EDITORIAL

Context is key for sequence data

Better annotation of marker gene sequences and environmental parameters in sequence databases is essential to maximize their use for researchers both today and in the future.

Scientists are keenly aware that understanding the context under which an experiment was carried out is key to interpreting the results in a given study. Consequently, data presented in a research paper are normally accompanied by an exhaustive description of the techniques employed, the organisms studied and the experimental parameters controlled for or varied. This enables others to compare the results with their own enquiries and to repeat an experiment to analyse any discrepancies. However, this level of description of experimental context is not applied equally in all fields. For example, most genomic, metagenomic and marker gene sequences stored in sequence data repositories are sparsely annotated with the information that would be needed to allow the integration of different data sets and comparative studies.

Although the explosion in sequencing capacity has meant that sequence databases have expanded rapidly over the past decade, they still rely on researcher annotation to enrich the value of the data sets. To ensure a richer description of experimental context in our sequence collections, the Genomic Standards Consortium (GSC) was established in 2005. The GSC proposed checklists for standardizing descriptions of the data — called the minimum information about a genome sequence (MIGS) and minimum information about a metagenome sequence (MIMS) checklists — that are to be filled in at the time of data submission. These checklists include information on the source of the sequence, the technical approach used to generate the data and the environment from which the sample was taken. The GSC now proposes several extensions to these standards, including checklists for phylogenetic and functional marker gene sequences and better environmental descriptions, which will help to ensure that even greater contextual information is added to all sequence data in our repositories[1].

To better catalogue the vast array of sequence data for phylogenetic and functional marker genes, the minimum information about a marker gene sequence (MIMARKS) checklist was developed by combining a survey of the parameters defined in over 80 publications with online surveys of researcher preferences for the core information that is needed to fully describe a marker gene sequence. The resulting checklist encompasses the information that is currently used in MIGS and MIMS checklists but is also extended to contain information such as PCR primer sequences, reaction conditions and target-gene names. In addition, 14 packages have been developed to provide an array of environmental and epidemiological data fields that will allow for a complete description of the environmental parameters for a data set. These packages were developed in collaboration with international research consortia (such as the Human Microbiome Project and the Terragenome consortium) to tailor individual packages to data sets from distinct environments. The GSC has also built an overarching framework that provides a single entry point into the MIGS, MIMS and MIMARKS checklists and the environmental packages, known as the minimum information about any (x) sequence (MIxS).

Why is it so important that completing these checklists when entering sequence data into a repository is adopted as standard practice? In a recent open letter to the microbial ecology community, the authors argued that "sequences without contextual data are like unlabelled cans in a supermarket — you don't know what you are purchasing until you open it and examine the contents" (REF. 2). Furthermore, the lack of contextual data risks wasting a hugely valuable resource that has in many cases been paid for with public funds. We support the call for all microbiologists who deposit sequence data to adopt the use of these checklists as standard practice. For their part, repositories such as the International Nucleotide Sequence Database Collaboration (INSDC) (which includes Genbank, the European Nucleotide Archive (ENA) and the DNA Data Bank of Japan (DDJB)) have already begun to adopt such standards, allowing for data submission in MIxS-compliant formats.

With minimum-information checklists being accessible to researchers and compatible with sequence database submission policies, there is little to stop their community-wide adoption, other than the additional time, effort and money needed to ensure that data sets are MIxS compliant. Although these are important factors to consider at a time when research budgets and researchers' time are under ever increasing pressures, we argue that not ensuring that our databases contain the correct contextual data would be a false economy that would cost the field dearly in the long term.

> "sequences without contextual data are like unlabelled cans in a supermarket"

1. Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotech.* **29**, 415–420 (2011).
2. Yilmaz, P. *et al.* The genomic standards consortium: bringing standards to life for microbial ecology. *ISME J.* 7 Apr 2011 (doi:10.1038/ismej.2011.39)