BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 11 2002
Pages 1544–1545

Poxvirus Orthologous Clusters (POCs)

*Angelika Ehlers, John Osborne, Stephanie Slack, Rachel L. Roper and Chris Upton**

Department of Biochemistry and Microbiology, University of Victoria, Victoria, BC V8W 3P6, Canada

Received on March 11, 2002; accepted on May 22, 2002

ABSTRACT

Summary: *P*oxvirus *O*rthologous *C*lusters (POCs) is a JAVA client–server application which accesses an updated database containing all complete poxvirus genomes; it automatically groups orthologous genes into families based on BLASTP scores for assessment by a human database curator. POCs has a user-friendly interface permitting complex SQL queries to retrieve interesting groups of DNA and protein sequences as well as gene families for subsequent interrogation by a variety of integrated tools: BLASTP, BLASTX, TBLASTN, Jalview (multiple alignment), Dotlet (Dotplot), Laj (local alignment), and NAP (nucleotide to amino acid alignment).

Availability: Direct access to the POCs database via the GENOME ANALYSIS link at the Poxvirus Bioinformatics Resource: http://www.poxvirus.org. The Software is available for download via HTTP at: http://athena.bioc.uvic.ca/pbr/POCs/pocs.html.

Contact: cupton@uvic.ca

Supplementary Information: Installation instructions, the User's Manual, screenshots, and examples are available at the POCs home page http://athena.bioc.uvic.ca/pbr/POCs/pocs.html. The software is free for non-commercial applications. For information on poxviruses see http://www.poxvirus.org.

The purpose of POCs is to provide a series of tools to organize and to permit a variety of comparative studies on the vast amount of data present in complete genome sequences. POCs is a Java based user-interface that connects to a Linux (RedHat V7.2) server and accesses a large mySQL database (about 1 megabyte per poxvirus genome). The POCs database stores: (1) the complete corrected data set provided in the GenBank file, for complete poxvirus genomes; and (2) annotations, ORFs, MW, PI, nucleotide and amino acid (aa) frequencies, codon use, BLASTP scores, and orthologous cluster membership data, for each gene. Compared to its predecessor, the Viral Genome DataBase (Hiscock and Upton, 2000), this database uses normalized tables, can handle spliced genes and bacterial genomes, includes administrator-added annotations and the interface integrates a variety of other programs to simplify and speed up data analysis.

The client program provides the user with two routes into the database. The 'Sequence Query' panel allows for defining a variety of constraints on the data including gene/ORF name, virus name, DNA/Protein sequence, MW, pI, DNA and aa composition to retrieve genes from the database. Second, the 'Gene Family Analyzer' panel provides an interface to find genes by family membership, virus ORF annotations, or BLASTP *E*-values. The results of a query, a list of genes or families, are displayed in a customizable table. The complete table (up to 122 columns) can be sorted by using any column as a constraint and the 'result table' columns can be ordered, enabled/disabled by the user; each user can store a personalized set of table layouts. Once one or more genes have been selected, the user can perform a variety of interactive analyses using the integrated applications: BLASTP, BLASTX, TBLASTN (Altschul *et al.*, 1997); BLASTZ, Laj, Lat (Schwartz *et al.*, 2000); Jalview (Clamp *et al.*, 1999); Dotlet (Junier and Pagni, 2000); NAP (Huang and Zhang, 1996). All of these analyses are performed by the POCs server with the exception of Jalview, which can use ClustalW either on the client machine or at the Jalview home site via a remote connection. POCs feeds all required data to the applications listed above, runs the application and provides the output in a separate window. The graphical user interface is intuitive and easy to use, and has been designed for molecular biologists rather than computer scientists. This interface makes it simple for the researcher to make a variety of otherwise complex SQL database queries (Table 1) very rapidly.

The administrative client program offers the same functionality as the client program plus functions to allow for the management of the database. These functions include: add a genome and all its genes into the database from a given GenBank file; delete or modify genomes and/or genes; assign genes to gene families based on user

*To whom correspondence should be addressed.

