BioSQL Reloaded: 1.0 Release, PhyloDB Module, and Future Features

Authors: Hilmar Lapp(1), Richard Holland(2), Aaron Mackey(3), William Piel(4), Mark Schreiber(5)
1 National Evolutionary Synthesis Center (NESCent), Durham, NC 27705, U.S.A. (hlapp@nescent.org)
2 EMBL—European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, United Kingdom
3 GlaxoSmithKline, Collegeville, PA 19426, U.S.A.
4 Peabody Museum of Natural History, Yale University, New Haven CT 06511, U.S.A.
5 Novartis Institute for Tropical Diseases, 138670 Singapore, Singapore

BioSQL (http://biosql.org) is a generic relational model for persistent storage of sequences, features, sequence and feature annotation, a reference taxonomy, and ontologies (or controlled vocabularies) in a way that is interoperable between the Bio* projects. While in its original incarnation (in 2001) conceived by Ewan Birney as a local relational store for GenBank, the project has since become a collaboration between the BioPerl, BioPython, BioJava, and BioRuby projects. The core schema of BioSQL has essentially been stable since November 2004, after it underwent substantial revisions at the 2002 and 2003 BioHackathons in Tucson, Cape Town, and Singapore. Here we report on a number of significant BioSQL developments that have taken place in the past 18 months.

Perhaps most notably, the 1.0 version of the core schema and supporting software was released in March 2008. While - intentionally - the 1.0 release does not introduce any structural changes to the model, for the first time of the project it defines a stable reference point that allows a roadmap for moving forward to be drawn. The 1.0 release was set in motion at the BioHackathon 2008 in Tokyo, which brought together a critical mass of leaders from several Bio* projects. Aside from starting the 1.0 release work, the concerted efforts at this event and those following it yielded for the first time a logo, a project wiki with information for both developers and users, a bug and feature request queue, and a language binding for BioRuby. The pre-release cleanup work, among many other things, also finally effected the change of the BioSQL license terms from the Perl-specific Artistic License to the widely used LGPL (Lesser GNU Public License). The future changes being charted at present range from better supporting generic object-relational mapping toolkits to versioning of sequence features, creating an audit trail, and supporting chimeric sequences.

Another significant development started more than a year ago at the Phyloinformatics Hackathon, which took place in December 2006 in Durham, North Carolina. At this event the BioSQL core schema was supplemented with a module, called PhyloDB, that extends the data types covered by BioSQL to phylogenetic trees (or networks). The module, which was considerably revised at the 2008 BioHackathon, follows the same design principles as the core schema, allowing arbitrary metadata attributes, typed by controlled vocabularies or ontologies, to be attached to nodes, branches, and trees. Tree nodes can be linked to sequences or taxa, and trees are scoped by a namespace similarly as databank entries. The module is accompanied by a number of common topological queries formulated in SQL. Furthermore, the PhyloDB module was the subject of a 2007 Google Summer of Code™ project under NESCent as the mentoring organization. The project resulted in a collection of stand-alone Perl scripts to import, export, manipulate, and query phylogenetic trees from or to the file formats supported by the Bio::TreeIO modules, which are part of BioPerl.

Aside from the significance of the accomplishments themselves, they continue the consistent history of BioSQL development sprints connected to events fostering intensive face-to-face collaboration. It is evidence for how an inherently collaborative cross-project effort such as BioSQL thrives, and possibly even relies on environments that allow deep and instant collaboration between key stakeholders and developers.