Volume 17 Number 10 1989 Nucleic Acids Research

Sequence errors described in GenBank: a means to determine the accuracy of DNA sequence interpretation

Stephen A.Krawetz

Department of Molecular Biology and Genetics and The Center for Molecular Biology, Wayne State University, 4th Floor MCHT, Laboratory 13, 2727 2nd Avenue, Detroit, MI 48201, USA

Received November 14, 1988; Revised and Accepted April 18, 1989

ABSTRACT

The accuracy of nucleic acid sequence data interpretation was determined by assessing and quantifying the discrepancies reported in the GenBank database. This permitted the calculation of an Error Rate (ER) for nucleic acid sequence determination. If one assumes that most entries (TB, Total Bases) were independently verified or those without reported discrepancies were correct, the ER is 0.368 errors per 1000 bases. However, if one assumes that only those sequences with reported discrepancies (TBIQ, Total Bases from entries In Question) were verified and are thus correct, the ER is 2.887 errors per 1000 bases. This establishes the first set of limit boundaries of the ER for sequence interpretation and sequence errors within the GenBank database and provides the foundation for future assessments and the monitoring of sequence data accumulation. In addition, the ER measure provides a basis to evaluate the efficiency and merit of present and future automated nucleic acid sequencing technologies which will have a direct impact upon the final outcome of the "Human Genome Initiative".

INTRODUCTION

DNA sequencing has become an integral requisite component that is utilized to address many of the questions molecular biology now poses. This can be directly attributed to the development of suitable technologies (1,2) to easily sequence the segment(s) of DNA of interest. Recently several approaches have been pursued to automate part(s) of this process (3,4,5,6,7 and references therein) in an attempt to increase the throughput and free the researcher from this labor intensive "error prone" task. This has essentially removed certain sources of procedural, but not interpretive errors (8). Although automation has minimized these sources of error, the accuracy rates reported for the automated sequencing machines varies from 94 % to 99 % (3,4,5,6) and is dependent on the optimum sensitivity threshold of detection (7). It is apparent that accuracy decreases markedly as attempts are made to interpret the sequence in the upper regions of the gel (4). This is not unexpected considering that the information content increases logarithmically with decreasing distance from the origin and the complex nature of the interpretive rule based system for nucleotide base assignment for both chain termination (1,9) and chemical (2) DNA sequencing. The minimal cost, ease

and rapidity of manual DNA sequencing lends itself to minimization of interpretive errors with the use of multiple repetitions of the same sequence determination employing slightly varied conditions. This has culminated in the exponential growth of the GenBank (10,11) database and the beginning of the "Human Genome Initiative". Although it has generally been believed that the accumulation of sequence data has proceeded in an accurate manner, the issue of accuracy of DNA sequence data interpretation and its impact on the GenBank (10,11) database and the "Human Genome Initiative" remains to be critically examined. To address these issues the Error Rate (ER) and types of errors were determined for GenBank (10,11) release 55. This release included new sequence data from release 13 of EMBL (12) and release 2 of the DDBJ (DNA Data Bank of Japan; 13) databases.

RESULTS

The comments sections (locus, definition, accession, keywords, segment, source, reference, comment, features, base count, origin) for sequence entries of GenBank (10,11) release 55 were initially searched for the keywords: conflict or revision or unsure, as given in the features key names (10,11), utilizing the Quest (14) program. This identified the vast majority of sequences which contained well-defined and documented discrepancies. However, review of the comments sections revealed that the terms corrections, differences and error were also utilized to identify sequences in which discrepancies were noted. In order to ensure that all sequences containing discrepancies were identified, GenBank release 55 was again searched for the keywords: corrections or differences or error, utilizing the Quest (14) program. In this case, sequences containing substantial revisions and transcription entry errors (identified by the search term error) were identified. The sequences thus identified were subsequently assessed to define and classify the discrepancy(ies). When sequence differences were identified as either a revision or conflict, the sequences were matched to produce an alignment with the minimum number of insertions or deletions or inconsistencies (mismatches) at each position. Each one of these positions (insertion or deletion or inconsistency) was considered a discrepancy. Thus the values for insertions or deletions and inconsistencies represent a minimum estimate. Results of this analysis are shown in Table 1.

Discrepancies were divided into the following four groups: i) insertions or deletions, ii) inconsistencies, iii) unsure, iv) errors, and in some cases their origin was documented in the corresponding comments section. As shown in Table 1, the majority of the discrepancies were insertions or deletions. These were often attributed to compressions arising from template secondary structure. The second largest group, inconsistencies (different bases assigned to the identical position which can be thought of as mismatches between two aligned sequences), were frequently ascribed to cloning

artifacts. In most cases these became apparent when cDNA and genomic sequences were compared. Variations between sequences were also identified in this group, however, these were usually attributed to polymorphic variation. Since this represents a natural phenomena, polymorphic variations were not considered discrepancies and were omitted from all ER calculations. Of the two largest groups, 39.6 % [(revisions x 100/ conflict + revisions) i.e. {Table 1: 1062 x 100/2682}] of the identified discrepancies can be attributed to sequence revisions from the laboratory of origin. The remaining 60.4 % discrepancies were conflicts between laboratories. The third group, unsure bases, frequently resulted from gels of poor quality, rendering data difficult to interpret. In this case, the worst scenario was assumed and these bases were considered in error. The fourth group, errors, which was identified by the search term error, contained the least number of discrepancies and primarily reflected typographical mistakes, noted by the annotators. These errors were either introduced at the source, or upon transcription into the database.

