Prophinder

Lima-Mendez, Gipsi; Van Helden, Jacques; Toussaint, Ariane; Leplae, Raphaël

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/btn043

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 6 2008, pages 863–865
doi:10.1093/bioinformatics/btn043

Genome analysis

Prophinder: a computational tool for prophage prediction in prokaryotic genomes

Gipsi Lima-Mendez*, Jacques Van Helden, Ariane Toussaint and Raphaël Leplae

Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe), Université Libre de Bruxelles, 1050 Bruxelles, Belgium

Received on December 10, 2007; revised on January 8, 2008; accepted on January 26, 2008

Advance Access publication January 30, 2008

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Prophinder is a prophage prediction tool coupled with a prediction database, a web server and web service. Predicted prophages will help to fill the gaps in the current sparse phage sequence space, which should cover an estimated 100 million species. Systematic and reliable predictions will enable further studies of prophages contribution to the bacteriophage gene pool and to better understand gene shuffling between prophages and phages infecting the same host.

Availability: Softare is available at http://aclame.ulb.ac.be/prophinder

Contact: gipsi@scmbb.ulb.ac.be

Supplementary information: Supplementary data is available on http://aclame.ulb.ac.be/Tools/Prophinder/evaluations_table.html.

1 INTRODUCTION

Phages are viruses infecting prokaryotes. The sub-group of temperate phages has the capability to remain in their host, in a latent stage, as prophages. Most of the prophages are found integrated in the host chromosome while some are established as plasmids. Whether functional or defective, prophages can recombine with other phages and/or prophages, a central mechanism in bacteriophage evolution (Hendrix *et al.*, 1999).

At present, the coverage of the phage sequence space remains very narrow and yet, phages are the most abundant organisms on Earth. The observed diversity gives us only a hint on the real variety of the phage population, estimated to a 100 million species (Rohwer, 2003). Previous studies support the view that a common gene pool is available to double stranded (ds) DNA tailed phages (Hendrix *et al.*, 1999). A significant portion of this pool resides in prophages, providing many recombination opportunities for other prophages residing in the same host or infecting phages. Detecting prophages in prokaryotic genomes will therefore largely expand the phage sequence space and facilitate studies on gene exchange within the phage population.

We present Prophinder, an algorithm that combines similarity searches, statistical detection of phage-gene enriched regions and genomic context for prophage prediction. A database with prophage predictions in sequenced prokaryotic genomes has been developed with a Web interface for browsing the results. Prophinder can also be accessed via a programmatic interface (Web services), ensuring interoperability with other software tools.

*To whom correspondence should be addressed.

2 ALGORITHM

Prophinder is written in Perl and uses the ACLAME database (Leplae *et al.*, 2004) as source of phage data for similarity searches, gene annotation and detection of conserved pairs of genes are found in phage genomes.

2.1 Input data

The algorithm takes as input, a prokaryotic genome sequence in GenBank format with annotated positions of genes and coding sequences (CDSs).

2.2 Detection of phage-like CDSs in the prokaryotic genome

Phage-like CDSs in a prokaryotic genome are identified by gapped BLASTP search (Altschul *et al.*, 1997) of all the translated CDSs from the input genome against all phage proteins in ACLAME.

2.3 Detecting phage-like dense regions (PGDRs)

Our method of prophage prediction is based on the detection of genomic segments statistically enriched in phage-like genes. Each set of n consecutive CDSs is modeled as succession of n trials (CDSs) that can each result in a success (phage-like) or a failure (i.e. not phage-like). The exceptionality of the enrichment is estimated with the binomial P-value, which represents the probability to observe by chance at least s phage-like CDSs in a set of n consecutive CDSs.

$$P\text{-value} = P(X \geq s) = \sum_{i=s}^{n} C_n^i p^i (1-p)^{n-i} \quad (1)$$

The probability of success p is estimated by computing the average density of phage-like genes, i.e. the number of phage-like CDSs divided by the total number of CDSs. The P-value can be interpreted as a risk of false positive, for a genome segment covering n CDSs, and starting at a given CDS. We compute this P-value for every segment starting at any CDS of

the genome, and for all n values in a given range of k window widths (e.g. $W_{min} = 5$ and $W_{max} = 300$). For a genome containing G genes, the number of segments considered is:

$$T = G \cdot k = G \cdot (W_{max} + 1) \quad (2)$$

Since the binomial test is applied for each segment, the *P*-value has to be corrected for multi-testing. For this, we compute the *E*-value, which indicates the expected number of false positives for a set of T tests. A logarithmic transformation defines the significance index (*sig*) of the segment.

$$sig = -\log(E - \text{value}) = -\log(P - \text{value} \cdot T) \quad (3)$$

The *sig* values of all segments are stored in a matrix M of m rows and k columns, i.e. one row per starting CDS and one column per window width. Negative *sig* values are not considered for further analysis. Each cell of the matrix containing a positive *sig* corresponds to a potential prophage starting at the CDS defined by the row and ending at the CDS defined by the column. The matrix is scanned to detect local maxima and the corresponding genome segments are selected as phage-like dense regions (PGDRs).

