The implications of errors in molecular sequence data

David J. States
National Center for Biotechnology Information
National Library of Medicine

Submitted to Trends in Genetics

Address correspondence to:

David J. States, M.D., Ph.D.
National Center for Biotechnology Information
National Library of Medicine
Building 38A, Room 8S806
Bethesda, MD...20894

Telephone: (301) 496-2475
FAX: (301) 480-9241

Summary

Molecular sequence is experimentally derived data and can be expected to contain errors at a finite rate as a result of diverse phenomena such as biological variation, molecular cloning artifacts, imperfect sequence determination, and data handling during contig assembly. The presence of errors will affect the reliability of database searches and sequence alignments, but their impact may be minimized by the use of analytic techniques which anticipate the presence of imperfect data.

Introduction

At a molecular level, a genome has a unique and absolute structure with a precisely defined sequence. That sequence may be probed through a variety of experimental techniques. Molecular sequence analysis has become an established technique in fields from medical genetics to forensic science, and it appears likely that large scale genome sequencing efforts will impact nearly all areas of biomedical science. The results of sequencing, like any experimental data, are subject to error. In order to properly interpret and utilize sequence data, an understanding of the frequency and characteristics of these errors is essential.

Sources of Error

Biological variation is perhaps the most important source of variability in molecular sequence data. While one may debate whether biological variability is “error”, sequences derived from two different individuals need not agree. For a species such as HIV where the polymerase misincorporation rate has been estimated at 1 in 1700 [1], the error frequency is comparable to the size of the genome. Since most mutations are lethal, any given virion may or may not contain a viable genome.These concerns are not limited to viruses or even lower organisms. RFLP studies suggest that the human genome carries polymorphisms at approximately 1 on 270 bases of non-coding sequences [2], and spontaneous mutation remains a major source of new cases in X-linked and autosomal dominant disorders such as Duchenne muscular dystrophy and tuberous sclerosis.

Sequencing is, in general, dependent on molecular cloning, a complex process with several steps where sequence errors might be introduced. Thermodynamics dictates that all polymerases will misincorporate bases at some finite rate. When used *in vitro*, polymerases do not have the benefit of the cellular error correcting apparatus, and the conditions for polymerization may be less than ideal. The retroviral reverse transciptases used to copy mRNA into cDNA for cloning typically have a misincorporation rate on the order of 1 in 17000 bases [3]. If the results of the *in vitro* polymerization reaction are cloned, errors present in the single parental molecule will become fixed and carried by all progeny molecules. Polymerase chain reaction (PCR) based cloning strategies are particularly error prone because the polymerization reaction is repeated many times *in vitro*. Since each cycle accumulates new errors, the molecules

generated by PCR may have error rates in excess of 1 in 1000 which will be evident when they are cloned. If the products of PCR amplification are analyzed directly, without cloning, (direct PCR sequencing or hybridization, for example) then the population averaging of errors avoids the problems of clonal selection and the correct *average* sequence is observed.

Errors may also be introduced in the process of cloning genomic DNA. Samples must be manipulated extensively *in vitro* in any cloning strategy and will necessarily be subject to mechanical shear, O_2 oxidation, photochemical damage, and the action of a variety of reactive chemicals such as phenol. It is also likely that foreign DNA will differ from the cloning host genome in base composition, methylation pattern, chromatin binding sites, transcription and replication control signals. Therefore, the possibility that there will be a selective pressure in favor of mutations which will "correct" these deviations must be anticipated.

Large scale rearrangements during molecular cloning are also possible, particularly in vectors carrying large sized inserts such as yeast artificial chromosomes (YACs) or cosmids. Olson has estimated that 2% of the YACs in his library are clonally unstable [4]. While the rearrangement frequencies for cosmids do not appear to be as high, numerous rearrangements in cosmid, plasmid, and phage vector inserts have been observed. Of particular concern are reports of "poison" sequences which prevent the propagation of an insert until modified or deleted [5]. In such cases, comparison of multiple independent clones may not be sufficient to establish the true genomic structure.Techniques such as Southern blotting and direct PCR sequencing, which do not depend on any cloning step, may be useful in verifying the results of sequence analysis on cloned DNA.

