BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 3 2003, pages 418–420
DOI: 10.1093/bioinformatics/btg004

IslandPath: aiding detection of genomic islands in prokaryotes

William Hsiao[1], Ivan Wan[2], Steven J. Jones[2] and Fiona S. L. Brinkman[1,*]

[1]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada V5A 1S6 and [2]Genome Sciences Centre, B.C. Cancer Agency, Vancouver, BC, Canada V5Z 4E6

Received on September 4, 2002; accepted on September 29, 2002

Downloaded from https://academic.oup.com/bioinformatics/article/19/3/418/258350 by guest on 21 August 2022

ABSTRACT

Summary: Genomic islands (clusters of genes of potential horizontal origin in a prokaryotic genome) are frequently associated with a particular adaptation of a microbe that is of medical, agricultural or environmental importance, such as antibiotic resistance, pathogen virulence, or metal resistance. While many sequence features associated with such islands have been adopted separately in applications for analysis of genomic islands, including pathogenicity islands, there is no single application that integrates multiple features for island detection. IslandPath is a network service which incorporates multiple DNA signals and genome annotation features into a graphical display of a bacterial or archaeal genome, to aid the detection of genomic islands.

Availability: This application is available at http://www.pathogenomics.sfu.ca/islandpath and the source code is freely available, under GNU public licence, from the authors.

Contact: brinkman@sfu.ca

Supplementary information: An online help file, which includes analyses of the utility of IslandPath, can be found at http://www.pathogenomics.sfu.ca/islandpath/current/islandhelp.html

Pathogenicity islands (PAIs), first discovered and named in uropathogenic *Escherichia coli* in the late 1980s (Hacker *et al.*, 1990), have been studied intensively since they represent an intersection of two notable biological phenomena, namely bacterial pathogenesis and horizontal gene transfer (HGT). As the name implies, genes associated with bacterial pathogenesis were found clustered in such islands and accumulating evidence suggests that such islands have horizontal origins (Blum *et al.*, 1994; Sullivan and Ronson, 1998). As more prokaryotic genome sequences became available, the concept of PAIs was extended to other genetic elements that share the general structure of PAIs but encode functions other than virulence. These genetic elements, collectively called genomic islands, encode genes involved in diverse cellular functions such as secondary metabolism (metabolism islands), antibiotic resistance (resistance islands), and secretion (secretion islands; Hentschel and Hacker, 2001). The prevalence of these elements in bacteria varies from an estimated 17% in *E. coli*, to none identified to date in some obligate intracellular species such as *Rickettsia* and *Chlamydia* (Ochman *et al.*, 2000).

Features commonly associated with genomic islands include the presence of flanking repeats, mobility genes (e.g. integrases, transposases), proximal transfer RNAs (tRNAs), and atypical guanine and cytosine content (%G+C; Hacker *et al.*, 1997). tRNAs are known phage integration sites (Reiter *et al.*, 1989), and they may also serve as integration sites for mobile genetic elements that become PAIs. %G+C and additional species-specific 'DNA signatures' (e.g. dinucleotide-bias and codon usage profile) have also been proposed to be useful in identifying islands (Lio and Vannucci, 2000; Karlin, 2001), which frequently exhibit distinct DNA signatures from the rest of the genome. While many of these features have been adopted separately in applications for analyzing genomic islands (Lio and Vannucci, 2000), there is no single computational tool that integrates multiple features for island detection. We hypothesize that by combining and overlapping these features, we can more easily and accurately identify genomic islands of interest. We therefore developed IslandPath, a web-accessible service, to graphically display island-associated features in full-genome context. The current version permits analysis of all currently available fully sequenced bacterial and archaeal genomes (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html) and the dataset is updated regularly.

Based on a literature survey of known islands, we first opted to display in IslandPath the %G+C of predicted open reading frames (ORFs), dinucleotide bias for gene-clusters

*To whom correspondence should be addressed.

 © Oxford University Press 2003; all rights reserved.

(see below), the location of known or probable mobility genes, and the location of tRNAs. These features are represented by different symbols in a compact graphical display at gene resolution (see online help file). Users can click on the graph to select a region of interest to view corresponding gene annotations of the ORFs (from the National Centre for Biotechnology Information (NCBI) ftp site) in table-format below the graph. A web-link to the NCBI record of each ORF is incorporated into the table to facilitate further analyses (e.g. BLAST, Pubmed) of the ORF. IslandPath provides a quick way to browse a genome for putative islands that may be of interest for further computational and/or 'wet-lab' characterization.

For %G+C and dinucleotide bias calculations, other programs frequently use a fixed sliding window as a basic calculation unit. IslandPath uses a single ORF and a cluster of ORFs as basic units for calculating %G+C and dinucleotide bias, respectively. This permits us to analyze potential gene-by-gene or gene-cluster variance in proposed coding sequences. For %G+C calculation, users can set their own cut-off values for 'high' or 'low' %G+C, with a default cut-off derived from analysis done with the chlamydial genomes (see IslandPath's online help file for details). Since *Chlamydia* species are noted for their apparent lack of HGT (Brinkman *et al.*, 2002; Read *et al.*, 2000), we hypothesize that the %G+C variance observed in their genomes represents variance due predominantly to other selective effects, such as gene expression level.

