Accessing and distributing EMBL data using CORBA (common object request broker architecture)

ABSTRACT

The CORBA interfaces to the EMBL database address some of the problems of traditional flat-file formats and provide an efficient means for accessing and distributing EMBL data. CORBA also provides a flexible environment for users to develop their applications by building clients to our CORBA servers, which can be integrated into existing systems.

INTRODUCTION

Background The EMBL (European Molecular Biology Laboratory) Nucleotide Sequence Database (often referred to as the EMBL database) is hosted at the European Bioinformatics Institute (EBI). It is a comprehensive database of DNA and RNA sequences that are directly submitted from researchers and genome sequencing groups, and collected from the scientific literature and patent applications. It is produced in an international collaboration with GenBank (NCBI, Bethesda, USA) and DDBJ (the DNA Data Bank of Japan, CIB, Mishima, Japan). Each of the three collaborating groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged daily. The amount of sequence data is growing exponentially. As our scientific understanding deepens, the complexity of the related information increases as well. As a result, the structure of the data also keeps changing. The EMBL database is managed and maintained using the relational database management system (DBMS) Oracle. It contains over 130 tables and 140 relationships, having around 80 Gigabytes (Gb) of data comprising nearly 10 million objects of primary data and millions of sub-objects called 'features'. Traditionally, the sequences and related information, which have been collected over a long period of time, are made available in flat-file format via ftp, CD-ROM, www tools, and so on. The queries through tools such as SRS (Sequence Retrieval System, a network browser for databanks in molecular biology) also return data in flat-file format. However, flat files have a number of shortcomings: the format may not be described formally; it is difficult to represent complex data and relationships, the meaningful units of information ('objects') are not represented or handled well; it is hard to retrieve objects separately; assembly of objects into bigger aggregates is difficult; elaborate parsing is often required; and so on. In general, the current availability of the resources is not matched by a flexible environment to meet individual researchers' needs. An industry standard, the Object Management Group's (OMG) common object request broker architecture (CORBA), provides platform-independent programming interfaces and models for portable distributed object-oriented computing applications. Its independence from programming languages, computing platforms and network protocols provides a solution for developing new applications for querying and distributing biological data, which can also be integrated into existing systems. Here we present a CORBA infrastructure developed at EMBL-EBI and show that the CORBA interfaces to the EMBL database address some of the limitations of the flat-file format and provide an efficient means for accessing and distributing EMBL data. CORBA also provides a flexible environment for users to develop application programs (for example, for sequence analysis or data mining).

CONCLUSION

Conclusions This paper presents a CORBA infrastructure developed at EMBL-EBI. The EMBL object model provides a basis to develop the CORBA server. Employing PersistenceTM maps the object model to the relational schema in the underlying Oracle database. To present Persistence with the right relations, views have been used to transform the vertically mapped tables to horizontal ones. Properly built loaders make use of the technique of 'live object caching' and enhance the performance. The evictor pattern is used for memory management. It has been demonstrated that the CORBA server addresses some problems of the flat-file format and provides a solution to accessing and distributing EMBL sequence data. It also provides a flexible and scalable environment for users to develop their applications by building clients. The future work will include migrating the implementation of the EMBL server to comply with the emerging standard - OMG standard for biosequences. By OMG rules, the EBI, as a co-submitter on the Biomolecular Sequence Analysis (BSA) standard, is obliged to implement the standard. As the BSA standard proposal is not fully compatible with the EMBL IDL specification currently used, care will have to be taken to make this transition as easy as possible for existing clients. Additional data The following additional data are included with the online version of this article: The EMBL Nucleotide Sequence Database object model and The EMBL IDL specification.