2.4 Selecting the putative prophages

All PGDRs are sorted by decreasing *sig* values. Mutually overlapping PGDRs are then compared on the basis of hierarchical rules: (1) PGDRs containing an integrase gene always take precedence over overlapping PGDRs lacking the integrase gene; (2) PGDRs with higher *sig* have the precedence over those of lower *sig*. The next step separates tandem prophages found in one single PGDR, based on the presence of integrase genes and/or instances of conserved gene pairs found in the ACLAME phage genomes.

2.5 Iterative process

Some PGRDs with lower *sig* values, often representing small prophages or prophage remnants, may escape the selection. These can be recovered through an iterative process, by running a new round of selection each time, on the same scoring matrix where PGDRs selected in the previous iteration are masked (*sig* scores set to −1). The iterative process stops either when no new PGDR is detected or when the number of iterations set by the user is reached.

2.6 Secondary search

For more exhaustive prophage detection in bacterial genomes, Prophinder can load a set of prophages and mask them when counting the phage-like CDSs. This lowers the *expected probability (p)* of phage-like CDSs in the genome, leading to additional putative prophages from the new significance matrix. This option may be useful for genomes with high average density of phage-like CDSs. Prophages predicted using this option need to be analyzed cautiously. The secondary search can be immediately run after completing the predictions using the predicted prophages as the source for masking the phage-like CDSs.

2.7 Consensus

Prophinder is by default executed with several maximum window sizes (300, 200, 100, 50 and 20 CDS). A consensus is then created by combining the results from all the predictions. The consensus is the default solution proposed to the users.

3 ASSESSMENT OF PROPHINDER PERFORMANCE

Prophinder predictions were evaluated against a collection of annotated prophages provided by S. Casjens [extension from (Casjens, 2003)]. The Supplementary Table (accessible at http://aclame.ulb.ac.be/Tools/Prophinder/evaluations_table.html) provides the evaluation procedure and results. This evaluation gives a sensitivity of 79% and a positive predictive value of 94% for Prophinder. For the sake of comparison, Phage_Finder (Fouts, 2006) features a sensitivity of 67% and a positive predictive value of 94% under the strict settings. However, it cannot be ruled out that with other combinations of parameters Phage_Finder can reach higher sensitivity than Prophinder. The two methods are thus capable of detecting a large number of prophages in bacterial genomes while producing very few false positives. Many false negatives in both methods consist of small prophages with only few genes similar to those of known phages. This limitation is expected for such methods, but is likely to be reduced when more phages will be annotated. Prophinder, as an additional asset, is capable of detecting tandem prophages as such. Moreover, the execution time for predicting prophages in a genome such as *Escherichia coli O157:H7 EDL933* (NC_002655) on a Pentium 4 2.6 GHz with 1 GB of memory is around 40 min for Prophinder and 130 min for Phage_Finder.

4 DATABASE AND WEB INTERFACE

A relational database has been developed to store all Prophinder predictions. A Web interface (http://aclame.ulb.ac.be/prophinder) allows for browsing the prediction results. A GenBank entry can be submitted to the server (http://aclame.ulb.ac.be/perl/Aclame/Prophages/prophinder.cgi) for running Prophinder and view the prediction results. Execution parameters can be set such as the BLASTP *E*-value threshold, the maximum window sizes to be used for screening the genome, the secondary search option, etc. All the options are documented on the web form. The Prophinder database is regularly updated with predictions from newly sequenced genomes using the default parameters.

5 WEB SERVICE

To facilitate the use of Prophinder in automated processes, such as an annotation pipeline, a Web service has been developed. The WSDL, which is the primary programmatic interface for the web clients, defines the services and is accessible at http://aclame.ulb.ac.be/prophinder/prophinder.wsdl. The Web service allows running remotely Prophinder on GenBank entries. The predictions stored in the Prophinder database can be retrieved

through the web service as well. Two Perl clients are available for download to use the Prophinder web service.

ACKNOWLEDGEMENTS

We are grateful to Sherwood Casjens for providing the manually annotated prophages data and to Olivier Sand and Morgane Thomas-Chollier for their help in the web service development. Our work is supported by ESA-PRODEX (contract C90254), the Fonds de la Recherche Fondamentale Collective (FRFC), the Actions de Recherche Concertés du Ministre de la Communauté Française de Belgique and the Université Libre de Bruxelles (ULB). G.L.M. was supported by the Fonds Xenophilia, ULB.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res*., **25**, 3389–3402.

Casjens,S. (2003) Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol*., **49**, 277–300.

Fouts,D.E. (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucl. Acids Res*., **34**, 5839–5851.

Hendrix,R.W. *et al*. (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl Acad. Sci. USA*, **96**, 2192–2197.

Leplae,R. *et al*. (2004) ACLAME: a CLAssification of mobile genetic elements. *Nucl. Acids Res*., **32** (Database issue), D45–D49.

Rohwer,F. (2003) Global phage diversity. *Cell*, **113**, 141.