Dinucleotide bias analysis was also incorporated as it represents an independent DNA signature that is not subject to the same influences as %G+C analysis. Our dinucleotide bias analysis method was adapted from formulas published by Samuel Karlin (for review see (Karlin, 2001)). Briefly, a genome is divided into 'ORF-clusters' of 6 consecutive ORFs. The average absolute dinucleotide relative abundance difference ($\delta^*(f, g)$) derived from the following formula was calculated for each ORF-cluster in the genome:

$$\delta^*(f, g) = \frac{1}{16} \sum |\rho^*_{xy}(f) - \rho^*_{xy}(g)|$$

where f (fragment) was derived from sequences in an ORF-cluster and their reverse complements and g (genome) was derived from all predicted ORFs and their reverse complements in the genome. The dinucleotide relative abundance, ρ^*_{xy}, was calculated from the formula ($\rho^*_{xy} = f^*_{xy}/f^*_x f^*_y$) where f^*_x denotes the frequency of the mononucleotide x and f^*_{xy} the frequency of the dinucleotide xy in each ORF-cluster or genome. The sampling of ORF-clusters was done using a sliding window, shifting by one ORF at a time. The mean $\delta^*(f, g)$ is calculated by averaging the results from all ORF-clusters in the genome, and regions with $\delta^*(f, g)$ greater than 1 standard deviation away from the mean are marked on the IslandPath graphical display (see online help file for details). We decided to use six ORFs as a cluster because single ORF dinucleotide bias is highly variable and previous codon based analysis has shown that a minimum cluster of genes of approximately 4.5 kb (corresponding to approximately 6–8 ORFs) is required for reliable estimation of nucleotide composition (Lawrence and Ochman, 1997).

Genome annotation features such as structural RNAs (tRNAs and ribosomal RNAs) and mobility genes (transposases and integrases) are part of the IslandPath genome view. RNA location information is obtained from NCBI and supplemented by tRNAscan-SE (Lowe and Eddy, 1997) if tRNA information is missing. Known or probable mobility genes are identified by keyword scanning of the NCBI genome annotations and supplemented with COG classification information (Tatusov *et al.*, 1997) if no annotation is available.

IslandPath has several advantages over other types of analyses. First, certain DNA signal based approaches by themselves can be poor indicators of HGT (Koski *et al.*, 2001) so IslandPath complements multiple DNA signal analyses with additional annotation features. Second, IslandPath provides an easily accessible web interface for convenient visualization and analysis of genomic islands by microbiology researchers. Lastly, its graphical view is relatively compact to allow visualization of putative islands in a full genome context. Of course, IslandPath can not be effective in detecting HGT of individual genes, or islands obtained from organisms with similar DNA signals. However, notably, many of the classic virulence genes are in genomic regions with unusual DNA signals (see IslandPath's online help file for examples). We propose that IslandPath may be a useful tool for facilitating the detection and analysis of genomic islands and, as more genomes become available, IslandPath, which can be updated automatically, will continue to be a resource for such island analysis for the research community.

ACKNOWLEDGEMENTS

We thank the NCBI (USA), particularly Tatiana Tatusov, for providing helpful files for IslandPath, and Francis Ouellette and Brett Finlay (UBC, Canada) for helpful suggestions. We also wish to acknowledge the efforts of the many genome sequencing projects that have made our analysis possible. This work was funded by the Peter Wall Institute for Advanced Studies.

REFERENCES

Blum,G., Ott,M., Lischewski,A., Ritter,A., Imrich,H., Tschape,H. and Hacker,J. (1994) Excision of large DNA regions termed pathogenicity islands from tRNA- specific loci in the chromosome of an *Escherichia coli* wild-type pathogen. *Infect. Immun.*, **62**, 606–614.

Downloaded from https://academic.oup.com/bioinformatics/article/19/3/418/258350 by guest on 21 August 2022

Brinkman,F.S., Blanchard,J.L., Cherkasov,A., Greberg,H., Av-Gay,Y., Brunham,R.C., Fernandez,R.C., Finlay,B.B., Otto,S.P., Ouellette,B.F., Keeling,P.J., Rose,A.M., Hancock,R.E. and Jones,S.J. (2002) Evidence that plant-like genes in *Chlamydia* species reflect an ancestral relationship between Chlamydiaceae, Cyanobacteria, and the Chloroplast. *Genome Res.*, **12**, 1159–1167.

Hacker,J., Bender,L., Ott,M., Wingender,J., Lund,B., Marre,R. and Goebel,W. (1990) Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal *Escherichia coli* isolates. *Microb. Pathog.*, **8**, 213–225.

Hacker,J., Blum-Oehler,G., Muhldorfer,I. and Tschape,H. (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol.*, **23**, 1089–1097.

Hentschel,U. and Hacker,J. (2001) Pathogenicity islands: the tip of the iceberg. *Microbes Infect.*, **3**, 545–548.

Karlin,S. (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol.*, **9**, 335–343.

Koski,L.B., Morton,R.A. and Golding,G.B. (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol.*, **18**, 404–412.

Lawrence,J.G. and Ochman,H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.*, **44**, 383–397.

Lio,P. and Vannucci,M. (2000) Finding pathogenicity islands and gene transfer events in genome data. *Bioinformatics*, **16**, 932–940.

Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.

Ochman,H., Lawrence,J.G. and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.

Read,T.D., Brunham,R.C., Shen,C., Gill,S.R., Heidelberg,J.F., White,O., Hickey,E.K., Peterson,J., Utterback,T., Berry,K. *et al.* (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.*, **28**, 1397–1406.

Reiter,W.D., Palm,P. and Yeats,S. (1989) Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic Acids Res.*, **17**, 1907–1914.

Sullivan,J.T. and Ronson,C.W. (1998) Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proc. Natl Acad. Sci. USA*, **95**, 5145–5149.

Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Downloaded from https://academic.oup.com/bioinformatics/article/19/3/418/258350 by guest on 21 August 2022