GENOMICS 14, 89–98 (1992)

The Accuracy of DNA Sequences: Estimating Sequence Quality

GARY A. CHURCHILL * AND MICHAEL S. WATERMAN †

*Biometrics Unit, 337 Warren Hall, Cornell University, Ithaca, New York 14853; and †Departments of Mathematics and Molecular Biology, University of Southern California, Los Angeles, California 90089-1113

Received February 7, 1991; revised March 20, 1992

In this paper we describe a method for the statistical reconstruction of a large DNA sequence from a set of sequenced fragments. We assume that the fragments have been assembled and address the problem of determining the degree to which the reconstructed sequence is free from errors, i.e., its accuracy. A consensus distribution is derived from the assembled fragment configuration based upon the rates of sequencing errors in the individual fragments. The consensus distribution can be used to find a minimally redundant consensus sequence that meets a prespecified confidence level, either base by base or across any region of the sequence. A likelihood-based procedure for the estimation of the sequencing error rates, which utilizes an iterative EM algorithm, is described. Prior knowledge of the error rates is easily incorporated into the estimation procedure. The methods are applied to a set of assembled sequence fragments from the human G6PD locus. We close the paper with a brief discussion of the relevance and practical implications of this work. © 1992 Academic Press, Inc.

1. INTRODUCTION

Central DNA sequence databases came into existence about 1982 and have undergone very rapid growth. Today the three major databases, DDBJ, EMBL, and GenBank, which contain about 50×10^6 nucleotides, are virtually identical for all practical purposes and are positioning themselves for the era of genomic sequencing. See Kahn and Cameron (1990) for a discussion of EMBL and Burks *et al.* (1990) for a discussion of GenBank. In Waterman (1990), a number of issues regarding genomic sequence databases are raised. The present paper is focused on one of these issues: DNA sequence accuracy. It is commonly understood that the existing data are not of uniform accuracy, but generally this is not apparent from the database entries. In fact, what is acceptable accuracy and how to estimate accuracy are the subjects of considerable recent debate (Roberts, 1990). Sequences appear as AGCTG . . . , for example, with no comments about their accuracy. In this paper we propose a method for estimating DNA sequence accuracy from the assembled fragments of a shotgun sequencing project or of any sequencing method that relies on redundant sequencing to achieve accurate finished sequence. The result of our analysis is an estimate of the likelihood of the base in each position or a simple estimate of an entire consensus sequence at a given level of confidence. While the reliability of the consensus or finished sequence usually exceeds that of an individual fragment, it is valuable to quantify this. Similar statistical approaches could be taken to determine the accuracy of other types of physical map data. See Michiels *et al.* (1987), Lander and Waterman (1988), Branscomb *et al.* (1990), and Balding and Torney (1991) for overlap likelihood estimates for physical mapping with fingerprints from clones.

In a shotgun sequencing project, a large DNA molecule is broken into a collection of fragments. The fragments are cloned into a suitable vector and sequenced individually. The fragment sequences are then assembled by determining their relative orientations and overlaps and aligned to form a column-by-column correspondence (Churchill *et al.*, 1990; Kececioglu and Myers, 1990; Staden, 1980). The accuracy of a finished sequence produced by this method will vary across sites. The depth of coverage varies for statistical and biological reasons (Lander and Waterman, 1988; Edwards *et al.*, 1990). It is intuitively obvious, and we shall make this notion precise, that increasing the depth of coverage should improve the accuracy of the finished sequence. The fragment assembly will inevitably contain columns with discrepancies due to errors in the fragment sequences. Sequencing errors will show some clustering due to chance, but errors can also be caused by properties of the sequence itself. Homopolymeric runs and repetitive sequences in general are prone to errors. Sequences with a tendency to form secondary structures are suspected of causing compression errors. Sometimes all the fragments covering a region will have the same orientation, making these strand-specific systematic errors hard to detect.

There are a number of reasons why it is important to assess the accuracy of DNA sequence data. The level of accuracy required may depend on the types of analyses for which the sequence will be used. Sequence compari-

0888-7543/92 $5.00

son algorithms that are robust to small numbers of errors exist. States and Botstein (1991) have studied the problem of locating homologous genes in a database when the query sequence has errors. However, some other aspects of sequence analysis are likely to be very sensitive to errors. The problem of finding coding regions is an important example. Current methods rely in large part on the ability to detect open reading frames (Fields and Soderland, 1990). The effect of a single insertion or deletion error is a frameshift that can mask an open reading frame. Posfai and Roberts (1991) have developed a technique to detect these errors.

The study of natural variation in DNA sequences is another area where accuracy must be considered. It will be important to distinguish genuine variation from sequencing errors. The frequency of polymorphic DNA sites, although it may vary between species and across genomic regions within species, may be on the order of 1 base in 250–2000 (Gusella, 1986). The level of accuracy should be known and accounted for in studies of sequence variation.

Sequence accuracy should be considered in the planning of experiments that require the design of specific oligonucleotides. Oligonucleotides selected from the most reliable regions of a sequence will maximize the probability of success. This probability could be calculated using the summary statistics we describe below. These considerations increase in importance for experiments that require a large number of specific hybridizations.

Finally, to design an effective strategy for genomic sequencing, we must consider the trade-off between accuracy and the cost of redundant sequencing. A reasonable strategy for a large project might be to produce the bulk of the sequence data rapidly and at low cost. Careful resequencing could then be focused on selected regions to attain a desired level of accuracy (E. Chen, personal communication).

