BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 8 2010, pages 1122–1124
doi:10.1093/bioinformatics/btq090

*Genome analysis*

Advance Access publication March 1, 2010

# An Ergatis-based prokaryotic genome annotation web server

Chris Hemmerich†, Aaron Buechlein†, Ram Podicheti, Kashi V. Revanna‡ and Qunfeng Dong*,‡

Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47405, USA

Associate Editor: Alex Bateman

**ABSTRACT**

**Summary:** Ergatis is a flexible workflow management system for designing and executing complex bioinformatics pipelines. However, its complexity restricts its usage to only highly skilled bioinformaticians. We have developed a web-based prokaryotic genome annotation server, Integrative Services for Genomics Analysis (ISGA), which builds upon the Ergatis workflow system, integrates other dynamic analysis tools and provides intuitive web interfaces for biologists to customize and execute their own annotation pipelines. ISGA is designed to be installed at genomics core facilities and be used directly by biologists.

**Availability:** ISGA is accessible at http://isga.cgb.indiana.edu/ and the system is also freely available for local installation.

**Contact:** qunfeng.dong@unt.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on December 21, 2009; revised on February 21, 2010; accepted on February 22, 2010

## 1 INTRODUCTION

The expansion of genomic sequencing operations from major sequencing centers to local genomic core facilities is a growing trend in biology. Particularly for microbial genomics, core facilities with next-generation sequencing technologies can sequence microbial genomes in days. Every newly sequenced genome must go through a series of computational steps for annotation. Since biologists often rely on the genomics core facilities where the raw data are generated to provide assistance with annotation and additional analyses, these centers must seek efficient computational tools to keep pace with the increasing sequencing throughput.

One increasingly popular prokaryotic genome annotation tool is Ergatis (Orvis *et al*., unpublished data; http://ergatis.sourceforge.net/), a flexible workflow management system that allows users to design and execute complex analysis pipelines by combining multiple bioinformatics tools. Ergatis has several features that are essential for genomics core facilities supporting the annotation requests of multiple biologists, such as: (i) it provides robust support for using distributed parallel computing results; (ii) it is highly fault tolerant in its ability to restart and/or clone failed pipelines; (iii) it can manage running multiple simultaneous bioinformatics pipelines through a single interface; and (iv) it has built-in commonly used utility components and an interface for building new pipelines. Ergatis is distributed with several existing pipelines, including a prokaryotic annotation pipeline that is used as the underlying annotation engine for the microbial annotation services at the J. Craig Venter Institute, the Institute for Genome Sciences and other institutions. Ergatis is under active development and has a growing user community.

However, similar to other stand-alone tools, the flexibility and power of Ergatis comes at a cost of complexity, and its installation and usage requires sophisticated computational skills beyond an ordinary biologist's domain. First, the installation of Ergatis is time consuming and requires specialized local computing resources (e.g. Linux/Unix computer system). Second, designing an Ergatis pipeline requires expertise with the command-line tools used, BSML (an XML specification for biological sequences) and the Ergatis method of running pipelines. Third, even simple modification of a previously defined pipeline (e.g. changing a parameter value of a bioinformatics program) must go through the same complex interface used for pipeline design. Fourth, its user interface does not handle uploading files to a pipeline or downloading results. Finally, Ergatis lacks the security features to prevent a novice user from accidentally viewing or interfering with another user's pipeline. Therefore, bioinformaticians must spend significant time interacting with Ergatis on behalf of biologists for every pipeline executed. Even a modest number of requests from biologists to prepare, customize, run, monitor and report results from pipelines can easily overwhelm a small bioinformatics team (as typically employed by core facilities)—a problem that will only worsen as sequencing work continues to accelerate.

Our solution is to extend the stand-alone Ergatis pipeline with a web-based prokaryotic genome annotation server, Integrated Services for Genomics Analysis (ISGA), which provides an intuitive web interface for biologists to easily upload their genome sequences (or multiple genomic contigs), execute pipelines, monitor the progress of those pipelines and download the results. ISGA also provides an optional, easy-to-use interface for choosing which programs to run in a pipeline and modifying parameter values for selected programs. That is, biologists can annotate genomes themselves, while bioinformaticians maintain the service, and monitor the status and performance of the running pipelines. Besides executing annotation pipelines through Ergatis, ISGA also integrates additional analysis tools such as a genome browser, database search and sequence comparison tools, so that the produced annotation results can be further analyzed dynamically.