Errors in sequence determination may be random or biased. Polymerase/ terminator sequencing methods depend on uniform incorporation of chain terminators. Significant advances have been made in identifying polymerases and reaction conditions where the incorporation of terminator is uniform and sequence independent [6] although residual template secondary structure artifacts may still be observed. Both polymerase/ terminator and chemical sequencing techniques depend on an electrophoretic size fractionation of the reaction products to read the sequence information. Electrophoretic

mobility is a decreasing function of fragment size, but the presence of secondary structures refractory to the denaturing conditions in the electrophoresis gel may distort this relationship, usually resulting in ambiguous regions during gel reading rather than occult errors. The inability to count homopolymer runs correctly (leading to single base insertions or deletions) and errors in the maintenance of lane to lane registration (leading to single base exchanges) are the tasks most sensitive to finite gel resolution and becomes the dominant error modes on long gel runs.

Sequence Assembly Process

Raw sequence data must be assembled into larger continuous regions of sequence. The assembly process may propagate and modify errors. In a region covered by 10 overlapping subclones each with a raw error rate of 0.5%, the directly assembled data will have an of average one discrepancy per 20 bases. Ideally, errors should be resolved during the assembly process and the assembled contig should contain fewer errors than any of the component segments, but this is difficult to achieve automatically and time consuming when performed manually. To accomplish this, information from both strands must be considered, and the relative reliability of different sequence runs must be weighed. Non-linearities in the electrophoretic mobility versus length curve may result from incompletely denatured secondary structure in fragments. This will result in regions of the sequencing gel which are ambiguous. Since such artifacts are dependent on local sequence, they tend to correlate in position even when the position of the sequencing primer is varied. Not infrequently, some segments of sequence remain ambiguous despite repeated attempts at conventional sequencing. In this situation, sequencing reactions using derivatized nucleotides such as ITP or d-aza-GTP may resolve the unreadable segments. Since data from multiple sources with varying reliabilities must be integrated in contig assembly, simple majority rule algorithms may be misleading.

Empirical Estimates of Error Rates in Existing Data

Several empirical estimates of sequence accuracy are available. Krawetz surveyed the GenBank database, and based on the frequency with which sequences have been revised or are in conflict, estimated that the current database contains errors at rates of 2.9 per 1000 [7]. Tolstoshev, at the National Library of Medicine has performed a comparison of the translated sequences for coding regions in GenBank with the

corresponding Protein Information Resource amino acid sequence entry. She finds that with some heuristic rules in the matching algorithm it is possible to perfectly match about 75% of the comparable entries, but that discrepancies remain in about 25% of cases. Since the average sequence is several hundred bases long, this again suggests an underlying error rate on the order of a few per thousand bases. Finally, in preparing the E. coli genetic map, Rudd et al have identified 161 segments of DNA which were sequenced independently in different laboratories. Of these, 66 disagree at least one site. This corresponds to a raw error rate of roughly one per thousand bases [8].

Errors and the Interpretation of Sequence

The impact of data errors on the usable information content of a sequence depends on the way in which the sequence is analyzed. There are theoretical limits to the impact of noise which are imposed by information theory, but the real effect of uncertainties in data is determined by the way the sequence is interpreted [9]. Translation of nucleic acid sequences into amino acid sequence is very sensitive to the presence of insertion or deletion errors. At 1% a insertion or deletion error rate most reading frames longer than 24 amino acids will be disrupted by at least one frameshifting error. Bayesian approaches combining translation and alignment offer a more robust interpretation scheme. As shown in figure 1, protein sequence alignments typically have regions of high sequence similarity separated by gaps and regions of little similarity. Errors occurring in one block of conserved sequence need not affect the alignment of other blocks making the alignment process much less sensitive to the presence of errors than is direct translation of open reading frames.