We consider DNA sequence accuracy as a function of the redundancy of coverage and the frequencies of random errors in the fragment sequences. We take the term accuracy to mean the probability that a given DNA base or sequence of bases in a finished sequence is identical to the corresponding base in the actual DNA molecule. We assume implicitly that there is a single true sequence and not a population of sequences being studied. The problems of errors that occur at the DNA preparation or cloning stages and of errors in data transcription are not addressed here. Systematic errors that occur in sequence determination are also beyond the current scope of these methods. Thus, the statistics described represent a bound on sequence accuracy. We have made a number of assumptions to facilitate the statistical analysis. As we emphasize, some of these assumptions can be relaxed without difficulty while others present more formidable obstacles to the statistical analysis. These assumptions, numbered A1 to A6, are described before we give an overview of our methods for estimating accuracy. These assumptions are never fully realized in practice, but it is worth emphasizing that if they were, the statistical methods are sure to be valid. Some are more critical to the method's success than others; we list them in order of decreasing effect. Moreover, some of the difficulties are more easily overcome than others. We consider this paper to be a starting point for future work on DNA sequence accuracy.

A1. The starting point of our analysis is the assembled fragment configuration. A crucial assumption is that *the fragment assembly is correct.* We realize that this will never be absolutely true. Some ambiguities will exist in the alignments and the final product may depend on the assembly algorithm and/or the judgment of the investigator. Still, with low error rates in the fragment sequences and moderate to deep coverage, it is possible to obtain highly reliable assemblies. We assume that all fragment sequences have been carefully examined prior to assembly and that all traces of vector have been removed. The investigator may wish to reexamine certain gels that are questionable and correct the fragment sequences. This is expected and acceptable. However, for the purposes of our statistical methods, it is not recommended that fragment sequences be altered to conform with an alignment after the assembly. Altering the fragment sequences in light of the alignment to conform with the "consensus" sequence can distort the estimated error probabilities and lead to unreliable estimates of accuracy.

This is easily seen with an example. If the column consists of AATA, the consensus letter is A. If the column is then "corrected" to AAAA, there is now no way to see that variation existed and no way to estimate the variation. Therefore if the correction of "errors" is based solely on the alignment, the error rates in raw fragments cannot be estimated. In light of the alignment, there will be two types of errors: those that create ambiguity and those that do not. We assume errors that do not create ambiguity are generated by the same mechanisms and occur at the same frequency (in unassembled fragments) as those that do create ambiguity. It is important that we obtain an unbiased estimate of this frequency. If "obvious" errors are corrected in light of the alignment the hidden errors will remain, and we will underestimate their frequency.

A2. We assume that *all regions within a fragment are equally reliable.* However, the resolution of small fragments on a gel is generally better than the resolution of large fragments. Hence, one end of a fragment sequence will be more reliable than the other. It may be possible to explicitly model the decay of accuracy that occurs as the gel is read out to greater lengths. This would allow one to include data that are still informative although less accurate without compromising the quality of the final sequence. A simple solution to this problem would be to use two sets of error rate parameters, one for the first, more reliable part of the sequence and a second set of rates for the more error-prone portions. Another approach would be to utilize a probabilistic representation of the bases in each fragment (see the discussion in Section 8). This is an area for future research.

A3. We assume that *all fragment sequences are equally reliable.* Any number of factors may influence the reliability of individual fragments. If these factors are known to the investigator, a subjective weighting scheme could be imposed on the fragment set. It also may be possible to detect "bad" fragments after the assembly and consensus computations. A quality-of-fragment statistic such as a χ^2 statistic could be computed on the basis of observed and expected numbers of errors in a fragment. However, a fragment sequence should not be removed from the assembly unless it is clearly anomalous, i.e., aligned incorrectly. A multiple testing problem is involved here and we refer the interested reader to Arratia and Gordon (1990) for a discussion of extreme values and the binomial distribution.

A4. We assume that *sequencing errors are independent of their local context,* i.e., that error probabilities depend only on the true base at a position and not on adjacent or more distant bases. This assumption can be relaxed somewhat at the cost of increasing the number of model parameters. Error rates that depend on the bases immediately to their 5′ side in the fragment can be introduced into the model. However, the true situation is likely to be more complex than a simple one-step dependence. For example, compression sites are a common source of errors that may be related to the formation of local secondary structures extending over several bases.

A5. We assume that the *sequencing error rates are constant across the entire sequence.* Constancy could be checked by a sliding window plot of the expected error frequencies. Some clustering of errors can be expected due to chance. The assembly itself may be incorrect in regions with a high rate of errors and should be reexamined. Otherwise, high error rates may be due to some unusual aspect of the local sequence.

A6. We assume that *the composition of the sequence is independent, both of adjacent bases and over large regions.* Markovian dependence between bases is well established, as are local variations in composition (Churchill, 1989). However, the effect of these assumptions on the inferred final sequence is likely to be minor. Additional modeling of the DNA sequence composition could be incorporated, again at the cost of additional parameters.

In the sections below, we begin with a description of the assembled fragments. Then we examine the problem of determining a consensus when the sequencing error rates are known. A consensus distribution that allows one to compute the most probable base at each position (Procedure A, Section 3) or a redundant set of bases that exceeds a specified level of probability (Procedure B, Section 3) is defined. The latter procedure is extended in Procedure C, Section 3, to produce a collection of sequences that constitutes a global confidence interval for the true DNA sequence. Next, we turn to the problem of estimating the sequencing error rates using the method of maximum likelihood. The result is a simple iterative solution (Procedure D, Section 4), which is a special case of the EM algorithm (Dempster *et al.,* 1977). Incorporation of prior knowledge about error rates into the estimation procedure is described. The methods are illustrated with an example from a large sequencing project (Chen *et al.,* 1990).

2. THE ASSEMBLED FRAGMENTS

A set of fragment sequences (fragments are indexed by $j = 1, \ldots, m$) is aligned by some procedure. The result of the alignment procedure is a matrix with m rows. Each row contains the ordered sequence of bases in a particular fragment written in either direct or reverse complemented orientation. Gaps may be inserted internally and each fragment is offset to produce a column-by-column correspondence among the entire set of fragments. The column index $i = 1, \ldots, n$ runs from the leftmost base in the assembly to the rightmost.

