Bioinformatics Open Source Conference 2005

Table of Contents

Schedule

Table 1. Schedule

	Thursday June 23rd	Friday June 24th
08:30am	call to order	call to order
08:40am	Stand Dong: Transforming Full-Text Literature to Formalized Facts	KeyNote: Open Bioinformatics Foundation
09:05am	Fabien Campagne: SigPath: Quantitative information management for cell signaling pathways and networks	Martin Senger: The BioMOBY project
09:30am	Gregg Helt: DAS/2: Next Generation Distributed Annotation System	Peter Rice: EMBOSS
10:00am	Coffee	Coffee
10:20am	Aaron Darling: mpiBLAST evolves - Success, Collaborations, and Challenges	Arek Kasprzyk: BioMart - Data Integration Made Easy
10:45am	Himanso Sahni: REMBRANDT: Building a Robust Translational Research Framework for Brain Tumor Studies	Tom Oinn: Taverna
11:10am	Xiaokang Pan: SynBrowse: A Synteny Browser for Comparative Sequence Analysis	Aaron Darling: Mauve - GPL'ed modules for multiple genome comparison
11:35am	Sumeet Muju: caArray data Management and Analysis Tools at the National Cancer Institute (NCI) Center for Bioinformatics	Warren L. DeLano: PyMOL: A commercially-supported open-source tool for bioinformatics
12:00	Lunch	Lunch
1:00pm	KeyNote Speaker: Jason Stajich: Building Bioperl: lessons for Open-Source and Bioinformatics	Peter A. Covitz: Interoperability Architecture of the Cancer Biomedical Informatics Grid
1:25pm		Richard Holland: DengueInfo
2:00pm	Lightning talks and Demos	Lightning talks and Demos
2:30pm	Coffee:	Coffee:
2:50pm	Lightning talks and Demos	Lightning talks and Demos
5:00pm	BOFs	BOFs

Lightning Talk Schedule

Table 2. Lightning Talk Schedule

Thursday June 23rd	Friday June 24th
Marcus Breese: BioNote	Christopher Bottoms: Accidentally becoming a bioinformatics programmer
Julien Gervais-Bird: Juscan sequence search	Allen Day: Biopackages.net
Toshiaki Katayama: BioRuby updates	Tetsuro Toyoda: ARTADE
Peng Chen: Bio.Mambo	Christopher Bottoms: Bioinformatics of protein bound water
Evgeny Kireev: VNTI Bioperl integration of Bioperl and VectorNTI API	Michael Heuer: BioJava updates
Martin Senger: Bionanny	Giridhar Pemmasani: CREAD
Andrew Dalke: EUtils client library for Biopython	Kazuharu Arakawa: G language Project in 2005
Paulo Nuin: InFASTA	Michael Heuer: Stack API for XML (StAX)
Carlos Rodriguez: Bioknoppix linux for the life sciences	Huhn-Kie Lee: DrosoPhylo
Further Speakers To Be Announced	Further Speakers To Be Announced

Welcome

Welcome to BOSC 2005! This is the 6th official Bioinformatics Open Source Meeting. We are very pleased to announce Jason Stajich, of BioPerl fame, as a keynote speaker this year. A second keynote presentation will be made by the Open Bioinformatics Foundation. We have an impressive group of speakers scheduled during each morning session, focusing on a broad range of topics ranging from user applications, to novice user education, to broader treatments of open source principals in the academic field. During the afternoon sessions we have scheduled Lightning Talks and Software demonstrations. Birds of a Feather (BOF) discussions will occur at the end of each day. Please take advantage of this time to attend discussions on specialized topics. If you would like to schedule a BoF see the signup chart that will be available in the mornings. If you have any questions or concerns about the conference please let one of the conference committee members know.

We hope you enjoy yourself, learn a lot, and most importantly get to know each other and become part of the community of open source development in the life sciences.

