
Title	Lowering barriers to publishing biological data on the web
Authors	Brad Chapman, BioSQL and Biopython communities
Author affiliation	Massachusetts General Hospital, Boston, MA
Contact	chapmanb@50mail.com
URLs	http://biopython.org http://www.biosql.org http://bcbio.wordpress.com http://biosqlweb.appspot.com
Code URLs	https://github.com/chapmanb/biosqlweb/tree
Licenses	Biopython License, GNU LGPL

Talk summary

Scientists making their primary research data available on the web face an abundance of options. Many public databases and formats exist to solve similar problems, reflecting the complex nature of biological data. Additionally, researchers will often wait until the time of publication to format their data for public access. These factors can result in critical data being available as ad-hoc supplementary materials.

Data reuse is facilitated by standard web based presentation tools. An underlying shared architecture allows programmers to provide access points in widely used formats. Scientists gain an advantage from the visualization and organizational structure provided by these tools; this encourages collection of data from the start of a project, and allows data to be made available to other researchers earlier during the research process.

This talk describes a web based interface deployable on cloud computing resources. An existing database representation developed by the BioSQL project is utilized to store sequence data along with associated annotations and features. This data model is ported to the Google App Engine infrastructure, providing a full development stack for rapidly building and deploying web applications. A web interface designed with jQuery and jQueryUI serves as a framework for quickly implementing custom display and editing widgets. Finally, flat file data export is available via Generic Feature Format (GFF3); the structured format facilitates manipulation in Excel or text editors while being amenable to automated parsing.

More generally, the talk will discuss reusing open source frameworks to move towards a cloud-based approach to open data sharing. As data production continues to expand with the growth of next generation sequencing, individual research labs can help share the burden of data organization and presentation. In addition to relying on central repositories, we each take ownership of our data, presenting it in standard ways that encourage reuse and reanalysis.

Open source developers can encourage movement in this direction by simplifying the deployment of existing data models and frameworks. While the example discussed in this talk utilizes the BioSQL and Biopython infrastructures, the goals could be employed with a wide variety of data models and programming libraries. Our focus should be structuring our own open source work to be increasingly amenable to rapid integration.