Let s_i denote the true state of the DNA sequence corresponding to column i in the fragment assembly. The true state may take any value in the set $\mathcal{A} = \{A, C, G, T, —\}$. The symbol — is included to allow for extra columns in the fragment assembly that do not correspond to any base in the true DNA sequence. The alphabet of redundant bases is defined to be the set of all nonempty subsets of $\mathcal{A}$. A partial notation for this set is given by the standard IUPAC DNA alphabet (Cornish-Bowden, 1985). When a redundant set includes one or more bases and —, we append an * to the IUPAC symbol, i.e., $R^* = \{A, G, —\}$. Redundant bases are used below in the definition of consensus sequences.

The elements of the fragment assembly matrix, denoted by x_{ij}, may take values in the set $\mathcal{B} = \{A, C, G, T, —, X, \phi\}$, where — denotes an internal gap and X denotes any ambiguous determination of a base. The set of ambiguous bases can be expanded, but the methods described below are essentially unchanged. The null symbol ϕ is used as a place holder for nonaligned positions beyond the ends of a fragment; in data files ϕ of course would be replaced by the blank character " ". The blank character is awkward in equations, hence the use of ϕ.

All sequenced bases x_{ij} are recorded in a standard orientation relative to the assembly. However, many (typically half) of the bases are sequenced on fragments that are reverse complemented relative to the standard orientation. It is necessary to keep track of the orientation because the specific sequencing errors made depend on orientation. An error of reading A instead of G may not have the same probability of reading T instead of C. For each fragment we will define the reverse complement indicator to be

$$r_j = \begin{matrix} 0 & \text{Fragment } j \text{ is direct} \\ 1 & \text{Fragment } j \text{ is reverse complemented.} \end{matrix} \qquad [1]$$

We use the notation a^c to denote the base complementary to a. Note that a^c is well defined even for redundant bases: for example, $A^c = T$, $—^c = —$, $\{A, G\}^c = \{C, T\}$.

The depth of coverage at position i is defined to be the number of fragments contributing sequence information (including internal gaps),

$$d_i = \sum_{j=1}^{m} \mathbf{1}(x_{ij} \neq \phi). \qquad [2]$$

The notation $\mathbf{1}(E)$ is used to denote the indicator function for the event E. This function takes the value one when E is true and zero otherwise.

If the fragment sequences could be determined without error, our knowledge of the true sequence would be exact and, except for the problem of correct assembly, there would be no need for inference. However, sequencing errors do occur and we propose to describe them by the following probability model. An error occurs when the true base $s_i = a$ is misread in fragment j to yield $x_{ij} = b, b \neq a$. Recall the assumptions that the error probabilities are constant across all positions, that sequencing errors occur independently within and across columns of the fragment assembly, and that the error rate depends only on the values of a and b. The sequencing error probabilities are denoted by

$$p(b|a) = \Pr(x_{ij} = b | s_i = a); \quad a \in \mathcal{A}, \quad b \in \mathcal{B}. \qquad [3]$$

It is also necessary to define a probability distribution for the composition of the DNA sequence. We have assumed that bases occur independently with identical distribution across the assembly. The composition probabilities are denoted by

$$p(a) = \Pr(s_i = a); \quad a \in \mathcal{A}. \qquad [4]$$

Note that our definition of sequence composition includes the gap frequencies. More complex models for the DNA base composition and the sequencing error probabilities could be incorporated at this stage.

3. CONSENSUS WITH CONFIDENCE

Our goal is to make the best possible determination of $s_1, \ldots, s_n$ given the data in the fragment assembly. We initially assume that the sequencing error probabilities and DNA base composition are known exactly. In Section 4, we address the practical problem of estimating these quantities while simultaneously estimating the consensus sequence.

The *consensus distribution* is a probability distribution over the set $\mathcal{A}$ defined by

$$\pi_i(a) = \Pr(s_i = a | x_{ij}, j = 1, \ldots, m). \qquad [5]$$

It is the probability that the true base is a given the fragment assembly data. These probabilities can be computed using Bayes' rule,

$$\pi_i(a) = \frac{p(a) \prod_{j=1}^{m} [(1 - r_j)p(x_{ij}|a) + r_j p(x_{ij}^c|a^c)]}{\sum_{b \in \mathcal{A}} p(b) \prod_{j=1}^{m} [(1 - r_j)p(x_{ij}|b) + r_j p(x_{ij}^c|b^c)]}, \qquad [6]$$

where we set $p(\phi|a) = 1$ for all $a \in \mathcal{A}$. See Feller (1968, p. 125) for a discussion of this widely used elementary formula. Note that $\pi_i(a)$ is defined to be the probability that the base $s_i = a$ given the fragment data x_{ij}, while Eq. [6] gives $\pi_i(a)$ in terms of the conditional probabilities of reading x_{ij} given the true base $s_i = a$. Bayes' rule allows us to reverse the conditioning. By studying the accuracy of the fragment sequencing process, we can estimate the accuracy of the assembly.

We present three different approaches to the definition of a consensus sequence. First, we compute the most likely values of s_i at each position. Second, we allow redundancy in the consensus and compute a subset of $\mathcal{A}$ such that the probability that s_i is a member of this set exceeds $1 - \alpha$ at each position. Finally, we describe the construction of a consensus sequence with redundant bases such that the probability that the entire sequence $s_1, \ldots, s_n$ or region of interest $s_a, \ldots, s_b$ is contained within the redundant consensus sequence with probability at least $1 - \alpha$.

