with less codon usage bias, fewer segments are assigned correctly, and the process is more sensitive to sequence errors.

The reading frame for a segment containing an insertion or deletion error is not defined. The probability of a segment of 90 bases suffering a frameshifting error is 84% for when the insertional error rate is 2% per base. The probability of such an error falls to 36% when the insertional error rate is 0.5% per base. Therefore, segmental assignment of reading frame remains relatively sensitive to the presence of insertion and deletion errors in the sequence data because definition of a correct reading frame depends on the absence of frameshifting errors in the segment. It remains to be determined whether there might be an optimum segment size smaller than 90 residues that would maximize assignment of coding regions and reading frames in the presence of high rates of frameshift errors.

DISCUSSION

We have shown that sequence alignments at the amino acid level are quite insensitive to substitution errors (up to 5%) and even frameshift errors (up to 1%) when one uses a Bayesian probabilistic approach, provided that the underlying similarity in the protein sequence is about 33% amino acid identity or more. Further, we have shown that prior knowledge of the location of the errors can be used to produce better alignment in the presence of even higher error rates. Finally, we have indicated that Bayesian algorithms including other prior knowledge (such as codon usage bias) can be used to find the proper reading frames even in the presence of large numbers of errors.

Table 1. Reliability of reading-frame assignment based on codon usage

	Error-free sequence				Sequence with 5% substitution errors			
	Yeast		Human		Yeast		Human	
Significance group	No. of segments	Fraction	No. of segments	Fraction	No. of segments	Fraction	No. of segments	Fraction
	Segments statistically assignable							
$P < 0.01$	3,563	0.65	5,198	0.42	12,070	0.55	18,552	0.37
$P < 0.05$	4,303	0.79	7,202	0.58	15,518	0.71	27,246	0.55
NS	1,158	0.21	5,280	0.42	6,326	0.29	22,682	0.45
	Accuracy of assignment							
$P < 0.01$	3,529	0.99	5,013	0.96	11,867	0.98	17,634	0.95
$P < 0.05$	4,230	0.98	6,741	0.94	14,974	0.96	24,695	0.91
NS	841	0.73	3,165	0.60	4,138	0.65	12,931	0.57
Total	5,071	0.93	9,906	0.79	19,112	0.87	37,626	0.75

For the indicated species, all of the coding and all of the intervening sequences in GenBank Release 63 were extracted and segmented in runs of 90 bases. The probability of drawing the 30 triplets in each reading frame of each segment was computed based on the assumption that the segment was a noncoding region or that the segment was a coding region and these codons were in the correct reading frame. The segment was assigned to a category (coding vs. noncoding, and if coding, to a reading frame) based on which category/reading frame yielded the most probable set of triplets. Presented are an analysis of error-free yeast and human gene sequence segments and an analysis of segments that have been subject to four trials of random substitution errors at 5% of the sites. In each case, the fraction of the sequence segments that can be assigned to a coding status is presented for various levels of confidence. $P < 0.01$ implies that the log odds score for one reading frame or noncoding status was >4.6 log units more favorable than any alternative; $P < 0.05$ implies that the log odds score for one assignment was >3 log units more favorable than any alternative. In the group labeled nonsignificant, no assignment was >3 log units more likely than any other. The actual accuracy of assignment is then presented for each of these significance groups. The number and fraction of the segments in the group that were correctly assigned are listed.

From the point of view of mathematics, none of these results is surprising. The substitution errors act simply to reduce the degree of similarity; it is not surprising that an additional 1–5% differences make little difference in sequences whose similarity is statistically secure at 33% identity or more. Frameshifting errors degrade the alignment score more quickly, but in an entirely predictable way. In the trypsin coding region (720 bases) a 1% frameshift error rate results in an average of 7 errors and thus 7 gap-opening penalties, degrading the score by about 80 points, close to the mean shift we observed. When the insertion/deletion error rate is increased from 1% to 2%, the mean alignment score for an unbroken segment becomes comparable to the gap-opening penalty, and this explains the dramatic loss of alignment significance that is observed.

Incorporation of prior knowledge into sequence analysis is also expected, on mathematical grounds, to improve the accuracy of the alignment process. We found significant improvement in discrimination between true positive and random alignments when prior knowledge of the possible locations of insertion/deletion errors was incorporated into the calculation. This result has a practical implication: sequence databases should, wherever practicable, incorporate information about the distribution of possible errors.

Prior knowledge can be used in other ways. We found that we could use codon usage bias to predict correctly coding frames in yeast, which has a highly unusual codon usage bias. This was less successful with human sequences, which have less bias and a smaller fraction of coding sequences in their genome. Nevertheless, the Bayesian principle might be used with other kinds of information, such as splice junction sequences in the case of coding sequence determinations.

In conclusion, our results suggest that relatively low-accuracy sequence data (i.e., up to 5% substitution and 1% insertion/deletion) are surprisingly useful for the detection of protein similarities even between quite distantly related proteins. This value needs to be measured against the cost of making the sequence more accurate. We suggest that low-accuracy sequence can also be used to open the door to higher accuracy sequencing through PCR and primer-directed sequencing strategies. Despite this, our view remains that the goal in molecular sequencing is as high an accuracy as can be practically and economically be achieved.

We thank Drs. David Lipman, David Landsman, and Mark Boguski for critical reviews of our manuscript and the DCRT/IBM AIX/370 project at the National Institutes of Health for generous use of computing facilities.

1. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
2. Church, G. M. & Kieffer-Higgins, S. (1988) *Science* **240,** 185–188.
3. Tabor, S. & Richardson, C. C. (1990) *J. Biol. Chem.* **265,** 8322–8328.
4. Drmanac, R., Labat, I., Brukner, I. & Crkvenjakov, R. (1989) *Genomics* **4,** 114–128.
5. Dayhoff, M. O., Barker, W. C. & Hunt, L. T. (1983) *Methods Enzymol.* **91,** 524–545.
6. Smith, T. F., Waterman, M. S. & Fitch, W. M. (1981) *J. Mol. Evol.* **18,** 38–46.
7. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195–197.
8. Sellers, P. H. (1974) *Appl. Math.* **26,** 787–793.
9. Sellers, P. H. (1974) *J. Comb. Theory Ser.* **16,** 253–258.
10. Craik, C. S., Choo, Q. L., Swift, G. H., Quinto, C., MacDonald, R. J. & Rutter, W. J. (1984) *J. Biol. Chem.* **259,** 14255–14264.
11. Sinha, S., Watorek, W., Karr, S., Giles, J., Bode, W. & Travis, J. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 2228–2232.
12. Newport, G. R., McKerrow, J. H., Hedstrom, R., Petitt, M., McGarrigle, L., Barr, P. J. & Agabian, N. (1988) *J. Biol. Chem.* **263,** 13179–13184.
13. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
14. Fickett, J. W. (1982) *Nucleic Acids Res.* **10,** 5303–5318.
15. Guthrie, C. & Abelson, J. (1982) *The Molecular Biology of the Yeast Saccharomyces, Metabolism and Gene Expression,* eds. Starthern, J. N., Jones, E. W. & Broach, J. R. (Cold Spring Harbor Lab., Cold Spring Harbor, NY).
16. Ikemura, T. (1982) *J. Mol. Biol.* **158,** 573–597.
17. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. (1981) *Nucleic Acids Res.* **9,** r43–r74.