Molecular sequence accuracy: analysing imperfect data

DAVID J. STATES

Molecular sequences are experimentally derived data that can be expected to contain errors as a result of diverse phenomena such as biological variation, molecular cloning artifacts, imperfect sequence determination, and data handling during contig assembly. Errors will affect the reliability of database searches and sequence alignments, but their impact may be minimized by the use of analytical techniques that anticipate that the data will be imperfect.

At a molecular level, a genome has a unique and precisely defined sequence. That sequence may be probed through a variety of experimental techniques. Molecular sequences, like any experimental data, are likely to contain errors, and obtaining data of the highest possible accuracy is both costly and time-consuming. Proposals have recently been made to sequence cDNA libraries rapidly but with low accuracy[1]. If sequence data are to be interpreted and used properly, it will be essential to understand the frequency and characteristics of these errors, as well as their impact on database search and sequence alignment algorithms. This is particularly critical when the techniques used to obtain the data are known to be error-prone.

Sources of error

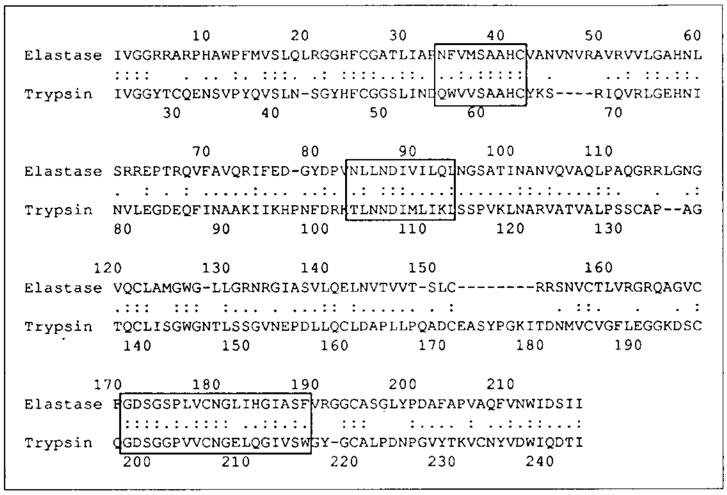
Biological variation is perhaps the most important source of variability in molecular sequence data. While one may debate whether biological variability is 'error', sequences derived from two different individuals need not agree. For an organism such as the human immunodeficiency virus (HIV), in which the polymerase misincorporation rate has been estimated at one in 1700 (Ref. 2), 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 to lower organisms. Restriction fragment length polymorphism (RFLP) studies suggest that the human genome carries polymorphisms at approximately one in 270 bases of noncoding sequences[3], 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 at which sequence errors might be introduced. Thermodynamics dictates that all polymerases will misincorporate bases. 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 transcriptases used to copy mRNA into cDNA for cloning typically have a misincorporation rate of about one in 17 000 bases[4]. 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. Cloning strategies based on the polymerase chain reaction (PCR) 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 one in 1000 which will be evident when they are cloned. If the products of PCR amplification are analysed 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 that will 'correct' these deviations must be anticipated.

Large-scale rearrangements during molecular cloning are also possible, particularly in vectors such as yeast artificial chromosomes (YACs) or cosmids that carry large inserts. Olson's group has estimated that 2% of the YACs in its library are clonally unstable[5]. 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 that prevent the propagation of an insert until they are modified or deleted[6]. 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 for which the incorporation of terminator is uniform and independent of sequence[7], although residual artifacts due to secondary structure of the template may still be observed. Both polymerase/terminator and chemical sequencing techniques depend on an electrophoretic size fractionation of the reaction products to enable the sequence information to be read. 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 regions of the gel where the sequence is ambiguous, rather than in occult errors. The difficulty of counting homopolymer runs correctly (leading to single base insertions or deletions) and errors in maintaining lane-to-lane registration (leading to single base exchanges) are the problems that are the most sensitive to limited gel resolution, and become the dominant types of error on long gel runs. Raw data error rates of 0.5% can be achieved with automated sequencers and error rates of at least this magnitude must be anticipated in cDNA sequencing surveys as they have been proposed[1].



```
                 10        20        30        40        50        60
Elastase IVGGRRARPHAWPFMVSLQLRGGHFCGATLIAHNFVMSAAHCVANVNVRAVRVVLGAHNL
         ::::   . .. :. :::. .: :::::.::. ..:.:::::  .     ..: ::.::.
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 FGDSGSPLVCNGLIHGIASFVRGGCASGLYPDAFAPVAQFVNWIDSII
          ::::.:.:::: ..::.:.   :::    :.... : ..:.::...:
Trypsin  QGDSGGPVVCNGELQGIVSWGY-GCALPDNPGVYTKVCNYVDWIQDTI
          200       210        220       230       240
```