Procedure A. The most likely value of s_i is the base a that maximizes $\pi_i(a)$ over a. Denote this value by c_i; then

$$\max_{a \in \mathcal{A}} \pi_i(a) = \pi_i(c_i). \qquad [7]$$

Procedure B. Now we wish to allow redundant bases into the consensus. Our goal is to find the character with minimum redundancy that has probability in excess of a specified level $1 - \alpha$. The probability that a redundant base includes the true base is the sum over all bases represented by the symbol c_i. The choice of consensus sequence is made so that at each position the probability $\pi_i(c_i)$ exceeds a specified level,

$$\sum_{a \in c_i} \pi_i(c_i) \geq 1 - \alpha. \qquad [8]$$

To find the best redundant set, we begin by choosing the most likely base at position i; call it a_i. If $\pi_i(a_i) > 1 - \alpha$, we stop. Otherwise, we add to the redundant set, the next most likely base; call it b_i. If $\pi_i(a_i) + \pi_i(b_i) > 1 - \alpha$, we stop. The process continues until the level exceeds $1 - \alpha$. It is guaranteed to stop because $\pi_i(\mathrm{A}) + \pi_i(\mathrm{C}) + \pi_i(\mathrm{G}) + \pi_i(\mathrm{T}) + \pi_i(—) = 1$.

Procedure C. A redundant sequence can be thought of as a collection of sequences with a different member for each unique expansion of its redundant characters. Formally this is a *cylinder set* in sequence space (Feller, 1968, p. 130). We wish to construct such a collection of sequences that will contain the true sequence with probability exceeding $1 - \alpha$. The probability that the true sequences $s_1, \ldots, s_n$ lies within a redundant sequence $c_1, \ldots, c_n$ is given by

$$\prod_{i=1}^{n} \Pr(s_i \subset c_i | x_{ij}, j = 1, \ldots, m), \qquad [9]$$

where c_i is taken from the redundant sequence alphabet and the event $s_i \subset c_i$ indicates that s_i is a member of the set represented by c_i.

A consensus sequence with global coverage probability $1 - \alpha$ can be constructed by the following procedure. Set a threshold level T and find the marginal consensus for this level according to Procedure B. Compute the global coverage probability for this consensus (Eq. [9]). If this probability is less than $1 - \alpha$, increase T. Otherwise T may be decreased until the probability of containment just exceeds $1 - \alpha$. Because the containment probability is a monotone function of T, we can solve the problem quickly using a bisection algorithm.

One practical problem with this procedure is that many positions may have equal distribution π_i and hence will all be set to the same level of redundancy. This problem could be avoided by randomization. We prefer to choose the smallest value of T such that the coverage exceeds $1 - \alpha$ and report the actual coverage achieved.

4. ESTIMATION OF THE SEQUENCING ERROR RATES

If the true DNA sequence were known, it would be trivial to estimate the composition and sequencing error rates. One would simply count the number of times that a base a occurs in the sequence

$$n_a = \sum_{i=1}^{n} \mathbf{1}(s_i = a) \qquad [10]$$

and the number of times the base a was recorded as the base b in a fragment

$$n_{ab} = \sum_{i=1}^{n} \sum_{j=1}^{m} [(1 - r_j)\mathbf{1}(x_{ij} = b)\mathbf{1}(s_i = a) + r_j \mathbf{1}(x_{ij}^c = b)\mathbf{1}(s_i^c = a)] \qquad [11]$$

for all $a \in \mathcal{A}$ and $b \in \mathcal{B}$. Maximum likelihood estimates of the base composition and error rate parameters are given by

$$\hat{p}(a) = n_a/n \qquad [12]$$

and

$$\hat{p}(b|a) = n_{ab}/n_a. \qquad [13]$$

This situation suggests the following algorithm for the simultaneous estimation of the error rates and the consensus distribution.

Procedure D.

1. Initialize the consensus distribution: Set $\pi_i(x) = 1.0$ where x is the most frequently occurring letter at column i.

2. Estimate $p(a)$ and $p(b|a)$ for all a, b: Set the counts n_a and n_{ab} equal to their conditional expected values

$$\hat{n}_a = \sum_{i=1}^{m} \pi_i(a), \qquad [14]$$

$$\hat{n}_{ab} = \sum_{i=1}^{n} \sum_{j=1}^{m} [(1 - r_j)\mathbf{1}(x_{ij} = b)\pi_i(a) + r_j \mathbf{1}(x_{ij}^c = b)\pi_i(a^c)] \qquad [15]$$

and estimate $p(a)$ and $p(b|a)$ as before (Eqs. [12] and [13]).

3. Recompute $\pi_i(a)$ for all i and a: according to Eq. [6], with $p(a)$ and $p(b|a)$ replaced by their current estimates.

4. Continue: If the changes in $\hat{p}(a|b)$ and $\hat{p}(a)$ are less than ϵ, for all a and b, stop. Otherwise go to Step 2.

This estimation procedure is a special case of the EM algorithm for a mixture of multinomial distributions (Dempster *et al.*, 1977).

5. USING PRIOR INFORMATION

When a sequencing project has been ongoing, a great deal of information will accumulate regarding the types and frequencies of sequencing errors (see Krawetz (1989), Posfai and Roberts (1991), States and Botstein (1991)). It is to our advantage to incorporate this information into estimates of the error rates. When error rates are small any one assembly will contain only a few errors and probabilities will not be estimated accurately from the data. Our revised estimate of $p(b|a)$ will be a mixture of the prior expectation $\mu(b|a)$ and the current maximum likelihood estimate $\hat{p}(b|a)$,

$$\tilde{p}(b|a) = \alpha\mu(b|a) + (1 - \alpha)\hat{p}(b|a), \qquad [16]$$

for some mixing proportion $0 \leqslant \alpha \leqslant 1$. In the extreme cases, for $\alpha = 1$ we use only prior information and for $\alpha = 0$ we use only the current data. This linear combination arises from our choice of the analytically convenient Dirichlet distribution as a representation of the prior information (Lindley, 1972, pp. 59). For each $a \in \mathcal{A}$ we have a prior distribution over $p(\cdot|a)$, which is proportional to $\prod_{b \in \mathcal{B}} p(b|a)^{\beta_{ab}}$. The parameters β_{ab} determine the prior means and the mixing proportions. The *weight* of the prior is defined as

$$k_a = \sum_{b \in \mathcal{B}} \beta_{ab}, \qquad [17]$$

and the prior means are

$$\mu(b|a) = \beta_{ab}/k_a. \qquad [18]$$

The error rate estimates corresponding to Eq. [16] are

$$\tilde{p}(b|a) = \left(\frac{k_a}{n_a + k_a}\right)\mu(b|a) + \left(\frac{n_a}{n_a + k_a}\right)\hat{p}(b|a). \qquad [19]$$

Thus, the prior information effectively adds $k = \sum_{a \in \mathcal{A}} k_a$ observations to the current data.