Sequence alignment may be viewed as the statistical problem of finding the most likely alignment between two sequences given their actual amino acid composition. The widely used PAM model for sequence alignment is, in fact, based on a model for the odds of exchanging one amino acid for another [10]. With David Botstein, I have shown that a Bayesian approach may be applied to probabilisitically combine translation and alignment calculations [11]. In this approach, the probability of an error occurring in the data is weighed against the probability of aligning (or misaligning) each section of the sequence. The second figure shows that it is possible, using this method, to distinguish true sequence alignments from alignments with random sequence even in the presence of 1% insertion deletion error rates and five per cent base substitution error rates. In this

example, the bovine α-lactalbumin mRNA sequence [12] with artificially introduced errors is aligned with the protein sequence of distantly related homolog, chicken lysozyme [13]. The distribution of true alignment scores is clearly distinct from the random sequence alignments.

The effect of sequence errors on homolog identification through database query may also be assessed empirically. Table 1. summarizes the results of protein sequence database searches performed with the true bovine α-lactalbumin mRNA sequence and with a copy containing substitution, insertion, and deletion errors at rates of 1% each. α-lactalbumin is a small protein with many sequenced homologs making it a good test case. Searches were performed using the BLASTX program which uses a computationally efficient algorithm and statistical scoring to identify significant ungapped sequence alignments between translations of all six possible reading frames of a query and a target protein sequence database [14]. With an accurate sequence query, 60 highly significant sequence matches were seen to α-lactalbumins and lysozymes, both of which are true homologs. Six of these matches fell below the significance threshold when errors were present in the query sequence, but 54 of the 60 (90%) were successfully identified even in the presence of sequence data errors. If repeated searches were performed with several sets of error prone data, this success rate would be even higher.

Some applications require, and have achieved, very high sequence accuracy. In medical genetics, sequence accuracies of better than one in 10^4 or better are needed to identify disease alleles. To reach this level of accuracy, raw sequence data for normal and diseased alleles are compared side by side. Wherever discrepancies are identified, comparative studies are pursued and through a process of repetitive sequencing, resequencing, and direct analysis of fresh genomic sequences with oligonucleotide hybridization or PCR sequencing to identify the true mutation site [15].

Formally estimating error rates would be a substantial undertaking and may not be feasible for many laboratories. Even if an attempt were made to determine the accuracy of data produced at a particular laboratory, the estimate would rapidly become invalid due to innovations in sequencing technology, protocols, and personnel turnover. However large scale sequencing laboratories might justify devoting the time and resources necessary to maintain formal estimates of error rates and characteristics.

Whether or not such formal accuracy estimates are made, the ultimate responsibility of the research community is, of course, to produce data that is as accurate as is practicable.

Bibliography

1 Roberts, J.D., Bebenek, K., Kunkel, T.A. (1988) The accuracy of reverse transcriptase from HIV-1. *Science* 242: 1171-3

2 Cooper, D.N., Smith, B.A., Cooke, H.J., Niemann, S., Schmidtke, J. (1985) An estimate of unique DNA sequence heterozygosity in the human genome. *Human Genetics* 69: 201-5

3 Roberts, J.D., Preston, B.D., Johnston, L.A., Soni, A., Loeb, L.A., Kunkel, T.A. (1989) Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis in vitro. *Mol. Cell.Biol.* 9: 469-76

4 Brownstein, B.H., Silverman, G.A., Little, R.D., Burke, D.T., Korsmeyer, S.J., Schlessinger, D., Olson, M.V. (1989) Isolation of single-copy human genes from a library of yeast artificial chromosome clones. *Science* 244: 1348-51

5 Brookes, S., Placzek, M., Moore, R., Dixon, M., Dickson, C., Peters, G. (1986) Insertion elements and transitions in cloned mouse mammary tumour virus DNA: further delineation of the poison sequences. *Nucleic Acids Res.* 14: 8231-45

6 Tabor, S., Richardson, C.C. (1990) DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. Effect of pyrophosphorolysis and metal ions. *J. Biol. Chem.* 265: 8322-8

