



ORFcurator: molecular curation of genes and gene clusters in prokaryotic organisms

Jeffrey A. Rosenfeld^{1,†}, Indra N. Sarkar^{2,†}, Paul J. Planet³,
David H. Figurski³ and Rob DeSalle^{1,*}

¹Division of Invertebrate Zoology, Central Park West, 79th Street, American Museum of Natural History, New York, NY 10024, USA, ²Department of Biomedical Informatics and ³Department of Microbiology, College of Physicians and Surgeons, Columbia University, New York, NY USA

Received on April 15, 2004; revised on June 2, 2004; accepted on July 16, 2004

Advance Access publication July 22, 2004

ABSTRACT

Summary: The ability to detect clusters of functionally related genes in multiple microbial genomes has enormous potential for enhancing studies on gene function and microbial evolution. The staggering amount of new genome sequence data presents a largely untapped resource for gene cluster discovery. To date, gene cluster analysis has not been fully automated, and one must rely on manual, tedious and time-consuming manipulation of sequences. To facilitate accurate and rapid identification of conserved gene clusters, we developed a database-driven web application, called ORFcurator. We used ORFcurator to find clusters containing any genes similar to those of the 14-gene Widespread Colonization Island of *Actinobacillus actinomycetemcomitans*. From 126 genomes, ORFcurator identified all 73 clusters previously determined by manual searching.

Availability: ORFcurator and all associated scripts are freely available as supplementary information.

Contact: desalle@amnh.org

Supplementary information: <http://www.genomecurator.org/ORFcurator/>

INTRODUCTION

Ordering, interpreting and annotating genes and clusters of genes across multiple genomes ('Molecular Curation') are the major goals of whole genome sequencing. Evolutionarily conserved genes and gene clusters across prokaryotic genomes may offer valuable insight into evolutionarily maintained biochemical processes and genome structure (Itoh *et al.*, 1999; Lathe *et al.*, 2000; Rogozin *et al.*, 2002; Xie *et al.*, 1999). In addition, analysis of gene clusters across multiple genomes can provide insight into their histories [e.g. the creation and dispersion of clusters (Itoh *et al.*, 1999)]. However, extracting

genes into easily interpretable formats, from both annotated ('complete') and unannotated ('incomplete') genomic sequences remains a significant task. Many tools have been created that can aid in these tasks [e.g. STRING (von Mering *et al.*, 2003) and the COG database (Tatusov *et al.*, 2003)]. Nonetheless, comparative analysis is time consuming and difficult even when genomes have been completed and fully annotated [e.g. available at GenBank (Benson *et al.*, 2004)].

Here we describe ORFcurator, a web application that simplifies and automates the discovery of putative genes and gene clusters for prokaryotic organisms. ORFcurator enables one to enter genes, either singly or as a cluster of genes from the same locus. The application searches across a database consisting of compiled sequences from complete and partially assembled genomes. The results are presented as both sequence data and graphical maps of the discovered clusters.

MATERIALS AND METHODS

An application was written that (1) identifies putative genes and (2) creates clusters from these identified genes. Using an alignment search tool, the application identifies regions of sequence similarity between every submitted 'query sequence' and selected genome sequences stored in a local MySQL database. All sequence information is updated on a weekly basis. The current implementation uses BLAST, supplemented with Apple/Genentech BLAST (ACG, 2002, <http://www.apple.com/acg/>). We have plans to incorporate more efficient alignment search tools such as BLAT (Kent, 2002).

Query sequences are aligned separately against each selected genome. The application selects sequence alignment regions using a specified criterion. Additional genome sequence is then retrieved upstream and downstream until the first in-frame stop codon is encountered. This sequence is referred to as the 'stop-alignment-stop' sequence. Within each stop-alignment-stop sequence, the application identifies the

*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.

