Nucleic Acids Research, 2006, Vol. 34, Database issue **D379–D381**
doi:10.1093/nar/gkj042

VMD: a community annotation database for oomycetes and microbial genomes

Sucheta Tripathy, Varun N. Pandey[2], Bing Fang[1], Fidel Salas[3] and Brett M. Tyler*

Virginia Bioinformatics Institute, [1]Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA-24061, USA, [2]Microsoft Corporation Ltd, Redmond, WA-98052, USA and [3]Omicia Corporation, CA, USA

Received August 13, 2005; Revised and Accepted October 3, 2005

ABSTRACT

The VBI Microbial Database (VMD) is a database system designed to host a range of microbial genome sequences. At present, the database contains genome sequence and annotation data of two plant pathogens *Phytophthora sojae* and *Phytophthora ramorum*. With the completion of the draft genome sequences of these pathogens in collaboration with the DOE Joint Genome Institute (JGI), we have created this resource to make the sequences publicly available. The genome sequences (95 MB for *P.sojae* and 65 MB for *P.ramorum*) were annotated with ~19 000 and ~16 000 gene models, respectively. We used two different statistical methods to validate these gene models, Fickett's and a log-likelihood method. Functional annotation of the gene models is based on results from BlastX and InterProScan screens. From the InterProScan results, we could assign putative functions to 17 694 genes in *P.sojae* and 14 700 genes in *P.ramorum*. We created an easy-to-use genome browser to view the genome sequence data, which opens to detailed annotation pages for each gene model. A community annotation interface is available for registered community members to add or edit annotations. There are ~ 1600 gene models for *P.sojae* and ~700 models for *P.ramorum* that have already been manually curated. A toolkit is provided as an additional resource for users to perform a variety of sequence analysis jobs. The database is publicly available at http://phytophthora.vbi.vt.edu/.

INTRODUCTION

Phytophthora species are oomycete pathogens infecting many host plant species. *Phytophthora sojae* is a soybean pathogen that has been extensively used as a model for this genus. *Phytophthora ramorum* is a newly emerged pathogen of woody shrubs and trees that is threatening California's coastal oak forests. The genome sequencing of *P.sojae* and *P.ramorum* was carried out in conjunction with the DOE Joint Genome Institute (JGI). The draft sequences, covering 9x for *P.sojae* and 7x for *P.ramorum* were completed in early 2004. An Annotation Jamboree was held at the JGI during August 2004, where community members manually annotated data that were hosted at the JGI (http://www.jgi.doe.gov/). To complement the JGI databases we have created a database, called the VBI Microbial Database (VMD), that has a community annotation web interface to enable ongoing editing and detailed annotation of the sequences by community members. The Genome Unified Schema (GUS; www.gusdb.org) (1), developed at the Computational Biology & Informatics Laboratory (CBIL), at the University of Pennsylvania, was chosen as the database system for storing the information. We have created a large number of software tools around the schema to perform various functions, such as data upload, data retrieval and user interfaces. We also extensively updated the GUS installation notes (http://www.gusdb.org/documentation/older/vbidoc.pdf and http://phytophthora.vbi.vt.edu/documents/gus.pdf).

VMD is an integrated resource with community annotation features, toolkits and a large number of other facilities to perform complex queries. We used Perl/CGI for most of the front end applications and Java servlets for storing query histories. For the community annotation interface we created a separate schema based on MySQL to temporarily store contributed annotations for review, before transfer to the GUS database. The database was released to the public in April 2005.

DATA RESOURCES

VMD currently contains the assembled genome sequences of two species, *P.sojae* and *P.ramorum*, that were generated by a

*To whom correspondence should be addressed. Tel: +1 540 231 7318; Fax: +1 540 231 2606; Email: bmtyler@vt.edu

random shotgun strategy. The assembly of paired end shotgun sequence reads was carried out with the JGI assembler, Jazz, at a coverage of 9x for *P.sojae* and 7x for *P.ramorum*. After trimming for vector and quality, over one million reads were assembled into 1810 scaffolds totaling over 86 Mb for *P.sojae* and 2576 scaffolds totaling 66.6 Mb for *P.ramorum*. Approximately half of the genome is contained in 54 scaffolds all at least 463 kb in length for *P.sojae* and 63 scaffolds all at least 308 kb in length for *P.ramorum*. The estimated genome size is 95 Mb for *P.sojae* and 65 Mb for *P.ramorum*. The initial gene calling was carried out at the JGI using the following four algorithms: FgeneSHAbinitio (2), FgeneSHhomology, GeneWise (3) and Synteny-based methods. We validated the four models using Fickett's (4) statistics and codon preference (5) methods and stored only the best model in VMD.

