Microbial-Pathogenesis.org, an Experimental System for Simultaneous Genomics Research and Data Sharing: Preliminary Report

WILLIAM C. RAY and ROBERT S. MUNSON JR.
Columbus Children's Research Institute and The Department of Pediatrics
The Ohio State University
700 Children's Dr. Columbus, OH 43205
USA
Email: ray.29@osu.edu  http://www.biosci.ohio-state.edu/~ray/

Abstract: Through publication of ray, the sharing of research results is the core reason for performing research; yet the process of online publication and data sharing is sometimes one of the most difficult aspects of sequence analysis projects. We have developed http://www.microbial-pathogenesis.org/ as an experiment in providing public and collaborative access to sequence data through public views directly into the same database in which active research and analysis is being performed. We hypothesize that sequence analysis and data sharing have many features in common with the widely-practiced collaborative authoring and management of online "web journals" and similar resources. Microbial-pathogenesis.org is therefore developed around the popular paradigm of a web-based Content Management System (CMS), optimized for managing sequence, and sequence-related data. We provide the curated annotations from the Haemophilus ducreyi HP35000 (GenBank[1] Accession# AE017143) and H. influenzae 86-028NP sequencing projects through this facility. A key feature is that the data visible on the system to the visiting public is identical in content to the research data being used for active analysis. Being database driven, updates to the research data are immediately reflected in any public views of that data with no effort required from the researcher. Further, all that is required to make a piece of data publicly visible, once it is deemed fit to publish, is to simply change its status in the database. Analyses and results for the supported projects include sequence-search access into the genomes and pre-generated BLAST[2] alignment results, as well as the results from custom local analyses for sequences in the database, literature and external resources links.

Key-Words: Database, Genomics, Sequence, Relational, LIMS, CMS, Publication.

1 Introduction

The microbial-pathogenesis.org system is an integrated database, data warehousing and analysis system, and online publication system. It was designed to simplify the tasks of analyzing genomic sequence data and providing the curated analysis results to the public. It is currently being used in the ongoing genomic and gene-expression analysis of H.d. 35000HP and H.i. 86-028NP, and for the online data-sharing repository for these analysis projects.

2 Rationale

Researchers involved in sequence analysis and publication often find a majority of their research time invested in data manipulations that are not directly related to the research they wish to conduct. Accessing sequence search tools, collecting and collating data, and archiving results frequently consume the majority of human resources applied to a sequence analysis project. Not only are the resources available for the actual analysis of the data limited by these impediments, but often, useful research insights can only be obtained after the expenditure of this initial, considerable, effort.

An additional common problem facing investigators involved in genomic research is the obligation to manage a variety of outside access needs for their data. Frequently, researchers are faced with situations where some data needs to be made public, some reserved for collaborators or lab members, and yet other data maintained privately. Managing these access restrictions is often cumbersome, and is a source of tension as researchers pit the opposing needs of reaping the full
This seemingly combined complication actually leads to a simplification of the process, if researchers are willing to accept a paradigm shift in the way they analyze and subsequently present their data. Specifically, a research project's published/publicly shared data and analyses may be thought of as simply a filtered view of all the data and analyses the project has collected. A single online database and analysis system used as a back-end, coupled with user and visitor permission management functions, and an appropriate collection of data and result filters that are matched to the user privileges can solve both the complications of data and analysis management, and simplify the task of data sharing.

3 Database Design

There have been three major design considerations taken into account in the development of the microbial-pathogenesis.org system.

First, the database needed to be as fully relational as possible with well-connected relations to support rapid access and presentation of all relevant information for any data item. Because of this design consideration, for example, sequences themselves are data objects, while their names, comments, and attached annotations are related values. This allows such typically imprecise manipulations as the uploading of an identical sequence under multiple different names, with different analyses carried out on each, to no-longer be an impediment to simultaneously understanding the complete body of the results. The system, as it knows the sequence itself as a single data entity, can present the researcher with all of the results, as carried out via any of the names for the sequence, when the sequence is accessed by any or all of the submitted names.
Second, the system needed to store sufficient data, and metadata regarding the data to encourage the research users to use it as their primary data and analysis storage system, rather than using the online system as a secondary adjunct to the traditional collection of sequences and results scattered across ZIP disks and email. This primary-level use is necessary to provide for the system's ability to keep shared versions of data consistent with the most recent data revisions.