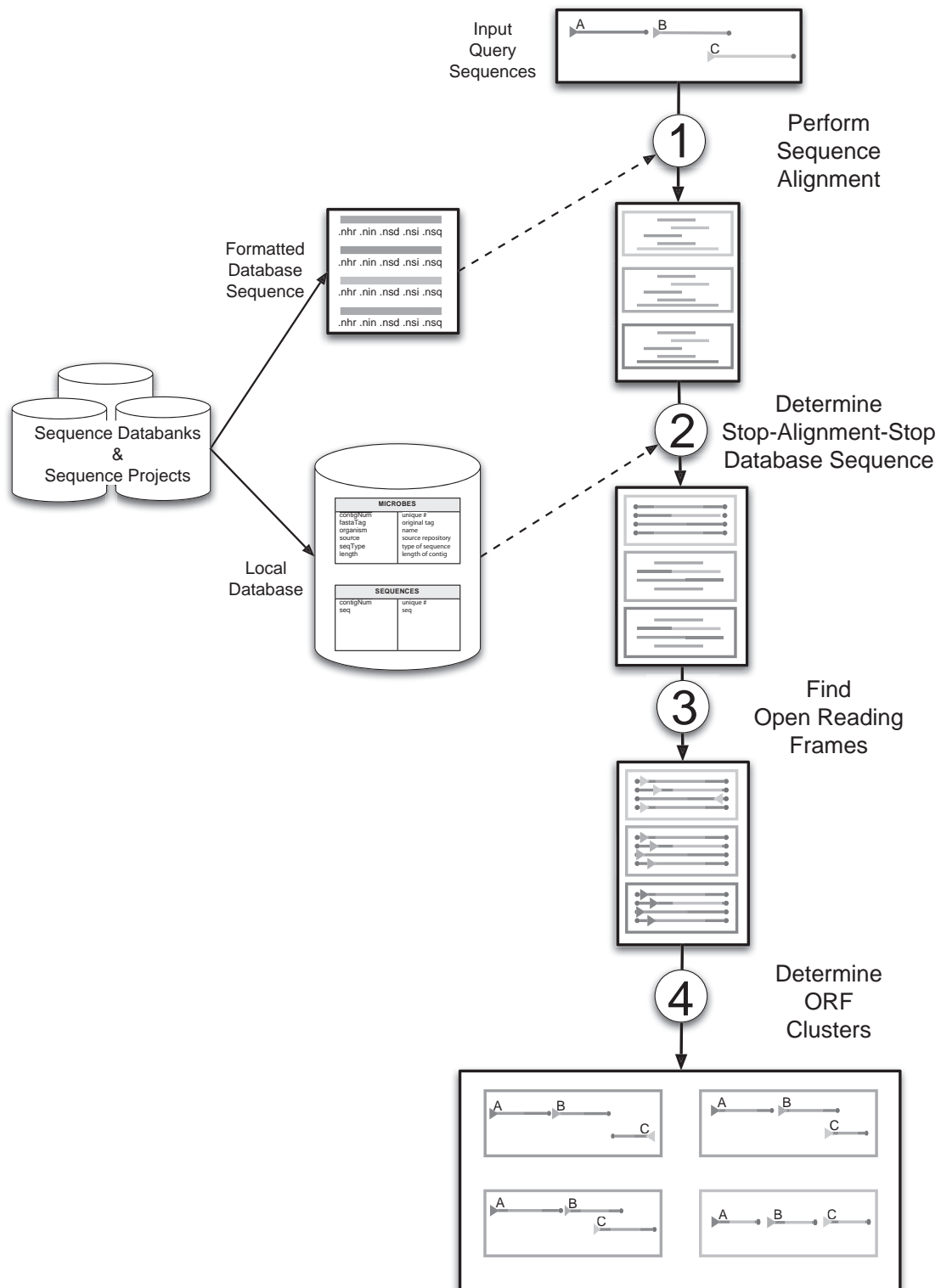


Fig. 1. ORFcurator is a database driven web application. The local database is populated with information from a number of sequencing centers and databanks. ORFcurator begins by first performing a sequence alignment for each query sequence individually with each selected genome (1); database sequence that includes the alignment and a single stop codon upstream and downstream is then retrieved from the local database (2); the database sequence is then searched for complete ORFs (3); and finally, the ORFs are clustered into loci based on a specified proximity threshold (4).

largest open reading frame (ORF). The largest ORF is defined as a sequence within the stop-alignment-stop sequence that begins with the first in-frame start codon and ends with a stop codon.

All complete ORFs are clustered based on sequence positions relative to each genome contiguously. Clustering is performed by iteratively identifying ORFs that are within a specified proximity criterion of each other. Every cluster is called a 'locus'. Within each locus, ORFs can be manually selected and stored in a Gene Finding Format (GFF), (Durbin and Haussler, 2000, http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml) file and used to generate graphical images using gff2ps (Abril and Guigo, 2000) and pstopdf (Apple, 2004). This entire ORF identification and clustering process, outlined in Figure 1, is termed 'ORFuration'.

Evaluation

The efficacy of ORFcurator was assessed using a gold standard created from 126 organisms that had been previously examined manually for a recently discovered genomic island consisting of multiple genes involved in bacterial adherence (collectively called the Widespread Colonization Island, WCI) (Planet *et al.*, 2003). Of these organisms, the WCI was identified in 73 organisms; 54 did not have the WCI. We used ORFcurator to search our database for these genes using the WCI gene cluster from two organisms, *Actinobacillus actinomycetemcomitans* and *Caulobacter crescentus*.

