

Identification and Characterization of Conserved Overlapping Genes in *Vibrio* Genomes

Yi-Feng Chang¹ and Chuan-Hsiung Chang²

Keywords: Genome comparison, Overlapping genes, *Vibrio* species

1 Introduction.

Overlapping genes are abundant in viruses, mitochondria, bacterial chromosomes and plasmids. There are two explanations to depict the existence of overlapping genes: 1) genome compactness [1] and 2) transcriptional or functional coupling [2]. However, a recent study had shown that the genome compactness did not correlate well with the overlapping frequencies [3]. Therefore, in this research, we tried to discover the conserved overlapping genes in microbial genomes for functional analysis and bidirectional promoter study. In previous studies, overlapping genes are usually defined as the adjacent coding sequences (CDS) that overlap partially and share one or more nucleotides [4]. Nevertheless, the complete gene structure should include both the transcription regulatory regions at the 5' upstream end and the termination region at the 3' downstream end of coding sequences. In addition, genes for ribosomal RNA (rRNA) and transfer RNA (tRNA) should also be identified as functional genes in genome sequences. Therefore, we extended the definition of overlapping genes to include coding sequences, ribosomal RNAs and transfer RNAs. Furthermore the overlapping region of coding sequence was expanded to allow longer spacing between two adjacent genes to include potential -10, -35 signals and transcription terminators. Under these constraints, we reinvestigated the properties of overlapping genes in completely sequenced *Vibrio* genomes. In addition, we extracted the overlapping gene pairs for similarity search to identify the conserved overlapping genes among all sequenced species.

2 Materials and Methods

Five completed *Vibrio* genome sequences from *Vibrio cholerae* O1 biovar eltor str. N16961, *Vibrio fischeri* ES114, *Vibrio parahaemolyticus* RIMD 2210633, *Vibrio vulnificus* CMCP6, and *Vibrio vulnificus* YJ016 (totally ten chromosomes) were downloaded from NCBI genome database in GenBank format. The overlapping genes were classified into three separate categories: 1) 'convergent' ($\rightarrow\leftarrow$), 2) 'unidirectional' ($\rightarrow\rightarrow$ or $\leftarrow\leftarrow$), and 3) 'divergent' ($\leftarrow\rightarrow$) [4]. To cover both the transcriptional initiation and termination sites of all the gene structures, we expanded the overlapping regions to allow adjacent genes be overlapped partially within 30 base pairs or share upstream or downstream regions in a length of 200 base pairs. All of the coding sequences, ribosomal RNAs, transfer RNAs and their positions in genome sequences and annotated functions (or products) were extracted, sorted and parsed into SQL database. Using both the start and stop positions of all neighboring genes, the overlapping and spacing distances were calculated. All of the gene pairs and the overlapping/spacing information were imported into SPSS for further statistical analysis. In addition, the nucleotide and amino acid sequences of three types of adjacent gene pairs were extracted and saved into FASTA format for further tblastx alignment and ClustalW-MPI analysis to identify all the conserved adjacent gene pairs. For bidirectional promoter

¹Institute of Bioinformatics, National Yang-Ming University, Taiwan, R.O.C.
E-mail: ian@gel.ym.edu.tw

²Institute of Bioinformatics, National Yang-Ming University, Taiwan, R.O.C.
E-mail: cchang@ym.edu.tw

analysis, only divergent adjacent genes were used. All these sequence alignment procedures were performed on an Apple Xserve G5 cluster.

3 Results

Based on the criteria used in Section 2, there were totally 17,026 gene pairs extracted. Figure 1 illustrated the size frequency distribution of overlapping and spacing sizes. The most frequently found overlapping sizes are 4 bps (908, 5.3%), 1 bps (555, 3.3%), and 8 bps (258, 1.5%). The most frequently discovered spacing sizes between neighboring genes are 2 bps (304, 1.8%), 9 bps (279, 1.6%) and 3 bps (236, 1.4%). In addition, Table 1 depicted the frequency distribution of the three different types of directional patterns. All of the gene pairs were used to identify the conserved overlapping genes. Furthermore, the 1,633 divergent gene pairs were used to predict all the bidirectional promoters.

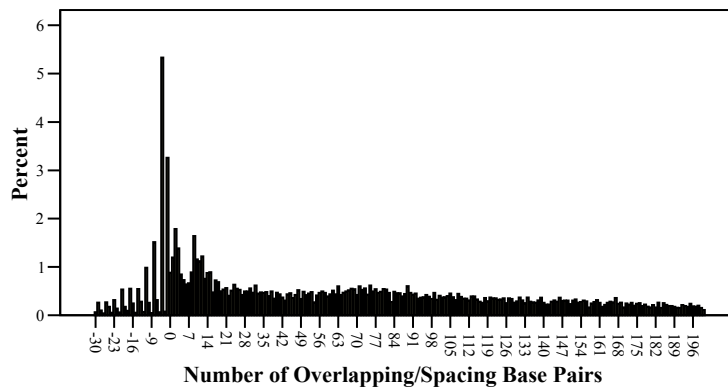


Figure 1: Size frequency distribution of the overlapping genes

	Frequency	Percent
Unidirectional ($\rightarrow\rightarrow$)	6071	35.7
Unidirectional ($\leftarrow\leftarrow$)	6233	36.6
Convergent ($\rightarrow\leftarrow$)	3059	18.0
Divergent ($\leftarrow\rightarrow$)	1663	9.8
Total	17026	100.0

Table 1: Distribution frequency of three directional pattern types

4 Acknowledgments

This work is supported by a NRPGM (National Research Program for Genomic Medicine) grant (NSC 93-3112-B-010-011) from the National Science Council of the R.O.C.

References

- [1] Krakauer, D.C., 2000, Stability and evolution of overlapping genes. *Evol. Int. J. Organ. Evol.* 54, 731-739.
- [2] Inokuchi, Y., Hirashima, A., Sekine, Y., Janosi, L., Kaji, A., 2000. Role of ribosome recycling factor (RRF) in translational coupling. *EMBO J.* 19, 3788-3798.
- [3] Johnson, Z.I., Chisholm, S.W., 2004. Properties of overlapping genes are conserved across microbial genomes. *Genome Research* 14, 2268-2272.
- [4] Fukuda, Y., Nakayama, Y., Tomita, M., 2003. On dynamics of overlapping genes in bacterial genomes. *Gene* 323, 181-187.