

Draft paper, February 2004

Argos, a replicable genome information system for FlyBase, euGenes and other databases.

Don Gilbert*, Joshua Goodman, Paul Poole, Hardik Sheth, Nihar Sheth, Vasanth Singan, Victor Strelets

Genome Informatics Lab, Department of Biology, Indiana University, Bloomington, IN 47405
USA

* To whom correspondence should be addressed

Running head: Argos replicable genome information system

ABSTRACT

Motivation: Genome data access is a growing need in many sectors of biosciences research and development. This includes issues of common standards, data federation, distribution and automated access to large and changing volumes of complex data. Monolithic database systems can impede use of new data and methods, especially for small informatics groups.

Results: Argos is a new genome information system structure for organizing a common tool set with Internet access, automatic replication, installation and updates of genome databases. It is component-based and loosely structured to allow adoption of new techniques readily. Example systems include FlyBase and euGenes genome databases.

Availability: Argos common components and implemented databases are available from <http://eugen.es.org/argos/>, <http://flybase.net/argos/>, <http://www.gmod.org/argos/>. Components are distributed with open source licenses governed by their originators.

Contact: argos@flybase.net

INTRODUCTION

Genome sequences, annotations and research literature are becoming a new reference collection that underpins research in genomics, gene expression, protein function and structure, cell and organismal biology. Needs for better access to these data are growing as bioscientists switch to data-mining to study thousands of genes rather than one. Web page scraping and bulk files, the current common practices for this, miss a middle ground for computable data access. Needed are programmable methods for Internet search and retrieval of genome objects distributed among many source. Data distribution alone is insufficient where the extraction of knowledge from it requires software designed to select and produce genome information objects, to handle clients' complex queries of this data, and provide reporting extracts of information relevant to the questions. Simple, flexible client program models should be supported in future genome webs or grids. These methods need to be efficient for high volumes where millions of objects and multi-gigabyte volumes are of interest.

Three genome data access building blocks are under construction to address in part these high-volume needs. *Argos* is a framework for distributing genome data systems built on common components. *LuceGene* provides rapid, Google-like searches with data and document retrieval to integrate a wide range of genome information. A *Genome Directory System* includes Web Services, Grid Services, and LDAP Internet standard interfaces to support genome data mining. We report here on Argos and its deployment in FlyBase and euGenes genome databases.

Following its first successful decade as *Drosophila* genome database and Internet information service, FlyBase [ref] is undergoing an evolution to the next generation genome information system. FlyBase Next Generation (FlyBase-NG) includes more changes "under the hood" than you see on its public web pages. These are part of an evolution to a next generation of genome databases and information systems. As FlyBase moves into its second decade, we want to ensure that the best new genome database methods from the collective wisdom of bioinformatics are added without major re-engineering costs, while retaining the best parts that have evolved over a decade of addressing needs and requirements of *Drosophila*, genomics and biosciences research communities.

FlyBase-NG can be copied, run on workstations, laptop computers or informatics center clusters, and is designed for automatic updates to keep it current. It works well on Mac OS X, Linux and Solaris Unix systems. The underpinning that makes this possible is Argos, a replicable

genome information system. Its uses for genome data access include load distribution, providing world-wide mirrors of the database system, as well as for institution/company systems with data integration projects and data mining. The euGenes multi-eukaryote genome database [ref] and a new Daphnia genome database also use this Argos infrastructure. Argos is being developed for general use as part of the Generic Model Organism Database (GMOD) project.

SYSTEM AND METHODS

A Replicable Genome infOrMation System. Argos is component-oriented, built from many open-source packages common to genome databases, web and information services, with new parts plugged in as needed to provide new services. Common parts include BLAST, Gbrowse genome maps, Apache and Tomcat Web servers, PostgreSQL and MySQL databases and informatics middleware. It includes libraries and support for Java and Perl components and middleware. It is designed for automatic distribution and updates to any Unix computer. Automation extends beyond providing software and data downloads: components are kept operating and up-to-date without human intervention. Customers can choose which packages to replicate, and these packages may be served from different source sites. Replication includes scripts, configurations, data, and Unix binaries for all needed programs except Perl, Java and Rsync, currently used as the primary distribution tool. FlyBase-NG and euGenes specific parts are kept separate from common genome system parts.

This system provides a common collection of genome informatics and information technology components and tools that are pre-packaged, installed and tested for a range of small and large organism genome projects. Sharing the benefits of “best of breed” genome tools is a goal of this project. Common parts are tested and maintained by a group of developers. A project’s needs specify the tool set, and each project specifies its own look and feel (user interface), web pages, contents, functions. This is in contrast to other genome data systems that are built with a single set of functions designed around comprehensive, complex databases.