To implement these estimates using the EM algorithm (Procedure D), augment Eqs. [14] and [15] as

$$\tilde{n}_a = \hat{n}_a + k_a \qquad [20]$$

$$\tilde{n}_{ab} = \hat{n}_{ab} + \beta_{ab}. \qquad [21]$$

These quantities are used to estimate $p(b|a)$ (Eq. [13]) but the estimates of $p(a)$ (Eq. [12]) are unchanged. A convenient choice for the prior distribution parameters is to take β_{ab} equal or proportional to $\tilde{n}_{ab}$ from previous sequence assemblies.

6. COVERAGE AND EXPECTED ACCURACY

Before a sequencing project is undertaken, the investigators should make some decision about the degree of accuracy desired. The size of the region to be sequenced and the purposes for which the sequence will be used should be considered. The need to quantify the relationship between cost of sequencing and accuracy was discussed by Waterston (Roberts, 1990). For example, if the target region is 10^6 bp and the maximum acceptable number of errors is 100, the probability of a sequencing error should be less than 10^{-4}. The desired degree of accuracy can be attained by adjusting the depth of coverage.

An approximate relationship between accuracy and the depth of coverage will be derived. An exact result depends on knowledge of the detailed structure of the error rates. We avoid this complication by the following assumption. Let

$$\sum_{\substack{b\in\mathcal{B}\\ b\neq a}} p(b|a) = \bar{p} \quad \text{for all } a \in \mathcal{A}. \qquad [22]$$

We think of $\bar{p}$ as the average error rate in sequenced fragments and we study the model where $p(a^c|a) = \bar{p}$ for all $a \in \mathcal{A}$.

Clearly, increasing depth increases the accuracy of columns with no discrepancies, e.g., a column with all A's. The probability that such a column is generated in error is approximately

$$\Pr(\text{error}) = 1 - \exp\left\{\left(\frac{\bar{p}}{1-\bar{p}}\right)^d\right\}, \qquad [23]$$

which goes to zero very rapidly as d gets larger. However, the probability that a column of depth d contains no discrepancies is $(1 - \bar{p})^d$, which decreases as d gets larger. Thus, the number of columns without discrepancies will decrease as depth of coverage increases. To show that increasing depth is desirable, we compute the expected accuracy, first for fixed depth d and then as an average over the distribution of d.

For a position covered to a depth d, we compute the probability that at most one-half of the observed bases are correct. This procedure for determining consensus is more conservative than procedure A and has the advantage of being much easier to compute. This is a binomial probability given by

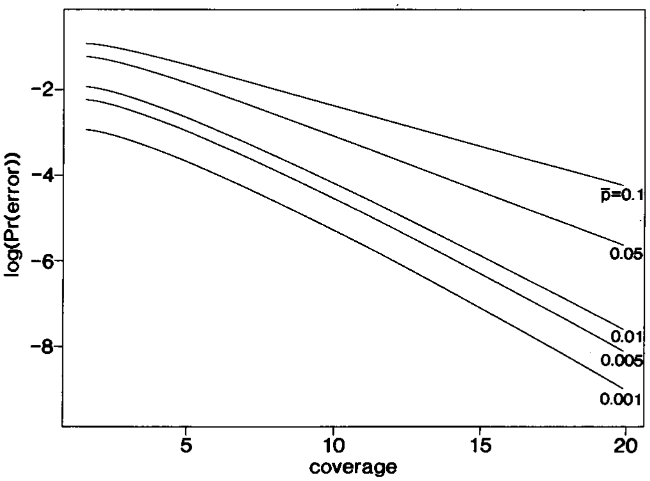


FIG. 1. The approximate probability of an error in one base of finished sequence (Eq. [25]) is shown as function of the mean depth of coverage. The scale for error probabilities is logarithm base 10. The five curves arise by varying the parameter $\bar{p}$ in Eq. [24].

$$\Pr(\text{error}|d) = \sum_{k=0}^{\lfloor d/2\rfloor} \binom{d}{k} (\bar{p})^{d-k}(1-\bar{p})^k. \qquad [24]$$

When sequence fragments are randomly located, there is some variation in depth of coverage. In fact the distribution of d in a shotgun assembly is approximately Poisson with some mean λ. We ignore positions not covered by any fragments. Thus the mean coverage is $c = \lambda/(1-e^{-\lambda})$. We can compute the average error probability conditional on the depth being at least 1 as

$$\Pr(\text{error}) = \frac{1}{1-e^{-\lambda}} \sum_{d=1}^{\infty} \frac{\lambda^d e^{-\lambda}}{d!} \Pr(\text{error}|d). \qquad [25]$$

This quantity is summarized in Fig. 1. Variations in the coverage decrease the accuracy relative to a uniformly covered sequence with the same mean coverage c. In either case, increasing depth of coverage increases the expected accuracy of the finished sequence.

For our simplified model, now we are able to study the depth of coverage required to attain a given level of accuracy in the finished sequence. If our goal is 10^{-4} errors per base, then for $\bar{p} = 0.01$ we require a mean coverage of $c = 9.4$. If the error rate in fragment sequences can be kept down to $\bar{p} = 0.001$, we need $c = 6.1$. To achieve 10^{-6} errors per base at $\bar{p} = 0.01$, we need $c = 15.3$ and at $\bar{p} = 0.001$ we need $c = 12$.

