Base-Calling of Automated Sequencer Traces Using *Phred.* II. Error Probabilities

Brent Ewing and Phil Green[1]

Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA

Elimination of the data processing bottleneck in high-throughput sequencing will require both improved accuracy of data processing software and reliable measures of that accuracy. We have developed and implemented in our base-calling program *phred* the ability to estimate a probability of error for each base-call, as a function of certain parameters computed from the trace data. These error probabilities are shown here to be valid (correspond to actual error rates) and to have high power to discriminate correct base-calls from incorrect ones, for read data collected under several different chemistries and electrophoretic conditions. They play a critical role in our assembly program *phrap* and our finishing program *consed.*

Read data from automated sequencers varies significantly in quality for a number of reasons (for review, see Ewing et al. 1998), and making the most effective use of such data requires having some measure of its reliability. Position-specific error probabilities (Lawrence and Solovyev 1994) are particularly useful for this purpose. In conjunction with appropriate assembly software they can: improve the accuracy and completeness of assembly by allowing better discrimination of repeats and by making it possible to use full read lengths; permit a more accurate consensus sequence to be derived; provide an objective criterion for guiding the finishing (deciding where additional data or editing are needed); and provide an objective measure useful in monitoring data quality and in setting the quality standard for the final sequence.

A number of developers of base-calling algorithms (Giddings et al. 1993; Golden et al. 1993; Berno 1996) have developed confidence measures for the base-calls, but do not report studies of their validity or discrimination power. The most thorough study of which we are aware is that of Lawrence and Solovyev (1994), who defined a large number of trace parameters and carried out an extensive discriminant analysis to determine the ones most effective at distinguishing accurate base calls from errors and assigning error probabilities. Here, we describe a different procedure for estimating error probabilities and investigate its properties. The main distinguishing features of our work are (1) a novel algorithm for deriving the probabilities, that does not require the multivariate normality distributional assumptions that are needed for discriminant analysis but are very far from being true in the case of the parameters we consider; (2) use of parameters computed from windows of the trace, that appear more effective at discrimination than the single-peak measures considered by Lawrence and Solovyev; and (3) an emphasis on optimizing discrimination ability in the high-quality part of the trace (error rates <0.01) rather than over the entire range, as it is this part of the trace that tends to be more important in practice.

An important technical aspect of our work is the use of log-transformed error probabilities rather than untransformed ones, which facilitates working with error rates in the range of most importance (very close to 0). Specifically, we define the quality value q assigned to a base-call to be

$$q = -10 \times \log_{10}(p)$$

where p is the estimated error probability for that base-call. Thus a base-call having a probability of 1/1000 of being incorrect is assigned a quality value of 30. Note that high quality values correspond to low error probabilities, and conversely.

METHODS

Overview of Error Probability Issues

Obvious requirements for the error probabilities are that they must be predictive, that is, assigned by use of a method that does not require knowledge of the

[1]Corresponding author.
E-MAIL phg@u.washington.edu; FAX (206) 685-7344.

 8:186–194 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

true sequence; and they must be valid in the sense that they correspond to observed error rates, that is, the set of bases assigned a particular error probability p should have an actual error rate equal to p.

Given these constraints, there are still many possible ways of assigning error probabilities to base-calls. For example, if a set of 1,000,000 base-calls contains 10,000 errors, then a method that assigns an error probability of 0.01 to each base-call is valid; but so is another method that identifies two sub-sets of 500,000 base-calls each, with the first set containing 9000 errors and the second set 1000 errors, and assigns an error probability of 0.018 to every base-call in the first set and 0.002 to every base-call in the second set. While the second method is no more valid than the first, it clearly does a better job at discriminating the less accurate base-calls from the more accurate.

We can formalize the notion of discrimination ability as follows. Intuitively, a method with high discrimination ability should spread out the error probabilities (or quality values) as much as possible. By that criterion, one natural measure of discrimination power is simply the variance of the quality value distribution. However, in practice, it is the most accurate read data that are most important, as those are the data used to derive the consensus sequence, and consequently we prefer to judge different methods by how well they perform at identifying a subset of the bases with a very low error probability. Specifically, given a set B of base-calls and a valid method of assigning an error probability $e(b)$ to each base-call b, then for any subset B of B, the expected number of errors in B is $\Sigma_{b\in B}e(b)$ (since the error probability for a base-call is also the expected number of errors for that call); and the expected error rate for B is the expected number of errors in B, divided by the number of base-calls in B. It is easy to see that for any given error rate, r, there is a unique largest set of base-calls, B_r, having the properties that (1) the expected error rate of B_r is $\leqslant r$, and (2) whenever B_r includes a base-call b, it includes all other base-calls whose error probabilities are $\leqslant e(b)$. The discrimination power at the error rate r is then defined to be

$$P_r = \left(\frac{|B_r|}{|B|}\right)$$

that is, the number of base-calls in B_r divided by the total number of base calls. P_r measures the effectiveness of the error probability assignments at extracting a subset of bases having a low error rate r. (There is a precise analogy here to the notions of validity and power of a statistical test.)

