Cataloguing our current genome collection

Our complete genome collection, the result of a significant investment of public funds, is one of our most valuable biological resources and also one of the most complex. The value of these data lies in the ability to view, compare and contrast the entire genetic complement of a wide range of organisms. The complexity of these data lies not only in the vast number of biological features within these genomes (i.e. genes, promoters, binding sites, etc.), but also in the complexity of the evolutionary relationships and ecological lifestyles – factors that play a major role in shaping genome features and content – of the organisms to which these genomes belong.

It is our opinion that the scientific community would benefit significantly from the establishment of a data standard to capture more of this complexity. This standard could be followed to provide electronic, machine-readable reports for submission to a public database at the time of publication of each genome (analogous to the submission of genome annotation files). In this way, the experts on each organism would be directly responsible for providing data to the wider community about the detailed features of the organism. This approach would not place undue burden on individual genome project consortia, but would tremendously increase the value of our collection of genomes.

While descriptive information about genomes is becoming increasingly available, it is spread out among many different locations and has not been unified to produce one set of data in a common format that can be easily maintained, searched, and disseminated. The completed genomes of all taxonomic groups (eukaryotes, bacteria, archaea, plasmids, organelles and viruses) should be included to maximize the ability to compare sets of co-evolving and interrelated sets of genomes.

This issue has been discussed in part previously. For example, Ward et al. (2001) suggested a standard set of practices be adopted for the preservation of all strains used for genome sequencing. Likewise, there are now projects that are collecting such metadata in a post hoc fashion through automated analyses and consultation with the literature and appropriate experts. TIGR’s ‘Genome Properties’ project within the Comprehensive Microbial Resource is perhaps the most comprehensive of these efforts but is restricted to bacteria (Haft et al., 2004).

It may appear upon casual consideration that all desirable data are easily available, but deeper examination shows that there is much to be done to capture even some of the most obvious fields of information. For example, about 10% of the first 150 published bacterial genomes do not have rRNA gene sequences annotated in their GenBank files (Ussery & Hallin, 2004). In an examination of 323 bacterial genome...
sequences, 31 contained no information about the identity (name) of the strain sequenced (Coenye & Vandamme, 2004).

To remedy this situation and create a new tool for discovery, the community should agree upon a standard and place the responsibility and credit for its collection with those most intimately responsible for the data. Aside from ensuring the long-term preservation of data integral to understanding these genomes, this coordinated activity would maximize the quantity and quality of the data.

MIGS – Minimal Information about a Genome Sequence

Here we describe seven general sections of information that we feel are important to capture within this proposed standard. This information would describe the salient features of each genome, place it in its proper organismal context through details of taxonomy and ecological niche, and maximize the potential to compare and contrast information across genomes. Ideally, details about the method of annotation, the origin of the isolate from which the genome was extracted, links to key electronic resources and the appropriate experts willing to be contacted for further information should also be captured. We suggest this standard be called MIGS, which stands for the ‘Minimal Information about a Genome Sequence’. Just as in the Minimal Information about a Microarray Experiment (MIAME) standard for transcriptomics (Brazma et al., 2001), each MIGS report would be built around a core set of community-determined fields that could be extended as necessary.

1. Genomic features. All genomes share core features that are possible to compare. Such features are perhaps the most easily captured pieces of information because they are a routine result of genome annotation. Many are captured in current annotation files. These features include genome size, G+C content, number of genes, percentage coding DNA, mean gene size, nucleic acid and whether a chromosome is circular or linear. A wide range of additional features common to the taxonomic subgroup could also be captured. For example, for bacteria it would be important to capture a variety of features not explicitly provided in current genome annotation files, like the number of chromosomes in the genome or the presence of plasmids, megaplasmids or phage. In essence this section would represent a concise distillation of the main genomic features reported in the primary publication (but standardized by adherence to the MIGS standard).