Third, the data structures stored were designed to be flexible and extensible so that new data or results types can be stored and accessed without requiring significant code revisions. The system is designed to allow the researcher to use it as a "notebook" to house and share any type of notational information they have collected regarding a sequence. The existing structures are sufficient to store, index and share almost any notational information that can be stored in a file, or indicated by a hypertext link. As the system is a research system, additional capabilities are developed as needed. Significantly, because the data, results and analysis capabilities provided to any user are entirely database driven, adding a new analysis type or new data relationship makes this feature available immediately for all data in all projects supported by the system.

Finally, while not strictly a database design issue, the JavaScript/HTML interface to the system and the database design are tightly related. The interface is designed to make use of advanced, but well-supported web-programming constructs such as HTML hidden-layer methods and user cursor-tracking to provide access to as many aspects of the data as possible in a seamless and responsive fashion.

The database schema used to drive www.microbial-pathogenesis.org is shown in Figure 1, while a screenshot from a representative data browsing session is shown in Figure 2.
This database is implemented in MySQL(mysql.com) and the database-driven Web interface, and back-end analysis capabilities are provided by a combination of PHP/php.net) and Perl[3] code. The web server is Apache(apache.org), and the system currently resides on a Linux/Intel system. The JavaScript and HTML features used are best supported in Microsoft Internet Explorer versions >= 5.0, and Netscape Navigator or Communicator versions >= 4.7.

4 The Data, and Access Patterns

Microbial-pathogenesis.org currently provides both research access to the active H. ducreyi and H. influenzae sequencing project members, and public views of the completed annotation for H. ducreyi HP35000 and of the ongoing sequencing results for H. influenzae 86-028NP. With the site having been in active use for approximately a year, usage is increasing roughly linearly. As of March 2004, it is serving approximately 8000 requests per month. Twenty percent of these are attributable to the sequencing project members, and the majority of the remainder are driven to the site by search-engine indexes of the publicly visible data.

The data on the site includes 15925 unique sequences, collected as 326 independent data sets. Interestingly, some of these 15925 sequences have been submitted multiple times, resulting in 26935 names. That so many sequences have been submitted two or more times with distinct names, strongly supports our initial hypothesis that manual management of sequence data is not typically handled efficiently by working researchers. 7934 of the sequences contained in the database are visible through public views.

The site also indexes 282150 unique archived BLAST search results and alignments containing information such as cross-comparisons between H.d. HP35000 and H.i. Rd. Of these, 4546 are visible through public views. The researchers have also created 38 searchable BLAST databases, 14 of which have been selected as publicly visible.

Curator notes and sequence-related literature links are part of the system and these will become available to the browsing public as the genomes (H.d. was submitted to GenBank as of July 2003, H. influenzae 86-028NP was at 8-fold coverage as of August 2003, and is expected to close in mid-2004) are formally published.

5 Future Developments

The microbial-pathogenesis.org system will continue to be used for analysis and data sharing for H.d. 35000HP and H.i. 86-028NP. The H.d. genome has been completed and is nearing publication as this paper is written, while the H.i. genome is nearing closure. The results of ongoing curation for each will be available from the system. Microarray expression experiments are underway for both organisms and results will be coupled to the stored sequence data as these become available. As this is an experimental system, additional features, such as automated searching and linking to external literature databases are being developed. We are working on developing links into additional external online sequence databases such as SwissProt[4]. We are also interested in the possibility of hosting other sequence analysis or publication projects on the system.

Acknowledgements:

This work was supported by National Institutes of Health grants R01 DC03915 (L. Bakaletz, PI), R01 AI45091 (RSM) and R01 DC005980 (RSM).

References: