Open source science literature review

For project: Open source science

Open source science literature review

Below is a chronological list of articles pertaining to Open Source Science in Software and Hardware.

Chemistry/ Biology / Medicine / Molecular Modeling


- Abstract

EMBOSS is "The European Molecular Biology Open Software Suite". EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology (e.g. EMBnet) user community. The software automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web. Also, as extensive libraries are provided with the package, it is a platform to allow other scientists to develop and release software in true open source spirit. EMBOSS also integrates a range of currently available packages and tools for sequence analysis into a seamless whole. EMBOSS breaks the historical trend towards commercial software packages.


- Abstract

Over the past few years the world of free and open source geospatial software has experienced some major changes. For instance, the website FreeGIS.org currently lists 330 GIS-related projects. Besides the advent of new software projects and the growth of established projects, a new organisation known as the OSGeo Foundation has been established to offer a point of contact. This paper will give an overview on existing free and open source desktop GIS projects. To further the understanding of the open source software development, we give a brief explanation of associated terms and introduce the two most established software license types: the General Public License (GPL) and the Lesser General Public License (LGPL). After laying out the organisational structures, we describe the different desktop GIS software projects in terms of their main characteristics. Two main tables summarise information on the projects and functionality of the currently available software versions. Finally, the advantages and disadvantages of open source software, with an emphasis on research and teaching, are discussed.

- Abstract

BioMOBY is an Open Source research project which aims to generate an architecture for the discovery and distribution of biological data through web services; data and services are decentralised, but the availability of these resources, and the instructions for interacting with them, are registered in a central location called MOBY Central. BioMOBY adds to the web services paradigm, as exemplified by Universal Data Discovery and Integration (UDDI), by having an object-driven registry query system with object and service ontologies. This allows users to traverse expansive and disparate data sets where each possible next step is presented based on the data object currently in-hand. Moreover, a path from the current data object to a desired final data object could be automatically discovered using the registry. Native BioMOBY objects are lightweight XML, and make up both the query and the response of a simple object access protocol (SOAP) transaction.


- Abstract

The flood of sequence data resulting from the large number of current genome projects has increased the need for a flexible, open source genome annotation system, which so far has not existed. To account for the individual needs of different projects, such a system should be modular and easily extensible. We present a genome annotation system for prokaryote genomes, which is well tested and readily adaptable to different tasks. The modular system was developed using an object-oriented approach, and it relies on a relational database backend. Using a well defined application programmers interface (API), the system can be linked easily to other systems. GenDB supports manual as well as automatic annotation strategies. The software currently is in use in more than a dozen microbial genome annotation projects. In addition to its use as a production genome annotation system, it can be employed as a flexible framework for the large-scale evaluation of different annotation strategies. The system is open source.


- Abstract

Recently, several authors have suggested that a new method of doing science called "open source biology" is about to emerge. However, very little has been written about how such an institution would differ from existing research institutions. Scientific databases provide a natural model. During the 1990s, scientists experimented with several new database initiatives designed to reconcile private support with the ideals of open science. Despite significant controversy, this paper argues that private/public transactions that unambiguously promote academic science should be encouraged. In principle, research communities can also organize database collaborations to pursue social and political goals. Examples include discouraging software patents, promoting "green" investment, and improving internet security. Finally, the new field of computational genomics blurs the traditional line between database creation and product development. This paper describes how traditional database institutions can be modified and extended to discover pharmaceuticals. The proposed institution ("open source drug discovery") would be particularly useful for combating Third World diseases. Success would demonstrate that the open source institution is not limited to computer science and can develop products other than software.

- great article - lots of good refs to 3rd world applications using drug discovery good lit review of open source databases

- Abstract

We have developed a toolbox and graphic user interface, EEGLAB, running under the crossplatform MATLAB environment (The Mathworks, Inc.) for processing collections of single-trial and/or averaged EEG data of any number of channels. Available functions include EEG data, channel and event information importing, data visualization (scrolling, scalp map and dipole model plotting, plus multi-trial ERP-image plots), preprocessing (including artifact rejection, filtering, epoch selection, and averaging), independent component analysis (ICA) and time/frequency decompositions including channel and component cross-coherence supported by bootstrap statistical methods based on data resampling. EEGLAB functions are organized into three layers. Top-layer functions allow users to interact with the data through the graphic interface without needing to use MATLAB syntax. Menu options allow users to tune the behavior of EEGLAB to available memory. Middle-layer functions allow users to customize data processing using command history and interactive 'pop' functions. Experienced MATLAB users can use EEGLAB data structures and stand-alone signal processing functions to write custom and/or batch analysis scripts. Extensive function help and tutorial information are included. A 'plug-in' facility allows easy incorporation of new EEG modules into the main menu. EEGLAB is freely available (http://www.sccn.ucsd.edu/eeglab/) under the GNU public license for noncommercial use and opensource development, together with sample data, user tutorial and extensive documentation.


- Abstract

The fragmented ownership of rights to intellectual property (IP) in agricultural biotechnology leads to situations where no single public-sector institution can provide a complete set of IP rights to ensure freedom to operate with a particular technology. This situation causes obstacles to the distribution of improved staple crops for humanitarian purposes in the developing world and specialty crops in the developed world. This Policy Forum describes an initiative by the major agricultural universities in the United States and other public-sector institutions to establish a new paradigm in the management of IP to facilitate commercial development of such crops.


- Abstract

'OpenSource' is a 20–40 year old approach to licensing and distributing software that has recently burst into public view. Against conventional wisdom this approach has been wildly successful in the general software market—probably because the openness lets programmers the world over obtain, critique, use, and build upon the source code without licensing fees. Linux, a UNIX-like operating system, is the best known success. But computer scientists at the University of California, Berkeley began the tradition of software sharing in the mid 1970s with BSD UNIX and distributed the major internet network protocols as source code without a fee. Medical informatics has its own history of OpenSource distribution: Massachusetts General’s COSTAR and the Veterans Administration’s VISTA software have been distributed as source code at no cost for decades. Bioinformatics, our sister field, has embraced the OpenSource movement and developed rich libraries of open-source software. OpenSource has now gained a tiny foothold in health care (OSCAR GEHR, OpenEMed). Medical informatics researchers and funding agencies should support and nurture this movement. In a world where open-source modules were integrated into operational health care systems, informatics researchers would have real world niches into which they could engraft and test
their software inventions. This could produce a burst of innovation that would help solve the many problems of the health care system. We at the Regenstrief Institute are doing our part by moving all of our development to the open-source model.


• Abstract

Widespread adoption of open-source software for network infrastructure, web servers, code development, and operating systems leads one to ask how far it can go. Will 'opensource' spread broadly, or will it be restricted to niches frequented by hopeful hobbyists and midnight hackers? Here we identify reasons for the success of open-source software and predict how consumers in drug discovery will benefit from new open-source products that address their needs with increased flexibility and in ways complementary to proprietary options.


• Abstract

Mothur aims to be a comprehensive software package that allows users to use a single piece of software to analyze community sequence data. It builds upon previous tools to provide a flexible and powerful software package for analyzing sequencing data. As a case study, we used mothur to trim, screen, and align sequences; calculate distances; assign sequences to operational taxonomic units; and describe the α and β diversity of eight marine samples previously characterized by pyrosequencing of 16S rRNA gene fragments. This analysis of more than 222,000 sequences was completed in less than 2 h with a laptop computer.


Abstract The Bioconductor project is an initiative for the collaborative creation of extensible software for computational biology and bioinformatics. The goals of the project include: fostering collaborative development and widespread use of innovative software, reducing barriers to entry into interdisciplinary scientific research, and promoting the achievement of remote reproducibility of research results. We describe details of our aims and methods, identify current challenges, compare Bioconductor to other open bioinformatics projects, and provide working examples.


• Abstract

This paper showcases that the current models of encouraging pharmaceuticals to research and develop drugs curing tropical diseases that affects poor people aren't working. These methods are 1) asking governments and NGOs to subsidize drugs rates for developed countries, and 2) to create non-profit venture capital firms. It proposes an open-source model for developing these drugs through a website (www.tropicaldisease.org). It describes how scientists could use chat pages and shared databases to make discoveries.

• The payment of scientists working on this database would not be monetary, but scientists would gain stature and enhance their reputation, as is similar to the motivations of the hacker community. The drugs would not be patented in order to ensure that retail costs remained low. Companies and universities would allow their workers to volunteer, and would even donate databases and resources because the value of their IP lies in North American and European medicines.

**Abstract**

This paper describes an open-source system for analyzing, storing, and validating proteomics information derived from tandem mass spectrometry. It is based on a combination of data analysis servers, a user interface, and a relational database. The database was designed to store the minimum amount of information necessary to search and retrieve data obtained from the publicly available data analysis servers. Collectively, this system was referred to as the Global Proteome Machine (GPM). The components of the system have been made available as open source development projects. A publicly available system has been established, comprised of a group of data analysis servers and one main database server.


**Background**

Biological pathways, including metabolic pathways, protein interaction networks, signal transduction pathways, and gene regulatory networks, are currently represented in over 220 diverse databases. These data are crucial for the study of specific biological processes, including human diseases. Standard exchange formats for pathway information, such as BioPAX, CellML, SBML and PSI-MI, enable convenient collection of this data for biological research, but mechanisms for common storage and communication are required.

**Results**

We have developed cPath, an open source database and web application for collecting, storing, and querying biological pathway data. cPath makes it easy to aggregate custom pathway data sets available in standard exchange formats from multiple databases, present pathway data to biologists via a customizable web interface, and export pathway data via a web service to third-party software, such as Cytoscape, for visualization and analysis. cPath is software only, and does not include new pathway information. Key features include: a built-in identifier mapping service for linking identical interactors and linking to external resources; built-in support for PSI-MI and BioPAX standard pathway exchange formats; a web service interface for searching and retrieving pathway data sets; and thorough documentation. The cPath software is freely available under the LGPL open source license for academic and commercial use.

**Conclusion**

cPath is a robust, scalable, modular, professional-grade software platform for collecting, storing, and querying biological pathways. It can serve as the core data handling component in information systems for pathway visualization, analysis and modeling.


**Summary**

ABNER (A Biomedical Named Entity Recognizer) is an open source software tool for molecular biology text mining. At its core is a machine learning system using conditional random fields with a variety of orthographic and contextual features. The latest version is 1.5, which has an intuitive graphical interface and includes two modules for tagging entities (e.g. protein and cell line) trained on standard corpora, for which performance is roughly state of the art. It also includes a Java application programming interface allowing users to incorporate ABNER into their own systems and train models on new corpora.

- **Abstract**

Camino is an open-source, object-oriented software package for processing diffusion MRI data. Camino implements a data processing pipeline, which allows for easy scripting and flexible integration with other software. This paper summarises the features of Camino at each stage of the pipeline from the raw data to the statistics used by clinicians and researchers. The paper also discusses the role of Camino in the paper "An Automated Approach to Connectivity-based Partitioning of Brain Structures".


- **Abstract**

We present the second and improved release of the TOUCAN workbench for cis-regulatory sequence analysis. TOUCAN implements and integrates fast state-of-the-art methods and strategies in gene regulation bioinformatics, including algorithms for comparative genomics and for the detection of cis-regulatory modules. This second release of TOUCAN has become open source and thereby carries the potential to evolve rapidly. The main goal of TOUCAN is to allow a user to come to testable hypotheses regarding the regulation of a gene or of a set of co-regulated genes. TOUCAN can be launched from this location: http://web.archive.org/web/20050914053334/http://www.esat.kuleuven.ac.be:80/~saerts/software/toucan.php.


- **Abstract**

IntAct is an open source database and software suite for modeling, storing and analyzing molecular interaction data. The data available in the database originates entirely from published literature and is manually annotated by expert biologists to a high level of detail, including experimental methods, conditions and interacting domains. The database features over 126,000 binary interactions extracted from over 2100 scientific publications and makes extensive use of controlled vocabularies. The web site provides tools allowing users to search, visualize and download data from the repository. IntAct supports and encourages local installations as well as direct data submission and curation collaborations. IntAct source code and data are freely available from http://www.ebi.ac.uk/intact.


- **Abstract**

The Chemistry Development Kit (CDK) provides methods for common tasks in molecular informatics, including 2D and 3D rendering of chemical structures, I/O routines, SMILES parsing and generation, ring searches, isomorphism checking, structure diagram generation, etc. Implemented in Java, it is used both for server-side computational services, possibly equipped with a web interface, as well as for applications and client-side applets. This article introduces the CDK's new QSAR capabilities and the recently introduced interface to statistical software.

- Abstract

PSI3 is a program system and development platform for ab initio molecular electronic structure computations. The package includes mature programming interfaces for parsing user input, accessing commonly used data such as basis-set information or molecular orbital coefficients, and retrieving and storing binary data (with no software limitations on file sizes or file-system-sizes), especially multi-index quantities such as electron repulsion integrals. This platform is useful for the rapid implementation of both standard quantum chemical methods, as well as the development of new models. Features that have already been implemented include Hartree-Fock, multiconfigurational self-consistent-field, second-order Møller-Plesset perturbation theory, coupled cluster, and configuration interaction wave functions. Distinctive capabilities include the ability to employ Gaussian basis functions with arbitrary angular momentum levels; linear R12 second-order perturbation theory; coupled cluster frequency-dependent response properties, including dipole polarizabilities and optical rotation; and diagonal Born-Oppenheimer corrections with correlated wave functions. This article describes the programming infrastructure and main features of the package. PSI3 is available free of charge through the open-source, GNU General Public License. © 2007 Wiley Periodicals, Inc. J Comput Chem, 2007


- Abstract

We have developed a freely available, open-source software system (OpenSim) that lets users develop models of musculoskeletal structures and create dynamic simulations of a wide variety of movements. We are using this system to simulate the dynamics of individuals with pathological gait and to explore the biomechanical effects of treatments. Dynamic simulations of movement allow one to study neuromuscular coordination, analyze athletic performance, and estimate internal loading of the musculoskeletal system. Simulations can also be used to identify the sources of pathological movement and establish a scientific basis for treatment planning. OpenSim provides a platform on which the biomechanics community can build a library of simulations that can be exchanged, tested, analyzed, and improved through a multi-institutional collaboration. Developing software that enables a concerted effort from many investigators poses technical and sociological challenges. Meeting those challenges will accelerate the discovery of principles that govern movement control and improve treatments for individuals with movement pathologies.


- Abstract

The FLR framework (Fisheries Library for R) is a development effort directed towards the evaluation of fisheries management strategies. The overall goal is to develop a common framework to facilitate collaboration within and across disciplines (e.g. biological, ecological, statistical, mathematical, economic, and social) and, in particular, to ensure that new modelling methods and software are more easily validated and evaluated, as well as becoming widely available once developed. Specifically, the framework details how to implement and link a variety of fishery, biological, and economic software packages so that alternative management strategies and procedures can be evaluated for their robustness to uncertainty before implementation. The design of the framework, including the adoption of object-orientated programming, its feasibility to be extended to new
processes, and its application to new management approaches (e.g. ecosystem affects of fishing), is discussed. The importance of open source for promoting transparency and allowing technology transfer between disciplines and researchers is stressed.


Abstract

Background There is a need for software applications that provide users with a complete and extensible toolkit for chemo- and bioinformatics accessible from a single workbench. Commercial packages are expensive and closed source, hence they do not allow end users to modify algorithms and add custom functionality. Existing open source projects are more focused on providing a framework for integrating existing, separately installed bioinformatics packages, rather than providing user-friendly interfaces. No open source chemoinformatics workbench has previously been published, and no sucessful attempts have been made to integrate chemo- and bioinformatics into a single framework.

Results Bioclipse is an advanced workbench for resources in chemo- and bioinformatics, such as molecules, proteins, sequences, spectra, and scripts. It provides 2D-editing, 3D-visualization, file format conversion, calculation of chemical properties, and much more; all fully integrated into a user-friendly desktop application. Editing supports standard functions such as cut and paste, drag and drop, and undo/redo. Bioclipse is written in Java and based on the Eclipse Rich Client Platform with a state-of-the-art plugin architecture. This gives Bioclipse an advantage over other systems as it can easily be extended with functionality in any desired direction.

Conclusion Bioclipse is a powerful workbench for bio- and chemoinformatics as well as an advanced integration platform. The rich functionality, intuitive user interface, and powerful plugin architecture make Bioclipse the most advanced and user-friendly open source workbench for chemo- and bioinformatics. Bioclipse is released under Eclipse Public License (EPL), an open source license which sets no constraints on external plugin licensing; it is totally open for both open source plugins as well as commercial ones. Bioclipse is freely available at http://www.bioclipse.net


Summary

Custom-made zinc-finger nucleases (ZFNs) can induce targeted genome modifications with high efficiency in cell types including Drosophila, C. elegans, plants, and humans. A bottleneck in the application of ZFN technology has been the generation of highly specific engineered zinc-finger arrays. Here we describe OPEN (Oligomerized Pool ENgineering), a rapid, publicly available strategy for constructing multifinger arrays, which we show is more effective than the previously published modular assembly method. We used OPEN to construct 37 highly active ZFN pairs which induced targeted alterations with high efficiencies (1%–50%) at 11 different target sites located within three endogenous human genes (VEGF-A, HoxB13, and CFTR), an endogenous plant gene (tobacco SuRA), and a chromosomally integrated EGFP reporter gene. In summary, OPEN provides an "open-source" method for rapidly engineering highly active zinc-finger arrays, thereby enabling broader practice, development, and application of ZFN technology for biological research and gene therapy.


Summary

BioJava is a mature open-source project that provides a framework for processing of biological data. BioJava contains powerful analysis and statistical routines, tools for parsing common file formats and packages for manipulating sequences and 3D
structures. It enables rapid bioinformatics application development in the Java programming language. BioJava is an open-source project distributed under the Lesser GPL (LGPL). BioJava can be downloaded from the BioJava website (http://www.biojava.org).


- Abstract

Background Mass spectrometry is an essential analytical technique for high-throughput analysis in proteomics and metabolomics. The development of new separation techniques, precise mass analyzers and experimental protocols is a very active field of research. This leads to more complex experimental setups yielding ever increasing amounts of data. Consequently, analysis of the data is currently often the bottleneck for experimental studies. Although software tools for many data analysis tasks are available today, they are often hard to combine with each other or not flexible enough to allow for rapid prototyping of a new analysis workflow.

Results We present OpenMS, a software framework for rapid application development in mass spectrometry. OpenMS has been designed to be portable, easy-to-use and robust while offering a rich functionality ranging from basic data structures to sophisticated algorithms for data analysis. This has already been demonstrated in several studies.

Conclusion OpenMS is available under the Lesser GNU Public License (LGPL) from the project website at http://www.openms.de.


- Abstract

The low number of novel therapeutics approved by the US FDA in recent years continues to cause great concern about productivity and declining innovation. Can open-source drug research and development, using principles pioneered by the highly successful open-source software movement, help revive the industry?

Brendan MacLean et al., (2010), Skyline: an open source document editor for creating and analyzing targeted proteomics experiments, Bioinformatics Volume 26, Issue 7, pp. 966-968

Summary: Skyline is a Windows client application for targeted proteomics method creation and quantitative data analysis. It is open source and freely available for academic and commercial use. The Skyline user interface simplifies the development of mass spectrometer methods and the analysis of data from targeted proteomics experiments performed using selected reaction monitoring (SRM). Skyline supports using and creating MS/MS spectral libraries from a wide variety of sources to choose SRM filters and verify results based on previously observed ion trap data. Skyline exports transition lists to and imports the native output files from Agilent, Applied Biosystems, Thermo Fisher Scientific and Waters triple quadrupole instruments, seamlessly connecting mass spectrometer output back to the experimental design document. The fast and compact Skyline file format is easily shared, even for experiments requiring many sample injections. A rich array of graphs displays results and provides powerful tools for inspecting data integrity as data are acquired, helping instrument operators to identify problems early. The Skyline dynamic report designer exports tabular data from the Skyline document model for in-depth analysis with common statistical tools.

Availability: Single-click, self-updating web installation is available at http://proteome.gs.washington.edu/software/skyline. This web site also provides access to instructional videos, a support board, an issues list and a link to the source code project.

- Abstract

The latest release of NWChem delivers an open-source computational chemistry package with extensive capabilities for large scale simulations of chemical and biological systems. Utilizing a common computational framework, diverse theoretical descriptions can be used to provide the best solution for a given scientific problem. Scalable parallel implementations and modular software design enable efficient utilization of current computational architectures. This paper provides an overview of NWChem focusing primarily on the core theoretical modules provided by the code and their parallel performance.


- Background

A chemist — and social entrepreneur — in Australia is launching an open-source research project to develop a more potent form of a front-line drug against the debilitating neglected tropical disease schistosomiasis.

Matthew Todd of the University of Sydney hopes to persuade research chemists across the world to share laboratory time and expertise in a collaborative effort to find a cheap and efficient synthesis of the drug praziquantel. All results will be published in almost real time on the project’s website — free of intellectual property restrictions — and later in journals, with substantial contributors becoming authors on any resulting papers. "My funded project is intended to be the kernel, to which anyone can add," Todd says. He hopes that the project will become a successful example of open-source science, and open-source 'wet lab' chemistry in particular, a concept that has been slow to take off.