Nucleic Acids Research, 2003, Vol. 31, No. 1 147–151
DOI: 10.1093/nar/gkg125

ASAP, a systematic annotation package for community analysis of genomes

Jeremy D. Glasner[1,2], Paul Liss[1,2], Guy Plunkett III[2], Aaron Darling[1], Tejasvini Prasad[1], Michael Rusch[1], Alexis Byrnes[1], Michael Gilson[1], Bryan Biehl[1], Frederick R. Blattner[2] and Nicole T. Perna[1,*]

[1]Animal Health and Biomedical Sciences, University of Wisconsin-Madison, 656 Linden Dr Madison, WI 53706-1581, USA and [2]Genetics Department, University of Wisconsin-Madison, 445 Henry Mall, Madison, WI 53706, USA

Received August 15, 2002; Revised and Accepted October 23, 2002

ABSTRACT

ASAP (a systematic annotation package for community analysis of genomes) is a relational database and web interface developed to store, update and distribute genome sequence data and functional characterization (https://asap.ahabs.wisc.edu/annotation/php/ASAP1.htm). ASAP facilitates ongoing community annotation of genomes and tracking of information as genome projects move from preliminary data collection through post-sequencing functional analysis. The ASAP database includes multiple genome sequences at various stages of analysis, corresponding experimental data and access to collections of related genome resources. ASAP supports three levels of users: public viewers, annotators and curators. Public viewers can currently browse updated annotation information for *Escherichia coli* K-12 strain MG1655, genome-wide transcript profiles from more than 50 microarray experiments and an extensive collection of mutant strains and associated phenotypic data. Annotators worldwide are currently using ASAP to participate in a community annotation project for the *Erwinia chrysanthemi* strain 3937 genome. Curation of the *E. chrysanthemi* genome annotation as well as those of additional published enterobacterial genomes is underway and will be publicly accessible in the near future.

INTRODUCTION

ASAP is designed to organize the data associated with a genome from the early stages of sequence annotation through genetic and biochemical characterization, providing a vehicle for ongoing updates of the annotation and a repository for genome-scale experimental data. Development was motivated by the need to more directly involve a greater community of researchers, with their collective expertise, in keeping the genome annotation current and to provide a synergistic link between up-to-date annotation and functional genomic data.

The current taxonomic focus of ASAP is enterobacterial genomes. *Escherichia coli* K-12 is the best studied member of this family and the annotation and experimental data from the MG1655 strain serves as an excellent example of the utility of the ASAP system. The annotated genome sequence was published in 1997 (1) and the GenBank deposition updated in 1998. Since then many additional gene functions were identified and genome-scale experiments performed. We have used ASAP to update the MG1655 annotation and distribute genome-wide transcript profiles, a mutant strain collection, and phenotypic assays, linked to annotated genome features.

The considerable effort involved with a one-time revision of the MG1655 annotation highlights the need for continuous updates with information derived from a large number of independent researchers, sometimes relevant to a small number of genome features, sometimes affecting the annotation of many genes. With this in mind ASAP was developed to facilitate distributed, assisted annotation of genomes via a web interface. The *Erwinia chrysanthemi* strain 3937 genome project, with more than 30 researchers annotating worldwide, served as our model for community annotation. The annotators use a web browser to add information about genome features using free-form descriptions as well as controlled vocabularies. Annotations are viewable by other users in the *Erwinia* group, and will become public when the genome is released. A single genome curator reviews all annotations and can prevent annotations containing errors from the public release. However, new annotations will be immediately available publicly with a clear indication of the supporting evidence, annotation author and uncurated status. A complete history of each annotation is maintained for future evaluation. We are annotating the *Erwinia* genome prior to completion of the sequence and describe use of ASAP to assist the migration of annotations between different versions of the sequence.

