

Absynte: a web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters

Arnaud Despalins, Souhir Marsit and Jacques Oberto*

Université Paris-Sud 11, CNRS, UMR8621, Institut de Génétique et Microbiologie, 91405 Orsay Cedex, France

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Summary: Absynte (Archaeal and Bacterial Synteny Explorer) is a web-based service designed to display local syntenies in completely sequenced prokaryotic chromosomes. The genomic contexts are determined with a multiple center star clustering topology on the basis of a user-provided protein sequence and all (or a set of) chromosomes from the publicly available archaeal and bacterial genomes. The results consist in a dynamic web page where a consistent color coding permits a rapid visual evaluation of the relative positioning of genes with similar sequences within the synteny. Each gene composing the synteny can be further queried interactively using either local or remote databases. Absynte results can be exported in .CSV or high resolution .PDF formats for printing, archival, further editing or publication purposes. Performance, real-time computation, user-friendliness and daily database updates constitute the principal advantages of Absynte over similar web services.

Availability: <http://archaea.u-psud.fr/absynte>

Contact: jacques.oberto@igmors.u-psud.fr

1 INTRODUCTION

The availability of more than 1400 completely sequenced archaeal and bacterial genomes at the National Center for Biotechnology Information (NCBI) repository provides a wealth of information to researchers involved in prokaryotic genetics. In particular, the comparison of the relative genetic positioning on microbial chromosomes is of special interest not only to measure evolutionary relationships between different species but also to deduce the function of uncharacterized proteins. This conservation of orthologous gene order is commonly referred as "synteny" even if in its original definition the term had a different meaning (Renwick, 1971). The extraction of syntenic information from sequenced genomes involves the impractical manipulation of large data files and the complexity of the task increases with the number of genomes that need to be compared. The comparative analysis of genome segments from prokaryotic organisms has been addressed by web services such as GeConT2 (Martinez-Guerrero, et al., 2008), PSAT (Fong, et al., 2008) and GCVView (Grin and Linke, 2011). Unfortunately, these web applications suffer from one or more restrictions as discussed below. In response to these limitations, we have developed Absynte (Archaeal and Bacterial Synteny Explorer) a web

tool which only requires a user-provided protein sequence to display the corresponding synteny from a daily updated list of archaeal and bacterial genomes. This interactive web application is executed in real time and is designed to extract, compare and predict orthologous gene clusters originating from any combination of sequenced prokaryotic organisms.

2 FEATURES

The Absynte workflow is initiated by the submission of a protein sequence which is first compared to itself using BLASTP (Altschul, et al., 1997) in order to determine the maximal bit score. The user then opts either to match this protein against all available completely sequenced archaeal and bacterial chromosomes in the database or to a selection of up to 50 individual replicons using TBLASTN (Altschul, et al., 1997). Normalization of the resulting scores with the maximal bit value allows to sort the chromosomes by decreasing score. Each matching chromosome determines a 15kilobase segment centered on the sequence similarity coordinates from which open reading frames and protein sequences are extracted according to GenBank annotations. The proteins from the highest ranking chromosome are compared to each other in order to detect potential duplicates/paralogs using the Smith-Waterman-Gotoh (SWG) algorithm (Gotoh, 1982). These proteins are then matched using SWG against the proteins of the remaining chromosomes. A consistent color code is assigned to matching proteins and synteny maps are then rendered with their genes drawn to scale according to the same color scheme. The alignment and coloring of all the genes composing the genomic context maps allow an immediate visual evaluation of the conservation of gene order across the various analyzed genomes (Fig. 1). This is an advantage over other synteny servers such as GCVView which will only trace user-defined syntenies using simple BLASTP searches. Absynte is designed to produce synteny maps in real-time in order to simplify daily database maintenance and update. The GeConT2 or PSAT web services rely instead on pre-computed databases which are rarely updated and lack recent genomes. Each newly added genome might indeed require the impractical regeneration of the whole pre-computed database. On the other hand, the real-time Absynte workflow is more computationally intensive. Several techniques were therefore elaborated in order to optimize the