Our goal in this study was to develop error probabilities that are valid and have high discrimination power at small values of r ($r \leqslant 0.01$). For this purpose, we used parameters computed from the processed trace from which the base-calls were made, focusing most on those parameters that appear to play a role in intuitive human assessments of data quality as data quality should be predictive of error probabilites. A number of different sets of such parameters were tested, using an algorithm (described below in Error Probability Calibration) that, given a set of parameters and a training set of reads for which it is known which base-calls are correct and which are errors, finds a way of associating parameter values to error probabilities that has (near) maximum discrimination power for small r.

Trace Parameters

Errors by the basecaller often are attributable to misinterpretation of peaks in a region of the trace, and, as a result, indications of the error may be present in the vicinity of the erroneous peak but not at the peak itself. Consequently the parameters most effective at detecting errors tend to be those that consider a window of the trace that includes several peaks flanking the one whose base-call is being assessed. The parameters that turned out to be the most powerful in our tests were all of this type. A by-product of using parameters computed from such a window is a smoothing of the parameter values (and hence of the error probabilities) from base to base.

The following four parameters were found to be particularly effective at discriminating errors from correct base-calls. In each case, smaller parameter values correspond to higher quality (more accurate sequence).

1. *Peak spacing.* The ratio of the largest peak-to-peak spacing, in a window of seven peaks centered on the current one, to the smallest peak-to-peak spacing. The minimum possible value of one corresponds to evenly spaced peaks.
2. *Uncalled/called ratio.* The ratio of the amplitude of the largest uncalled peak, in a window of seven peaks around the current one, to the smallest called peak; if there is no uncalled peak, the largest of the three uncalled trace array values at the location of the called base peak is used instead. [An uncalled peak is a peak in the signal that was not assigned to a predicted location by *phred* (Ewing et al. 1998) and thus does not result in a base call.] If the called base is an *N*, *phred*

assigns a large value of 100.0. Note that this is not what is sometimes called the signal to noise ratio, as uncalled peaks may be true peaks missed by the base-calling program rather than noise in the conventional sense. The minimum parameter value is 0 for traces with no uncalled peaks.

3. Same as 2, but using a window of three peaks.
4. *Peak resolution.* The number of bases between the current base and the nearest unresolved base, times −1 (to force the parameter to have the right direction). (A base is unresolved if it is called as *N* or if for at least one of its neighboring bases, there is no point between the two corresponding peaks at which the signal is less than the signal at each peak). The minimum possible parameter value is half the number of bases in the trace, times −1, and the maximum value is 0.

Error Probability Calibration

Given a set of parameters that can be computed from the trace for each base-call, a set of threshold values for those parameters is said to be optimal for a particular error rate *r,* if the set *B* consisting of all bases whose parameter values are less than the threshold values has an error rate $\leq r$, and no other set of thresholds yields a larger set with error rate $\leq r$. We describe below a simple greedy algorithm that finds a nearly optimal set of thresholds for small error rates. With respect to linear discriminant analysis (which has similar goals), our method has the advantages of not assuming multivariate normality (which is very far from being true in our case) or other distributional properties for the parameters, of allowing parameters that take on nonnumerical values, and of not requiring that parameters be transformed to normality or that outliers receive any special treatment. It is compute-intensive, however, and, in the form described, can only be used with a relatively small number of parameters simultaneously. The only significant assumption about the parameters is that their values should be ordered such that small values tend to correspond to more accurate base-calls and large values to less accurate base-calls.

The algorithm produces a lookup table consisting of a set of lines, each line containing a set of parameter thresholds, together with the error probability and quality value corresponding to those thresholds (there can be multiple lines having the same quality value). Although we have only applied the algorithm to generate a single error probability for each base (combining all types of error—substitution, deletion, and insertion), it can easily be adapted to produce separate probabilities for each error type.