*To whom correspondence should be addressed. Tel: +1 6082620728; Fax: +1 6082627420; Email: perna@ahabs.wisc.edu

KEY ASPECTS OF ASAP

- Each feature qualifier (gene name, product, function, etc.) entered includes a reference to evidence used to support that annotation and is tagged with the annotator's identity and contact information.
- Annotation is a two-tiered system with a group of annotators contributing data but a single curator systematically reviewing all annotations for consistency.
- Annotation data is decoupled from sequence coordinates providing a mechanism for persistence of annotations even when the sequence itself changes.
- Annotation vocabulary is constrained wherever possible to ensure that features with similar functions are annotated the same way.
- The database is flexible and permits storage and retrieval of many kinds of genomic data.
- Guests, annotators and curators can access ASAP from most typical computer platforms using any standard internet browser and the entire system is portable to other servers.

STRUCTURE OF ASAP

The ASAP system includes a MySQL database designed to organize predicted features of a genome, evidence available as a basis for further characterization of that feature, the annotators' evaluation of that evidence and an expert curator's assessment of each contributed annotation as well as experimental data associated with annotated genome features.

The database design underlying ASAP is generalized to accommodate annotation of different feature types predicted from multiple sources. The schema supports integration of any type of International Nucleotide Sequence Databank (INSD: DDBJ, EMBL and GenBank) defined feature predicted using any available algorithm. Text parsers were developed to process the output of two popular gene prediction programs: GLIMMER2 for protein-coding genes (2), and tRNAscan-SE for tRNAs (3) into ASAP tables. Other types of predicted features can be entered manually or entered in bulk from a tab-delimited spreadsheet. The MG1655 data described below was entered *en masse* from pre-existing annotations and updated with information from a variety of sources.

Once sequence features are entered, ASAP includes accessory stand-alone programs to extract DNA and protein sequences, as multiple or single files as input for routine analyses, like sequence similarity searches. BLAST (4) searches are the single most common approach to identifying putative gene functions, and ASAP includes support for processing and displaying BLAST output to users as they view or annotate features. This approach is general and allows incorporation of additional types of evidence that might assist in characterization of a feature, such as protein motif searches, structural characteristics, and gene expression data. Currently, BLAST search results against the NCBI "nr" nucleotide database, Genpept, an MG1655-specific protein database, and Pfam (5) protein family search results are available for *Erwinia* annotators. These searches were conducted locally. In addition, interactive links from each annotation page will launch BLAST at NCBI against the most current nonredundant nucleotide or protein databases.

Annotators and annotation viewers are presented with basic information about each feature, such as its location in the genome and type (protein-coding gene, rRNA, tRNA, etc.) and any available evidence as described above. Any annotations that have been entered for this feature are also displayed, subdivided into appropriate categories, like gene name, product or function and coupled with supporting evidence and author's identity. Annotators have additional functionality through the same interface, namely the ability to enter new data using INSD qualifiers. Annotators can also characterize the predicted function, location and role of a feature according to MultiFun, a hierarchical scheme designed to fully describe the predicted genes from prokaryotes (6). The functional classification interface is driven by a series of pull-down menus that constrain annotators to a controlled vocabulary. This is increasingly recognized as a critical element of truly useful annotation (7). A single designated curator for each genome can review each contributed annotation and mark the records for deletion from the database or acceptance.

All genome features and annotations in the database are tagged and searchable based on their dates of creation and modification providing a natural mechanism for designating versions of sequence annotation. ASAP supports the analysis of projects that contain thousands of sequences, such as genome sequence scans or EST sequences as well as completed genome sequences and can be used to move annotations between versions of a sequence or even closely related genome sequences.