Minimal information technology expertise is expected of those who copy an Argos genome database, with no requirement to compile software or manage details of the system operation. The system works well on common Unix, Linux and Mac OS X workstations and laptop computers. ‘Live’ updates with Rsync keeps servers current, for local clusters to support high-volume traffic. Options for data access security with password or hostname protected sections, collaborative editing and updates via WebDAV protocol are included.

Base packages. Argos uses a package structure for its components, with configurations that identify the reference Internet source for each package. Base packages are ones that provide a fully usable system, and are listed in Table 1. The base packages all require additional common packages.

Table 1 Argos base packages

Package	Description	Disk use	Requires
argos-root	Argos root server	50 Mb	argos-common
argos-sources	Argos software sources	100 Kb	none
centaurbase	Test server	50 Mb	argos-common centaurbase-data
eugenenes	multi-eukaroyte database	10 Gb	argos-common eugenenes-data
flybase	Drosophila genome database	4 Gb	argos-common flybase-data flybase-blast

Common components. A design goal in Argos is to use available open-source software with minimal changes, to facilitate updating to the best tool set available. To this end, most software components are installed as per their standard instructions, with adjustments made to configuration and file system directories using symbolic links and local configuration files. Operating system specific binaries are separately located and linked back as needed, for the servers, Perl and Java native library portions. Table 2 shows the range of these common components. The only common components that have required source code revisions are certain Apache web server additional modules. Otherwise, common components are matched to genome server needs with Java and Perl middleware additions, with configuration files and file system structuring.

Table 2 Argos common components

Section	Components
Java	Chado database tools, GnoSeq genome sequence reports, Lucene text search, Ant build system, database interfaces, XML tools, Tomcat web server, Axis Web Services, and others
Perl	BioPerl, GBrowse, Chado database tools, database interfaces, HTML and Web tools, XML tools, and others
Servers	BLAST (NCBI), Apache web server, PostgreSQL, MySQL and BerkeleyDB databases
Systems	Compiled portions for supported operating systems
Install	Argos instructions, installation scripts and usage
ROOT	Common directory, configurations and web server

Requirements and Installation. A design goal in Argos is to keep prerequisites to a minimum, by incorporating needed components within the distribution system. Although this duplicates components that may be already installed in a computer, it is necessary to ensure tested compatibility among components. The necessary software components are compiled and tested for Apple Mac OS X (v10.2 tested), Intel Linux (Redhat v8, 9 tested), and Sun Solaris (v8 & v9 tested). Prerequisite packages, commonly preinstalled on these operating systems, are Perl v5.6 or later, Java v1.3 or later, rsync v2.5 or later [ref]. A minimal number of steps are required to install Argos: download a Perl installation script; bootstrap the install process; configure for local system; download and install full genome data and software; check and run web servers; update as desired (including automatic daily updates).

The configuration process has default options that can be overridden: as the web servers and database software in Argos often need to co-exist with system default installations, all of the file

paths and port settings chosen as defaults are different from standard settings. There is a web interface to help choose best configuration for a system, which writes the desired local configuration file. Generally each organism project has its own web port numbers, and acts as a virtual host in common root Argos web server with separate project configurations. The separate projects are accessible from the root web page.

Centaurobase, an Argos playpen. The Centaurobase play database provides mythical genome data for species of *Centaur* (the man-horse), to show the features of Argos with minimal data, and to form a starter kit for new genome databases. It currently uses LuceGene for search and retrieval of Medline and other XML data, GenBank sequences, HTML documents. Sample sequences can be BLASTed. It has its own web look and feel, with WebDAV editing and uploading of new documents.

Chado genome database. FlyBase started data management with, and took part of its name from, the SyBase commercial relational database. For this next generation, the FlyBase project wanted a reusable database shared with other genome projects. This new database is named Chado (after "the Way of Tea" tea ceremony). It includes a new schema for structuring genome information, works with the free PostgreSQL database package, and includes a Chado XML exchange format and tools for this. Significantly, a larger group of bioinformaticians is sharing development and use of these parts, in the GMOD group (<http://www.GMOD.org/>). Initial work with Chado has been a migration of genome annotations from its previous database. At the close of December 2003, we provided the first public release of these Chado annotation data for the *Drosophila* genome. Over the coming year, more genome data will move into Chado, with more options added for public database access.