2. Taxonomy. Understanding the evolutionary relationships between genomes is essential to the accurate interpretation of comparative genomic data. This section should therefore capture taxonomic designations from the top levels (i.e. domain, division/phylum, etc.) to the levels which distinguish strains or isolates (i.e. subspecies, serovar, biotype). These entries would be based on the recommendations of the appropriate international taxonomic authorities (Garrity et al., 2001; van Regenmortel et al., 2000) and the widely used NCBI taxonomy (Benson et al., 2000).

3. Ecology. It is essential to understand the lifestyle of an organism when interpreting genomic information. Increasingly, links between ecology and genome features are being established (Huang et al., 2004; Tekaia et al., 2002). For bacteria, such data could include information on habitat type, type of interactions (pathogen, mutualist, symbiont), trophic level, general aspects of metabolism (e.g. preferred carbon source) and other features such as growth rate, oxygen, cell size, cell shape. This is the section of the report that would benefit most from the development of a controlled vocabulary.

4. The annotation process. The annotation process should be recorded when annotations are to be compared. This section is intended to allow the methods and comprehensiveness of annotation to be assessed and compared at a glance, and would list software and key parameters used.

5. Origin and availability of strain. Details about the salient features of an index strain and why it was selected is a useful but often overlooked type of information that can have important implications on the interpretation of results. For example, how long an isolate has been passaged in culture is relevant to understanding the extent to which the isolate is potentially ‘lab-adapted’ versus ‘wild-type’. This is of special importance since it is clear that sequencing projects are targeting a large number of non-index strain isolates (Coenye & Vandamme, 2004; Ward et al., 2001). Information about the geographical origin of the strain and any other details relevant to its provenance and availability (Ward et al., 2001) should also be captured.

6. Related publications and electronic resources. While the primary publication of each genome is of obvious importance, there is also value in recording year of date of publication and chronological order of genomes. Genome annotations are state-dependent analyses resulting from the application of particular methods to a given data set at one point in time. This fact should be considered when interpreting these data as the size of available databases and our understanding of the best analysis methods may have grown significantly since publication. The newest analyses are often to be found in associated electronic databases. This section should provide links to such resources, including accession numbers within NCBI/EMBL.

7. Contact information. Usually the only contact person for a genome is the corresponding author on the primary genome publication. This section would contain contact people with various domains of expertise and could include the bioinformaticians responsible for the annotation as well as experts on various aspects of the organism’s biology.

Formulating a MIGS Specification through the establishment of a working group

We suggest that the details of the proposed MIGS specification should be developed within the auspices of an international working group. Members of this working group should include scientists involved in a range of research areas, such as researchers responsible for initiating genome sequencing projects for various taxa, representatives of the primary sequence databases (GenBank/EMBL/DDJB) and genome sequencing centres, taxonomy experts, members of
taxonomic authorities, ecologists, evolutionists, computational biologists wishing to use the data in comparative genomic studies, data standards and ontologies experts and participants from allied projects, for example the Sequence Ontology Project (SO) (http://song.sourceforge.net/) and the Gene Ontology Project (GO) (http://www.geneontology.org/). With this combined expertise, the working group should be able to devise a specification that is of a high quality and simultaneously realistic (despite the obvious challenges involved) because it would be based on an empirical understanding of the steps involved in generating and analysing genomic data. This would in turn improve the chances of its wider acceptance by the community as a valuable international standard.

The proposed working group would also act to provide an infrastructure for the development and maintenance of this standard modelled after the successful outreach activities of initiatives like Microarray Gene Expression Data (MGED) Society (http://www.mged.org/) (reviewed in Quackenbush, 2004) or the Open Biology Ontologies (OBO) initiative (http://obo.sourceforge.net/). This infrastructure should provide a mechanism for the working group to consider suggestions made by the community for changes and extensions in the specification and tools that simplify the completion of new MIGS reports. This infrastructure should include at a minimum a website containing information about this initiative, a master copy of the specification and documentation on its use. This web portal could also supply easy to use web forms for generating new reports, and perhaps access to other software. For example, implementing the specification as an XML schema would allow the use of Pedro, a rapid prototyping tool that is already in use in proteomics and which can easily be adapted for use in other domains (Garwood et al., 2004).