True positives (TPs) were identified as instances when ORFcurator found a previously identified cluster. False positives (FPs) were noted when clusters were found in organisms that were known not to contain the cluster. True negatives (TNs) were characterized by instances when ORFcurator did not identify a cluster in an organism that did not contain the cluster. False negatives (FNs) were noted when a cluster was not found in an organism that was known to have a cluster. The positive predictive value [PPV = TP/(TP + FP)], sensitivity [TP/(TP + FN)] and specificity [TN/(TN + FP)] were calculated.

ORFcurator identified all the known gene clusters from the gold standard. PPV values improved when setting a lenient *E*-value (e.g. all clusters were retrieved at an *E*-value of 0.1). When using no *E*-value cutoff in our evaluation, the PPV was 0.92 and sensitivity was 0.89. Although imposing a more stringent *E*-value threshold increased the PPV, it impacted the sensitivity of the searches. Thus, while ORFcurator simplifies identifying putative gene clusters, manual inspection is still required to validate found clusters.

SUMMARY

A fundamental operation in microbial genetics and genomics is identifying conserved gene clusters. As the availability of sequence data multiplies, this task is increasingly dependent on automated tools. By centralizing disparate sequence

information into a single database, researchers can use these tools to search across various sequence repositories and centers, regardless of whether or not the sequences have been fully annotated. In addition, such tools can also exploit the well-known association between physical proximity of genes on the chromosome and function (Jacob *et al.*, 1960) to predict function of the many 'hypothetical' genes from genome projects. Tools such as ORFcurator will play a valuable role in identifying putative genes and gene clusters, which may lead to subsequent functional hypotheses.

ACKNOWLEDGEMENTS

The authors thank Kurt Lienau and Mark Siddall for their valuable insights towards the development of ORFcurator. R.D. and J.A.R. thank the Louis and Dorothy Cullman Program for Molecular Systematics at the AMNH. NIH/NLM Grant LM-07079-09 supports I.N.S. The Columbia University MSTP program supports P.J.P. This work was funded in part by NIH/NIGMS Grant R01-GM62351. The authors acknowledge the use of genetic sequence from the following sources: Genoscope-Centre National de Séquençage, The US Department of Energy Joint Genome Institute, University of Illinois, Urbana-Champaign, University of Minnesota Microbial Genome Project, The BBRP Sequencing Group at LLNL, The National Center for Biotechnology Information, Ohio State University, University of Oklahoma's Advanced Center for Genome Technology, Microbial Sequencing group at the Sanger Institute, Stanford Genome Technology Center and The Institute for Genome Research.

REFERENCES

- Abril, J.F. and Guigo, R. (2000) gff2ps: visualizing genomic annotations. *Bioinformatics*, **16**, 743–744.
- ACG (2002) AGBlast. Cupertino, Apple Computer.
- Apple (2004) pstopdf (pre-installed on OS X 'Panther' systems). Cupertino, Apple Computer.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.I., Ostell, J. and Wheeler, D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32** (Database issue), D23–D26.
- Durbin, R. and Haussler, D. (2000) GFF Protocol Specification.
- Itoh, T., Takemoto, K., Mori, H. and Gojobori, T. (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.*, **16**, 332–346.
- Jacob, F., Perrin, D., Sanchez, C. and Monod, J. (1960) Operon: a group of genes with the expression coordinated by an operator. *C. R. Hebd Seances Acad. Sci.*, **250**, 1727–1729.
- Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
- Lathe, W.C., III, Snel, B. and Bork, P. (2000) Gene context conservation of a higher order than operons. *Trends Biochem. Sci.*, **25**, 474–479.
- Planet, P.J., Kachlany, S.C., Fine, D.H., DeSelle, R. and Figurski, D.H. (2003) The widespread colonization island of *Actinobacillus actinomycetemcomitans*. *Nat. Genet.*, **34**, 193–198.

- Rogozin,I.B., Makarova,K.S., Murvai,J., Czabarka,E., Wolf,Y.I., Tatusov,R.L., Szekely,L.A. and Koonin,E.V. (2002) Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res.*, **30**, 2212–2223.
- Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
- von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
- Xie,G., Brettin,T.S., Bonner,C.A. and Jensen,R.A. (1999) Mixed-function supraoperons that exhibit overall conservation, albeit shuffled gene organization, across wide intergenomic distances within eubacteria. *Microb. Comput. Genomics*, **4**, 5–28.