The basic idea of the algorithm is as follows. One starts with a small number of parameters (in our case, four). First, a finite number of threshold values for each parameter is selected; we allow 50 different thresholds, a number large enough to avoid significantly sacrificing resolution but small enough that it is computationally feasible to consider each possible threshold for each parameter in subsequent steps. Then each 4-tuple of parameter thresholds (one for each parameter) is considered in turn, and the empirical error rate is computed for the set of bases defined by those thresholds. The 4-tuple for which the empirical error rate is smallest is selected (in the event of ties, the largest set having a given error rate is taken) and defines the first line of the lookup table. The set of bases defined by these thresholds is then eliminated, and the process is repeated using the remaining bases. Iteration of this procedure produces the desired lookup table.

Note that by construction the first set of thresholds found by this procedure is nearly optimal for its error rate, in the sense defined above (it may fail to be strictly optimal because only a finite set of thresholds are considered for each parameter). Subsequent sets of thresholds also tend to be nearly optimal, but in this case the effect of eliminating bases at earlier steps is another factor which may cause strict optimality to fail. This effect tends to be small in the early steps when error rates are small, but may become more significant later.

We now give a more precise description of the algorithm, in the case where four different parameters *r, s, t,* and *u* are used. In the following, if *p* denotes a parameter, then $p(b)$ indicates the value of the parameter for a particular base-call *b*. First, for each parameter *p* we find values $p_0 < p_1 < \ldots < p_{49} < p_{50}$ such that all $p(b)$ lie between p_0 and p_{50}, and the number of bases *b* satisfying $p_{i-1} < p(b) \leq p_i$ is approximately the same for all *i*.

We define a cut to be a 4-tuple (i, j, k, m) where $i, j, k,$ and m range between 1 and 50. There are 50^4, or 6.25 million cuts in all; although this is large relative to the number of data points, we will avoid overfitting by imposing a monotonicity condition. Each cut has a corresponding set of parameter thresholds r_i, s_j, t_k, and u_m, defined as above. For each cut (i, j, k, m), define $err_{(i,j,k,m)}$ to be the total number of erroneous base calls *b* below the cut, that is, satisfying $r(b) \leq r_i$, $s(b) \leq s_j$, $t(b) \leq t_k$, and $u(b) \leq u_m$. Similarly, let $corr_{(i,j,k,m)}$ be the total number of correct base-calls below the cut.

The error rate below the cut, $e_{(i,j,k,m)}$, is defined by

$$e_{(i,j,k,m)} = \frac{1.0 + err_{(i,j,k,m)}}{1.0 + corr_{(i,j,k,m)} + err_{(i,j,k,m)}}$$

and the corresponding quality value by

$$q_{(i,j,k,m)} = -10 \times \log_{10}(e_{(i,j,k,m)})$$

Here 1.0 is a small-sample correction added to ensure that both the numerator and denominator are positive; it produces a slight upward bias in e, that is most pronounced for very small e. We round $q_{(i,j,k,m)}$ to the nearest integer. The following two steps are now iterated to create the lookup table.

1. Find the cut (i,j,k,m) for which $q_{(i,j,k,m)}$ is largest. In the event of ties, take the one for which $corr_{(i,j,k,m)} + err_{(i,j,k,m)}$ is largest; if there is more than one of these, take the one for which the sum of the indices is highest. Output $q_{(i,j,k,m)}$, $e_{(i,j,k,m)}$ and the parameter values r_i, s_j, t_k, and u_m. Delete (i, j, k, m) from the list of cuts.
2. For each remaining cut (i', j', k', m'), adjust the counts $err_{(i',j',k',m')}$ and $corr_{(i',j',k',m')}$ by deleting bases below the removed cut (i, j, k, m), and recompute $e_{(i',j',k',m')}$ and $q_{(i',j',k',m')}$ using the new values. If *err* and *corr* are 0 for all remaining cuts, stop. Otherwise go to step 1.

Note that, by construction, the error probability e output in step 1 is the (small-sample corrected) error rate for the set of bases defined by the property that their parameter values are less than or equal to the parameter thresholds output in this step, but not less than or equal to the thresholds output in any previous line. As a result, given the parameter values for a base, one can find an appropriate marginal error probability for that base by looking through the table until the first line is found in which the parameter thresholds equal or exceed the parameter values of the base in question and then reading the error probability on that line.