While other existing web-based prokaryotic genome annotation servers (e.g. Aziz *et al*., 2008; Bocs *et al*., 2003; Lee *et al*., 2009;

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
‡Present address: Department of Biological Sciences, Computer Science and Engineering, University of North Texas, Denton, TX 76203-5017, USA.

 

Markowitz *et al*., 2008; Meyer *et al*., 2003; Moriya *et al*., 2007; Van Domselaar *et al*., 2005) focus on providing a centralized service on their web sites, ISGA also emphasizes its portability. In addition to hosting our own web server, ISGA was developed as a portable package for others (e.g. bioinformaticians in genomics core facilities) to download and install at their local institutions for better performance and flexibility. We believe that any public web server will ultimately suffer deteriorating performance as user demand increases for limited resources. For example, it takes several weeks for the popular IMG web server (Markowitz *et al*., 2008) to return microbial annotation results to users. Therefore, a local installation of ISGA is the most sustainable annotation solution for genomic facilities. ISGA is distributed under an open source license, and all third-party components and prerequisites are also freely available. For facilities that have adequate resources (e.g. skilled bioinformaticians and computer equipment), installation of ISGA and all its dependencies is a standard procedure. The recently published DIYA package (Stewart *et al*., 2009) was also designed to meet the demand from genomics core facilities for a portable annotation. But DIYA is a command-line-based pipeline as opposed to being web based. To our best knowledge, the only other web-based prokaryotic annotation systems currently supporting a locally installable version are GenDB (Meyer *et al*., 2003) and AGMIAL (Bryson *et al*., 2006). Unlike ISGA, however, both GenDB and AGMIAL focus their development efforts on providing interface for expert manual annotations instead of annotation pipelines.

## 2 IMPLEMENTATION

ISGA provides user functionality while using an underlying Ergatis installation for pipeline execution (Supplementary Figure S1). While Ergatis pipelines are composed of low-level components, ISGA organizes those individual Ergatis components into high-level functional modules (i.e. a set of bioinformatics tools and utility programs that together perform a biological analysis). This abstraction simplifies the pipeline presentation by preserving information relevant to biologists while hiding the engineering details of the system. For example, 64 raw Ergatis components are grouped into the six modules of the ISGA annotation pipeline. The configuration of these modules as well as the type of input and output is stored in a relational database. For each input dependency, we identify whether or not it is required, and if users can supply their own input file. For each program, we identify any parameter of interest to biologists and define them, along with documentation and validation checks in a YAML (http://www.yaml.org/) configuration file, which is used to dynamically generate the corresponding web form users interact with.

Many bioinformatics tools are already part of the default Ergatis prokaryotic annotation pipeline, such as Glimmer3 (Delcher *et al*., 2007) for predicting protein-coding genes, RNAmmer (Lagesen *et al*., 2007) and tRNAscan-SE (Lowe and Eddy, 1997) for RNA-coding genes, BLAST (Altschul *et al*., 1997) for sequence similarity search and HMMPFAM (Eddy, 1998) for protein domain search. We have added additional bioinformatics tools to the pipeline, including MAST (Bailey and Gribskov, 1998) for scanning predicted promoter regions against RegTransBase (Kazakov *et al*., 2007) to detect transcription factor binding sites, Asgard (Alves and Buck, 2007) for metabolic pathway reconstruction and OligoPicker (Wang and Seed, 2003) for designing microarray oligonuleotide probes based on predicted gene sequences. Popular databases are also used for the pipeline, e.g. GenBank (Benson *et al*., 2009) and COG (Tatusov *et al*., 2003). Briefly, for each bioinformatics tool added to the default Ergatis pipeline, we have created a corresponding Ergatis configuration file to define parameters, input and output formats, as well as how the tool is executed with respect to other tools in the workflow.

ISGA, developed in object-oriented Perl, displays the annotation pipeline as a clickable graph (Supplementary Figure S2A). In the graph, each node represents a functional module, and edges indicate the flow of data through the pipeline in the form of output files. By clicking on a graph node, users can select or deselect that functional module for execution (Supplementary Figure S2B). For example, a user can skip the gene prediction portion of the pipeline and instead supply a file of predicted genes to be used by later modules (e.g. protein domain search). Users can also modify the parameter values of each selected bioinformatics program (Supplementary Figure S2C). By enabling users to customize their pipelines, ISGA provides transparency, flexibility and control to users, whereas many other web annotation servers limit users with a fixed set of programs and/or default parameter values. These customizations are saved in a PostgreSQL database and used to build the corresponding Ergatis configuration files. Users are required to provide necessary input data, e.g. genome sequence or multiple contig sequences (*.gz* compressed format is allowed to speed up file upload) and a basic description of the organism (Supplementary Figure S2D). After the pipeline is submitted, users can track detailed execution status (Supplementary Figure S2E). Once the computation is completed, the user is notified via e-mail and may access the annotation results in several popular formats (e.g. GenBank and GFF3). The raw output of each module is also available for download (Supplementary Figure S2F). ISGA implements a secure account system to ensure the privacy of users' submitted data and results. Users can browse their previously submitted pipelines (Supplementary Figure S2G). Any customizations a user makes to a pipeline are also saved as part of their account information, and can be used to perform the same analysis on additional datasets. Occasionally, hardware problems cause pipelines to fail. ISGA notifies administrators by e-mail when this happens, who then resolve the underlying problem and restart the pipeline with Ergatis. Additional administrative tools include the ability to cancel or duplicate users' pipelines for problem diagnosis.

