S 6-5

A Computational Approach for Identifying Pathogenicity Islands in Prokaryotic Genomes

Sung Ho Yoon, Cheol-Goo Hur, Ho-Young Kang, Yeoun Hee Kim, Tae Kwang Oh[1] and Jihyun F. Kim

Genome Research Center and [1]21C Frontier Microbial Genomics and Applications Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), 52 Oun-dong, Yuseong, Daejeon 305-333, Korea

Pathogenicity islands (PAIs) are distinct genetic elements of pathogens encoding various virulence factors such as protein secretion systems, host invasion factors, iron uptake systems, and toxins (1,2). PAIs are a subset of genomic islands (GIs) which have been transferred by horizontal gene transfer event and confer virulence upon the recipient. PAIs can be identified by features such as the presence of virulence genes, biased G+C content and codon usage, carriage of mobile sequence elements, and/or association with tRNA genes or repeated sequences at their boundaries (1). Identification of pathogenicity islands (PAIs) is essential in understanding the development of disease and the evolution of bacterial pathogenesis (2). Up to now, attempts to identify PAIs (3-5) have been made by detecting genomic regions which only differ from the rest of the genome in their base composition and codon usage. Considering that a PAI is a GI encoding virulence factors, compositional criteria such as G+C content and codon usage is not sufficient for identifying PAIs because genomic approaches can only lead to the identification of GIs (2).

In this work, we designed a computational method for identifying PAIs in sequenced genomes by combining a homology-based method and detection of abnormalities in genomic composition (6). To do this, we first collected 207 GenBank accessions containing either part or all of the reported PAI loci. In sequenced genomes, strips of PAI-homologs were defined based on the proximity of the homologs of genes in the same PAI accession. Overlapping or adjacent genomic strips were then merged into a large genomic region. Among the defined genomic regions, PAI-like regions were identified by the presence of homolog(s) of virulence genes. Also, GIs were postulated by calculating G+C content anomalies and codon usage bias. Of 148 prokaryotic genomes examined, 23 pathogenic and 6 non-pathogenic bacteria contained 77 candidate PAIs (cPAIs) that partly or entirely overlap GIs. Supporting the validity of our method, included in the list of cPAIs were thirty four PAIs previously identified from genome sequencing papers. Furthermore, in some instances, our method was able to detect entire PAIs for those only partial sequences are available.

Our method used for finding PAI-homologous regions is reminiscent of the sequence-assembly procedure and proven to be an efficient method for demarcating the potential PAIs in our study. The function and origin of a cPAI can be inferred by investigating the known PAIs comprising it. The detection

capacity of our approach can be easily increased by adding more data to the query data set. We are currently improving the detection scheme and are developing a database for cPAIs in sequenced genomes. With the availability of rapidly increasing complete genome sequences as well as PAI data, the proposed method will be useful in identifying potential PAIs in microbial genomes.

Acknowledgements

This work was supported by the 21C Frontier Microbial Genomics and Applications Center Program (grant MG02-0402-001-1-0-0), Ministry of Science and Technology, Republic of Korea

References

1. Hacker, J. and Kaper, J.B. (2002) Pathogenicity islands and the evolution of pathogenic microbes. Springer-Verlag, Berlin.
2. Schmidt, H. and Hensel, M. (2004) Pathogenicity islands in bacterial pathogenesis. *Clin. Microbiol. Rev.*, 17, 14-56.
3. Karlin, S. (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol.*, 9, 335-343.
4. Lio, P. and Vannucci, M. (2000) Finding pathogenicity islands and gene transfer events in genome data. *Bioinformatics*, 16, 932-940.
5. Tu, Q. and Ding, D. (2003) Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. *FEMS Microbiol. Lett.*, 221, 269-275.
6. Yoon, S.H., Hur, C.-G. Kang, H.-Y., Kim, Y.H., Oh, T.K., and Kim, J.F. "A computational approach for identifying pathogenicity islands in prokaryotic genomes" submitted.