Cosmid Sets

For these studies we used the ABI-processed trace data from four sets of cosmids (Table 1): two training sets consisting of 9 mammalian cosmids from L. Rowen (L. Hood's laboratory, University of Washington), and 9 *Caenorhabditis elegans* cosmids from the Washington University Genome Sequencing

Table 1. Cosmid Set Descriptions

				Dye primer reads			Dye terminator reads		
Set	**Cosmids[a]**	**% GC**	**Total reads**	**aligned reads**	**aligned bases**	**errors (%)**	**aligned reads**	**aligned bases**	**errors (%)**
				Training sets					
1	9	43	8240	6527	3258752	140901 (4.3)	143	60461	2963 (4.9)
2	9	37	13448	10307	4741753	220395 (4.6)	279	113398	8019 (7.1)
				Test sets					
3	22	39	26091	17973	8931830	324737 (3.6)	3516	1848671	107716 (5.8)
4	36[b]	43	27184	21417	17379770	732303 (4.2)	1541	1338434	73070 (5.5)

[a]Cosmid set 1 GenBank accession nos.: AE000663 (cosmid 0742C), AE000665 (cosmid 82C), U66059 (cosmids A14, G54, K26, K35, and X21B), AF029308 (cosmids X13A and X224). Cosmid set 2 accession nos.: U23454, U39645, U23529, U39742, U29535, U23518, U29381, U28732, and U29536. Cosmid set 3 accession nos: U88311, AF016443, AF003740, AF016447, AF040643, AF036695, AF040649, AF026210, AF040648, AF040653, AF016678, U97001, AF040654, U80848, AF026211, U97550, AF040655, U41017, AF014940 (cosmids C11D2, C45G7, and R12E2 were not submitted yet). Cosmid set 4 accession nos: AC000099, AC000123, AC000109, AC000110, AC000354, AC000361, AC000362, AC000363, AC000364, AC000355, AC000356, AC000124, AC000125, AC000357, AC000126, AC000127, AC000358, AC000359, AC002495, AC002424, AC000373, AC000365, AC000366, AC000367, AC002113, AC002114, AC002497, AC002083, AC002084, AC000369, AC000370, AC002057, AC000374, AC000371, AC000372, and AC002498.

[b]Two of these are cosmid fragments 4.2 and 8.9 kb long.

Sequencing reactions used *Taq* polymerase (sets 1 and 2), Sequitherm or TaqFS (set 3), or TaqFS (set 4). Sets one and two were generated almost entirely on ABI 373 sequencing machines running short (34 cm) gels. Set three was generated on ABI 373 machines running short gels (50%) or on ABI 377 machines running short (36 cm) or long (48 cm) gels (21%). (Information about the remaining 29% of the data in this set was not recorded.) Set four was generated mostly on ABI 373 machines running long (48 cm) gels (2.2% of set 4 was sequenced on ABI 377 machines). ABI analysis software was used for lane tracking and processing, and *phred* v. 0.961028 was used to call bases and assign quality values.

Center (R. Waterston), and two test sets consisting of 22 *C. elegans* cosmids from the Washington University Genome Sequencing Center, and 36 human chromosome 7 cosmids from the University of Washington Genome Sequencing Center (M. Olson). *Phred* basecalls (Ewing et al. 1998) were generated for each trace, and reads were then screened for sequencing vector and aligned against the finished cosmid sequence using *cross_match* as described previously (Ewing et al. 1998). Each read base in the alignable part of the read was classified as correct (matching the final sequence) or erroneous (discrepant with the final sequence); deletion errors (i.e. cases where one or more bases in the cosmid sequence are missing from the read) were assigned randomly to one of the two read bases adjacent to the deletion. Unaligned bases are ignored in the following.

Quality Assignment

Application of the error probability calibration algorithm, with the parameters described above, to a single combined training set (containing all reads in cosmid sets 1 and 2) produced a lookup table with 2011 lines. This table is used to assign quality values to the base-calls from a particular read, as follows. For each base call, *phred* computes the four parameter values, and then searches the lookup table line by line, in order, until it finds a line in which each of the four parameter values is at least as large as the corresponding parameter value for the base-call. The quality value associated to that line is then assigned to the base. If no such line is found, the base-call is assigned a quality of 0.