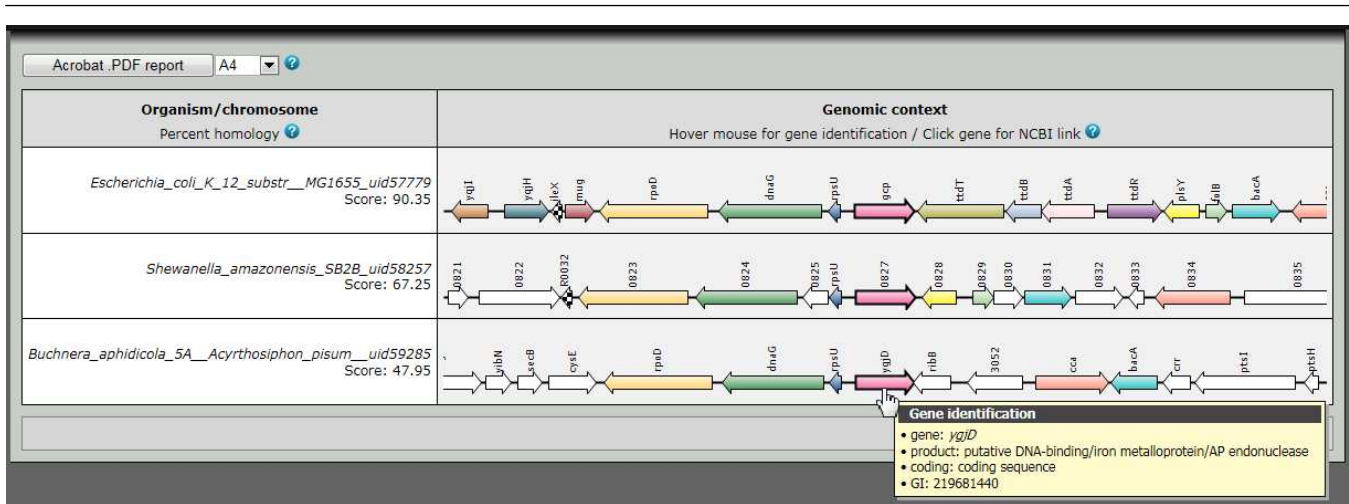


Fig. 1. Absynte report for *Escherichia coli* protein YgiD in three bacterial genomes. The gene corresponding to the protein under analysis is drawn in bold at the center of each context map. The consistent genetic color-coding allows immediate visualization of the synteny. Additional genetic information, displayed in a tooltip, is available for all the genes shown in the genomic contexts.

overall performance. Among these, the development of parallel processing routines permitted to fully exploit multi-core processors. In addition, a particular interest was devoted to the implementation of routines such as SWG, able to process data directly from faster live memory instead of relying on slower temporary disk files. An additional benefit of Absynte is constituted by its "multiple center star" gene clustering topology which allows the detection of duplicates/paralogs even in the highest ranking chromosome. These duplicates/paralogs would be overlooked by the aforementioned web services which use instead a star network topology. Absynte provides also full access to all archaeal genomes contrarily to PSAT and GeConT2 which offer respectively partial access and no access to these genomes. To maximize intuitive user experience, the Absynte web service was designed with an uncluttered user interface to insulate the user from the complexity of the underlying data processing. The results are presented in a rich and interactive graphical form, providing additional information and external web links for each individual gene. The results can be exported in .PDF and .CSV formats for printing, imaging or further processing purposes

3 DATABASES

The Absynte database is located on the web server, it contains all the bacterial and archaeal genomes provided by the NCBI repository. It is used to calculate synteny maps and to provide contextual gene information. The database update is fully automated and synchronizes daily at 7:00 GMT with the NCBI repository using the FTP protocol. Since they share the same database, Absynte and the previously described BAGET and FITBAR servers form a complementary set of web tools (Oberto, 2008; Oberto, 2010). Additional taxonomic and specific gene coding information is available through hypertext links to the corresponding NCBI services.

4 CONCLUSIONS

The analysis of synteny constitutes the criteria of choice to establish the orthology of genomic regions in different species and more

importantly allows inference of important functional relationships between genes. We presented here Absynte, a web tool able to explore conservation of prokaryotic gene sequence similarity by the means of a multiple center star clustering algorithm. The initial protein comparisons are executed with TBLASTN instead of BLASTP to avoid potential inaccuracies in genomic annotations. The computations are executed in real-time, from live memory, on the full set of daily updated archaeal and bacterial genomes, allowing independence from pre-calculated databases more difficult to maintain. We believe that the rapid identification of synteny provided by Absynte might be of wide interest for researchers dealing with prokaryotic genetics and could constitute a valid complement to phylogenetic analyses of gene clusters.

ACKNOWLEDGEMENTS

The authors wish to thank the "Agence Nationale pour la Recherche" and the "Centre National pour la Recherche Scientifique" for financial support.

REFERENCES

Altschul, S.F., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389-3402.
 Fong, C., *et al.* (2008) PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes, *BMC Bioinformatics*, **9**, 170.
 Gotoh, O. (1982) An improved algorithm for matching biological sequences, *J Mol Biol.*, **162**, 705-708.
 Grin, I. and Linke, D. (2011) GCView: the genomic context viewer for protein homology searches, *Nucleic Acids Res.* [Epub ahead of print]
 Martinez-Guerrero, C.E., *et al.* (2008) GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways, *Nucleic Acids Res.*, **36**, W176-180.
 Oberto, J. (2008) BAGET: a web server for the effortless retrieval of prokaryotic gene context and sequence, *Bioinformatics*, **24**, 424-425.
 Oberto, J. (2010) FITBAR: a web tool for the robust prediction of prokaryotic regulons, *BMC Bioinformatics*, **11**, 554.
 Renwick, J.H. (1971) The mapping of human chromosomes, *Annu Rev Genet.*, **5**, 81-120.