Conference Committee

Darin London (chair)	European Bioinformatics Institute
Ewan Birney	European Bioinformatics Institute
Andrew Dalke	Dalke Scientific Software
Hilmar Lapp	GNF
Nomi Harris	University of California, Berkeley
Amonida Zadissa	Otago University, New Zealand

Abstracts

Keynote: Building Bioperl: lessons for Open-Source and Bioinformatics

Jason Stajich, Department Molecular Genetics and Microbiology, Duke University

Transforming Full-Text Literature to Formalized Facts

Stan Dong, Department of Genetics, Stanford University School of Medicine

Thursday June 23rd, 8:40am

Scientific research literature provides data to support and contradict hypotheses, but the data is not constrained to formats suitable for the extraction of information using an automated system. The Saccharomyces Genome Database (SGD; http://www.yeastgenome.org) and the Tetrahymena Genome Database (TGD; http://www.ciliate.org) are model organism databases (MOD) that provide gene and protein information for their respective model organisms. The bulk of this information is curated from published literature, and requires manual knowledge extraction by scientific curators. Manual literature based curation is precise but time consuming. Given the limitations of human resources, text-mining techniques can identify the most relevant literature that should be reviewed by scientific curators for inclusion in a MOD. An additional benefit is that the same textmining techniques can be used by the larger scientific community to identify literature that contains the appropriate information. SGD and TGD are building an automated pipeline for collecting full-text documents of literature relevant to their respective scientific communities. So far more than 15,000 full-text documents in either PDF or HTML format have been archived. Of these, 64% have already been reviewed by scientific curators and can serve as a training set for existing and novel text-mining algorithms. As our first attempt to incorporate full-text literature searching as a resource for scientific curators and the scientific community, SGD and TGD implemented Textpresso, a vocabulary-based information retrieval and extraction system developed by Muller and Kenny at WormBase. Textpresso allows users to perform custom queries of full-text journal articles based on keywords and/or ontology-based categories of terms. We report software modifications that increased the speed of Textpresso ontology markup. We also report database-specific expansions and modifications to the underlying ontologybased categories used in the processing of papers. Textpresso is a valuable tool that can help users and curators identify relevant information within a large body of literature. Other text mining strategies and techniques under development will also be presented. SGD is funded by the US National Human Genome Research Institute. TGD is funded by the National Institute of General Medical Sciences.

Textpresso is open-source software and is part of the Generic Model Organism Database (GMOD; http://www.gmod.org) effort.

Reference: Muller H, Kenny EE, Sternberg PW. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PLoS Biol. 2004 November; 2(11): e309.

SigPath: Quantitative information management for cell signaling pathways and networks

Fabien Campagne, Weill Medical College of Cornell University

Thursday June 23rd, 9:05 am.

Understanding complex protein networks within cells requires the ability to develop quantitative models and to numerically compute the properties and dynamical behavior of the networks. To carry out such computational analysis, it is necessary to use modeling tools and information management systems (IMSs) where the quantitative data, associated to its biological context, can be stored, curated, and reliably retrieved. The SigPath project focuses on the biochemical computation of cellular interactions and develops an IMS that stores quantitative information on the cellular components and their interactions. This information can be used to construct pathways and eventually large-scale networks. Yet, assembling the information to achieve this goal will require the active collaboration of many experimental and computational labs. The SigPath IMS aims to serve as a resource for these labs, where information can be stored electronically and shared across the internet, to enable collaborative data gathering and biochemical modeling studies. The talk will present the goals, approaches and road-map of the SigPath project. SigPath is distributed under the GNU General Public License. We invite and welcome others interested in joining this project.

Project URL: http://www.sigpath.org

License: GPL

References: http://www.biomedcentral.com/content/pdf/1471-2105-6-5.pdf http://www.biomedcentral.com/content/pdf/1471-2105-6-5.pdf

DAS/2: Next Generation Distributed Annotation System

Gregg Helt, Affymetrix, Inc.

