Poster B-42

Evidence handling in EMBL Third Party Annotation: a two-tiered structure.

Authors:
Cochrane G.R. (*EMBL-EBI*)
Bates K. (*EMBL-EBI*)
Faruque N. (*EMBL-EBI*)
Kanz C. (*EMBL-EBI*)
Kulikova T. (*EMBL-EBI*)
Zhu W. (*EMBL-EBI*)
Apweiler R (*EMBL-EBI*)

Short Abstract: The EMBL Third Party Annotation database (EMBL TPA) recruits and presents peer-reviewed annotation and re-annotation of primary nucleotide sequence available in the public nucleotide sequence databases and trace archives. A recent broadening of scope of EMBL TPA increases coverage to include submitted annotation derived from non-experimental and indirect experimental evidence.

Long Abstract:

The EMBL Nucleotide Sequence Database established EMBL TPA in 2002 [1]. In collaboration with partners in the International Nucleotide Sequence Database Collaboration (INSDC), EMBL has since worked towards the creation of a comprehensive database of submitted third party annotation and re-annotation of existing primary nucleotide sequence. Since its launch, the project has attracted annotation from a large number of researchers. There are currently 5,065 publicly available TPA records from 401 studies and submissions continue to be received on a regular basis. The steady growth seen so far is expected to increase further as recently developed large-scale submission tools for EMBL TPA become established and as awareness grows of the need to represent annotation information in a centralized searchable resource.

The principal role of the EMBL Nucleotide Sequence Database remains the archiving of newly generated sequence and its corresponding annotation, where the original data submitters retain responsibility for the accuracy of the sequence, and for the biological content of the annotation. Increasingly, though, high quality biological studies are revealing further insights into the biological role of nucleotide sequences, without themselves generating novel sequence data. While primary data cannot be updated with the new information without the compliance of the original submitters, EMBL TPA allows presentation of the new information in the form of novel annotation, alongside the original primary records, as distinct database entries.

EMBL TPA entries are clearly marked (in Keyword, KW, and Description, DE, lines), such that users are aware of the source of the annotation that they are viewing. Precise references to the contributing primary entry/entries are given in the form of a table (TPA Assembly, AS, lines) where defined sequence spans from primary entries (cited by primary accession and sequence version number) are linked to defined spans from the sequence of the EMBL TPA entry. While primary entries remain largely untouched, they are given cross-references (Database Reference, DR, lines) to any EMBL TPA entries in which they

are cited.
From the outset, emphasis has been placed on the presentation of quality annotation. In contrast to primary data, EMBL TPA are not made available to the public until such time as the annotation included has been specifically discussed in a peer-reviewed publication.
Until recently, presentation of data in EMBL TPA was restricted to those researchers who had generated direct experimental evidence for the annotation in their entries. While this ensured a high degree of reliability of the annotation, it did not provide a repository for those submitters who were generating high quality annotation where evidence was non-experimental or inferred from indirect experimentation. In a recent broadening of the scope of the project, third party annotation data from researchers with inferred and non-experimental evidence is now welcomed from submitters [2]. Accordingly, the EMBL TPA dataset has been divided into two tiers, experimental and inferential, to represent these differences to users; all EMBL TPA entries belong to one of the two tiers, indicated by the inclusion of keyword ‘TPA:experimental’ or ‘TPA:inferential’. EMBL TPA continues to disallow annotation derived solely from high-throughput analysis pipelines where no manual validation is provided.
EMBL TPA entries can be submitted through a number of routes. Most submissions are directed through the online tool, Webin (www.ebi.ac.uk/embl/Submission/webin.html). Based on information gathered early in a Webin session, users are directed through one of a number of routes designed to minimise submission time and provide optimal annotation; for submission of a single EMBL TPA entry, for example, users can add annotation feature by feature; for submitters of many or all members of a gene family to EMBL TPA, submitters can choose to describe only those fields that vary between entries (such as sequence, primary spans, gene symbol, exon and CDS locations); for submitters of complete genome re-annotations, users can upload pre-prepared feature tables from a range of annotation tools that they may have used. For each EMBL TPA submission, in house curation ensures that newly published EMBL TPA records comply with quality standards. Quality standards are detailed in the INSDC TPA Submission Guidelines Document (www.insdc.org/TPAguidelines.html).
EMBL TPA data are presented alongside EMBL Nucleotide Sequence Database primary entries in a variety of access tools at the EBI. Retrieval by accession number is available from the EBI Dbfetch service (www.ebi.ac.uk/cgi-bin/emblfetch). More complex queries based on term search and logical combinations of search results are available through the Sequence Retrieval System, SRS (srs.ebi.ac.uk). EMBL TPA are also made available to homology search tools (www.ebi.ac.uk/Tools/similarity.html). Retrieval and comparison of current and previously published sequence versions of EMBL TPA records are available at the EMBL Sequence Version Archive (www.ebi.ac.uk/cgi-bin/sva/sva.pl). EMBL TPA data are also made available on the EBI FTP site (ftp.ebi.ac.uk/pub/databases/embl/) and are included in the quarterly EMBL Nucleotide Sequence Database release. Data are exchanged on a nightly basis with INSDC partners to ensure synchrony and exhaustive coverage. Finally, specialist data sets and queries, when not available through access tools already offered, are prepared, where possible, to users' specifications.
[1] Stoesser G., Baker W., van den Broek A., Garcia-Pastor M.P., Kanz C., Kulikova T., Leinonen R., Lin Q., Lombard V., Lopez R., Mancuso R., Nardone F., Stoehr P., Tuli M.A., Tzouvara K. and Vaughan R. (2003) The EMBL Nucleotide Sequence Database: major new developments. Nucleic Acids Research, 2003, Vol. 31, No. 1 17-22
[2] Cochrane G., Bates K., Apweiler R., Tateno Y., Mashima J., Kosuge T., Karsch Mizrachi I., Schafer S. and Fetchko M. (2006) Evidence standards in experimental and inferential

INSDC Third Party Annotation data. ‘Omics in press