Connectivity between different databases is important for wide-spread access to information and the means to add value to annotations by taking advantage of the strengths of various systems. Most sequence data collected to date has been disseminated in the flat file data format of the INSD. Although ASAP uses an expanded feature type definition list, all feature types used in ASAP are mapped to appropriate standard types used by the INSD and the data in ASAP can be readily converted to this format for submission in the public databanks. This allows us to subdivide heterogeneous INSD categories like '/misc_feature' to facilitate querying meaningful subsets of features. No single data exchange standard exists for sequence information and related experimental data although useful suggestions for standards have been proposed such as the XML specifications of the Distributed Annotation System (DAS) (8) for sequence annotations and numerous approaches for standardization of microarray data (9). Although not currently compatible with these systems, ASAP is extensible through additional scripts to provide data in a variety of alternative formats. All tables viewed through ASAP are currently available for download as tab-delimited text files.

The goal of the ASAP system is to provide support for extensive community input of annotation information, rapid public dissemination of this information as well as to maintain some control over the quality and consistency of annotations by indicating their approval by the database curator. Permission to perform database operations is controlled by a user's status as a guest, an annotator or a curator. Guests can view genomes

and experiments that are labelled for public release in the database. Guests can also comment on the annotation of genome features by typing a note to the curator attached to each feature. These comments are not viewable by the public until reviewed and added to the annotation by the curator. Anyone willing to directly input data can request annotator status from the curator who will provide them with a username and password. The aim is to ensure that all annotations are attributed to a recognized user, not to restrict individuals from contributing information. Annotators are permitted to add annotations to features using a large number of pre-defined feature qualifiers, such as adding product names or gene functions for coding regions. The qualifier list, like the feature type definition is an expanded version of the INSD list. Annotators are not able to directly delete annotations to preserve a record of annotation history. Annotators can recommend changes to the existing features and annotations, such as changing the start coordinate of an open reading frame or deleting an incorrect annotation, to the curator who evaluates the recommendation and makes changes where appropriate. The addition of new features to the database, such as the addition of a new ORF, is also currently in the hands of the curator. As described below we have developed a simple interface that allows curators to add and modify features and annotations. Future development of the system will include additional scripts that improve the communication of changes between annotators and curators and speed the curators' ability to implement the changes. We are exploring the feasibility of allowing annotators to directly add features.

ASAP also includes a set of server-side scripts (PHP) that generate the web-based interface that all guests, annotators and curators use to access the database. The modular nature of these scripts has allowed the ASAP interface to be developed in stages. Among existing modules are interfaces for:

- Guests to query and view the annotation of genomes marked for release.
- Guests to add a note to the curator regarding a predicted feature.
- Annotators to query the complete set of features predicted in a genome.
- Annotators to retrieve an assignment from the curator.
- Annotators to enter data (e.g. gene names, products, etc.) for predicted features.
- Curators to assign features to annotators.
- Curators to review, reject and accept annotators' contributions.
- Curators to add, delete or alter the coordinates of predicted features.
- All users to download DNA and protein sequences of predicted features.
- All users to query and download whole genome transcript profiles.
- All users to query and download growth assay data.
- All users to query and request strains from a large collection of mutants.

COMMUNITY ANNOTATION OF THE *ERWINIA CHRYSANTHEMI* GENOME

Erwinia chrysanthemi is a plant pathogen within the family Enterobacteriaciae. Sequencing of its genome is underway and we have used ASAP to collect sequence annotations from an international group consisting of more than 30 *Erwinia* biologists. Annotation of the sequence began with an initial assembly of random sequence data representing approximately six-fold coverage of the genome and containing several hundred contigs. Genes were predicted using GLIMMER2 and tRNAscan-SE and assigned feature identifiers in ASAP. Features were distributed to annotators at this initial stage to ensure that all predicted features were examined by at least one annotator. Annotation is ongoing and at present nearly every predicted feature has been examined.