7 Krawetz, S.A. (1989) Sequence errors described in GenBank: a means to determine the accuracy of DNA sequence interpretation. *Nucleic Acids Res.* 17: 3951-7

8 Rudd, K.E., Miller, W., Ostell, J., Benson, D.A. (1990) Alignment of Escherichia coli K12 DNA sequences to a genomic restriction map. *Nucleic Acids Res.* 18: 313-21

9 Shannon, C.E., and Weaver, W. (1949) The Mathematical Theory of Communication University of Illinois Press, Urbana IL

10 Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1979) in “Atlas of Protein Sequence and Structure” (M.O. Dayhoff, ed.), Volume 5, Suppl. 3, p. 345. National Biomedical Research Foundation, Washington, D.C

11 States, D.J., and Botstein, D. (1991) *Proceedings of the National Academy of Sciences, USA*, in press

12 Hurley,W.L. and Schuler,L.A. (1987) “Molecular cloning and nucleotide sequence of a bovine alpha-lactalbumin cDNA.” *Gene* 61, 119-122

13 Jung, A., Sippel, A.E., Grez, M., Schutz G.(1980) Exons encode functional and structural units of chicken lysozyme *Proc. Nat. Acad. Sci. USA* 77:5759-5763

14 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.* 215: 403-410

15 Levedakou, E.N., Landegren, U., Hood, L.E. (1989) A strategy to study gene polymorphism by direct sequence analysis of cosmid clones and amplified genomic DNA. *Biotechniques* 7: 438-42

16 Pearson, W.R., Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85: 2444-8

Figure1.

The first figure shows the sequence alignment of two distantly related proteins, human neutrophil elastase and rat trypsin generated by the FASTA program [16]. Sites where the aligned sequences are perfectly conserved are shown with a ":" between the sequences and sites substitutions are conservative are shown with a "." between the sequences. The boxed regions contain the residues composing the catalytic triad of the active site.

Figure2.

The second figure shows the effects of sequence errors on alignment scores. 100 trials were performed in which errors were introduced into the bovine lactalbumin mRNA sequence with 5% of the bases randomly substituted and insertion and deletion errors present at 1% of the bases. These "mutated" sequences were then aligned against either the correct chicken lysozyme protein sequence or a randomly sequence of the same amino acid composition. The distribution of alignment scores against the true chicken lysozyme protein sequence is shown in solid, and the distribution of alignment scores against random sequence are shown in cross hatch. All alignments were based on the PAM250 model [10] using a simultaneous translation and alignment algorithm [11]. The score for the lysozyme alignment in the absence of any errors is indicated.

Figure 1.