DATA PROCESSING AND ANALYSIS

For each gene model, we have stored the Fickett (4) and log-likelihood (5) validation data. The BLASTX (6) outputs of the sequence scaffolds against 22 different databases belonging to various taxonomic groups are also stored in the database. The predicted coding and non-coding regions in the *P.sojae* and *P.ramorum* sequences were compared to each other and the results displayed in the browser. Similarity of a coding region in one species to a non-coding region in the other is a useful guide to possible mis-called gene models. Expressed sequence tag (EST)-derived unigenes from *P.sojae* and *Phytophthora infestans* were aligned to the *P.sojae* scaffolds using BLAT (7) and stored in the database. Alignments similar to the *P.ramorum* sequence will soon be added. InterProScan (8) implements Pfam (9), Prosite (10), Prodom (11), Fingerprint (12), Interpro (13), Smart (14) and TigrPfam (15) searches. The outputs from InterProScan were parsed and stored in supplementary tables outside GUS.

Other data that have been uploaded to the database include, but are not limited to, external resources such as the Gene Ontology dump files from GenBank, nr data from GenBank, external databases and their release versions.

Analysis data were uploaded to the database using the Perl object layer of the GUS system. Additional scripts were written for data that did not fit into the GUS relational model.

USER INTERFACE DESIGN AND IMPLEMENTATION

We have created an easy-to-use interface for the database, mostly based on CGI written in Perl running on an Apache web server. Our user interface has several components, the most important ones are as follows:

(i) browser page,
(ii) query page,
(iii) edit annotation page and
(iv) toolkit page.

Apart from these there are several other utilities such as Boolean searches, History, News and a Statistics page.

Browser page

We have created an easy-to-use Perl GD-based browser which is distinct from Gbrowse (16). The browser currently shows the gene models, the non-coding regions of scaffolds, gaps and the EST-to-genome sequence alignments (Figure 1A). The view can be zoomed in and out to increase clarity and visibility

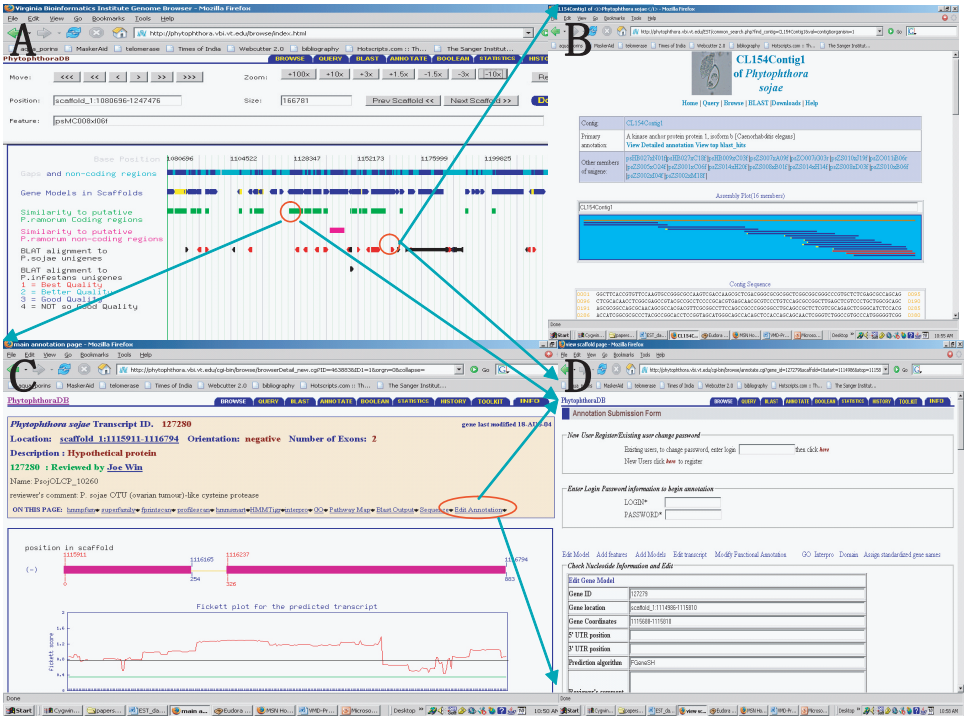

Figure 1. (**A**) The genome browser at 10× zoom. (**B**) When clicking a gene model, the main annotation page opens with a detailed description of the gene model. (**C**) When clicking the EST to genome alignment, the link opens to EST database with more detailed information. (**D**) The community annotation interface.

of data. Each gene model or EST unigene is made clickable to view detailed annotation information (Figure 1B). The main annotation page is the central part of the user interface. It summarizes details of the current annotation and also provides functions for community members to add or modify existing annotations. This page notes if the gene has already been manually annotated.