Our results are sensitive to the definition of $P(\text{error}|d)$. If $P(\text{error}|d)$ is defined to be the probability that at least one-half of the observed bases are correct, the $P(\text{error})$ changes noticeably. With this definition, and a goal of 10^{-4} errors per base, for $\bar{p} = 0.01$ we require a mean coverage of $c = 6.8$. If the error rate in fragment sequences can be kept down to $\bar{p} = 0.001$, we

TABLE 1

EM Estimates from G6PD Data (see Section 7) with N = 11865 Aligned Positions

	c	*t*	*a*	*g*	Δ	*x*
			(*i*) Estimated error counts = $\hat{n}_{ab}$			
c	11918	6	4	2	65	3
t	9	10401	3	2	44	8
a	6	4	11035	3	30	5
g	5	1	11	12010	58	12
Δ	81	87	61	90	377	3
			(*ii*) Estimated error rates $\hat{p}(b\|a)$			
c	0.993304	0.000504	0.000356	0.000178	0.005405	0.000252
t	0.000843	0.993742	0.000266	0.000188	0.004239	0.000721
a	0.000496	0.000402	0.995662	0.000266	0.002735	0.000438
g	0.000443	0.000086	0.000931	0.992712	0.004832	0.000996
Δ	0.116051	0.123898	0.087734	0.128809	0.539908	0.003599
			(*iii*) Estimated sequence composition $\hat{p}(a)$			
	0.258896	0.225850	0.239146	0.261041	0.015067	

need $c = 3.8$. To achieve 10^{-6} errors per base at $\bar{p} = 0.01$, we need $c = 12.5$ and at $\bar{p} = 0.001$ we need $c = 4.2$. These numbers are much smaller than those in the previous paragraph.

7. THE HUMAN G6PD LOCUS

Recently Chen *et al.* (1990) presented the sequence of 20,114 nucleotides of human DNA, which includes the human glucose-6-phosphate dehydrogenase gene (G6PD). Defective G6PD genes can cause hemolytic anemia but they also offer partial protection against malaria. Consequently many variants exist and over 300 have been described. Chen *et al.* (1991) sequenced 15,860 nucleotides of the transcribed region from the mRNA start site to the polyadenylation site. There are 13 exons, and the 20,114 nucleotides are about 25% *Alu* sequence.

The total sequence was determined from three *Eco*RI fragments. Chen *et al.* (1990) kindly provided us with their assembly of the largest fragment, 11,791 nucleotides in length. The fragment was isolated in λ clones, subcloned in pUC18, and sequenced in fragments randomly subcloned in M13. The assembly has 165 fragments with mean length 283. The average depth of coverage is therefore about 4. The fragments are very accurate with small error probabilities. We ran Procedure D on the fragment set and estimated π_i and $p(a|b)$ (see Table 1). To summarize the analysis we present two graphs. In Fig. 2a is a graph of depth of coverage vs nucleotide position. In addition to depth, we graph entropy e_i in Fig. 2b to show the variability of π_i. The entropy of the distribution π_i is

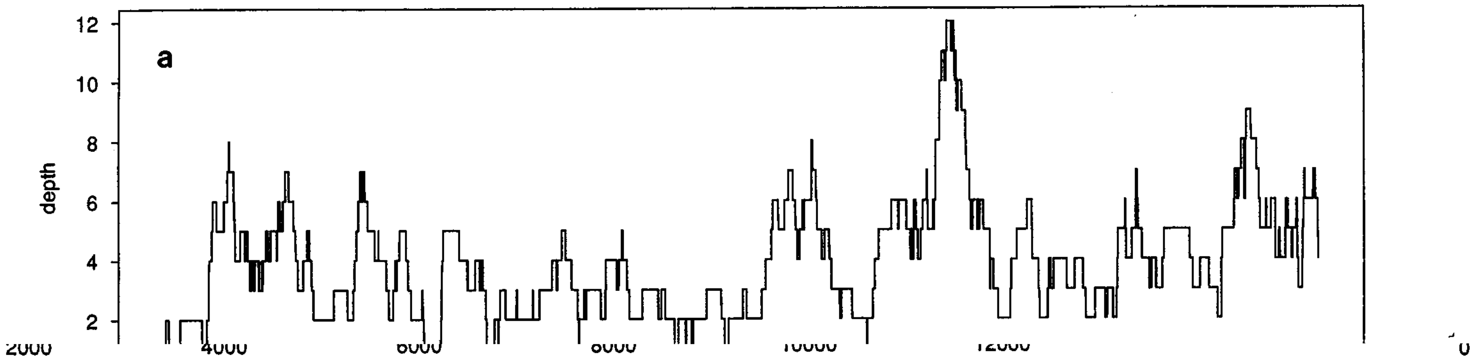
$$e_i = -\sum_{a \in \mathcal{A}} \pi_i(a)\log_2(\pi_i(a)). \qquad [26]$$

When $\pi(a) = 1$ for some a, $e = 0$, the minimum value e can attain. The function e is maximized when $\pi = (\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$: then $e = \log_2(5)$. Therefore uncertainty in the consensus distribution is indicated by large values of e_i. Statistical entropy was introduced in 1948 by C. E. Shannon who founded the modern field of information theory. That paper remains one of the best introductions to the subject and is reprinted in Shannon and Weaver (1964).

As an illustration of Procedure C, we have constructed a series of cylinder sets around the 605-bp region containing the *Alu* cluster 8^R98^L (see Chen *et al.*, 1990). This region was selected because of its intrinsic interest and because the depth of coverage is 1 at 165 positions and ranges up to 5 with a mean coverage $c = 2.94$. There are 11 columns with discrepancies involving insertion/deletion of bases and only two mismatched bases. Two portions of this region are shown in Fig. 3 A cylinder set with containment probability ⩾95% comes at the high price of 255 ambiguous bases, leaving 350 bases well defined. This result should not be surprising, as the precise determination of a large number of independent quantities is a very strong inference. By reducing the containment probability, we can increase the number of well-defined bases and/or reduce the ambiguity of the remaining bases. Note, for example, the progression from * to X to R to A in the depth 1 portions of the sequence as the containment probability is decreased. The tightest cylinder set computed contains a single ambiguous base corresponding to a C,T mismatch in the alignment. This cylinder contains the true sequence with 12% probability.

