EMBO reports

analysis

Righting the wrongs

DNA and protein sequence databases are increasingly useful research tools. But to maximize their potential, the errors in them need to be addressed

The past decade has seen an unprecedented explosion in the number of genomes being sequenced. These bacterial and eukaryotic sequences offer a unique view into the specifics of individual organisms and, by their comparison, an opportunity to analyse life on a larger scale. Many publicly accessible databases have been created to store and distribute this information in support of research. For this sequence information to be useful, it must be both accurate and reliable, and accessible in a clear and consistent manner. But, as with any human endeavour, the creation and maintenance of sequence databases is prone to error. The extent of these errors, and their impact on the use of sequence information, has widespread implications for research in academia and the biotechnology industry.

Almost every aspect of a sequence database is subject to error. The format of the files contained in the database, the complementary information characterizing the sequence, or the sequence itself might contain scientific, syntactic or typographical errors. Among these, the incorrect functional annotation of proteins—additional information describing the sequence in more detail—is responsible for many of the errors in public sequence databases. This information can include the protein's or the gene's identification, the pathway or reaction in which a protein is involved, its active-site residues or cofactor-binding sites and any other information that helps to describe its cellular, biochemical or molecular function. These annotations can then be applied to newly discovered proteins; it is often easier and quicker for scientists to determine a protein's function or structure by comparing it to homologous sequences rather than relying on experimental methods for characterization. But if the original annotation is erroneous, this mistake can trickle through databases and spread to other sequences until a flood of incorrect information has been generated. Indeed, many now recognize error propagation by annotation transfer as a dangerous problem for public sequence databases (Kyrpides & Ouzounis, 1999; Gilks *et al*., 2003).

An inspection of functional annotations for the *Mycoplasma genitalium* genome estimated that the prevalence of errors could be as high as 8% (Brenner, 1999). There is no clear understanding of how extensive database errors really are, despite the attention that this problem has received, because, so far, no large-scale study has assessed the number of errors in public sequence databases or the rate of error propagation. Nor is it clear how to deal with them. As Claire Fraser, President and Director of the Institute for Genomic Research in Rockville (MD, USA), said, "We know that these errors exist—it is hard to say at what level. But we don't have a good solution for how and who will fix them and where the funding for this will come from."

For each complete genome sequence released, there is a subsequent onslaught of publications that detail mistakes in the initial annotation. If anything, this highlights the ongoing nature of protein annotation. Sequences are rarely deposited in a 'mature' state; as with all scientific research, protein annotation is a continual process of learning, revision and correction. Jim Ostell, Chief of Production Resources at the National Center for Biotechnology Information (Bethesda, MD, USA), which oversees GenBank, likened the database to primary scientific research literature. In both cases, each submission represents the view of the submitting author, not the views of the editors. The editors require that the article is internally consistent, follows certain formats, provides new primary experimental results, and is of appropriate scope and quality. But they cannot confirm that the experiments done or the conclusions reached are correct beyond the information contained in the article itself. "If, a year later, the weight of scientific evidence indicates that some of the conclusions of the paper may be incorrect or incomplete, that is not surprising nor would it necessarily be considered an 'error' in the paper. At the time the paper was published, reasonable conclusions were reached. With later information, other conclusions may be reached," Ostell pointed out.

> …if the original annotation is erroneous, this mistake can trickle through databases and spread to other sequences until a flood of incorrect information has been generated

The problem of errors is further exacerbated by the fact that in many nucleotide sequence databases, including GenBank, the EMBL Data Library in the UK and the DNA Databank of Japan, sequence data and annotations can only be modified by the scientist who originally submitted the sequence. By contrast, the SwissProt protein sequence database is continuously revised by database curators to reflect updated sequence annotations. Amos Bairoch, head of the SwissProt group at the Swiss Institute of Bioinformatics, acknowledged that "there is almost no entry where there's not something corrected."