Thursday June 23rd, 9:30am

The Distributed Annotation System (DAS) is a specification for sharing distributed annotations of biological sequences, first introduced in 2000. DAS allows researchers to integrate biological information from many different sources via standardized URL queries and XML responses over HTTP. Although DAS has enjoyed some success it also has a number of problems, and there have been many discussions on the DAS mailing list of how best to improve it, ranging from minor modifications to complete overhauls. It has become clear that a major revision of the DAS protocol is needed. We have integrated many of the suggestions from the DAS developer and user community to produce a new version of DAS, DAS/2.

The preliminary DAS/2 specification is now ready for review by a wider audience. Enhancements are currently centered on two major categories. First is revising retrieval protocols to support more flexible queries and responses. This includes the ability for servers to support arbitrary formats for annotation retrieval, and for clients to choose which format they prefer. Nested hierarchies of annotation are supported, and annotations are typed based on ontologies. Extensive filtering is also supported in the annotation request. The second major category of enhancement is development of a new writeback protocol to support creation of new biological data objects and editing of existing ones. Future plans include adding support for DAS/2 server registry and discovery.

In addition to the DAS/2 specification itself, we have developed preliminary implementations of a DAS/2 server, a DAS/2 client, and a DAS/2 validation suite. The server implementation has a plugin architecture to support different backends including GMOD databases. The client implementation is part of the Integrated Genome Browser (IGB), an application for genome visualization and exploratory

analysis. The standalone validation suite is used to verify that client requests and server responses follow the specification.

Client, server, and validation suite are all available under open source licenses.

This work is funded in part by NIH grant R01HG003040.

mpiBLAST evolves - Success, Collaborations, and Challenges

Aaron Darling, Dept. of Computer Science, Dept. of Animal Health & Biomedical Science Univ. of Wisconsin-Madison

Thursday June 23rd, 10:20am

mpiBLAST has become a widely used and widely critiqued open-source parallelization of NCBI BLAST. Many of its users (and developers!) have a love/hate relationship with it. Early implementations of mpiBLAST's database segmentation method promised tremendous potential for super-linear speedup on compute clusters. As mpiBLAST has evolved it has gained considerable flexibility and applicability--originally providing only blastn functionality with approximate E-values whereas it now provides all of {t}blast{n,p,x}, all 12 result output formats, exact E-value statistics, and load-balancing. As mpiBLAST evolved, its efficiency and scalability met pitfalls that were overcome largely in part to feedback and contributions from the bioinformatics open-source community. The first part of this talk describes challenges during mpiBLAST's collaborative development and how they were overcome through community contributions. In the second part of the talk, I will assess the current state of mpiBLAST (its scalability and stability) and discuss future development and maintainence ideas. Audience feedback heartily encouraged.

LICENSE: GPL

PROJECT URL: http://mpiblast.lanl.gov

REMBRANDT: Building a Robust Translational Research Framework for Brain Tumor Studies

Himanso Sahni, National Cancer Institute Center for Bioinformatics (NCICB)

Thursday June 23rd, 10:45am

The mission of the National Cancer Institute Center for Bioinformatics (NCICB) is to provide informatics infrastructure and scientific applications that support advanced translational research in cancer biology and medicine. REpository for Molecular BRAin Neoplasia DaTa (REMBRANDT) is a robust bioinformatics knowledgebase framework that leverages data warehousing technology to host and integrate clinical and functional genomics data from clinical trials involving patients suffering from Gliomas. The knowledge framework will provide researchers with the ability to perform ad hoc querying and reporting across multiple data domains, such as Gene Expression, Chromosomal aberrations and Clinical data. Scientists will be able to answer basic questions related to a patient or patient population and view the integrated data sets in a variety of contexts. Tools that link data to other annotations such as cellular pathways, gene ontology terms and genomic information are embedded within this system.