Once the specification is implemented, it would be hoped that the primary databases (GenBank, EMBL and DDJB) would be willing to host the MIGS catalogue for the international community. For example, NCBI has already begun an initiative to collect supplemental data to describe genome projects, including all strain information (http://www.ncbi.nlm.nih.gov/ genomes/static/gprj_help.html). The implementation of the catalogue should maximize the ease with which data can be exchanged and accessed in analysis pipelines. For example, each report should be assigned a Life Science Identifier (LSID) (http://www.i3c.org/wgt/ta/resources/lsid/docs/index.asp) and the web interface to the catalogue should be developed so that users can access it with web browsers and workflow tools such as Taverna (Oinn et al., 2004). There should be easy access for viewing and downloading MIGS reports and facilities to allow the download of subsets or all of the data. Over time, members of the wider community would hopefully help contribute tools to manipulate, search and analyse this large dataset. Ideally, a metadata exchange standard would also be recommended by the working group and its adoption would provide a further mechanism for metadata from a variety of computed and curated sources to be easily exchanged, disseminated and analysed.

The long-term success of such an initiative will depend on the level of awareness of its existence. Likewise, it will depend on the willingness of researchers working on individual genome projects to complete a MIGS report. If the specification is well constructed, well documented and made easily accessible through easy to use web forms or custom software, this should require minimal extra effort within the scope of any given project. Ideally, to maximize the usefulness and completeness of this catalogue, the primary databases would make the MIGS report a requirement for deposition of all new genomes, and journals would require a completed MIGS report prior to publication (just as many high profile journals now require MIAME compliance).

We accept that there would be challenges in developing and maintaining such a standard given that aspects of genomic annotation and analysis are still debated and for a wide range of methodologies are used. For example, how exactly should a ‘gene’ be defined? Likewise, the issue of handling updates must be solved. How should re-annotation, the development of new methods of annotation, or changes in taxonomic designation of relevant species be handled? In short, we suggest that these and related issues could be made manageable by well thought out definitions of each field of data. Likewise, we suggest that there should be the flexibility for multiple MIGS reports to be successively generated for a single genome by different groups. In addition, those responsible for generating the original genomic MIGS should take responsibility for making sure the primary entry is up to date (since the amount of information, unlike an annotation, would be relatively small). Wherever possible, this initiative should draw upon the experiences of other data standards initiatives and related projects. This is especially true when developing the necessary controlled vocabularies to go with the specification. A primary role of the working group would be to co-ordinate and oversee such interactions.

**Long-term benefits for the genomic and computational biology communities**

The implementation of such a standard would produce a new dataset of immense value and protect valuable information about each genome in a uniform way for posterity. The primary use of this completed catalogue would be as a powerful research tool. In addition, such reports could be linked to the outputs of downstream ‘omic’ experiments (transcriptomic, proteomic, metabolomic, etc.) derived from a given genome. While there is a significant number of existing genomes to catalogue retrospectively – more than 3000 – this number is dwarfed by the number of genomes yet to be completed. Furthermore, an increasing number of these future genomes will come directly from the environment, further underscoring the need to place genomes (and data derived from metagenomic projects) into wider context through direct association with a rich set of metadata. For example the Moore Foundation’s Marine Microbiology Microbial Genome Project (http://www.moore.foundation/microgenome/microb_list.asp)
and The Venter Institute’s Sorcerer II Expedition (http://www.sorcerer2expedition.org/version1/HTML/main.htm) will provide vast quantities of genomic information about the microbes of marine environments. The task of developing such a standard and implementing may seem daunting today, but we argue that the task will only become more difficult as time passes. The value of such a catalogue, especially after another decade of genome sequencing produces a collection of thousands of taxonomically and ecologically diverse genomes, appears unquestionable.

Dawn Field1 and Jennifer Hughes2

1Molecular Evolution and Bioinformatics Section, Oxford Centre for Ecology and Hydrology, Mansfield Road, Oxford UK OX1 3SR
2Dept. of Ecology and Evolutionary Biology, Brown University, USA

Correspondence: Dawn Field (dfield@ceh.ac.uk)


DOI 10.1099/mic.0.27914-0