Of the 13 divisions in GenBank (10,11) only 12 are shown in Table 1. The most expansive division, labeled unannotated (3,921,518 bases), was not included since this division contains the most recent sequences entered and as such, independent verification was not expected. In addition, the lack of any significant annotation (comments sections) precluded the detection of any discrepancies utilizing the approach described in this paper.

As expected, the total number of discrepancies were directly proportional to the total number of bases entered in that division. In most cases this relationship remianed constant and thus it did not dramatically effect the TB (Total Bases) or the TBIQ (Total number of Bases from entries In Question) ER (Error Rate). It should be noted that, the extent and nature of some discrepancies were not described in the comments sections of the GenBank database and were excluded when calculating the TB ER and TBIQ ER. It must be emphasized that if all bases from these sequences were considered questionable, 88,806 (VRL, 20,580 + PLN, 1,255 + INV 2,851 + ORG 32,333 + VRT 2,194 + RNA 1,823 + ROD 18,532 + PRI 9,238) bases (4.2 % of the total) would be added to the total number of bases in question and would affect the individual and average TBIQ ER accordingly.

The highest ER was observed in the RNA (structural RNA) division which had the lowest total number of bases. In this instance, template secondary structure would present difficulty in sequencing and interpreting data. As expected when the sequence was known the lowest ER was observed (i.e. SYN, synthetic division; chemically or enzymatically synthesized or constructed). If one eliminates these two data points, with the supposition that they are not representative because of the limited number of sequence entries (RNA and SYN division) and that the SYN division sequences were

Division	Total Bases (TB)	Total Bases from Entries in Question (TBIQ)	Inconsistencies	Insertions or Deletions	Unsure	Errors	Total Discrepancies	TB Error Rate	TBIQ Error Rate
VRL	2,519,726	428,686[a]	552	285	184	4	1,025	0.41	2.39
PLN	1,502,452	141,672[a]	117	176	144	5	442	0.29	3.12
BCT	2,096,651	319,346	279	328	24	3	645[b,c]	0.31	2.02
INV	1,096,431	72,821[a]	98	124	38	3	263	0.24	3.61
ORG	1,048,359	132,401[a]	96	113	249	4	466	0.44	3.52
MAM	523,751	58,279	69	64	30	0	163[d]	0.31	2.80
VRT	713,361	103,076[a]	106	119	30	5	260	0.36	2.52
RNA	97,268	5,471[a]	109	56	36	6	267[b]	2.74	48.80
SYN	116,001	5,104	6	1	0	1	8	0.07	1.57
ROD	2,415,457	300,336[a]	530	696	373	4	1,603	0.66	5.34
PRI	2,727,863	412,328[a]	461	452	297	0	1,248[b]	0.46	3.03
PHG	377,164	144,288	40	32	3	0	75	0.20	0.52

known a priori, the average TB ER is 0.368 per 1000 bases [Table 1: (Sum of TB Error Rates - RNA TB Error Rate - SYN TB Error Rate) / 10 divisions]. Alternatively, assuming that only those sequences with reported discrepancies were verified, the average TBIQ ER is 2.887 errors per 1000 bases. It must be emphasized that these ER values are estimates based on very different assumptions. The first approximation assumes that all other sequences were correct while the second approach does not consider entries without a reported discrepancy correct or incorrect. This being the case, one would assume that the actual ER is within these boundaries, suggesting that the estimate of the average TBIQ ER approximates the upper limit boundary of the ER of data interpretation.

DISCUSSION

It is apparent that the TB ER approximates that of the viral reverse transcriptases (15,16) but is higher than that of the DNA polymerases (17,18,19). This contrasts the TBIQ ER, as it was significantly higher (ER of 2.887 errors per 1000 bases). Given the assumption that in any sequencing project the sequence that is reported is derived from a consensus of many independent determinations (9) one would expect to eliminate any random errors, thus these ERs are higher than expected. This would support the view