Annotation of a genome before completion of sequencing allows researchers more rapid access to genome information but results in wasted effort if it is not possible to associate annotated features with coordinates in an updated version of the genome sequence assembly. We address this problem by mapping genome features from an earlier version of a sequence to the newer version based on a complete alignment of the two sequence versions. This alignment provides an automated method for identifying sequence features that have not changed between the versions and assigning them coordinates in the new version. Features that are affected by changes in the sequence are also rapidly identified and can be reassigned to annotators to reevaluate their annotations. This approach was employed by some of the ASAP developers during the annotation of the *E. coli* O157:H7 EDL933 genome (10). ASAP facilitates moving annotations between sequence versions by decoupling storage of feature coordinates from storage of feature annotations that are not dependent on the coordinates. A complete list of current coordinates can be downloaded, mapped to a new sequence coordinate system using a genome-scale aligner and uploaded into ASAP without disrupting the existing annotations if the sequence of that particular feature remains intact. The ability to rapidly move annotations between sequence versions and retain a complete revision history is a key aspect that differentiates ASAP from other annotation systems that rely either on reannotation of the entire sequence or do not present users with a record of changes that may impact their research.

UPDATED ANNOTATION OF THE MG1655 SEQUENCE

All users logging on to ASAP have access to an updated version of the annotation for *E. coli* K-12 strain MG1655. A large number of gene functions have been experimentally determined for K-12 and this genome serves as a model for the annotation of additional enterobacterial genomes. ASAP contains all annotations present in the 1998 submission to GenBank (accession number U00096) including features subsequently dropped by NCBI, plus extensive updates including the assignment of many new gene functions, changes in gene names and revisions in gene content and boundaries (11–13). After logon users select the genome they want to view followed by the version of the sequence. After selecting a

genome and sequence version, users can choose among sequence features available for further investigation. Choosing CDS for example would return a list of all MG1655 coding regions. Selecting an individual feature launches a new web page with detailed annotations. For MG1655 protein coding features this information includes gene names, encoded products, classification in the MultiFun system, access to sequences, BLAST search results and links to external databases. A link to a graphical display of the chromosomal region is also provided. Some updated annotations reflect changes in the coordinates of a feature in the genome and it is possible to obtain a history of changes for these annotations.

In addition to coding regions the database contains information about RNA-encoding genes, operons, promoters, regulatory sites and other miscellaneous features. The results of comparing the *E. coli* K-12 and O157:H7 (10) genomes can be obtained by selecting the feature type 'islands' to return a list of segments that found in the K-12 but not the O157:H7 sequence or 'conserved_segments' to provide a list of regions conserved between the genomes. Physically related features are displayed in an interactive table explaining the relationship. This can be particularly useful for understanding features such as islands or operons, that often contain other features of interest such as genes and promoters.

The MG1655 genome in ASAP provides a mechanism for continued annotation of the sequence. Users are encouraged to sign up to add annotations reflecting their particular knowledge. Periodic collection of these annotations can then be used to generate updates of the annotated sequence file in other databases (this includes GenBank since some of the ASAP developers are the original submitters of the MG1655 sequence).

DATASETS OF FUNCTIONAL CHARACTERIZATION

The completion of genome sequences and the development of high-throughput experimental approaches to study genes generate large datasets that are logically linked to sequence annotations. ASAP currently supports diverse types of experimental information from a genome strain or a characterized derived mutant strain. We accommodate experiments yielding data about a strain in general and experiments resulting in data that can be linked directly to an annotated feature. Constraints are enforced to ensure that all experiments are associated with a valid strain and genome in the database and all feature-specific data is interactively connected to the updatable annotation. Mutations in derived strains are linked to corresponding feature(s) in the genome strain.

A large collection of *E. coli* strains can be searched to identify strains with characteristics such as mutations in particular genes, or to download a list of the entire collection. Detailed information about the nature of the mutations is reported. Links to request the strains from the depositors are contained within the results of these queries. Users can query the experiments in ASAP for those using a particular strain.

