Sequence analysis

Augur—a computational pipeline for whole genome microbial surface protein prediction and classification

A. Billion, R. Ghai, T. Chakraborty and T. Hain*

Institute of Medical Microbiology, Justus-Liebig-University, Frankfurter Strasse 107, D-35392 Giessen, Germany

Received on July 26, 2006; revised on August 25, 2006; accepted on August 28, 2006

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: The analysis of protein function is a challenge and a major bottleneck towards well-annotated and analysed microbial genomes. In particular, bacterial surface proteins present an opportunity for pharmacological intervention and vaccine development. We present Augur, an automatic prediction pipeline that integrates major surface prediction algorithms and enables comparative analysis, classification and visualization for gram-positive bacteria on a genomic scale.

Availability: http://bioinfo.mikrobio.med.uni-giessen.de/augur
Contact: Andre.Billion@ mikrobio.med.uni-giessen.de
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Several gram-positive bacteria are the cause for common, severe illnesses e.g. Mycobacterium tuberculosis, Staphylococcus aureus and Streptococcus pyogenes. Major efforts are constantly underway to develop preventive and treatment regimens and to identify microbial proteins that may be attractive drug targets for pharmacological agents or can yield important vaccine candidates. In this regard, bacterial surface proteins, that are crucial to interaction with the environment, especially the host immune response, are extremely promising candidates. In the last decade a large number of microbial genomes, pathogenic and non-pathogenic, have been sequenced. Despite this, protein function assignment remains a major challenge, as a large number of proteins remain with insufficient or no annotation at all. In addition, surface protein prediction using various methods and algorithms (e.g. PROSITE and HMMs) often requires adjustment of several parameters, which currently is a time consuming process. Tools that integrate predictions are available, but tools that allow cross comparisons across genomes are not. In order to facilitate the prediction of bacterial surface proteins, and to allow for interesting genome-wide comparisons between pathogenic and non-pathogenic bacteria, we have developed Augur, an automatic pipeline to run multiple tools, integrate predictions and visualize results across all gram-positive genomes.

Augur uses the best available methods to predict seven different types of surface protein motifs such as signal peptides, lipobox (for lipoproteins), LPXTG motif, GW modules, NlpC/P60 domain, LysM motif and transmembrane helices. In addition, it can also annotate genomes by the COG functional classification (Tatusov et al., 2000) and the SCOP protein structural classification. The resulting output may be viewed in a web browser as an illustrative graphical bar chart, or as table view in which genes are linked to supporting information (Fig. 1). Currently, Augur’s database consists of all 90 publicly available gram-positive genome sequences.

2 IMPLEMENTATION

The calculation module of Augur’s pipeline was implemented Java 1.5. The web interface runs under the Apache web server and is built using HTML, PHP and JavaScript. Data is stored in a MySQL database with up to 21 tables. All required software packages are usually installed on Linux and Solaris systems, but can also be freely downloaded. The calculation module works only on Unix-based systems. We recommend using a PC with ~1.5 GB RAM and 2500–3000 GHz CPU speed. The software package requires ~200 MB disk space. To access Augur’s interface, the user needs only a standard web browser.

3 PIPELINE DESCRIPTION

Augur extracts basic sequence information from standard EMBL files. These files are parsed and data is stored into the database. An all-versus-all BLASTP (Altschul et al., 1997) is run to identify orthologs. Ortholog relations calculated using BLASTP are based on: (1) protein identity ≥50%, (2) e-value ≤0.001 and (3) both coverages between 75 and 125% (alignment-length ¸ 100/protein length). After BLASTP, the surface prediction module runs signal peptide prediction via SignalP, both by HMM and neural networks (Nielsen et al., 1997); transmembrane helices with TMHMM (Krogh et al., 2001); GW modules with HMM Model 22279 from Superfamily (Madera et al., 2004); and HMMs from Pfam (Bateman et al., 2000), LysM domains (PF01476), NlpC/P60 domain (PF00746), and LRR region (PF00877), LPXTG sorting signals (PF00746), and LRR region (PF00877). In addition, a pattern match (‘[MV].{0,13}[^DERKQ]{6 ,20}[^LIVMFESTAG][AV][LV]’ a) (Sutcliffe et al., 2002) is also run for the lipoprotein motif.

