Errors, Inconsistencies and Contamination in Public Sequence Databases

Philipp Pagel
MB&B 452a: Genomics and Bioinformatics

December 12, 2001

Sequence databases Over approximately the past decade sequence databases have become one of the most important tools in biological research. The information content of the databases has been growing exponentially increasing their usefulness constantly. Scientist use them to identify transcripts cloned out of libraries and as templates for PCR and sequencing primer design. They are the basis for phylogenetic studies and codon usage analyses. People use ESTs (Expressed Sequence Tags) to assemble putative ORF's. Mutated alleles are compared to their wild-type relatives in order to pinpoint mutations causing hereditary diseases. The correctness of the data stored in the databases appears to be a prerequisite for most of the before-mentioned applications—but how reliable is the stored data, really?

Public sequence collections like GenBank [2] are open to submissions by essentially anyone. Guidelines and quality standards have been put in place to assure the quality and reliability of databases. Nevertheless the actual quality of the data varies considerably depending on the source. Some databases like SwissProt [20] have more rigorous rules and are more stringently curated than GenBank. Information in those repositories is generally considered significantly more reliable. But despite the efforts of curators and contributing scientists errors, misinformation, inaccuracies, inconsistencies and contamination keep making their way into the databases.

Sequencing errors The most obvious kind of problems are sequencing errors. The first step in a sequencing project is usually the isolation of genomic DNA or mRNA from the target organism (maybe followed by fragmentation of the DNA). In order to generate sufficient amounts of template DNA for the sequencing reaction the fragments are cloned into a plasmid and amplified in E. coli. In the case of mRNA reverse transcription is performed to obtain dsDNA before cloning. In the process, the DNA samples are exposed to mechanical shear stress, oxygen, photochemical damage and various chemicals [19] all of which have the potential to damage the DNA molecules. In many cases the inserts are generated by PCR which introduces another source of error: *in vitro* error rates of commonly used DNA polymerases range in the order of one mismatch in 1000 bases [19]. Direct sequencing of the PCR product reduces this problem since the signal on a sequencing gel is determined by the majority of copies but sequencing of cloned PCR products yield the sequence of a single randomly chosen copy.

The sequencing reaction itself is also prone to errors: The mobility of a DNA fragment is determined by its length, but if the molecule maintains secondary structure despite the denaturing conditions of the sequencing gel the mobility changes resulting in compressions, ambiguous bases and flipping of bases. Estimated error rates for sequencing machines vary from 0.2 % to 5 % [12, 4]. Although some of these estimates are based on older models one has to keep in mind that much of the sequences in the database have been obtained using those (now obsolete) instruments. The error rate of sequences is obviously dependent on the coverage: Multiple coverage will greatly reduce the rate of random errors in the sequence. Of cause this is not true for systematic sequencing problems like highly repetitive regions that will yield errors in independent sequencing reactions. While multiple coverage is good practice for individually cloned genes ESTs are generated in a single sequencing reaction in the context of high throughput sequencing efforts. When working with ESTs, one needs to remember that they don't fulfill the same quality standards as conventional gene sequences.

Several groups have attempted to estimate the overall error rate in GenBank and came up with quite divergent numbers in the range of 0.37-35 in 1000 bp [12, 13, 19]. Whatever the actual error rate turns out to be, it is important to take errors into account when working with the data. Clark and Whittam have performed an in-depth analysis of the impact of sequencing errors on molecular evolution analysis and point out that the the estimate for nucleotide diversity in humans is close to the sequencing error rate of the most careful laboratories [8].

Annotation Another source of problems is sequence annotation. In order to keep up with the rapid generation of sequence data for whole genomes researchers had to resort to automated annotation techniques. These programs assign functional annotations to new genes based on similarity to previously annotated genes (e.g. in other organisms). This approach bears a few problems: If the original annotation is incorrect this mistake is propagated to a whole family of misannotated genes [6, 7]. Even when the original annotation is correct the transfer by similarity remains problematic. How can the computer decide if a similar gene codes for an ortholog or a paralog? In addition, database entries are not always updated on a regular basis (or at all!). I.e. mistakes in an annotation may long be known but have simply not been incorporated in the respective entry. Some proteins introduce additional trouble by having multiple unrelated functions depending on the context [5]. On top of that, functional annotation is often highly unsystematic which makes automatic analysis even more difficult.