Power to the data miners. FlyBase's main web server has seen 60% growth in 2003, peaking over 100,000 hits per day, accelerating over prior 20% growth/year. This growth includes increases in commercial use (from 12% to 24%), robots, data miners and other high-volume users. It reflects a growing importance of genome information in biotechnology research and development. While not as busy as the NASA Mars web sites on landing day, FlyBase-NG can use similar methods of distributing use among servers around the world to meet demands for genome data.

The Argos underpinning for the new FlyBase server provides a general method for making it robust to high volume usage: compute intensive calls are passed to any number of cloned servers, transparent to clients who know and see only one main URL. This is now done with computed data reports and BLAST sequence searches, and a note *Run on computer xxx* now graces the bottom of many FlyBase web pages. The euGenes service can now offer BLAST searches of several eukaryote genomes with the same ability to use several computers in the background. Load balancing, cluster and grid computing software all offer usable ways to implement distributed compute services for web-based services. We chose a web load distribution system (`mod_backhand` [ref]) for proxying client requests to a cluster of cloned servers, returning results as if run on the primary server.

One obvious benefit is that computed web pages appear faster. Customers now see gene reports about 5 times faster (1.2 seconds now versus 6.5 seconds for the previous FlyBase generation). As usage increases these will be kept running fast by adding more clone servers.

A small percentage of these clients are over-eager robots and data-miners, and misbehaving web browsers (commonly Microsoft Explorer web-archive spidering and other malfunctions). When a single client misbehaves with compute intensive database functions, it can drive an unprotected server to its knees. Components of the Argos system include protections against these misbehaving clients (`mod_throttle` [ref] and others), to keep their use in bounds. Combined with

distribution of compute intensive calls for BLAST and genome reports, these methods keep the primary server responsive under a wide range of traffic.

Reviews and Previews. Some people want old data, some prefer the newest data. It is common practice in sciences and industry to review old experiments. The clonability of Argos systems, including all software and data, makes it easier to create and maintain frozen copies, for companies and others with needs for establishing provenance of data. For those with a desire for the newest, we clone a database server, adding pre-release subsets. Database developers and students clone and test on their own copy, folding in updates to the primary replication server. Our current paradigm is to separate replication servers from public access servers. A replication server can then be updated with standard settings, newer and older data sets, and software still in test modes, without affecting quality-checked public access servers.

What remains to be done. The ability to add new genome database tools to this framework is a major reason for its design, and that will continue. Improvements are needed to streamline the configuring, installation and updating of packages, with hopes to reach a 'one-click' installation process. Data mining support remains a basic design emphasis, with plans to integrate common search and retrieval tools with Web Services and Grid Services interfaces, as well as the simpler tabular bulk data retrieval options that many scientists find practical. We plan to add more GMOD tools, and mod-perl for Apache web server. As new genome data functions are developed for specific services, these can be generalized for common use in other services. File synchronization methods are useless for PostgreSQL database updates; we will add methods based on transactions such as the erServer database replication tool.

DISCUSSION

Argos joins EnsEMBL in providing a reusable system for multi-organism genome databases and web services that is open-source and freely copyable. It is not as comprehensive in scope as EnsEMBL, and it focuses on other aspects: being automatically replicable with all necessary components pre-built and tested; being service-centric and componentized, rather than a tightly co-developed database system. Each organism genome functions and contents take precedence in Argos, using common genome system components as desired.

Argos uses a replication system that is focused on the joint needs for volume data updates and software configured to use this data. This centers on remote file synchronization provided by Rsync [ref], as a standard program that has been used widely for automation of large volume file system updates. A number of software replication methods are available, notably bundled with operating systems such as the Redhat Package Manager [ref] and related Linux and Unix package updaters, and the Mac OS X system updater. These software replication systems are more mature than Argos's but don't meet needs for volume data updates. Grid systems includes replication tools such as Globus Grid package management [ref], and PacMan [ref] also have useful features, including multi-operating system targets, and also offers binary program replication, installing on remote systems, and system configuration options. However at this point, Grid tools for data replication are immature and less functional than existing FTP, Rsync or Web file replication and mirroring.

Improving genome data distribution. Argos goes beyond current genome software and database systems by doing away with much of the *install-make-update* cycle that an engineer has to do manually, and a non-engineer finds daunting. Current genome informatics practices do not well address question of timely, automated data and software updates that keep genome information current. The Argos design will enable bioscientists with limited computing expertise

to take advantage of local copies of useful data, and permit further steps to integrate laboratory data with reference genome information.