ASAP currently contains experimental data from gene expression analyses of *E. coli* cultures grown under diverse conditions and Biolog phenotypic microarray studies of single-gene mutant strains. Choosing either type of data within ASAP returns a list of available experiments in the database. Currently the gene expression databank consists of RNA abundance estimates from more than 50 *E. coli* microarray hybridization experiments. Database users can download entire datasets, request data associated with a particular feature across multiple experiments or view details about a particular experiment. Selecting an experiment a user can, for example, obtain the estimated RNA abundance for every genome feature represented in the experiment in a table that contains links to the most up-to-date annotations of those features. The dataset of phenotypic characterizations consists of results of growth of greater than 200 mutant strains on 5×96 different phenotype microarray assays. The data is searchable by mutant strain or by individual compounds in the phenotypic assay.

The import of experimental data into ASAP is facilitated by an interface that guides researchers through the uploading of tab-delimited text files containing results. Data is checked before entry to ensure the referential integrity of identifiers that link experimental information with genome annotations. A significant advantage of storing experimental data in ASAP is that it will be attached to updateable annotations of genome features. The annotation can be altered in light of new experimental data and interpretation of experimental results is facilitated by rapid access to existing annotation.

IMPLEMENTATION OF ASAP

The hardware and operating system platform that ASAP runs on is a dual-processor AMD Athlon™ system with 2 GB of RAM and a RAID storage array of 160 GB, running SUSE Linux (http://www.suse.com). ASAP uses Apache (http://www.apache.org) for web services, OpenSSL (http://www.openssl.org) for encryption, MySQL (http://www.mysql.com) for database services, PHP (http://www.php.net) and Perl (http://www.perl.com) for scripting, and Sun's Java platform (http://www.sun.com/java) for visualization. ASAP uses commodity hardware and open-source software to reduce costs and licensing burdens and promote portability of the system.

All information traffic between the user's internet browser and ASAP is encrypted using the Secure Sockets Layer protocol (SSL) with an encryption key length of 128-bits to ensure privacy. Annotators and curators are given a username and a password to use to access private data. The username and password allow access to data within the database based upon an access control list granting specific rights (e.g. annotator, curator, submitter) to the data for a specific genome project and sequence version.

ACKNOWLEDGEMENTS

We would like to thank Tim Durfee, Connie Kang, Mingzhu Liu and Yu Qiu for access to unpublished *E. coli* functional genomic data and Margrethe Hauge Serres and Monica Riley for providing their updated *E. coli* annotations. We are grateful to the International *Erwinia* Consortium for their participation in a community annotation of ECH3937 using ASAP as it develops. This work was supported by USDA grant

2001-52100-11316 and NIH grants GM62994-02 and GM35682-15A1.

REFERENCES

1. Blattner,F.R., Plunket,G.III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
2. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.
3. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
4. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
5. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
6. Serres,M.H. and Riley,M. (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb. Comp. Genomics*, **5**, 205–222.
7. The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
8. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The Distributed Annotation System. *BMC Bioinformatics*, **2**, 7.
9. Hayes,A. (2000) The second international meeting on Microarray Data Standards, Annotations, Ontologies and Databases. *Yeast*, **17**, 238–240.
10. Perna,N.T., Plunkett,G.III, Burland,V., Mau,B., Glasner,J.D., Rose,D.J., Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A., Posfai,G., Hackett,J., Klink,K., Boutin,A., Shao,Y., Miller,L., Grotbeck,E.J., Davis,N.W., Lim,A., Dimalanta,E., Potamousis,K., Apodaca,J., Anantharaman,T.S., Lin,J., Yen,G., Schwartz,D.C., Welch,R.A. and Blattner,F.R. (2001) Genome sequence of enterohemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–533.
11. Riley,M. (1998) Genes and proteins of *Escherichia coli* K-12. *Nucleic Acids Res.*, **26**, 54.
12. Riley,M. and Serres,M.H. (2000) Interim report on genomics of *Escherichia coli*. *Annu. Rev. Microbiol.*, **54**, 341–411.
13. Serres,M.H., Gopal,S., Nahum,A., Liang,P., Gaasterland,T. and Riley,M. (2001) A functional update of the *Escherichia coli* K-12 genome. *Genome Biol.*, **2**, 35.1–35.7.