Another aspect of public sequence databases has the potential to seriously affect those who use them. As Peter Karp, Director of the Bioinformatics Research Group at SRI International (Menlo Park, CA, USA) explained, "A big problem that I see […] is

> "People are smart enough to be able to interpret the wrong information in the wrong fields, and rearrange things. Computers aren't smart enough to do that."

the sequence databases accepting submissions in formats that violate their own standards. I think that's just as big a problem as the errors in the databases." Although they are not technically errors, these inconsistencies can make it difficult to use sequence data effectively. Each sequence database has specific templates for submitting information, from the type of information required to the text format that should be used. Difficulties arise when depositors provide incorrect information, put information in the wrong place, or simply include too much information. Most database users would therefore like to see a stricter enforcement of the standards that are in place at present. GenBank, from which SwissProt obtains most of its sequence data, often comes under fire for including entries that do not conform to its own rules. In most cases, this is due to genome sequencing centres submitting large quantities of data in variable formats. "The way they provide the information is completely heterogeneous. It's not that it's erroneous at all, it's more that each of them have a style, and it means that we have to adapt to a lot of different styles," Bairoch explained. Ostell agreed, noting, "I think the observation that heterogeneity of annotation and historical artefacts make the database challenging to use properly is absolutely true, but I don't see any magic bullets." He laments that making changes to the database format to satisfy some database users will inevitably dissatisfy others. However, Ostell noted that the format of database entries is continually improved to structure the information more clearly.

Other types of database error are less common. The most fundamental are mistakes in DNA or protein sequences themselves. But thanks to improvements in experimental techniques and technology, sequencing errors are mostly a thing of the past. Current technology aims to reduce error rates to as low as 1 base in 10,000. Bairoch noted, "the issue is not really sequence quality but the quality of everything around the sequence." In addition, the supplementary text information in databases can include typographical mistakes, such as misspelled species names or nonexistent enzyme classification numbers. The potential for these errors is easily reduced by using systems that allow sequences to be annotated with consistent terminology. One such system is Gene Ontology, developed by an international consortium of scientists to describe gene-product attributes such as molecular function, the biological process to which it belongs, and cellular components with which it is involved (Gene Ontology Consortium, 2000). The attention focused on addressing these errors suggests that they too may soon disappear, or at least become uncommon.

Database errors and inconsistencies have several implications. Obviously, the most serious is their impact on subsequent analyses. In some cases, a single error in a database has led to incorrect conclusions. Sekyere *et al*. (2003) identified a new melanotransferrin gene in GenBank, which was later found to be missing from the validated version of the human genome sequence. Although they recognize the importance of having early access to genomic information, their experience is an unfortunate reminder that preliminary sequence information is often unavoidably erroneous. Willerslev *et al*. (2002) were similarly unlucky when they found a sequence in human genome data that they were sure had been laterally transferred from prokaryotes to humans. Only after extensive analysis was it confirmed that the sequence was in fact a contaminant.

But the research that will be affected most by database errors is probably large-scale studies that use extensive portions of the sequence database. Peter Forster, a geneticist at the University of Cambridge, UK, found that more than half of the mitochondrial DNA sequencing studies published contain obvious errors (Röhl *et al.*, 2001; Forster, 2003). These errors led researchers from deCODE Genetics (Reykjavik, Iceland) to incorrectly describe the genetic diversity of Icelanders. An extensive re-analysis identified anomalies in the data as the source of deCODE's mistake (Árnason, 2003). As Karp pointed out, "People are smart enough to be able to interpret the wrong information in the wrong fields, and rearrange things. Computers aren't smart enough to do that. So it's really the people doing high-throughput type research that are in trouble." Fraser agreed: "Sometimes this is a minor annoyance that we can devise a fix around—other times it is more serious and wastes a great deal of time trying to extract what we need."

Ironically, these errors also have implications for improving computational methods for analysing protein and DNA sequences, because new algorithms are tested on current sequence information. Without knowing how often errors and inconsistencies occur in the databases, it becomes very difficult to improve these methods. "You can't develop [the next generation of functional annotation systems] unless you know where the errors are being made by current systems," Karp said.

Efforts to address the problem of erroneous and inconsistent data and to find ways to fix them are hampered by disagreement over who is ultimately responsible: the database curators, or database users. As a database user, Karp believes "It's both groups' responsibility, but ultimately the databases are the gatekeepers." Fraser disagrees, saying "the sequence depositors should be responsible", whereas SwissProt database developer Bairoch thinks "everyone has to feel responsible". Regardless of responsibility, database users and curators both seem to agree that most scientists underappreciate the problem of database errors. But Fraser confirmed that the existence of annotation errors is "considered serious by most genome centers and by many bioinformaticists."