The Rembrandt application is designed using object oriented methodology and implemented using Java 2 Enterprise Edition, a Data Warehouse schema, and various other open source technologies. It is designed to conform to an n- tiered architecture that includes several layers: A web-based graphical user interface is built using Apache Struts application framework and runs on Apache Jakarta Tomcat, a business object layer, server components that process the queries and manipulate the result sets, a data access layer using the Apache ObjectRelationalBridge project (OJB), and

a robust data warehouse. Graphical Reports are generated using the Krysalis jChart and the jFreeChart packages. For tabular reports, the application first converts the result set into an XML document and then applies various XSL templates using JDOM to display it in various contexts such as Gene Expression, Copy Number or Clinical report.

One of major challenges when dealing with any kind of gene expression or genomic data is that the volume of such data grows exponentially even when dealing with a small number of patient samples. In order to create efficient queries that give us best performance, we designed a parallel query mechanism that executes queries within the OJB framework using Java threads, An extensive caching mechanism using Ehcache allows users to access previous results within the user session while errors and exceptions are logged using Log4j.

The 0.51 version of the Rembrandt application can be accessed from http://rembrandt-db.nci.nih.gov and is available for free access (To obtain login information, please register with NCICB Application support, ncicb@pop.nci.nih.gov). The source code will be made available later this year during 1.0 release. This effort is supported by funds from National Institute of Neurological Disorders and Stroke, Neuro-Oncology Branch, National Cancer Institute and the National Cancer Institute, Center for Bioinformatics and further information about the project can be accessed at http://rembrandt.nci.nih.gov.

SynBrowse: A Synteny Browser for Comparative Sequence Analysis

Xiaokang Pan, Department of Genetics, Development and Cell Biology, Iowa State University

Thursday June 23rd, 11:10am

We have developed SynBrowse, a synteny browser for visualizing and analyzing genome alignments both within and between species. It is intended to help scientists visualize and analyze macro-synteny, micro-synteny and homologous genes between sequences. It can also aid with identification of uncharacterized genes, putative regulatory elements, and novel structural features of a newly sequenced species by comparison with a well-annotated reference sequence. SynBrowse is a GBrowse (the Generic Genome Browser) family software tool that runs on top of the open source Bioperl modules Bio::Graphics and Bio::DB::GFF. It consists of two components: a web-based front end and a set of relational database back ends. Each database stores pre-computed protein and nucleotide alignments from a focus sequence to reference sequences in addition to its genome annotations, such as gene models and EST spliced alignments. The user interface lets end users select a key comparative alignment type and search for syntenic blocks between two sequences and zoom in to view the relationships among the corresponding genome annotations in detail. Like GBrowse, SynBrowse provides system administrators with simple installation, flexible configuration, convenient data input, and easy integration with other components of a model organism system.

The project page and the demonstration of the software configured for plant cross-species comparisons are available at http://www.synbrowse.org.

Source code of SynBrowse will be publicly available at http://www.gmod.org/ in June, 2005.

caArray data Management and Analysis Tools at the National Cancer Institute (NCI) Center for Bioinformatics

Sumeet Muju, National Cancer Institute (NCI) Center for Bioinformatics

Thursday June 23rd, 11:35am

caArray database is a standards based open source data management system, version 1.0 was released in January 2005. caArray features MIAME 1.1 compliant data annotation forms, controlled vocabularies (MGED ontology), and MAGE-ML import and export. caArray also provides open interfaces for programmatic access to microarray data.

The caArray database and analysis tools were developed to be consistent with caBIG compatibility guidelines that highlight use of controlled vocabularies, CDEs, well documented APIs and UML models. caBIG is a new initiative coordinated by NCI in partnership with other members of the cancer research community. caBIG seeks to create a network that links organizations, institutions, and individuals to enable the sharing of cancer research infrastructure, data, and interoperable tools. It is an open-access, open-source activity that promises to expedite progress in cancer research.

caArray datasets and open source tools are publicly available, and can be accessed at http://caArray.nci.nih.gov, caArray source code is available for local installations at http://ncicb.nci.nih.gov/download under an open source license.

