Nucleic Acids Research, 2004, Vol. 32, Database issue ***D281–D283***
DOI: 10.1093/nar/gkh116

The ORFanage: an ORFan database

Naomi Siew[1,2,*], Yaniv Azaria[2] and Daniel Fischer[2]

[1]Department of Chemistry and [2]Bioinformatics Group, Department of Computer Science, Ben Gurion University, Beer-Sheva 84105, Israel

Received July 13, 2003; Revised October 8, 2003; Accepted October 15, 2003

ABSTRACT

As each newly sequenced genome contains a significant number of protein-coding ORFs that are species-, family- or lineage-specific, many interesting questions arise about the evolution and role of these ORFs and of the genomes they are part of. We refer to these poorly conserved ORFs as singleton or paralogous ORFans if they are unique to one genome, or as orthologous ORFans if they appear only in a family of closely related organisms and have no homolog in other genomes. In order to study and classify ORFans we have constructed the ORFanage, an ORFan database. This database consists of the predicted ORFs in fully sequenced microbial genomes, and enables searching for the three types of ORFans in any subset of the genomes chosen by the user. The ORFanage could help in choosing interesting targets for further genomic and evolutionary studies. The ORFanage is accessible via http://www.bioinformatics.buffalo.edu/ORFanage.

INTRODUCTION

Large-scale genome sequencing projects enable us to gain different perspectives on nature, not possible before. One interesting finding that has become clear is that a large percentage of each newly sequenced genome contains protein-coding ORFs that do not resemble any other sequence in the databases. This phenomenon is by now well established, and seems to be an intrinsic part of the genomic material, independent of the growth in the number of new genomes that are sequenced (1–3). These species-specific unique sequences have been referred to as orphan ORFs or ORFans (4,5), and, in particular, singleton ORFans (1,6). The percentage of singleton ORFans in each newly sequenced genome can be as high as 60% (7), and they can be found also among strains of the same organism (8). In addition to these unique ORFans, a large fraction of ORFs in each genome has homologs only in the same genome or in closely related genomes. We refer to these ORFs as paralogous and orthologous ORFans, respectively (1,6).

The existence of so many species-, family- and lineage-specific ORFs in each newly sequenced genome raises many interesting questions about their origin, evolutionary mechanisms, roles and functions (1,2,6). In addition, many ORFans could be attractive targets for further studies because of their uniqueness.

In order to study the ORFan phenomenon and its dynamics as new genomes are sequenced, we have constructed the ORFanage, an ORFan database. This database consists of all predicted ORFs in fully sequenced microbial genomes, and allows searching for singleton, paralogous and orthologous ORFans. Since orthologous ORFans are defined as appearing in only a small subset of organisms, the database allows the user to specify the subset of genomes on which to perform the search. Currently, our database includes ORFs from the first 84 fully sequenced microbial genomes (strains not included), containing 31 850 singleton ORFans out of a total of 248 992 ORFs (13%). Out of these 31 850 singleton ORFans, 20 237 (63.5%) are shorter than 150 residues. Concerns have been raised as to whether short ORFans are coding sequences at all, and it has been assumed that their abundance may be due to errors, wrongly annotated genes or random distribution of nucleotides (9–13). Our and other studies have shown that this indeed could be the case for many of the short ORFans; however, not for all of them (1,11,14,15). Therefore, our database lists all ORFans found in the genomes, along with their length.

Lastly, we would like to emphasize that the ORFans listed are only ORFans with respect to our database. That is, homology searches in larger databases such as nr (NCBI non-redundant database) may result in finding matches in other genomes that are not currently listed in the ORFanage. However, analyses we have conducted on families of paralogous and orthologous ORFans identified by the ORFanage show that most of these ORFans are still ORFans in the nr database (our unpublished results). Therefore it seems that sequences that are poorly conserved in 84 genomes are poorly conserved by nature. Since complete genome sequences are being deposited in the databases at a fast rate, updates of the ORFanage will be carried out periodically.

SOURCES OF GENOMIC DATA AND METHODS

The amino acid sequence files of complete genomes were downloaded, in chronological order of their publication,

*To whom correspondence should be addressed. Tel: +1 716 849 6719; Fax: +1 716 849 6749; Email: nomsiew@cs.bgu.ac.il
Present address:
Daniel Fischer, Buffalo Center of Excellence in Bioinformatics, 901 Washington Street, Suite 300, Buffalo, NY 14203-1199, USA

The authors wish it to be known that, in their opinion, all authors should be regarded as joint First Authors

Table 1. Summary of a query on three members of the *Bacillus* family: ORFans in each genome and orthologous ORFans

Genome	ORFs[a]	Singleton ORFans[b]	Paralogous ORFan families (ORFs)[c]	Orthologous ORFan families (ORFs)[d]
Bacillus subtilis	4100	494	24 (53)	
Bacillus halodurans	4066	494	13 (34)	
Oceanobacillus iheyensis	3496	333	8 (16)	
B.subtilis–B.halodurans				46 (94)
B.subtilis–O.iheyensis				27 (56)
B.halodurans–O.iheyensis				29 (58)
All three genomes				51 (175)

ORFans in the *Bacillus* family of closely related organisms. The numbers of singleton, paralogous and orthologous ORFans is shown. Numbers of singleton and paralogous ORFans are listed for each individual genome; numbers of orthologous ORFans are listed for groups of at least two organisms. The number of paralogous and orthologous families is listed, and the total number of ORFs in each type of ORFan family appears in parenthesis. All numbers correspond to the 84-genome version of the database. An electronic version of the table in an extended form can be found at the ORFanage main page, under 'NAR supplementary material'.