Argos follows in the footsteps of what we have been using successfully in FlyBase since 1996, to avoid remote site administrators having to spend management time updating the system. Some design and attention is required setting up Argos reference services, so that they are self-contained and path-independent. The needed Java and Perl modules are included, and Perl native library modules are built in a path-independent fashion, contrary to default Perl installation methods.

Genome data federation, searches and reporting. xxx.

Genome data mining, web services, and data grids. xxx.

Evolution of genome information systems. xxx.

[? Drop this] Public data access to FlyBase [ref] database have used an information system approach to document and data search and retrieval since its inception in 1993. Practical concerns of providing efficient access to a diverse set of genome object-oriented data, using a small bioinformatics staff, have driven and validated this design, originally using WAIS [ref], then SRS [ref], with new work focusing on Lucene. Compared with the costs for designing efficient object-oriented use of complex data in relational database structures, this approach has allowed FlyBase to provide good integrated public access to a diverse range of genome data, including literature, sequence, expression data, anatomical and functional data. The euGenes genome summary database is built from many of the same component parts used in FlyBase, with a goal of extending these methods to any group of organisms' genome information. Bioinformatics centers are welcome to replicate FlyBase and euGenes for local and regional users.

ACKNOWLEDGEMENTS

[FlyBase project; NIH GMOD project; NSF – IUBio]. This work is supported in part by NIH grant 1R01HG002733-01 and NSF grant 0090782 to D. Gilbert.

REFERENCES

- Bio-Mirror project, 1997. A public service for distribution and access to biosequence and bioinformatics data. URL: <http://www.bio-mirror.net/>
- EBI. The European Bioinformatics Institute. URL: <http://www.ebi.ac.uk/>
- Ensembl. The Ensembl project. URL: <http://www.ensembl.org/>
- FlyBase Consortium, 1999. The FlyBase database of the Drosophila Genome Projects and community literature. *Nucleic Acids Res.*, 27: 85-88.
- Gilbert, D.G., 2002. euGenes, a eukaryote organism genome information service. *Nucleic Acids Res.*, 30, 145-148 URL: <http://iubio.bio.indiana.edu/eugenesis/>
- Gilbert, D.G. (2002) Directories of Bio-data. <http://iubio.bio.indiana.edu/biogrid/directories/>
- GO. The Gene Ontology Consortium. URL: <http://www.geneontology.org/>
- IUBio. The IUBio Archive. URL: <http://iubio.bio.indiana.edu/>
- McLoughlin, L. (1998) Mirror perl package. URL: <http://sunsite.org.uk/packages/mirror/>
- NCBI. The National Center for Biotechnology Information. URL: <http://www.ncbi.nih.gov/>
- OSGA-DAI (2002) Open Grid Services Architecture: Data Access and Integration project. URL: <http://www.ogsadai.org.uk/>
- Tridgell, A. (2002) rsync, remote file synchronization system. URL: <http://rsync.samba.org/>
- XDDBJ. XML Central of DDBJ. URL: <http://xml.nig.ac.jp/>
- XEMBL. EMBL Nucleotide Sequence data in XML. URL: <http://www.ebi.ac.uk/xembl/>
- Zdobnov, E.M., Lopez,R., Apweiler,R., Etzold, T. (2002) The EBI SRS server – recent developments. *Bioinformatics*, 18, 368-373.
- ////
- Stein, L. 2003. Integrating Biological Databases. *Nature Reviews Genetics*. 4: 337-345. doi:10.1038/nrg1065
- Clamp, M., et al. (2003). Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Research*, 31: 38–42. doi: 10.1093/nar/gkg083
- Hubbard, T. et al. (2002). The Ensembl genome database project. *Nucleic Acids Research*, 30: 38-41
- EnsMart (URL: <http://www.ensembl.org/EnsMart/>)
- Valencia, A. (2002). Search and retrieve: Large-scale data generation is becoming increasingly important in biological research. But how good are the tools to make sense of the data? *EMBO Reports* 3(5): 396 - 400. doi: 10.1093/embo-reports/kvf104
- Ware, D. H. et al. (2002). Gramene, a Tool for Grass Genomics. *Plant Physiology*, 130: 1606 - 1613. doi: 10.1104/pp.015248

Pennacchio, L.A. and Rubin, E.M. (2003). Comparative genomic tools and databases: providing insights into the human genome. *J. Clin. Invest.* 111:1099–1106. doi: 10.1172/JCI200317842.

Add: caCORE cancer informatics report, *Bioinformatics*, 19(18): 2404-2412. doi: Dec 2003.