The BioMOBY project

Martin Senger, European Bioinformatics Institute

Friday June 24th, 9:05am

Since its inception in September, 2001, the BioMOBY project has been exploring solutions to Web Service interoperability. The project goal was to create a community-backed specification that addressed the critical data access needs of biologists and informaticians, and yet was simple for host institutions to implement. Among the architectures tested, the MOBY Services (MOBY-S) architecture has been the most widely implemented. MOBY-S limits the archetypal Web Service paradigm by defining valid data structures in an end-user-extensible ontology, and making the Web Service registry (MOBY Central) aware of this ontology. As of April, 2005, there were four MOBY Central registries deployed worldwide (Canada, Germany, Philippines, Australia), with the primary public registry holding more than 240 interoperable services from over 80 service providers. The data-type ontology has been extended by MOBY users, and now exceeds 120 data classes. The focus of MOBY has now shifted to creation of powerful client programs, including enhanced support for MOBY Services in the Taverna software from the myGrid project. In this presentation we will discuss the MOBY API, the successes and failures of the project over the past 4 years, and the lessons we have taken from these experiences.

Project URL: http://biomoby.org

EMBOSS

Peter Rice: European Bioinformatics Institute

Friday June 24th, 9:30am

EMBOSS is a joint development by the institutes on the Hinxton Genome Campus. EMBOSS is licensed under GPL, includes over 300 applications and has comprehensive libraries for rapid code development. EMBOSS has been installed at more than 10,000 sites, and is used by thousands of registered users at the Rosalind Franklin Centre for Genomics Research.

EMBOSS is aimed at providing all the common sequence analysis functions, and related applications, using local and remote sequence data sources. Sequence databases can be shared with other software, as EMBOSS includes access methods for most common sequence database formats. Databases and other resources are defined in site-wide and personalised user resource files.

A key feature of EMBOSS is the use of "ACD files" to completely define the interface between the code and the user. By building an ontology around these ACD files, and by converting them into various alternative representations, EMBOSS has been automatically transformed into web interfaces, GUIs, CORBA, and web services (a total of more than 30 known interfaces). The ACD "standard" has also been extended to define various third party applications and packages.

The new release of EMBOSS (3.0.0) provides advanced features for application integration into GUI and web interfaces, and for web service and grid service providers.

The focus of EMBOSS now extends beyond sequence analysis to include phylogenetics, structural biology and other non-sequence datatypes. These will continue to expand in future releases.

EMBOSS is licensed under GPL/LGPL

Project URL: http://www.emboss.org">http://www.emboss.org

BioMart - Data Integration Made Easy

Arek Kasprzyk, European Bioinformatics Institute

Friday June 24th, 10:20am

BioMart is a simple and robust data management system that can readily be adapted to any type of data. It provides a range of sophisticated and highly configurable query interfaces. The system makes it possible to optimse querying of large databases and facilitates data federation between different types of data available anywhere on the network. It can be readily deployed as a standalone perl-based website and/or java-based GUI and text-based interfaces.

BioMart uses an XML-based configuration system which contains meta-data describing data semantics and the properties of all it's interfaces. The configuration system is managed by MartEditor - a java based configuration editor. This tool offers a variety of configuration options including automatic 'de-novo' configuration generation, validation and updates.

BioMart has been integrated with variety of third party software. It is an integral part of the Taverna workflow system and Ensembl website. It has been integrated with the Bioconductor suite and provides support for the ProServer DAS server.

All components of the BioMart system can be installed 'out of the box' without the need for any additional programming. The software has now been installed and adapted for external data in a variety of academic and comercial institutions.

All BioMart software is free, licenced under LGPL and freely available to everyone

Project URL: http://www.ebi.ac.uk/biomart

Taverna

Tom Oinn, European Bioinformatics Institute

Friday June 24th, 10:45am

Taverna is a graphical workflow construction toolkit and playground targeted at the life sciences (bioinformatics and cheminformatics). The last year has shown a rapid acceptance and increase in understanding of service oriented technologies within this community and Taverna has evolved to suit.