This procedure was used to assign a quality value to each base-call in the two test sets of cosmids (Table 1). Initial examination of errors and quality values in the test data sets showed more errors than expected among bases with predicted quality values of 40 and above. On inspection, we found five kinds of spurious contribution to the error counts:

1. *Chimeras.* Chimeric reads contain two or more noncontiguous segments that are incorrectly juxtaposed, as a result of a cloning artifact (a chimeric subclone) or a gel lane tracking error. In such cases, cross_match finds an alignment involving one of the two pieces, but occasionally the alignment extends spuriously for a few bases into the adjacent piece because of fortuitously matching nucleotides and includes spurious high quality discrepancies in this extension.
2. *Contaminant or Misassembled Reads.* Most data sets have a small number of contaminant reads arising either from sample mistracking or from low levels of contamination of the subclone libraries. If the contaminating read contains a repeated sequence, it may spuriously appear to match the cosmid sequence with multiple high-quality discrepancies. Similarly a noncontaminant read lying in a near-perfect repeat in the cosmid may be aligned against the wrong copy of the repeat by *cross_match* if errors in the low-quality part of the read result in a higher Smith–Waterman score against the wrong copy. Such cases are generally revealed by examining all matches of the read to the same and other cosmids.
3. *Subclone Mutation.* Some errors in high-quality trace regions appeared as a single inserted or deleted base in a long mononucleotide run, or as a single deleted unit of a microsatellite repeat. In several such cases another read obtained from the same subclone showed the same discrepancy. These examples appear to represent spontaneous subclone mutations. Another type of subclone mutation seen was a larger deletion, apparently mediated by recombination involving an imperfect direct repeat in the subclone. In this case the alignment found by *cross_match* included the part of the read lying to one side of the deletion, but extended spuriously into the region on the other side of the deletion (often for a significant distance) because of the direct repeat, resulting in multiple high-quality discrepancies.
4. *Unremoved Vector.* Occasionally multiple errors in the sequencing vector part of a read prevented *cross_match* from finding the match between it and the vector sequence, so that it remained unmasked; and the alignment of the insert against the cosmid sequence found by *cross_match* continued into this unmasked sequence because of a few fortuitous base matches at the end of the vector sequence, and the extension included spurious high-quality discrepancies.
5. *Misalignment.* In some cases *phred* deleted the first base in a mononucleotide run of three or more bases, when imperfect mobility correction shifted the run close to the preceding peak, but (owing to the details of the Smith–Waterman algorithm) the alignment constructed by *cross_match* instead indicated the site of the deletion to be the last base in the run. The called bases on each side of the deleted peak received low-quality values because of bad peak spacing and signal to noise, but the last base in the run received a high-quality value.

To identify and remove as many of the above cases as possible, we examined each read having two or more discrepancies with quality values of 40 or more, using *consed* (Gordon et al. 1998). Reads that appeared to be unambiguous instances of one of the first four types above were eliminated; in all, 10 reads (from a total of 26091) were removed from cosmid set 3, and 35 (from a total of 27184) were removed from cosmid set 4. It is likely that a small number of spurious high-quality discrepancies of the above type occurred only once in a read and thus escaped detection by this procedure and that others occurring at quality values below 40 were also not detected. Instances of high quality discrepancies occurring in mononucleotide runs of two or more bases were also inspected and (in cases where the wrong base was flagged) the quality value altered to that of the base actually deleted by *phred.* This resulted in changing 10 quality values in cosmid set 3, and 30 in set 4. Following these corrections to the data set, the observed quality values were recalculated.

RESULTS AND DISCUSSION

Initial inspection of traces and quality values using *consed* (Gordon et al. 1998) indicated that the quality designations obtained using the procedure described in Methods correspond fairly well to quality judgments by a human reviewer of the data. The quality values tend to decrease later in the trace because of deterioration in the peak resolution and uncalled/called parameter values as peaks become wider, smaller, and noisier. Similarly, noisy or weak signal regions earlier in the trace, as in the vicinity of the dye primer peak or peak dropouts, or in traces from failed sequencing reactions, tend to have their quality values reduced due to a poor uncalled/called ratio. In higher quality parts of the trace, compressions tend to receive lower quality values because of uneven peak spacing, poor resolution, and/or poor uncalled/called ratio (in cases where a shifted peak is dropped by the base caller), except in rare instances (see below).

For a more rigorous examination we analyzed the quality values in two test sets of cosmids (Table 1), considering the dye primer and dye terminator reads separately. *Phred* was used to assign quality values to each base-call as described (Methods, Quality Assignment). Note that the test sets differed in important respects (sequencing polymerase and chemistry, sequencing machines and running conditions) from the training sets (Table 1), so the results to some extent indicate how robust the quality values are to changes in sequencing methods.