8. DISCUSSION

We have described a method for assessing the accuracy of DNA sequences and quantified the relationship between accuracy and the depth of coverage. By introducing a model of the sequencing error process we can


a
depth
12
10
8
6
4
2
2000
4000
6000
8000
10000
12000
0

a

```
           R gagcaagactccatctcaaaaaaaagaaaaaacaaaaattagctggatgtggtggcaggcacctggaatcccagctactcaggaggctaaggcagg
           R                                    aaattagctggatgtggtggt
           D                                                                tggaatcccagctactcaggaggctaaggcagg
           R                                                                   aatcccagctactcaggaggctaaggcagg
           R                                                                           ctactcaggaggctaaggcagg
           D                                                                                        taaggcagg
             ------------------------------------------------------------------------------------------------

Depth        111111111111111111111111111111111112222222222222222222221111111222333333334444444444444555555555

0.9562 (255) ***********************************aaa**a*c***a*******************aatcccagctactcaggaggctaaggcagg
0.9153 (210) ***********************************aaa**agc*gga*g*gg*gg*********ggaatcccagctactcaggaggctaaggcagg
0.8771 (170) ***********************************aaattagctggatgtggtgg********tggaatcccagctactcaggaggctaaggcagg
0.6727 (170) *x**xx*x****x****xxxxxxxx*xxxxxx*xxaaattagctggatgtggtgg*x***x**tggaatcccagctactcaggaggctaaggcagg
0.6524 (170) *r**rr*r****r****rrrrrrrr*rrrrrr*rraaattagctggatgtggtgg*r***r**tggaatcccagctactcaggaggctaaggcagg
0.4372 (122) *a*xaa*ax*xxa*x*xaaaaaaaa*aaaaaaxaaaaattagctggatgtggtgg*a**xaxxtggaatcccagctactcaggaggctaaggcagg
0.3105  (73) gagyaagay*yya*y*yaaaaaaaagaaaaaayaaaaattagctggatgtggtgg*aggyayytggaatcccagctactcaggaggctaaggcagg
0.1215   (1) gagcaagactccatctcaaaaaaaagaaaaaacaaaaattagctggatgtggtggxaggcacctggaatcccagctactcaggaggctaaggcagg
```

b

```
           R tct-gggcgtggtggctcatgcctgtaa-ccccagcactttgggaggctgaggagggtggatcacctgaggtcaggagttcgagaccagcctggcaaacatggtgaaaccccgtctctact
           R tct-gggcgtggtggctcatgcctgtaa-ccccagcactttgggaggctgaggagggtggatcacctgaggtcaggagttcgagaccagcctggcaaac
           R tctggggcgtggtggctcat-cctgtaacccccagcactttgggaggctgaggagggtggatcacctgaggtcaggagttcgagaccagcctggcaaacatggtgaaaccccgtctctact
           R        cgtggtggctcatgcctgtaa-ccccagcactttgggaggctgaggagggtggatcacctgaggtcaggagttcgagaccagcctggcaaacatggtgaaaccccgtctctact
           D                                  agcactttgggagg-tgaggagggtggatcacctgaggt-aggagt-cgagaccagcctgg-aaacatggtgaaaccc-gtctcg

Depth        3333333444444444444444444444444445555555555555555555555555555555555555555555555555555555555555555555555554444444444444444444333

0.9562 (255) tct*gggcgtggtggctcat*cctgtaa-ccccagcactttgggagg*tgaggagggtggatcacctgaggt*aggagt*cgagaccagcctgg*aaacatggtgaaaccc*gtctc*act
0.9153 (210) tct*gggcgtggtggctcat*cctgtaa-ccccagcactttgggagg*tgaggagggtggatcacctgaggt*aggagt*cgagaccagcctgg*aaacatggtgaaaccc*gtctc*act
0.8771 (170) tct*gggcgtggtggctcat*cctgtaa-ccccagcactttgggaggctgaggagggtggatcacctgaggtcaggagttcgagaccagcctggcaaacatggtgaaaccc*gtctc*act
0.6727 (170) tct*gggcgtggtggctcat*cctgtaa-ccccagcactttgggaggctgaggagggtggatcacctgaggtcaggagttcgagaccagcctggcaaacatggtgaaaccc*gtctc*act
0.6524 (170) tct*gggcgtggtggctcat*cctgtaa-ccccagcactttgggaggctgaggagggtggatcacctgaggtcaggagttcgagaccagcctggcaaacatggtgaaaccc*gtctc*act
0.4372 (122) tct*gggcgtggtggctcat*cctgtaa-ccccagcactttgggaggctgaggagggtggatcacctgaggtcaggagttcgagaccagcctggcaaacatggtgaaaccc*gtctc*act
0.3105  (73) tct*gggcgtggtggctcat*cctgtaa-ccccagcactttgggaggctgaggagggtggatcacctgaggtcaggagttcgagaccagcctggcaaacatggtgaaaccc*gtctc*act
0.1215   (1) tct-gggcgtggtggctcatgcctgtaa-ccccagcactttgggaggctgaggagggtggatcacctgaggtcaggagttcgagaccagcctggcaaacatggtgaaaccccgtctctact
```

FIG. 3. Two segments of the fragment assembly from the region of *Alu* cluster 8R98L. Symbols in the left-hand column indicate the orientation of the sequenced fragments: R, reversed complement, D, direct. The depth of coverage is indicated. A series of cylinder sets is shown at the bottom. Containment probabilities for the entire 605-bp *Alu* cluster are given at the left. The numbers of ambiguous bases are shown in parentheses. Positions denoted by $*$ are ambiguous including both a gap ($-$) and a base.

should be adjusted to give the correct overlap of the data sets. It may be necessary to add columns internally to allow for insertions in one sequence relative to the other. The cost of tracking accuracy is an increase in storage space requirements to store one integer and five floating point numbers for each sequence position. If accuracy information were to be stored in a central database, simple sequence data could be distributed to the majority of users and the more detailed accuracy information could be available upon request.