The EST unigene alignment to the genome uses four colors to indicate the quality of the alignments, similar to the NCBI Blast output viewer. Clicking on a unigene icon opens a more detailed view (Figure 1C) of the unigene from the EST database (S. Tripathy and B. M. Tyler, unpublished).

Query page

The query page has several different combinations of queries, including advanced queries, which open to the main annotation page. The query can take either gene_ids (if known), EST ids, or key words from the primary annotation or protein domain annotations. All records retrieved by the query are linked through the gene_ids to the main annotation page.

Edit annotation page

This page opens from the main page or from any other location through the annotation button (Figure 1D). Registered users can directly edit information by providing their user name and passwords. New users can register and then edit the information. The data from this page goes to a temporary MySQL database, where the curator reviews the information before it is finally transferred to the main schema. Once in the main schema the color of a manually annotated gene model appears yellow on the main browser.

Toolkit page

The toolkit page provides a combination of several web-based programs including the EMBOSS (17) suite, Blast services and pairwise sequence alignment using bl2seq. Additionally, there are other utility programs written by us (S. Tripathy and B. M. Tyler, unpublished) based on Fickett statistics and codon usage to validate potential gene models on the fly. The EMBOSS suite has more than 160 sequence analysis tools that can perform a range of jobs. The blast interface is provided with graphical outputs with links to the internal databases. On clicking a sequence id, the user can view a detailed annotation of that particular gene or EST sequence.

CONCLUSION AND FUTURE DIRECTIONS

We will soon be uploading information on orthology and paralogy relationships among *P.sojae* and *P.ramorum* genes, as well as gaps in the genome sequences closed using BAC end sequencing. Our next release will contain a viewer to link BAC physical map data to the sequence scaffolds via BAC end sequences. In the next year we expect to add genome sequences for the fungal pathogen of *Arabidopsis, Alternaria brassicicola* and the oomycete pathogen of *Arabidopsis*, *Hyaloperonospora parasitica* to the database. In addition we plan to implement support for proteomic and microarray data. The microarray data have already been stored in the database, and we will create links between the functional genomic data and the genome sequences.

ACKNOWLEDGEMENTS

We thank Tejal Karkhanis for programming assistance, the CBIL group at the University of Pennsylvania for valuable advice and Margaret Gabler for assistance in manuscript preparation. This work was supported by grants from the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service (2002-35600-12747 and 2004-35600-15055) and from the National Science Foundation (MCB-0242131, EF-0412213 and DBI-0211863). Funding to pay the Open Access publication charges for this article was provided by the Virginia Bioinformatics Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Davidson,S.B., Crabtree,J., Brunk,B., Schug,J., Tannen,V., Overton,G.C. and Stoeckert,J.C.J. (2001) K2/Kleisli and GUS: experiments in integrated access to genomic data sources. *IBM Syst. J.*, **40**, 512–531.
2. Salamov,A.A. and Solovyev,V.V. (2000) *Ab initio* gene finding in *Drosophila*. *Genomic DNA*, **10**, 516–522.
3. Birney,E., Clamp,M. and Durbin,R. (2004) GeneWise and GenomeWise. *Genome Res.*, **14**, 988–995.
4. Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.*, **10**, 5303–5318.
5. McLachlan,A.D. (1984) A method for measuring the non-random bias of a codon usage table. *Nucleic Acids Res.*, **12**, 9567–9575.
6. Altschul,S.F., Madden,T.L., Schaffer,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.
7. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
8. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
9. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S. and Sonnhammer,E.L. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
10. Hulo,N., Sigrist,C.J., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.
11. Bru,C., Courcelle,E., Carrère,S., Beausse,Y., Dalmar,S. and Kahn,D. (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.*, **33**, D212–D215.
12. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K. and Taylor,P. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.
13. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P. and Cerutti,L. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
14. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copeley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
15. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
16. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. and Lewis,S. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
17. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.