Without further research, it is difficult to quantify the effect that errors have on database usefulness. However, as yet, no concerted effort has been made to specifically analyse public sequence databases. "We're investing huge amounts of money in the sequence databases, the entire scientific community relies on them, and yet we don't know some very basic things about their properties," Karp said. An assessment of database accuracy and reliability would also go some way towards educating the community about errors and would encourage debate about the problem. But this is likely to require additional funding, and the source of this funding is not clear. As Karp pointed out, "In some sense this is perhaps even a lower priority because it's not actually spending money on curating databases, it's spending money to check up on the people who curate the databases." Perhaps the first step in addressing the problem should be to educate the scientific community and encourage a greater collaboration in maintaining error-free resources. In their survey of quality-control procedures in archival databases, the CODATA Task Group on Biological Macromolecules concluded that the only possible solution is a dynamic annotation process, with the workload distributed among database curators and specialists (CODATA Task Group on Biological Macromolecules and Colleagues, 2000). Ultimately, database developers could find this specialist knowledge by appealing to the altruism of users. "Making people aware of errors is good and great; making people aware that they're responsible also for correcting errors is even greater," Bairoch said.

> "Making people aware of errors is good and great; making people aware that they're responsible also for correcting errors is even greater."

If one thing is certain, it is that the number of sequences in public sequence databases will continue to increase exponentially for the foreseeable future—as will the errors, most likely. As these databases constitute the foundations for advanced research in biology, their ability to maintain this role effectively has implications not just for bioinformatics and genomics, but for all fields of scientific research. If the strength of these foundations is not tested now, extracting useful information from databases may become even more difficult. "I think it's going to get much worse before it gets better. There's going to be an explosion in terms of heterogeneity of resources and of people not finding what they want. Then people will complain and things will slowly get better," Bairoch predicted.

REFERENCES

Árnason, E. (2003) Genetic heterogeneity of Icelanders. *Ann. Hum. Genet.*, **67**, 5–16.

Brenner, S.E. (1999) Errors in genome annotation. *Trends Genet.*, **15**, 132–133.

CODATA Task Group on Biological Macromolecules and Colleagues (2000) Quality control in databanks for molecular biology. *Bioessays*, **22**, 1024–1034.

Forster, P. (2003) To err is human. *Ann. Hum. Genet.*, **67**, 2–4.

Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.

Gilks, W.R., Audit, B., De Angelis, D., Tsoka, S. & Ouzounis, C.A. (2003) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641–1649.

Kyrpides, N.C. & Ouzounis, C.A. (1999) Whole-genome sequence annotation: going wrong with confidence. *Mol. Microbiol.*, **32**, 886–887.

Röhl, A., Brinkmann, B., Forster, L. & Forster, P. (2001) An annotated mtDNA database. *Int. J. Legal Med.*, **115**, 29–39.

Sekyere, E., Dunn, L.L. & Richardson, D.R. (2003) The double-edged nature of using genetic databases: melanotransferrin genes and transcripts. *FEBS Lett.*, **547**, 233.

Willerslev, E., Mourier, T., Hansen, A.J., Christensen, B., Barnes, I. & Salzberg, S.L. (2002) Contamination in the draft of the human genome masquerades as lateral gene transfer. *DNA Seq.*, **13**, 75–76.

Caroline Hadley

doi:10.1038/sj.embor.embor932

Less is more

Research into anti-angiogenic therapies for treating cancer has finally had its first breakthroughs. But it may also influence the way in which classical chemotherapy is used for cancer treatment

After many promises and failures, anti-angiogenic drugs finally seem to be making progress towards their clinical use. In May 2003, researchers at the meeting of the American Society of Clinical Oncology announced the first positive results with an anti-angiogenic therapy in a phase III cancer trial. The randomized, double-blind study of more than 900 patients with metastatic colon cancer showed that Avastin™, an antibody against vascular endothelial growth factor (VEGF), combined with chemotherapy, extended overall survival beyond that achieved with chemotherapy alone and had a significantly improved response rate and duration of response. The US Food and Drug Administration reacted to the trial's positive result by granting Avastin™ 'fast track' drug-review status in June this year.

"This is an enormously important study and represents a very exciting step forward," said Leonard Saltz, a physician at New York City's Memorial Sloan–Kettering Cancer