Table 1. Examples of POCs queries. Queries are performed in either the Sequence Query or Gene Family Analyzer window of POCs

Window	
Sequence Query	Find genes whose name fulfills a given condition
Sequence Query	Find genes whose DNA contains a given sequence of characters
Sequence Query	Find proteins whose pI is within a given interval
Sequence Query	Find proteins whose predicted MW lies within a given interval
Sequence Query	Find proteins whose serine content is greater than 13%
Sequence Query	Find genes whose cytosine content is less than 25%
Gene Family Analyzer	Find the gene family containing gene *X* from a specific genome
Gene Family Analyzer	Find a specific gene family, or group of gene families
Gene Family Analyzer	Find genes matching a query protein with a BLAST *E*-value of >50
Gene Family Analyzer	Find genes with a specific user-annotation
Gene Family Analyzer	Find gene families present in genome *Y* but absent from genome *Z*

assigned BLASTP *E*-values; edit, delete or modify user notes. The server program is the database connector for the client programs. Any request to retrieve data from the database by the client programs is sent to the server, which then executes the request and sends the answer to the client via TCP. This makes it possible to offer access to a mySQL database and to provide UNIX only applications (like BLASTZ or NAP) to a client program running on Macintosh (OS X) or Windows computers; the only system requirement for the POCs client program is Java Runtime Environment 1.3.1 or higher.

The client program is simply downloaded as a jar file and started as an application. If the local system supports Java Web Start, the client program can be run via Java Web Start from the web via the POCs home page without manually downloading the software. In each case the client connects (by default) to the poxvirus database at the University of Victoria. This database currently contains 21 complete poxvirus genomes with 4194 ORFs. These 4194 predicted genes have been grouped into 337 families and named by function (if known). 52 families contain seventeen genes, one from each of the 21 genomes indicating that these genes are most likely absolutely essential for poxvirus function. Currently, 15 families contain more than 21 genes because of the presence of paralogs or gene fragments. One of the key functions of POCs is to permit the continual analysis of genome data and provide a simple process to update the database as new relationships between proteins are determined or predicted ORFs are confirmed as functional genes or discovered to be inactive gene fragments. To this aim, future versions of the database will have data elements that allow us to mark gene fragments as distinct from complete genes within a family; we do not wish to delete the fragmented genes from the database because they are important for comparative studies.

For researchers wishing to create their own database, a local 'POCs server' and an empty 'POCs database' can be installed on a Linux machine via a shell script. The administration client permits the addition of several types of genomes to this database; we have tested the import of bacterial, herpesvirus and baculovirus genomes successfully.

ACKNOWLEDGEMENTS

This work was supported by funds from NSERC OPG0155125-01 and NIH U01-AI48653-02. The authors are grateful to Matt Boone and Paul Ripley for their contribution to the design of a POCs prototype.

REFERENCES

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.

Clamp,M.E., Cuff,J.A. and Barton,G.J. (1999) Jalview—a java multiple sequence alignment viewer and editor. http://www.compbio.dundee.ac.uk/.

Hiscock,D. and Upton,C. (2000) Viral Genome DataBase: a tool for storing and analyzing genes and proteins from complete viral genomes. *Bioinformatics*, **16**, 484–485.

Huang,X. and Zhang,J. (1996) Methods for comparing a DNA sequence with a protein sequence. *CABIOS*, **12**, 497–506.

Junier,T. and Pagni,M. (2000) Dotlet: diagonal plots in a Web browser. *Bioinformatics*, **16**, 178–179.

Schwartz,S., Zhang,Z., Frazer,K.A., Smit,A., Riemer,C., Bouck,J., Gibbs,R., Hardison,R. and Miller,W. (2000) PipMakerA web server for aligning two genomic DNA sequences. *Genome Res.*, **10**, 577–586.