Name conventions Ideally the name of a gene or gene product would be a unique identifier. Unfortunately, this is frequently not the case. Often a gene was discovered independently by more than one group, each of which assigned a name to the gene. The process of the scientific community to agree on one of them can take years or in some cases is never resolved [6]. Examples include E. coli *hns* which is known under no less than eight additional acronyms!

Even worse, sometimes totally unrelated genes share the same name—like *MRF1* which is either the *mitochondrial peptide release factor 1* or the *mitochondrial respiratory function protein 1* depending on who you talk to...

In some cases relevant database entries are missed because of spelling mistakes, differences in US vs. UK spelling or problems of representation of characters absent from the ASCII code (german umlauts, french accents, chinese names etc.) [6].

Recombination While sequencing errors usually affect only a few base pairs maybe resulting in a different amino acid or introducing a frame shift while preserving the overall structure of the sequence, recombinations that occur in plasmids, BACs or YACs can disrupt the inserts and/or artificially splice together parts of a genome in a new way. The severity of this problem correlates with the size of the piece of sequence carried by the respective vector: plasmids are relatively small and thus not very prone to recombination while YACs and their even larger relatives the megaYACs have been found to be chimeric in up to 80 % [1]!

Contamination Finally a sequence can contain foreign pieces of sequence intentionally or accidentally introduced at various steps of the cloning procedure or by recombination in yeast or bacteria. The literature is full of reports of such contaminants "secretly" making their way into sequence databases.

Wenger and Gassmann found mitochondrial sequence as a contaminant in genes of nuclear origin [22]. Although translocation of mitochondrial genes to the nuclear genome during evolution is a known phenomenon, for most of the cases it is much more likely that we are looking at an artifact. Preparations of total RNA from cells often are the starting point of a cloning/sequencing project. Those RNA preparations routinely contain a certain amount of mitochondrial RNA some of which ends up being cloned along with the mRNA.

Gonzalez and Sylvester describe rRNA and rDNA derived contaminant sequences in databases [11]. They propose several different mechanisms: pseudogenes of rRNA in other genomic locations, mRNA derived pseudogenes residing in rDNA, cDNA derived from rRNA, cDNA derived from transcripts of the rDNA intergenic spacer and genomic DNA contamination of RNA preparations.

Other "popular" contaminants include DNA sequence from host organisms such as E. coli insertion sequences (highly mobile sequences transposing themselves into chromosomes and plasmids) [23, 3], yeast

genomic DNA [1] or bacterial DNA [9].

The most common foreign sequence found in databases is derived from cloning vectors. Usually these sequences are found at the 5' and/or 3' ends of database sequences. They are remnants of the vector(s) the gene of interest was cloned into for maintenance and sequencing and the submitters of the sequence forgot to remove it or were not aware of it being present in their sequence. Surveys of popular databases reveal contamination of most popular cloning/sequencing vectors [23, 14, 9, 15]. Vector contamination has not only been observed at the 5' and 3' ends but also *in* the insert [14]. It is believed that this is a result of recombination events and/or ligation artifacts.

What can be done? For sequencing errors the obvious answer is to assure sufficient coverage to guaranty reasonable quality. In the case of ESTs high quality is not the goal and it remains the responsibility of the user to stay aware of the very nature of those kinds of sequence tags. Although it can not be the responsibility of the database curators to enforce the quality standard in sequencing they could fill the important role of communicating suspected problems to the original submitter.

Annotation problems are hard to solve on the bioinformatics side—the situation can only improve gradually as researchers learn more about functions of genes and their relations. Bioinformatics people on the other hand have the responsibility of making the source of an annotation obvious to the user who will then be aware of potential misannotations.

Redundancy and confusion of gene names is an issue that needs to be addressed by the whole scientific community. Not only do we need less ambiguous names, we should also push towards the improvement of systematic, hierarchical classification of genes and proteins in order to facilitate integration of different databases (sequence, structure, expression, function, ...).

Recombination is hard to address other than the use of the most stable system suitable for the task and statistical analysis of the data obtained.

Vector sequence contamination is probably the easiest to remedy. Using available alignment programs like BLAST, FASTA and others vector sequence and other contaminants can be detected and even automatically removed. Even without alignments the presence of a multiple cloning site is often revealed by the presence of an unusual accumulation of restriction sites. Many databases offer vector screening programs as part of the sequence submission package [21, 10]. Some groups have developed specialized programs to detect and/or remove such contaminants [16, 18]. Other contaminants as described above are harder to deal with since they behave less predictable and come from more different sources.