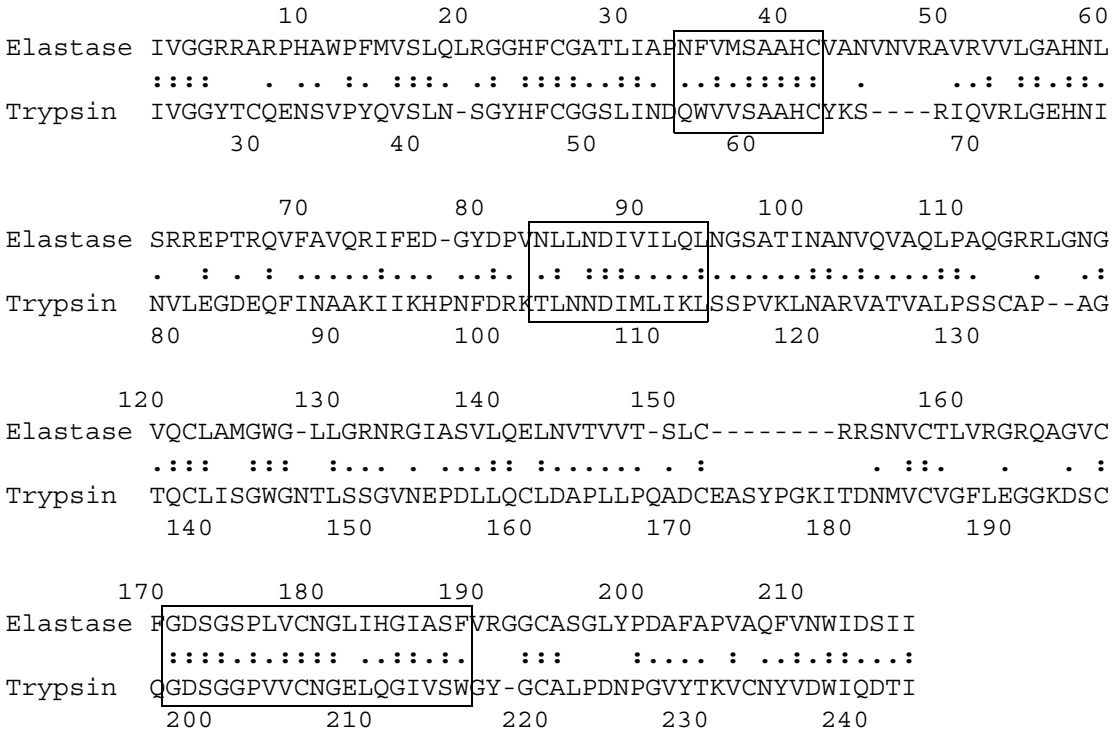
10 20 30 40 50 60
Elastase IVGGRRARPHAWPFMVSLQLRGGHFCGATLIAPNFVMSAAHCVANVNVRAVRVVLGAHNL
Trypsin IVGGYTCQENSVPYQVSLN-SGYHFCGGSLINDQWVVSAAHCYKS----RIQVRLGEHNI
30 40 50 60 70
70 80 90 100 110
Elastase SRREPTRQVFAVQRIFED-GYDPVNLLNDIVILQLNGSATINANVQVAQLPAQGRRLGNG
Trypsin NVLEGDEQFINAAKIIKHPNFDRKTLNNDIMLIKLSSPVKLNARVATVALPSSCAP--AG
80 90 100 110 120 130
120 130 140 150 160
Elastase VQCLAMGWG-LLGRNRGIASVLQELNVTVVT-SLC--------RRSNVCTLVRGRQAGVC
Trypsin TQCLISGWGNTLSSGVNEPDLLQCLDAPLLPQADCEASYPGKITDNMVCVGFLEGGKDSC
140 150 160 170 180 190
170 180 190 200 210
Elastase PGDSGSPLVCNGLIHGIASFVRGGCASGLYPDAFAPVAQFVNWIDSII
Trypsin QGDSGGPVVCNGELQGIVSWGY-GCALPDNPGVYTKVCNYVDWIQDTI
200 210 220 230 240

Figure 2.