The presentation would focus on two main areas - firstly the underpinning technologies that Taverna makes use of to allow workflow construction including how it relates to other open source projects such as BioMart and Soaplab as well as standards such as SOAP. Secondly it will present the abstractions and user interface metaphors which specialize it as a bioinformatics workflow platform as opposed to a generic

scientific or business orchestration tool. Such abstractions are particularly important when trying to provide a bridge between communities as distinct as the distributed computing / grid world and that of the typical domain scientists.

Taverna is part of the myGrid project, is available under the LGPL and has active Linux, Windows and Mac OSX binary releases alongside the source and a comprehensive user manual. It has attracted contributions from a significant number of other projects and developers.

Project URL: http://taverna.sf.net.

Mauve - GPL'ed modules for multiple genome comparison

Aaron Darling, Dept. of Computer Science, Dept. of Animal Health & Biomedical Science Univ. of Wisconsin-Madison

Friday June 24th, 11:10am

With a glut of genome sequences being determined from all domains of life, comparative genomics has entered prime time. Bacterial and eukaryotic genomes have complex structures, frequently involving genome rearrangement, segmental duplication, and mutations derived from other types of recombination. Mauve performs multiple alignment and comparative visualization of complex genomic structures using a set of open-source, documented C++ and Java libraries.

Mauve has two top-level components--the mauveAligner implemented in C++ and the visualization environment implemented in Java. mauveAligner uses an anchored alignment algorithm to rapidly align rearranged genomes. The first step of the method identifies high scoring local alignments. Next, some local alignments (presumably either spurious matches or paralogs) are filtered and remaining local alignments are grouped into collinear regions called Locally Collinear Blocks (LCBs). Each LCB is a region of the chromosome without significant rearrangement. Finally, Mauve aligns each LCB using a progressive alignment method.

The Mauve visualization environment reads alignments in eXtended Multi-FastA format in addition to any available annotation in GenBank format. The visualization environment displays a similarity profile based on average alignment column entropy. Unlike other visualization tools, Mauve displays genomic structure by showing all genomes rather than a single reference genome. Finally, Mauve leverages other open-source tools like BioJava to read and display sequence data and features.

Project URL: http://gel.ahabs.wisc.edu/mauve

LICENSE: GPL

PyMOL: A commercially supported Open Source tool for Bioinformatics

Warren L. DeLano, DeLano Scientific LLC

Friday June 24th, 11:35am

The PyMOL molecular viewer enables communication of molecular structure information in live, animated, and published formats. It provides molecular visualization on any OpenGL-compatible computing platform and supports automation and integration through its native Python interface. Because PyMOL is open-source, community-driven, and widely used, bioinformaticians can feel comfortable about making a long-term investment in learning the program and depending upon it for a variety of routine visualization tasks. Also, because PyMOL is commercially supported, companies and institutions can adopt the package with confidence that there is an accountable party standing behind the code. Here we provide an overview of the package and demonstrate some of the recent new features that increase its suitability for broad sharing of molecular structures and associated information. High-

lights include a sequence browser, 3D visualization of aligned molecules, surface property display, animated scene transitions, and portable session files.

Project Homepage: www.pymol.org

Open Source License: BSD-like (based on the Python 1.5.2 license).

Releases: PyMOL has been released continuously since April 2000.

Interoperability Architecture of the Cancer Biomedical Informatics Grid

Peter A. Covitz, National Cancer Institute Center for Bioinformatics

Friday June 24th, 1:00pm

The Cancer Biomedical Informatics Grid (caBIG) program was launched by the U.S. National Cancer Institute (NCI) to meet the challenge and need for a more highly coordinated approach to informatics resource development, management and dissemination. The caBIG program includes a large federation of participants from cancer centers, government agencies, bioinformatics companies, and patient advocate groups that define and build interoperable, reusable systems for cancer research information.