Validity Tests

For each predicted quality value q, the numbers $corr_q$ of correct and err_q of incorrect base-calls having that quality in the test sets were counted and the observed quality value was computed as

$$q_{\text{obs}}(q) = -10 \times \log_{10}\left(\frac{err_q}{corr_q + err_q}\right)$$

Figure 1 displays the observed quality values as a function of the *phred* predicted values. (Table 2 indicates the number of bases in each quality range used to determine the points in Fig. 1.) The slope one line corresponds to perfect agreement; points above the line (i.e., observed quality value higher than predicted) indicate an observed error rate less than *phred* predicts, while points below the line indicate an observed error rate greater than predicted.

Most points lie on or very near the slope one line, indicating that the error probabilities are reasonably valid. In the quality range below 20 (pre-

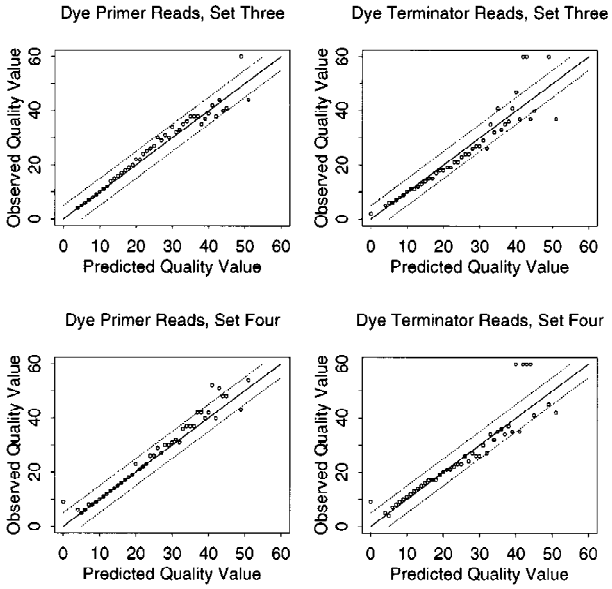

Figure 1 *Phred* quality value validity. The solid line corresponds to perfect agreement between observed and predicted, and the dotted lines indicate a deviation of ±5 (or roughly a factor of 3 in the corresponding error probabilities). Predicted quality values with no errors among the aligned bases are assigned observed quality value 60. Quality values 1, 2, 3, 46, 47, 48, and 50 are unused because the calibration procedure did not produce lookup table lines at these values.

Table 2. Base Counts for Data Used in Fig. 1.

Quality value range	Dye primer reads		Dye terminator reads	
	set 3	set 4	set 3	set 4
0–5	53994	28251	1739734	110436
6–10	1608428	466490	2445065	318662
11–15	810982	211506	1074522	117609
16–20	576383	126763	995229	85179
21–25	728923	145538	1005002	97460
26–30	1002981	141353	972248	86930
31–35	1575967	235986	1628439	148591
36–40	1552397	260893	2507051	182187
41–45	724021	167783	2427413	127885
46–50	23793	25826	391942	28218
51	249206	36241	2118720	28322

dicted error rates above 1%), agreement is very good. The majority of calls in this range occur where resolution is poor or signal-to-noise is low, particularly at the ends of reads or in low-quality traces.

However, there are some (generally small) systematic biases for quality values above 20. In general the error rates tend to be slightly overpredicted in the dye primer data and underpredicted in the dye terminator data, with the trends becoming more pronounced at higher quality levels (the effect is less apparent at the highest values because of higher dispersion resulting from small sample sizes). Such a pattern is consistent with a small systematic excess of errors in the dye terminator test sets relative to the training set, and a small systematic excess of errors in the training set relative to the dye primer test sets. Such excesses would be expected to become increasingly magnified towards the high end of the quality scale as they would represent a larger fraction of the errors as the absolute number of errors decreases.

To investigate this possibility we examined a selection of erroneous base calls having a predicted quality value $\geqslant$30. For dye primer reads we examined all errors with quality $\geqslant$40 in set 4; apart from a few apparent subclone mutations undetected by our previous screen (Methods, Quality Assignment), essentially all of these were CC or GG compressions, in which two adjacent peaks of the same base merge into a single peak, without appreciably disturbing the peak spacing. These are difficult to detect by eye, essentially the only clue being an increase in the size of the merged peak relative to surrounding peaks. For dye terminator reads we examined all errors with quality $\geqslant$30 in 11 of the set 3 cosmids. These were mostly attributable to peak dropouts reflecting poor incorporation of a particular dye-labeled dideoxy base by the polymerase in certain sequence contexts; in such cases the base-caller may call a small noise peak instead of the correct base, and if that peak is in the expected location and there are no other noise peaks nearby, the parameter values may still all be in the good range. About 85% of the high-quality dye terminator errors resulted from a missing G peak following an A (Lee et al. 1992; Parker et al. 1996), or a missing A following a T; a missing T following an A occurred five times, but in only one cosmid of the 11 examined. There was also a missing T peak following a G peak, and an apparent AA compression.