[a]Number of ORFs in the genome.

[b]Number of singleton ORFans in the genome.

[c]Number of paralogous ORFan families; in parenthesis: the total number of ORFs in these families.

[d]Number of orthologous ORFan families; in parenthesis: the total number of ORFs in these families.

mainly from the NCBI FTP server, with several genomes being downloaded from TIGR's site and the Kasuza Institute. Only one strain of each organism was downloaded. A table listing the download site, date and strain for each genome can be found at the ORFanage website.

After the addition of each new genome to our database, gapped BLAST (16) was run for each ORF in the new genome, against our growing database of fully sequenced genomes. A match between this ORF and the other sequences was considered significant if the first BLAST hit had an E-value of $<10^{-3}$ (or 10^{-5} for alignments of <80 residues) (6). An ORF without significant matches was labeled as an ORFan, and this label could be changed later to non-ORFan if matches with ORFs in a newer genome were later found.

THE DATABASE

The ORFanage can be reached via http://www.bioinformatics.buffalo.edu/ORFanage and is constructed of two sections. The first one is dedicated to singleton ORFans. It contains a list of all the genomes in our database as well as information about the percentage of ORFans in each genome at the time it was added to the database and after 60 and 84 genomes. Separate information is given for long and short ORFans (6). Each genome's name is clickable, giving a list of the singleton ORFans in the genome.

The second, and main, section of the database is the ORFan searcher. Here the user can choose a subset of genomes to perform the search upon. The genomes are listed alphabetically, according to kingdom. The results of the submitted search are mailed to the user.

The resulting page is constructed of several ORFan tables. For each individual genome there are two tables, of paralogous and singleton ORFans. In addition, if the chosen subset of genomes contains sequences that are shared only among these genomes and are found in no other genome in the database, a table of orthologous ORFans is also produced. Thus, if two genomes that are closely related are chosen, five tables will be produced (see Table 1 for examples).

The orthologous and paralogous ORFan tables list the sequence families that were found, each family consisting of at least two ORFs. Clicking on an ORF's name will display its amino acid sequence. The tables can be sorted by the number of sequences in each family, the length of the shortest or longest sequence in each family, or the average length of each family.

Since the ORFanage enables searching for sequences that are common to only several of the genomes, it enables studying species-specific or family-specific proteins, or searching for potential horizontally transferred genes among genomes that are not phylogenetically related. Thus, this database can be of valuable aid in choosing interesting and important targets for further studies.

DATABASE ACCESS

The ORFanage is freely available through http://www.bioinformatics.buffalo.edu/ORFanage.

ACKNOWLEDGEMENTS

We are grateful to Samir Genaim for help in constructing the web interface. This joint project was partially supported by Grant No. 81948101 from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel. N.S. is supported in part by grants from the Ministry of Science, Israel and from the Kreitman Foundation Fellowship.

REFERENCES

1. Siew,N. and Fischer,D. (2003) Twenty thousand ORFan microbial protein families for the biologist? *Structure*, **11**, 7–9.
2. Siew,N. and Fischer,D. (2003) Unraveling the ORFan puzzle. *Comp. Funct. Genomics*, **4**, 432–441.

3. Kunin,V., Cases,I., Enright,A.J., de Lorenzo,V. and Ouzounis,C.A. (2003) Myriads of protein families and still counting. *Genome Biol.*, **4**, 401–402.
4. Fischer,D. and Eisenberg,D. (1999) Finding families for genomic ORFans. *Bioinformatics*, **15**, 759–762.
5. Fischer,D. (1999) Rational structural genomics: affirmative action for ORFans and the growth in our structural knowledge. *Protein Eng.*, **12**, 1029–1030.
6. Siew,N. and Fischer,D. (2003). Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins*, **52**, 241–251.
7. Gardner,M.J., Hall,N., Fung E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S. *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498–511.
8. Boucher,Y., Nesbo,C.L. and Doolittle,W.F. (2001) Microbial genomes: dealing with diversity. *Curr. Opin. Microbiol.*, **4**, 285–289.
9. Dujon,B. (1996) The Yeast Genome Project: what did we learn? *Trends Genet.*, **12**, 263–270.
10. Tatusov,R.L., Galperin,M.Y., Natale D. A. and Koonin,E.V. (2001) The COG Database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
11. Andrade,M.A., Daruvar,A., Casari,G., Schneider,R., Termier,M. and Sander,C. (1997) Characterization of new proteins found by analysis of short open reading frames from the full yeast genome. *Yeast*, **13**, 1363–1374
12. Skovgaard,M., Jensen,L.J., Brunak,S., Ussery,D. and Krogh A. (2001) On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.*, **17**, 425–428.
13. Mackiewicz,P., Kowalczuk,M., Gierlik,A., Dudek,M. and Cebrat,S. (1999) Origin and properties of non-coding ORFs in the yeast genome. *Nucleic Acids Res.*, **27**, 3503–3509.
14. Basrai,M.A., Hieter,P. and Boeke,J.D. (1997) Small open reading frames: beautiful needles in the haystack. *Genome Res.*, **7**, 768–771.
15. Ochman,H. (1996) Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes. *Trends Genet.*, **18**, 325–327.
16. 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.