Published online 20 December 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D517–D518***
doi:10.1093/nar/gkm886

AlterORF: a database of alternate open reading frames

Inti Pedroso[1], Gustavo Rivera[1], Felipe Lazo[2], Max Chacón[2], Francisco Ossandón[1], Felipe A. Veloso[1] and David S. Holmes[1,*]

[1]Center for Bioinformatics and Genome Biology, Life Science Foundation, MIFAB and Andrés Bello University, Santiago, Chile and [2]Department of Informatics, University of Santiago, Santiago, Chile

Received August 17, 2007; Revised and Accepted October 2, 2007

ABSTRACT

AlterORF is a searchable database that contains information regarding alternate open reading frames (ORFs) for over 1.5 million genes in 481 prokaryotic genomes. The objective of the database is to provide a platform for improving genome annotation and to serve as an aid for the identification of prokaryotic genes that potentially encode proteins in more than one reading frame. The AlterORF Database can be accessed through a web interface at www.alterorf.cl

INTRODUCTION

A DNA sequence contains six potential open reading frames (ORFs), three on one strand and three on the reverse strand. However, typically only one of the six is actually expressed because it is associated with appropriate genetic signals that specify the DNA strand and the reading frame to be transcribed and translate. Exceptions occur in which more than one open reading frame is translated into a protein, as has long been observed in the case of viral genes, where it was suggested that this property permitted a high packing density of information (1). However, analysis of the coding potential of 481 prokaryotic genomes revealed the surprisingly high frequency of alternate ORFs of annotated genes especially in high G + C rich genomes, where almost every annotated ORF exhibits an alternative ORF that could potentially encode a protein of 100 amino acids or more (2).

The frequency of alternate open reading frames in high G + C genomes gives rise to the possibility that this property could be exploited to evolve novel genetic information and it is important to be able to detect this potential. However, this high frequency also provokes serious problems of gene annotation, where the incorrect ORF may inadvertently be mis-annotated as the coding sequence. This potential for error is especially problematic when automatic gene prediction programs are used to annotate genomes, but errors can also slip by human annotators. The problem is exacerbated if an alternative ORF is mis-annotated and the error is propagated in subsequent genome annotations.

AlterORF provides a searchable database of all possible alternative ORFs in sequenced prokaryotic genomes that are potentially capable of encoding proteins of 100 amino acids or more. The objectives are 2-fold: to improve genome annotation by indicating possible errors in ORF identification and, perhaps more important in the long term, to predict instances of genes that potentially could give rise to more than one protein.

DATABASE CONSTRUCTION

Annotated protein coding genes were extracted from completely sequenced prokaryotic genomes in the Genome Database of NCBI. All alternative ORFs, potentially encoding 100 amino acids or more, were extracted from each gene sequence using Perl scripts and the BioPerl Application Programming Interface (API) (3). Using the standard genetic code, the *in silico* translated amino acid sequence of each alternative ORF was searched for similarity in completely sequenced prokaryotic genomes (4) and for conserved domains and motifs using CDD (5), PFAM (6), COG (7), KOG (8), SMART (9) and UniProt. (10). Hierarchical clustering using the software hcluster_sg developed as part of the TreeFam project (11) was used to build sequence families with the alternate ORFs. BLAST *e*-values were normalized from 0 to 100 (with 100 corresponding to *e*-value 0.0). The resulting information was stored in a relational database built with Microsoft SQL Server 2005.

DATABASE CONTENTS

Release 1.0 (September 2007) contains approximately 1.5 million annotated genes from 481 organisms and

*To whom correspondence should be addressed. Tel: +56 2 239 8969; Fax: +56 2 237 2259; Email: dsholmes2000@yahoo.com

about 3 million alternate ORFs. Of these 942 856 (33%) occur in frame −1, 621 306 (21%) in frame −2, 322 284 (11%) in frame −3, 350 805 (12%) in frame +2 and 675 525 (23%) in frame +3. The following are provided for each alternate ORF sequence: (i) conserved domains and motifs including CDD (5), PFAM (6), COG (7), KOG (8), SMART (9) and UniProt. (10) and (ii) BLAST results with annotated sequences in completely sequenced prokaryotic genomes and alternate ORFs identified in AlterORF. The cross genera conservation of some alternate ORFs suggests that they might represent new protein families or domains and hierarchical clustering (11) was used to build sequence families from conserved alternate ORFs.

WEB INTERFACE AND SERVICES

The AlterORF database can be accessed through a simple and easy to use web interface at www.alterorf.cl. The database can be searched by protein ID (derived from NCBI), by organism and by sequence using a sequence search service. In addition, an option is provided to analyze complete genome sequences not present in the database. Searching by protein ID: a protein ID can be used to recover the original annotated gene that appeared in the database (e.g. GenBank), and also any alternate ORF(s) associated with that gene. If alternate ORFs are detected, tables providing information regarding domains, motifs and protein family are displayed with links to further information. Searching by organism: the user can select an organism from a pulldown menu or index for a pre-analyzed list of annotated protein coding genes with alternate ORFs. Searching by protein sequence: a search using a protein sequence can be carried out against all sequences stored in AlterORF using WU-BLAST (blast.wustl.edu/). Downloading data: all data in the AlterORF database can be freely downloaded by ftp. Additional information on the use of AlterORF can be found in the FAQs and Tutorial sections.

ACKNOWLEDGEMENTS

The database is supported in part by a Microsoft Sponsored Research Award and Fondecyt 1050063. Funding to pay the Open Access publication charges for this article was provided by the above sponsors and by Andrés Bello University.

Conflict of interest statement. None declared.

REFERENCES

1. Fiddes,J.C., Seeburg,P.H., DeNoto,F.M., Hallewell,R.A., Baxter,J.D. and Goodman,H.M. (1979) Evolution of the three overlapping gene systems in G4 and phi X174. *J. Mol. Biol.*, **133**, 19–43.
2. Valdes,J., Veloso,F., Jedlicki,E. and Holmes,D. (2005) Large-scale, multi-genome analysis of alternate open reading frames in bacteria and archaea. *OMICS*, **9**, 91–105.
3. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
4. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zheng,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSIBLAST—A new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
5. Marchler-Bauer,A., Anderson,J.B., Derbyshire,M.K., DeWeese-Scott,C., Gonzales,N.R., Gwad,M., Hao,L., He,S., Hurwitz,D.I. *et al.* (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.*, **35**, D237–D240.
6. Finn,R.D., Mistry,J., Schuster-Böckler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
7. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
8. Tatusov,R.L., Fedorova,N.D., Jackson,J.J, Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
9. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
10. UniProt Consortium. (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
11. Li,H., Coghlan,A., Ruan,J., Coin,L.J., Hériché,J.K., Osmotherly,L., Li,R., Liu,T., Zhang,Z. *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, **34**, D572–D580.