In light of these observations, we interpret the slight deviations between the observed and predicted error rates as follows. First, there is a slight tendency to overcount errors in the training set relative to the test sets owing both to the small sample correction (which was applied to the training set but not the test set) and to spurious error contributions of the types discussed in Methods, Quality Assignment (which we attempted to remove from the test set, but not the training set). This produces a slight upward bias in the predicted error rates, which should be very small at the lower quality levels (where the number of errors is high), but somewhat larger at higher levels. This presumably accounts for the tendency to overestimate error probabilities at higher quality values for the dye–primer data.

The bias toward underestimating error rates in the dye terminator test data is probably attributable to the fact that the training data set consists almost entirely of dye primer reads. As a result, peak dropouts with good parameter values of the sort occasionally seen in terminator reads were largely absent from the training data, so that the error rate in the terminator test data at each quality value is higher than predicted from the training set. This effect apparently outweighs the opposing bias that comes from overcounting errors in the training set. A similar effect does not occur with compression errors in the dye primer test set data, because the training set contains compression errors at a similar rate and the error probability calibration therefore reflects them.

It is important to point out that both the error probability calibration procedure and the tests described above used only the alignable parts of the reads, that is, the parts accurate enough to align against the final sequence given the specifics of the alignment algorithm (that being the only part

where one can reliably distinguish correct and incorrect base-calls). In assigning quality values to a new read, however, it is not known in advance which part is alignable and which is not, so all bases are assigned quality values. The unaligned part is generally less accurate and does indeed tend to receive substantially lower quality values; but strictly speaking the quality estimates are not expected to be valid there. In particular one occasionally finds islands in this part consisting of a few high-quality bases, which on inspection are clearly not accurate, surrounded by low-quality bases.

Discrimination Power

Figure 2 indicates the distribution of the *phred* quality values for the alignable bases in the test sets. (Alignable bases here refers to the bases in each read that could be aligned against the final sequence using *cross_match*, with the same criteria used in *phrap* assembly. This corresponds to the part of the read usable in assembly, and includes the portion of the read up to the point where, roughly, the discrepancy rate starts to exceed 30%.)

The spread of the distribution suggests that the quality values are doing a reasonable job of discriminating accurate from erroneous calls, and in particular that a significant fraction of basecalls are singled out as having a high accuracy. The discrimination power as defined earlier cannot be read directly from the graphs, but is calculated to be as follows:

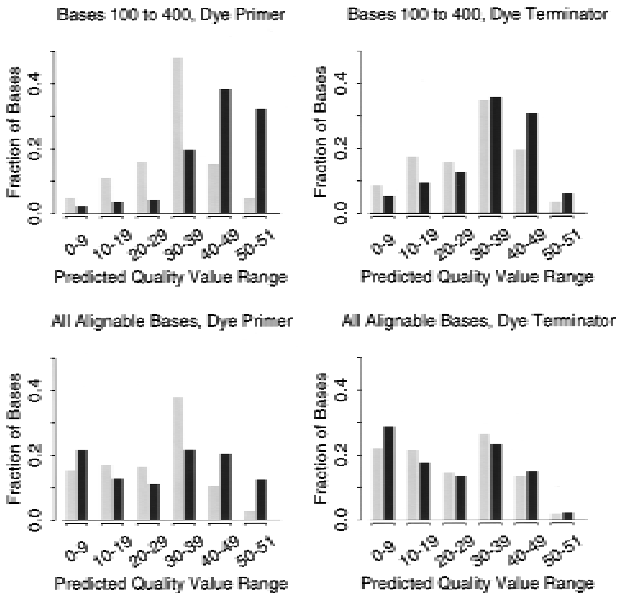

Figure 2 Quality value distributions for cosmid sets three (gray bars) and four (black bars).

16% of the alignable bases in set 3, and 52% of the alignable bases in set 4, can be discriminated as having an error rate <1 per 10 kb; 65% of the alignable bases in set 3, and 65% of the alignable bases in set 4, can be discriminated as having an error rate <1 per kb; 80% of the alignable bases in set 3, and 78% of the alignable bases in set 4, can be discriminated as having an error rate <1 per 100 bases.