Table 1. Frequency and Types of Errors Identified in the GenBank Database. A search of the comments sections of GenBank (10,11) release 55 was carried out utilizing the Quest (14) program for the following keywords: conflict or revision or unsure or corrections or differences or error. Sequences that were identified were subjected to further analysis to define and determine the nature and extent of the discrepancy. Inconsistencies are the sum of the bases within that division that could not be resolved as insertions or deletions. Insertions or deletions are the sum of additional or missing bases within that division. Unsure bases are the sum of the bases within that division for which identity was not defined. Errors are the sum of the bases that were incorrectly entered and since corrected within that division. Total discrepancies are the sum of the bases identified as inconsistencies plus insertions or deletions plus unsure plus errors within that division. TB (total number of bases of all the entries) ER (Error Rate) is the total number of discrepancies x 1000/TB in that division. TBIQ (total number of bases from entries in question) ER is the total number of discrepancies x 1000/TBIQ in that division. The divisions are annotated utilizing the three letter GenBank (10,11) identifier (VRL, viral; PLN, plant; BCT, bacterial; INV, invertebrate; ORG, organelle; MAM, other mammalian; VRT, other vertebrate; RNA, structural RNA; ROD, rodent; PRI, primate; PHG, phage).[a]Discrepancies not reported due to their extensive nature and thus not included in the calculation of TBIQ. In the case of the VRL division, this accounted for 20,580; PLN, 1,255; INV 2,851; ORG 32,333; VRT 2,194; RNA 1,823; ROD 18,532; PRI 9,238 bases.[b]Discrepancies were reported but not identified (BCT, sequence ECOMOTAB; RNA, sequence ECORRG1; PRI,sequence HUMIGFIG3) and could not be categorized but were included in the calculation of total discrepancies. [c]Sequence ECOTRP was previously corrected but those corrections were not noted and could not be categorized or included in the calculation of total discrepancies. [d]Conflicts for sequence BOVGH were resolved amongst the laboratories and most differences were not reported. These could not be categorized or included in the calculation of total discrepancies.

that in the case of manual sequencing, the human interpreter contributes significantly to the introduction of sequence errors at the level of interpretation. Comparatively, even with automation, the available machine rule based systems still impart a high level of errors (10-60/1000;3,4,5,6).

Examination of how the discrepancies once identified were resolved, revealed that it often involved the repetitive task of resequencing that segment of DNA several times. To address this need and to increase throughput, automated devices have become available (3,4,5,6,7). Unfortunately all interpret data with significantly higher ERs (3,4,5,6), approaching or above that of the TBIQ for structural RNA. This suggests that although these automated devices represent a considerable technological advance, before they become a truly viable alternative to manual technologies, their interpretive component will require significant refinement.

It is evident that collation and deposition of nucleic acid sequence data into GenBank has proceeded in a relatively accurate manner (Table 1, column 7 Errors). The estimates of ERs that are provided by this analysis are based on a large sample size that is independent of sequence or acquired knowledge base, thus providing a relatively unbiased measure against which automated sequencing technologies and the accuracy of sequence interpretation can be judged.

ACKNOWLEDGEMENT

Support for this research was from an establishment grant from the Department of Molecular Biology and Genetics and The Center for Molecular Biology, Wayne State University. Support from the AHFMR is gratefully acknowledged. Computer resources used to carry out this study was provided by the BIONET National Computer Resource for Molecular Biology which is funded by the Biomedical Research Technology Program Division of Research Resources, National Institutes of Health, Grant Number P41RR01685.

REFERENCES

1) F. Sanger, S. Nicklen, and A.R. Coulson, Proc. Natl. Acad. Sci. U.S.A. 74, 5463 (1977).
2) A.M. Maxam, and W. Gilbert, Proc. Natl. Acad. Sci. U.S.A. 74, 560 (1977).
3) L.M. Smith, et al., Nature 321, 674 (1986).
4) J.K. Elder, D.K. Green, and E.M. Southern, Nuc. Acids Res. 14, 417 (1986).
5) L. Johnston-Dow, et al., BioTechniques 5, 754 (1987).
6) C. Connell et al., BioTechniques 5, 342 (1987).
7) T.P. Keenan and S.A. Krawetz, CABIOS 4, 203 (1988).
8) B. Baum, Nature 305, 90 (1983).
9) J. Hindley, laboratory techniques in biochemistry and molecular biology DNA sequencing (Elsevier Biomedical Press, New York, 1983)
10) GenBank (1988) The Los Alamos National Laboratory Release 55; H.S. Bilofsky and C. Burks, Nuc. Acids Res. 16, 1861 (1988).

11) C. Burks et al., CABIOS 1, 255 (1985).
12) G.N. Cameron, Nuc. Acids Res. 16, 1865 (1988).
13) DDBJ, DNA Data Bank of Japan, National Institute of Genetics, Teikyo University, Tokyo, Japan.
14) Bionet, IntelliGenetics Inc. Core program Quest version 5.1 (1988).
15) G.F. Gerard, Focus 8, 12 (1986).
16) J.D. Roberts et al., Science 242, 1171, (1988).
17) L.A. Loeb and M.E. Reyland, In F. Eckstein and D.M.J. Lilley (eds.) Nucleic Acids and Molecular Biology-Fidelity of DNA Synthesis, Springer-Verlag, Berlin, pp 157-173 (1987).
18) T.A. Kunkel, L.A. Loeb, Science 213, 765 (1981).
19) T.A. Kunkel, K.R. Tindall, Biochemistry 27, 6008 (1988).