FIG 1

The sequence alignment of two distantly related proteins, human neutrophil elastase and rat trypsin, generated by the FASTA program[17]. Colons indicate sites where the aligned sequences are perfectly conserved, and full points indicate sites of conservative substitutions. The boxed regions contain the residues comprising the catalytic triad of the active site.

Sequence assembly

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 ten overlapping subclones, each with a raw error rate of 0.5%, the directly assembled data will have an average of 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. Information from both strands must be considered, and the relative reliability of different sequence runs must be weighed. Nonlinearities in the relationship between electrophoretic mobility and length may result from incompletely denatured secondary structure in fragments. This will result in regions of the sequencing gel that are ambiguous. Since such artifacts are dependent on local sequence, they tend to occur in similar positions 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, on the basis of 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 (Ref. 8). At the National Library of Medicine we have compared the translated sequences for coding regions in GenBank with the corresponding Protein Information Resource amino acid sequence entry. With some heuristic rules in the matching algorithm it is possible to match perfectly about 75% of the comparable entries, but discrepancies remain in about 25% of cases. Since the average sequence is several hundred bases long, this again suggests an underlying error rate of about a few per 1000 bases. Finally, in preparing the *E. coli* genetic map, Rudd *et al.* have identified 161 segments of DNA that were sequenced independently in different laboratories. Of these, 66 disagree at one or more sites. This corresponds to a raw error rate of roughly one per 1000 bases[9].

Errors and the interpretation of sequence

The impact of errors on the usable information content of a reported sequence depends on the way in which the sequence is analysed. Information theory imposes theoretical limits on the impact of noise, but the real effect of uncertainties in data is determined by the way the sequence is interpreted[10]. Translation of nucleic acid sequences into amino acid sequence is very sensitive to the presence of insertion or deletion errors. At an insertion or deletion error rate of 1% 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 Fig. 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[11]. With David Botstein, I have shown that a Bayesian approach may be applied to combine translation and alignment calculations probabilistically[12]. 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. Figure 2 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 5% base substitution error rates. In this example, the bovine α-lactalbumin mRNA sequence[13] with artificially introduced errors is aligned with the protein sequence of a distantly related homolog, chicken lysozyme[14]. The distribution of true alignment scores is clearly distinct from the random sequence alignments.