One significant implication of the above results is that the highest quality data from a single trace, even in the absence of confirming reads on the opposite strand, can often ensure an error rate of <1 per 10 kb. This has important consequences for finishing (Gordon et al. 1998), since even if the target error rate for the cosmid is 1 per 10 kb it is unnecessary to obtain opposite strand coverage, or do any editing, for such a region.

Possible Improvements

Further improvements to the error probabilities are certainly possible, with respect to both validity and discrimination power. As with the accuracy of basecalling (Ewing et al. 1998) the high-quality part of the trace is the most important in practice and is therefore where improvements will have the greatest impact.

Regarding validity, the studies above suggest that the error probabilities are generally valid independent of sequencing chemistry and machine running conditions; preliminary studies of data generated on ABI 377 machines (which represents a significant fraction of cosmid set 3 reads), and/or processed using our program *plan* (B. Ewing and P. Green, in prep.) instead of the ABI lane-processing software, suggest that the error probabilities retain their validity in this context. Nonetheless some improvement at the high quality end of the scale can be gained by having separate training sets for dye primer and dye terminator traces, and by removing the various spurious contributions to the error counts in the training set.

An important additional issue is whether the error probabilities remain valid when the $G + C$ content is significantly higher than in the cosmids studied here. In dye–primer data the frequency with which compressions occur tends to correlate with $G + C$ content since $G + C$-rich strands are more likely to form the stable hairpin structures which cause compressions. Thus one might expect that in more $G + C$-rich sequence high quality compression errors occur at a significantly higher rate than would be predicted from our current training set. If so, it will be necessary to recalibrate the error prob-

ability lookup table on such data. It should also be possible to improve the accuracy of base-calling through compressions and improve discrimination as discussed below, either of which would tend to reduce the magnitude of the bias.

Improvements in discrimination power should also be possible. Because the main types of high quality base-calling errors are compressions in dye primer traces and peak dropouts in dye terminator traces, improved discrimination may come from using trace parameters that are more sensitive to these types of errors. Both cases involve peaks of abnormal size (large in the case of the dye primer compressions, small in the case of the dye–terminator peak dropouts), so use of relative peak size as a parameter may improve discrimination power. Moreover, because in both cases, sequence context plays an important role, discrimination may be improved by including relevant features of the read sequence as parameters.

Program Availability

C source code for *phred* is available at no charge to academic researchers for research purposes, and by commercial license from the University of Washington to other users; contact Brent Ewing at bge@u.washington.edu.

ACKNOWLEDGMENTS

This work was partly supported by grants from the National Human Genome Research Institute. We thank several people for helpful suggestions and/or data sets, in particular LaDeana Hillier, Bob Waterston, Shawn Iadonato, and Lee Rowen.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked ''advertisement'' in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Berno, A.J. 1996. A graph theoretic approach to the analysis of DNA sequencing data. *Genome Res.* 6: 80–91.

Ewing, B., L. Hillier, M. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using *Phred.* I. Accuracy assessment. *Genome Res.* (this issue).

Giddings, M.C., R.L. Brumley Jr., M. Haker, and L.M. Smith. 1993. An adaptive, object oriented strategy for base calling in DNA sequence analysis. *Nucleic Acids Res.* 21: 4530–4540.

Golden, J.B., D. Torgersen, and C. Tibbetts. 1993. Pattern recognition for automated DNA sequencing: I. On-line signal conditioning and feature extraction for basecalling. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology* (ed. L. Hunter, D. Searls, and J. Shavlick), pp. 136–144. AAAI Press, Menlo Park, CA.

Gordon, D., C. Abajian, and P. Green. 1998. *Consed:* A graphical tool for sequence finishing. *Genome Res.* (this issue).

Lawrence, C.B. and V.V. Solovyev. 1994. Assignment of position-specific error probability to primary DNA sequence data. *Nucleic Acids Res.* 22: 1272–1280.

Lee, L.G., C.R. Connell, S.L. Woo, R.D. Cheng, B.F. MacArdle, C.W. Fuller, N.D. Halloran, and R.K. Wilson. 1992. DNA sequencing with dye-labeled terminators and T7 DNA polymerase: Effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments. *Nucleic Acids Res.* 20: 2471–2483.

Parker, L.T., H. Zakeri, Q. Deng, S. Spurgeon, P.-Y. Kwok, and D.A. Nickerson. 1996. AmpliTaq DNA polymerase, FS dye-terminator sequencing: Analysis of peak height patterns. *BioTechniques* 21: 694–699.

Received December 12, 1997; accepted in revised form February 3, 1998.