After surface predictions are complete, the COG module assigns as many genes as possible in COG functional groups. Lastly, the SCOP prediction module uses up to 8500 hidden Markov models (HMMs) to classify all proteins in the database into ~1500 super families. Augur implements the CRC64 checksum algorithm to extract all external database entries for a sequence from a uniprot flat file. All results for each protein are stored in Augur’s database, and each gene has its own gene page. The entire pipeline can be controlled by a graphical user interface.

*To whom correspondence should be addressed.
Fig. 1. A schematic of the pipeline for protein prediction, classification and annotation. Some possible graphical outputs are also shown.

4 EVALUATION

A new HMM was constructed for lipoprotein prediction. From several gram-positive bacteria (e.g., *Bacillus sp.*, *Streptococcus sp.*, *Lactococcus sp.*) 68 experimentally validated lipoproteins were used to construct the HMM. The HMM was tested on 28 experimentally validated lipoproteins from *Listeria monocytogenes* (by personal communication of Uwe Kaerst and Lothar Jaensch, HZI Braunschweig.). All 28 proteins were successfully detected, and in addition, 30 new lipoproteins were predicted in *L. monocytogenes*. We have also evaluated and compared our results with previously published results for *L. monocytogenes* and *L. innocua* (Cabanes et al., 2002). The lipoprotein and LPXTG predictions were also tested using results from a study of surface proteins of *S.pyogenes* (Rodriguez-Ortega et al., 2006) and found to be consistent with experimental data. Twenty additional lipoproteins were also predicted in *S.pyogenes*. Our predictions are slightly more conservative than those provided by PSORT (Nakai and Horton, 1999) used by the authors for lipoproteins and LPXTG motif prediction. All of the other HMMs in the Augur pipeline have been previously described and evaluated by the respective authors.

5 USAGE

The web interface of Augur allows users with no special programming background or bioinformatics experience to examine the results. The user is able to view the distribution of any motif e.g. all LPXTG motif harbouring proteins in a pathogenic bacterium, and compare these results to a non-pathogenic bacterium; or retrieve all predicted surface proteins from any genome. Cross-organism comparisons are possible with all surface protein predictions, SCOP and the COG classification. From one to many ortholog, paralog and specific genes as well. It is possible to filter results based on the HMM raw score or the e-value. Running in a time course. All results may be viewed in tabular form or as customizable bar charts that can be saved as a PNG image. An option to compute an ‘ortholog core’ from a set of organisms is also provided. It is also possible to include manual or experimental annotation in the database, e.g. if users have experimentally verified predictions, these can be marked and stored. In this manner, the database also helps to maintain a record of manually verified motifs and experimentally verified predictions. Detailed illustrative tutorials and sample data are provided on the web site and also in the accompanying manual.

Augur provides an integrated view of high quality surface protein predictions for microbial genomes and the possibility to perform inter-genome comparisons. The pipeline and the database may also be downloaded and installed locally in case users are interested to add more genomic data (e.g. unpublished genomes) and for faster access. Future versions of Augur will incorporate additional gram-negative and eukaryotic genomes into its database.

ACKNOWLEDGEMENTS

We thank M. Maier for excellent technical support. T.C. acknowledges support from the Bundesministerium für Bildung und Forschung, Germany (NGFN 01GS0401), T.C. and T.H. from Pathogenomics (PTJ-Bio/03U213B) and R.G. from the Graduate College of Biochemistry of Nucleoprotein Complexes (GK370), Justus-Liebig-University, Giessen, Germany.

Conflict of Interest: none declared.

REFERENCES