Conclusion Despite all the problems discussed in this paper sequence databases remain extremely valuable research tools especially when data from different sources is integrated [17]. Many of the issues can and will be addressed by the database curators and the scientific community. In my opinion the most important conclusion from the mentioned problems is to use common sense when analyzing the data and not to assume that the database is always right—a strategy that most scientists have been using most of the time, anyway.

References

[1] C. Anderson. Genome shortcut leads to problems. *Science*, 259(5102):1684–7, Mar 19 1993.

[2] D. A. Benson, M. Boguski, D. J. Lipman, and J. Ostell. Genbank. *Nucleic Acids Res*, 22(17):3441–4, September 1994.

[3] M. Binns. Contamination of dna database sequence entries with escherichia coli insertion sequences. *Nucleic Acids Res*, 21(3):779, Feb 11 1993.

[4] Applied Biosystems. Specifications for abi prism377 sequencer.

[5] P. Bork. Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Res*, 10(4):398–400, April 2000.

[6] P. Bork and A. Bairoch. Go hunting in sequence databases but watch out for the traps. *Trends Genet*, 12(10):425–7, October 1996.

[7] G. Casari, M. A. Andrade, P. Bork, J. Boyle, A. Daruvar, C. Ouzounis, R. Schneider, J. Tamames, A. Valencia, and C. Sander. Challenging times for bioinformatics. *Nature*, 376(6542):647–8, Aug 24 1995.

[8] A. G. Clark and T. S. Whittam. Sequencing errors and molecular evolutionary analysis. *Mol Biol Evol*, 9(4):744–52, July 1992.

[9] M. Dean and R. Allikmets. Contamination of cdna- libraries and expressed-sequence-tags databases. *Am J Hum Genet*, 57(5):1254–5, November 1995.

[10] Emvec. http://www2.ebi.ac.uk/blastall/-vectors.html.

[11] I. L. Gonzalez and J. E. Sylvester. Incognito rrna and rdna in databases and libraries. *Genome Res*, 7(1):65–70, January 1997.

[12] S. A. Krawetz. Sequence errors described in genbank: a means to determine the accuracy of dna sequence interpretation. *Nucleic Acids Res*, 17(10):3951–7, May 25 1989.

[13] T. Kristensen, R. Lopez, and H. Prydz. An estimate of the sequencing error frequency in the dna sequence databases. *DNA Seq*, 2(6):343–6, 1992.

[14] E. D. Lamperti, J. M. Kittelberger, T. F. Smith, and L. Villa-Komaroff. Corruption of genomic databases with anomalous sequence. *Nucleic Acids Res*, 20(11):2741–7, Jun 11 1992.

[15] K. W. Liao, Y. W. Chang, and S. R. Roffler. Presence of cloning vector sequences in the untranslated region of some genes in genbank. *J Biomed Sci*, 7(6):529–30, Nov-Dec 2000.

[16] C. Miller, J. Gurd, and A. Brass. A rapid algorithm for sequence database comparisons: application to the identification of vector contamination in the embl databases. *Bioinformatics*, 15(2):111–21, February 1999.

[17] A. Pandey and F. Lewitter. Nucleotide sequence databases: a gold mine for biologists. *Trends Biochem Sci*, 24(7):276–80, July 1999.

[18] G. A. Seluja, A. Farmer, M. McLeod, C. Harger, and P. A. Schad. Establishing a method of vector contamination identification in database sequences. *Bioinformatics*, 15(2):106–10, February 1999.

[19] D. J. States. Molecular sequence accuracy: analysing imperfect data. *Trends Genet*, 8(2):52–5, February 1992.

[20] SwissProt. Expasy molecular biology server: http://www.expasy.ch/.

[21] Vecscreen. http://www.ncbi.nlm.nih.gov/-vecscreen/vecscreen.html.

[22] R. H. Wenger and M. Gassmann. Mitochondria contaminate databases. *Trends Genet*, 11(5):167–8, May 1995.

[23] T. Yoshikawa, A. R. Sanders, and S. D. Detera-Wadleigh. Contamination of sequence databases with adaptor sequences. *Am J Hum Genet*, 60(2):463–6, February 1997.