Introduction

The International Nucleotide Sequence Database, a collaboration between the EMBL (Kulikova et al., 2004), Genbank (Benson et al., 2003) and DDBJ (Miyazaki et al., 2004) nucleotide sequence databases, is a repository for DNA and RNA sequence and annotation. Since the advent of whole genome sequencing in the mid 1990s, the sequences of over 170 organisms have been completely determined and deposited in the public repositories. The rate of deposition of such sequences is still increasing, with over 60 genomes sequenced between March 2003 and March 2004. Although the determination of mammalian genomes has attracted attention, the majority of sequenced organisms continue to be bacteria (140 sequenced organisms in March 2004) and archaea (18 sequenced organisms in March 2004).

Submitters retain ownership of data in the nucleotide sequence repositories and in consequence annotation is often not standardized or updated. Various problems can result from this: annotation may be poor; information may be represented irregularly; and as much annotation for predicted genes is inferred by similarity from other sequences, it becomes out-of-date as new sequences are annotated. Additionally, the integration of theoretical annotation inferred from sequence may not be integrated with data from laboratory experiments. Many of these issues are addressed in curated databases such as UniProt/Swiss-Prot; but annotation improvements/updates implemented in these resources cannot be incorporated in the archive record.

As an illustration of this problem, there are currently 4356 entries in UniProt that together define a non-redundant proteome set for the Gram-negative bacterium *Escherichia coli* K-12, representing the products of 4396 genes. Of these, some 1045 proteins (24 %) have been assigned a protein sequence by UniProt curators other than that originally predicted in the EMBL entry derived from the original genome submission (Blatter et al., 1997), which was last updated in 1998.
The NCBI RefSeq project (Pruitt et al., 2003) aims at providing compatible records for DNA, RNA and protein sequence for all complete genomes. RefSeq entries for bacteria frequently contain more regular annotation than the primary submissions with, for example, a standardized representation of gene names. However, little additional information is imported into RefSeq entries. Additionally, there is limited compatibility with external resources (for example, CDS features in RefSeq are not supplied with the common identifiers used by EMBL, Genbank and DDBJ, but instead use their own identifier space).

Therefore, we have launched Genome Reviews, a new project to make annotated genome sequence available in an EMBL-compatible format but with standardized, up-to-date annotation, derived from curated data sources such as UniProt, InterPro and GOA.

**Methods and Algorithms**

EMBL entries describe (nucleotide) sequences, features (annotated regions of sequence) and feature qualifiers (individual annotations attached to a feature). Additionally, there is also some annotation that is attached to the database entry itself as opposed to the sequence (for example, the database accession number). The ‘CDS’ (CoDing Sequence) feature is used to identify sub-sequences within the overall sequence that correspond with the sequence of nucleic acids in a protein (proteins are the most widely annotated biological entity); the ‘/db_xref’ feature contains cross-references to entries in other databases. The ‘/protein_id’ qualifier uniquely identifies each CDS.

There are therefore 3 ways in which an entry in another database can be identified as referring to the same biological entity as a given EMBL (CDS) feature: if the EMBL feature cross-references that entry, if that entry cross-references the EMBL feature, or if an entry in a third database cross references both other entries. Through tracking identifiers between databases, additional annotation belonging to a feature can be identified. UniProt (Apweiler et al., 2004), a well-annotated protein knowledgebase in which redundant submissions are merged, is a particularly useful hub database for retrieving annotation and cross-links to further resources. For each type of annotation, a particular preferred source is nominated; and annotation of that type from that is imported into corresponding features in Genome Reviews either as a supplement to or a replacement for the original annotation. Overlapping sets of annotations (e.g., gene names derived from different sources) are case-standardized and merged.

The annotation attached to other types of features (for example, non-coding RNAs) has also been standardized, and redundant or rarely used features and feature qualifiers removed. In addition, new features have been added (for examples, mature peptides produced after cleavage) by propagating features on protein sequences described in UniProt back onto the corresponding DNA. Sequences are first compared to ensure that the co-ordinate systems used for features are compatible.

**Implementation and Results**

Software for producing Genome Reviews has been implemented using the Java programming language (Java 2 Standard Edition v1.4.2), in particular utilizing tools provided by the BioJava project ({{http://www.biojava.org}}). A persistence layer has been implemented in a relational database management system (Oracle 8i Enterprise Edition release 8.1.7). This database, used to build each release, contains manually curated information used in making each entry (for example, corrected literature citations and taxonomy) and a representation of the relationships between genes, transcripts and proteins for all complete genomes.

Release 0.4 of Genome Reviews, made publicly available on 29th March 2004, contained files describing 256 chromosomes and plasmids from 153 bacterial and archaearal species. Some details of the change in the quantity of annotation are shown in the Table.
The Table shows how in Genome Reviews we have standardized annotation, reducing the number of features and feature qualifiers actually in use, but increasing the total quantity of annotation. The creation of additional features, and the addition of extra qualifiers to existing features, is illustrated for the ‘mat_peptide’ feature type and the ‘/db_xref’ feature qualifier respectively. The increased use of the ‘/locus_tag’ feature qualifier indicates the proper use of this qualifier to indicate the systematic gene names (such information is found variously, if at all, in the primary EMBL submissions).

The introduction of evidence tags into Genome Reviews entries represents the most significant change to EMBL flat file format in Genome Reviews. Evidence tags provide a simple record of the source of each piece of information in a Genome Reviews entry. Tags are applied to the feature qualifiers (indicating why the information contained in a qualifier has been applied to the corresponding feature). If a tag is applied to the ‘/evidence’ qualifier, this indicates why the feature has been attached to the sequence.

**Discussion**

In databases such as Ensembl (Birney et al., 2004), consisting wholly of predictions on a genomic sequence, a set of complete theoretical genes, transcripts and proteins are provided. In the case of many bacterial genomes, however, it is known that many of these theoretical gene predictions are wrong (Skovgaard et al., 2001), and annotation is not regularly updated. In Genome Reviews, we are attempting to standardize and update annotation in line with experimental results.

In addition to tracking identifiers as described, it is also possible to co-identify equivalent biological entities described in different databases through their common co-ordinates on a shared reference sequence; to identify features unannotated in the genome; or identify disagreements in sequence between entries known (through use of common identifiers) to represent the same entity. In future releases of Genome Reviews, we will add new feature qualifiers to describe disagreement at the sequence level, to support the consistent representation of genomic DNA and experimentally verified protein sequence.

**Acknowledgements**

This work has been funded by the award of grant number QLRI-CT-2001000015 from the European Union under the RTD program “Quality of Life and Management of Living Resources”.

**References**