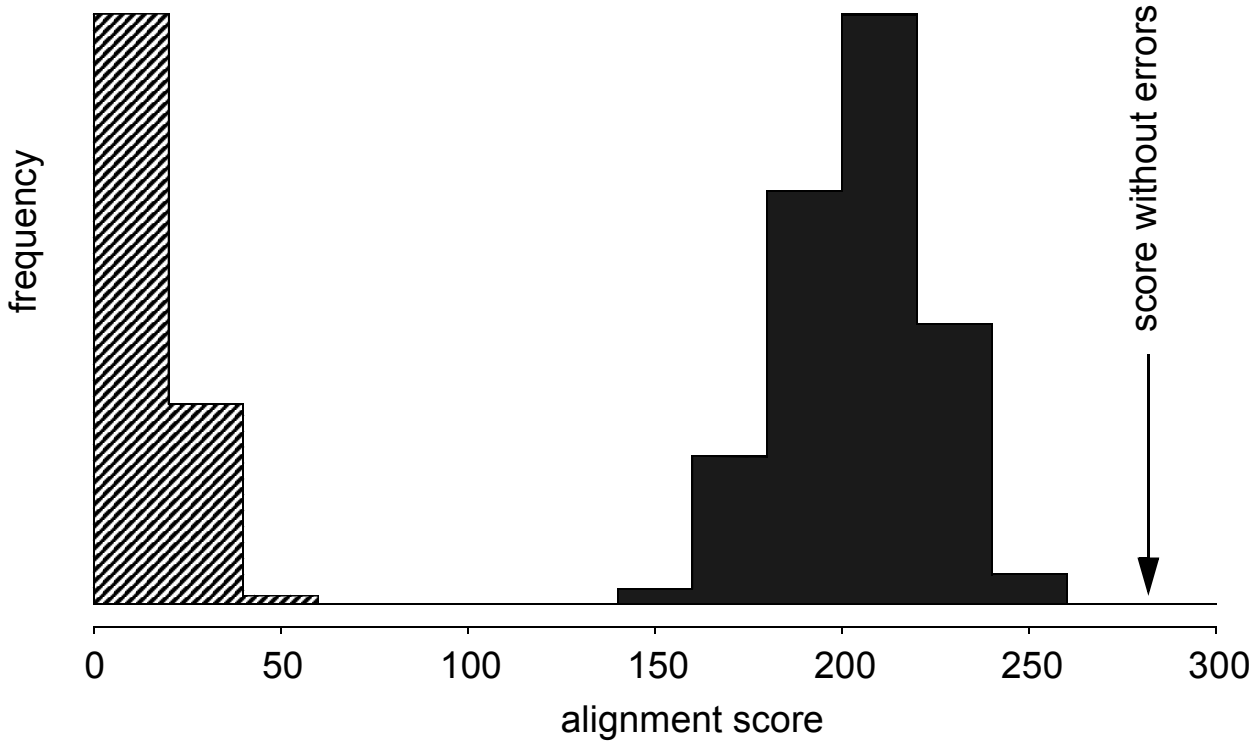
frequency
score without errors
0
50
100
150
200
250
300
alignment score

Table 1.

BLASTX matches with the true bovine α-lactalbumin mRNA sequence:

>A27360 (PIR) α-lactalbumin precursor - Bovine
Frame = +1, Score = 792, Expect = $5.1*10^{-117}$

(17 other α-lactalbumin matches)

>LZBO (PIR) Lysozyme c 2 - Bovine #EC-number 3.2.1.17
Frame = +1, Score = 178, Expect = $5.0*10^{-20}$

(14 other mammalian Lysozyme matches)

>LZPY (PIR) Lysozyme c - Pigeon #EC-number 3.2.1.17
Frame = +1, Score = 159, Expect = $4.7*10^{-17}$

(26 other α-lactalbumin or Lysozyme matches with P<0.01)

BLASTX matches with a "mutated" bovine α-lactalbumin mRNA sequence:

>S02332 (PIR) α-lactalbumin precursor - Bovine
Frame = +2, Score = 352, Expect = $2.2*10^{-47}$
Frame = +1, Score = 278, Expect = $9.3*10^{-38}$

(14 other α-lactalbumin matches)

>LZRT (PIR) Lysozyme - Rat #EC-number 3.2.1.17
Frame = +2, Score = 144, Expect = $9.6*10^{-15}$
Frame = +1, Score = 56, Expect = 0.25

(2 other mammalian Lysozyme matches)

>LZPY (PIR) Lysozyme c - Pigeon #EC-number 3.2.1.17
Frame = +2, Score = 142, Expect = $2.0*10^{-14}$

(35 other α-lactalbumin or lysozyme alignments with P<0.01)

Table 1. shows the effect of sequence errors on a database search performed using the BLASTX program to translate all six reading frames of the bovine α-lactalbumin mRNA sequence and to search the Protein Information Resource (PIR) database. The values labeled "Expect" show the probability with which an alignment of equal score would be expected in searching a database of random sequence and size equal to the PIR database. The "mutated" query sequence was generated by randomly substituting 1% of the bases in the native sequence, then randomly deleting 1% of the bases, followed by inserting random bases at 1% of the sites in sequence. Output from one typical search with a "mutated" query are shown.