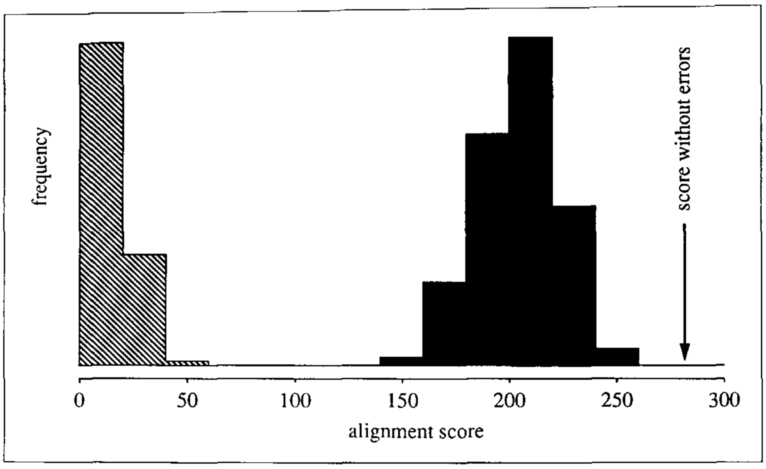


FIG **2**

The effects of sequence errors on alignment scores. A total of 100 trials were performed in which errors were introduced into the bovine lactalbumin mRNA sequence: 5% of the bases were randomly substituted, and insertion and deletion errors were introduced at 1% of the bases. These 'mutated' sequences were then aligned against either the correct chicken lysozyme protein sequence or a random 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 is shown in cross-hatch. All alignments were based on the PAM250 model[11], using a simultaneous translation and alignment algorithm[12]. The score for the lysozyme alignment in the absence of any errors is indicated.

The effect of sequence errors on homolog identification through database searches may also be assessed empirically. Figure 3 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 sequence and a target protein sequence database[15]. With an accurate query sequence, 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 errors in the sequence data. If repeated searches were performed with several sets of error-prone data, this success rate would be even higher.

Many applications require, and have achieved, very high sequence accuracy. In medical genetics, sequence error rates as low or lower than one in 10^4 are needed to identify disease alleles. Similarly, the use of conceptually translated sequences to analyse protein families requires highly accurate nucleic acid sequence data. To reach this level of accuracy, raw sequence data for normal and disease alleles are compared side by side. Wherever discrepancies are identified, comparative studies are pursued and the true mutation site identified through a process of repetitive sequencing, resequencing, and direct analysis of fresh genomic

BLASTX matches with the true bovine α-lactalbumin mRNA sequence:	**BLASTX matches with a 'mutated' bovine α-lactalbumin mRNA sequence:**
>A27360 (PIR) α-lactalbumin precursor – bovine	>S02332 (PIR) α-lactalbumin precursor – bovine

Frame = +1, Score = 792, Expect = 5.1×10^{-117}	Frame = +2, Score = 352, Expect = 2.2×10^{-47}
	Frame = +1, Score = 278, Expect = 9.3×10^{-38}

(17 other α-lactalbumin matches)	(14 other α-lactalbumin matches)
>LZBO (PIR) lysozyme c 2 – bovine EC number 3.2.1.17	>LZRT (PIR) lysozyme – rat EC number 3.2.1.17

Frame = +1, Score = 178, Expect = 5.0×10^{-20}	Frame = +2, Score = 144, Expect = 9.6×10^{-15}
	Frame = +1, Score = 56, Expect = 0.25

(14 other mammalian lysozyme matches)	(2 other mammalian lysozyme matches)
>LZPY (PIR) lysozyme c – pigeon EC number 3.2.1.17	>LZPY (PIR) lysozyme c – pigeon EC number 3.2.1.17
Frame = +1, Score = 159, Expect = 4.7×10^{-17}	Frame = +2, Score = 142, Expect = 2.0×10^{-14}
(26 other α-lactalbumin or lysozyme matches with $p<0.01$)	(35 other α-lactalbumin or lysozyme alignments with $p<0.01$)

FIG **3**

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 of the same size as 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. The output from one typical search with a 'mutated' query is shown.

sequences with oligonucleotide hybridization or PCR sequencing[16].

Low accuracy sequence can be a step towards acquiring highly accurate data, by its use for probe generation, primer-directed strategies, or PCR sequencing. Low accuracy (1% error rate) cDNA sequence databases could also play a useful role in the identification of coding regions by homology search. Extending the length of the sequence and improving its accuracy will be necessary if such cDNA sequence databases are to be used as the basis for the analysis of conceptually translated protein sequence. The usefulness of sequence databases for purposes such as this will be determined by the quality of the data they contain.