As a further refinement to sequence accuracy statistics, we recommend that automated sequencing devices output a probability distribution reflecting the accuracy of individual base determinations. The methods described here can be readily modified to handle such data. The base values in fragment sequences are not simply single letters $x_{ij} \in \mathcal{B}$ but instead are probability distributions over the set $\mathcal{A}$, denoted $p_{ij}(x)$ ($i = 1, \ldots, m$; $j = 1, \ldots, n$; $x \in \mathcal{A}$). The error probabilities $p(b|a)$ are still well defined. The new version of the consensus distribution is

$$\pi_i(a) \propto p(a) \prod_{j=1}^{m} \Big(\sum_{b\in\mathcal{A}} p(b|a)p_{ij}(b)\Big), \qquad [28]$$

where the denominator and the reversed-complement bookkeeping have been suppressed to simplify the notation. Given data of this type, assumption A2 would no longer be necessary. There may be additional advantages to this approach. For example, fragment assembly algorithms could be modified to accept such input and would be less likely to get stuck trying to fit "bad" data. It may also be possible to read gels beyond the current limits of accurate resolution. One could extract information that is currently discarded without compromising the accuracy of the final sequence.

The currently achievable error rates in manually sequenced fragments are more than adequate to achieve high-quality finished sequence. Post-assembly editing of sequences is a labor intensive task that is essential to maintaining this high quality. The methods described here can provide some degree of automated assistance by flagging areas of high uncertainty. However, it would be a mistake to abandon the careful rechecking of sequences at this stage of technological development. As fully automated sequencing systems are developed, manual rechecking will become impractical. Automated quality checking will become essential but realistically cannot be expected to match the quality of a manual approach in the near future.

ACKNOWLEDGMENTS

This study was supported by grants from the National Institutes of Health (2R01GM36230-06) and the National Science Foundation (DMS90-05833, DMS87-20208).

REFERENCES

Arratia, R. A., and Gordon, L. (1990). Tutorial on large deviations for the binomial distribution. *Bull. Math. Biol.* **51:** 125–131.

Balding, D. J., and Torney, D. C. (1991). Statistical analysis of DNA fingerprint data for ordered clone physical mapping of human chromosomes. *Bull. Math. Biol.* **53:** 853–879.

Branscomb, E., Slezak, T., Pae, R., Galas, D., Carrano, A. V., and Waterman, M. S. (1990). Optimizing restriction fragment fingerprinting methods for ordering large genomic libraries. *Genomics* **8:** 351–366.

Burks, C., *et al.* (1990). GenBank: Current status and future directions. *In* "Methods in Enzymology" (R. F. Doolittle, Ed.), Vol. 183, pp. 3–22. Academic Press, San Diego.

Chen, E., *et al.* (1991). Sequence of human glucose-6-phosphate dehydrogenase cloned in plasmids and a yeast artificial chromosome (YAC). *Genomics* **10:** 792–800.

Churchill, G. A. (1989). A stochastic model for heterogeneous DNA sequences. *Bull. Math. Biol.* **51:** 79–94.

Churchill, G. A., Burks, C., Eggert, M., and Waterman, M. S. (1990). "Fragment Assembly Methods for DNA Sequencing," Manuscript.

Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in DNA sequences: Recommendations 1984. *Nucleic Acids Res.* **13:** 3021–3030.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. B* **39:** 1–38.

Edwards, A., *et al.* (1990). Automated DNA Sequencing of the Human HPRT Locus. *Genomics* **6:** 593–608.

Feller, W. (1968). "An Introduction to Probability Theory and Its Applications," Vol. I, 3rd ed. Wiley, New York.

Fields and Soderland (1990). gm: A practical tool for automating DNA sequence analysis. *CABIOS* **6:** 263–270.

Gusella, J. F. (1986). DNA Polymorphism and human disease. *Ann. Rev. Biochem.* **55:** 831–854.

Kahn, P., and Cameron, G. (1990). EMBL Data Library. *In* "Methods in Enzymology" (R. F. Doolittle, Ed.), Vol. 183, pp. 29–31. Academic Press, New York.

Kececioglu, J., and Myers, E. (1990). "A Robust Automatic Fragment Assembly System," Preprint.

Krawetz, S. A. (1989). Sequence errors described in GenBank: A means to determine the accuracy of DNA sequence interpretation. *Nucleic Acids Res.* **17:** 3951–3957.

Lander, E. S., and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2:** 231–239.

Lindley, D. V. (1972). "Bayesian Statistics, A Review." SIAM, Philadelphia, PA.

Michiels, F., Craig, A. G., Zehetner, G., Smith, G. P., and Lehrach, H. (1987). Molecular approaches to genome analysis: A strategy for the construction of ordered overlapping clone libraries. *CABIOS* **3:** 203–210.

Posfai, J., and Roberts, R. J. (1991). Error detection in DNA sequences. In preparation.

Roberts, L. (1990). Large-scale sequencing trials begin. *Science* **250:** 1336–1338.

Shannon, C. E., and Weaver, W. (1964). "The Mathematical Theory of Communication." Univ. of Illinois Press, Urbana.

Staden, R. (1980). A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res.* **8:** 3673–2694.

States and Botstein (1991). Molecular sequence accuracy and the analysis of protein coding regions. *Proc. Natl. Acad. Sci. USA* **88:** 5518–5522.

Waterman, M. S. (1990). Genomic sequence databases. *Genomics* **6:** 700–701.