Participants are organized into workspaces that tackle the various dimensions of the program. A diverse number of open source projects are underway, including systems for clinical trials data collection; serious adverse event reporting; tissue bank management; MIAME-compliant microarray data management; microarray and proteomics data analysis; and biomolecular pathway modeling. Given the diversity, interoperability and standards are heavily emphasized. Two cross-cutting workspaces one for Architecture the other for Vocabularies and Common Data Elements - govern syntactic and semantic interoperability requirements.

caBIG has defined compatibility categories for information models, common data elements, vocabularies, and programming interfaces. These categories stem from previous work by the caCORE project in the area of semantics, metadata, and model-driven architecture. Compatibility in each category is defined with differing levels of stringency, labeled in ascending order as caBIG Bronze, Silver and Gold. The Silver level is quite stringent, and demands that systems adopt and implement standards for information modeling and data typing, metadata registration, and programming interfaces. The Gold level adds a service-oriented data and analysis grid architecture named "caGrid" that future caBIG systems will register with and plug into.

caGrid leverages caCORE to provide the necessary semantic typing, modeling, and metadata structures that define and constrain the contents of services in caGrid. A caGrid service registry supports advertising, discovery, query, and analysis workflow use cases. caGrid specifies common interface and syntax requirements for registering and using grid services. A universal data object identifier strategy based upon the Life Science Identifier (LSID) specification is being explored. caGrid facilities are all built and extended from the Globus, Open Grid Services Architecture-Data Access Integration (OGSA-DAI), and Mobius toolkits.

caCORE and caGrid software components are available under the open source caBIO license that can be found at http://ncicb.nci.nih.gov/NCICB/core/caBIO/technical_resources/core_jar/license. This license allows for both nonprofit and for-profit usage.

caBIG Compatibility Guidelines are at http://cabig.nci.nih.gov/guidelines_documentation.

The caCORE project page is http://ncicb.nci.nih.gov/core

The caGrid project is currently run out of the caBIG Architecture Workspace: http://cabig.nci.nih.gov/workspaces/Architecture.

Code can be found on the caBIG CVS site: http://cabigcvs.nci.nih.gov/viewcvs/viewcvs.cgi and on the NCICB Download site: http://ncicb.nci.nih.gov/download.

DengueInfo

Richard Holland, Genome Institute of Singapore

Friday June 24th, 1:25pm

DengueInfo is a collaboration between the Novartis Institute for Tropical Diseases and the Genome Institute of Singapore. It serves as a repository for dengue virus genomic and clinical information. The database currently holds all the dengue genome sequences from Genbank and also, where available, the associated clinical information. The web frontend also serves as a link to global information on dengue, eg. WHO, CDC, PubMed, Google News etc. A structured but logical query mechanism allows sequences to be searched on any number of attributes or annotations, and in any combination. This is one of the key features of the system, making it just as easy to search for a single accession number as it is to search for all genomes of a particular strain from a particular country within a given date range of isolation. At present, sequences from search results can be displayed as Fasta (truncated to any range and/or translated in any frame), as Genbank-style records, as part of a table of clinical information, or annotated as a group. The sequences are also compiled into a searchable database using the BioJava implementation of the SSAHA algorithm, allowing users to perform similarity searches across the entire collection. DengueInfo is written as a web application using Java Struts and JSTL, and makes extensive use of the new features Introduced with Java 1.5. It relies heavily on BioJava and BioSQL for sequence retrieval and presentation, implementing a custom annotation ontology to enable us to store clinical information inside the standard BioSQL schema. The custom ontology is well-defined using an XML XSD, and Castor is used to import/export the entire ontology on-the-fly. The system can also synchronize itself with Genbank by using the NCBI Entrez Utilities Web Service (via Apache Axis) to identify and import any sequences of interest that match given search criteria. The database and frontend are designed to be entirely portable and customisable for individual applications, making it easy to adapt for use with other species or existing BioSQL-based sequence databases. The tiered structure of the application also makes it possible for users to substitute any other database schema (BioSQL) or data access interface (BioJava) of their choice.

LGPL version 2.1, February 1999. To be released on request by email to the authors, and/or (hopefully) by anonymous read-only access to a Subversion source code repository.