References

1 Adams, M.D. *et al.* (1991) *Science* 252, 1651

2 Roberts, J.D., Bebenek, K. and Kunkel, T.A. (1988) *Science* 242, 1171–1173

3 Cooper, D.N. *et al.* (1985) *Hum. Genet.* 69, 201–205

4 Roberts, J.D. *et al.* (1989) *Mol. Cell. Biol.* 9, 469–476

5 Brownstein, B.H. *et al.* (1989) *Science* 244, 1348–1351

6 Brookes, S. *et al.* (1986) *Nucleic Acids Res.* 14, 8231–8245

7 Tabor, S. and Richardson, C.C. (1990) *J. Biol. Chem.* 265, 8322–8328

8 Krawetz, S.A. (1989) *Nucleic Acids Res.* 17, 3951–3957

9 Rudd, K.E., Miller, W., Ostell, J. and Benson, D.A. (1990) *Nucleic Acids Res.* 18, 313–321

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

11 Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1979) in *Atlas of Protein Sequence and Structure* (Vol. 5, Suppl. 3) (Dayhoff, M.O., ed.), p. 345, National Biomedical Research Foundation, Washington DC

12 States, D.J. and Botstein, D. (1991) *Proc. Natl Acad. Sci. USA* 88, 5518–5522

13 Hurley,W.L. and Schuler, L.A. (1987) *Gene* 61, 119–122

14 Jung, A., Sippel, A.E., Grez, M. and Schutz, G. (1980) *Proc. Natl Acad. Sci. USA* 77, 5759–5763

15 Altschul, S.F. *et al.* (1990) *J. Mol. Biol.* 215, 403–410

16 Levedakou, E.N., Landegren, U. and Hood, L.E. (1989) *DNA Biotechniques* 7, 438–442

17 Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl Acad. Sci. USA* 85, 2444–2448

D.J. States is in the National Center for Biotechnology Information, National Library of Medicine, Building 38A, Room 8S806, Bethesda, MD 20894, USA.

The secret life of the hair follicle

MARGARET H. HARDY

The mammalian hair follicle is a treasure waiting to be discovered by more molecular geneticists. How can a tiny cluster of apparently uniform epithelial cells, adjacent to a tiny cluster of uniform mesenchymal cells, give rise to five or six concentric cylinders, each of which is composed of cells of a distinctive type that synthesize their own distinctive set of proteins? There is now evidence that several growth factors, cell adhesion molecules and other molecules play important roles in the regulation of this minute organ.

The first hair follicles are formed from the ectoderm, an epithelial layer that will give rise to the epidermis, and the underlying mesoderm, a mesenchymal layer that will form the dermis. Figure 1 indicates the main stages of follicle development as seen on the back of a mouse, but is representative of hair follicles in most mammals. The numbers assigned to stages will be used in this review. Melanoblasts, of neural crest origin, are usually present among the epithelial cells at the beginning of this period, and will differentiate into melanocytes in the base of the follicle and transfer pigment to the hair. Since they have relatively little influence on follicle development, they are not included in this review.

In most mammals the follicles that produce the pelage hairs – which form the coat of fur, hair or wool – begin to form in the skin during prenatal life at one location, such as the crown of the head, and extend in a wave-like manner over the body surface. In animals with a dense coat this first wave may be followed by further waves of smaller follicles producing smaller hairs. These follicles are arranged in regular patterns, usually in groups with one large primary follicle flanked by two slightly smaller ones, and a group of secondary follicles associated with each trio[1,2]. Whisker (vibrissa) follicles develop earlier, in restricted locations, and have their own regular pattern, for example, in rows on the upper lip.

Since both pelage hair and vibrissa follicles can form *de novo* from organ cultured fragments of embryonic skin, and hairs are produced from them[3], this process does not require ongoing neural or humoral signals. The early pattern of the hair follicles formed *in vitro* corresponds with the pattern observed *in vivo*. Similar development of feathers and/or scales can occur in cultures or grafts of embryonic skin of birds or lizards[4]. Developmental biologists have taken advantage of these features to analyse the roles of epithelium and mesenchyme in the formation of skin appendages[5]. Experiments in which the epidermis and dermis were separated in skin taken from different body regions of embryonic mice, chicks and lizards, and recombined in various ways as explants or chick chorioallantoic membrane grafts, revealed a great deal about the content of the messages that pass from one tissue to the other. However, the physical or chemical signals that convey the messages were unknown.

Figure 2 summarizes the messages that have been discovered for hair follicle differentiation in mammals. The dermal mesenchyme, separated from the back of