ISGA also provides a bioinformatics 'toolbox' for further analysis of annotation results, including displaying predicted genes along the annotated genome with GBrowse (Stein *et al*., 2002) (Supplementary Figure S2H) and searching annotation results by keyword (Supplementary Figure S2I). Each predicted gene is accompanied with detailed annotation information. Users can also perform dynamic BLAST searches against their uploaded genome sequence, gene prediction results and other databases (Supplementary Figure S2J).

Currently, we have limited our implementation to prokaryotic genome annotation because of the increasing demands for annotating microbial genomes by genomics core facilities. The modular design of ISGA and Ergatis allows the system to be easily extended. Besides annotation tools, we plan to incorporate genome assembly programs into the pipeline. Eukaryotic genome annotation and other functional genomics pipelines can also be included in the future development.

## 3 CONCLUSION

ISGA provides a new and distinct option for the research community when choosing from the variety of available genome annotation tools. Through its intuitive web interface for the popular Ergatis workflow system, ISGA enables professional bioinformaticians to better and more efficiently serve the annotation needs of biologists. Unlike many other web-based centralized annotation servers, ISGA also emphasizes its support of local installation. In particular, ISGA may benefit genomics core facilities that adopt Ergatis to provide a genome annotation service. By installing a local copy of ISGA, genomics core facilities are able to allow their biologist users to easily invoke the underlying complicated Ergatis annotation pipelines by themselves. Additional bioinformatics tools are also integrated for biologists to dynamically analyze the annotation results.

## ACKNOWLEDGEMENTS

We thank Jon Burgoyne and Phillip Steinbachs for system support, Rajesh Gollapudi for participating in the early development of ISGA and Kayce Reed-Buechlein for improvements to the user interface. We also thank many microbiologists at Indiana University for providing invaluable feedback.

*Funding*: This work was supported in part by the Indiana Metabolomics and Cytomics Initiative of Indiana University, funded in part through a major grant from the Lilly Endowment, Inc.

*Conflict of Interest*: none declared.

## REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Alves,J.M. and Buck,G.A. (2007) Automated system for gene annotation and meta bolic pathway reconstruction using general sequence databases. *Chem. Biodivers.*, **4**, 2593–2602.

Aziz,R.K. *et al.* (2008) The RAST Server: rapid annotations using subsystems Technology. *BMC Genomics*, **9**, 75.

Bailey,T.L. and Gribskov,M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.

Benson,D.A. *et al.* (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.

Bocs,S. *et al.* (2003) AMIGene: Annotation of MIcrobial Genes. *Nucleic Acids Res.*, **31**, 3723–3726.

Bryson,K. *et al.* (2006) AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. *Nucleic Acids Res.*, **34**, 3533–3545.

Delcher,A.L. *et al.* (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Kazakov,A.E. *et al.* (2007) RegTransBase–a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res.*, **35**, D407–D412.

Lagesen,K. *et al.* (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.

Lee,D. *et al.* (2009) WeGAS: a web-based microbial genome annotation system. *Biosci. Biotechnol. Biochem.*, **73**, 213–216.

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.

Markowitz,V.M. *et al.* (2008) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res.*, **36**, D528–D533.

Meyer,F. *et al.* (2003) GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.*, **31**, 2187–2195.

Moriya,Y. *et al.* (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, **35**, W182–W185.

Stein,L.D. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

Stewart,A.C. *et al.* (2009) DIYA: a bacterial annotation pipeline for any genomics lab. *Bioinformatics*, **25**, 962–963.

Tatusov,R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Van Domselaar,G.H. *et al.* (2005) BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res.*, **33**, W455–W459.

Wang,X. and Seed,B. (2003) Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics*, **19**, 796–802.