

Fusion genes as putative microbial drug targets in *H. pylori*

Meena Kishore Sakharkar
Nanyang Technological University,
Singapore
mmeena@ntu.edu.sg

Kishore R. Sakharkar
NUMI, National University of Singapore,
Singapore
phskrs@nus.edu.sg

Vincent T.K. Chow
Programme in Infectious Diseases,
National University of Singapore,
Singapore
micctk@nus.edu.sg

Abstract

Fusion genes have been reported as a means of enabling the development of novel or enhanced functions. In this report, we analyzed fusion genes in the genomes of two Helicobacter pylori strains (26695 and J99) and identified 32 fusion genes that are present as neighbours in one strain (components) and are fused in the second (composite), and vice-versa. The mechanism for each case of gene fusion is explored. All the genes identified as fusion products in this analysis were reported as essential genes in this bacterium in the previously documented transposon mutagenesis of H. pylori strain G27. This observation suggests the potential of the products of fusion genes as putative microbial drug targets. These results underscore the utility of bacterial genomic sequence comparisons for understanding gene evolution and for in silico drug target identification in the post-genomic era.

Keywords: Drug targets, Essential genes, Fusion genes, *Helicobacter pylori*, Intergenic DNA, Overlapping genes

1. Introduction

The event of bringing together two separate genes into a single gene (gene fusion) has long been identified as a potentially important evolutionary phenomenon [1-2]. Gene fusion events have been proposed to represent a valuable "Rosetta stone" information for the identification of potential protein-protein interactions and metabolic

or regulatory networks [3-4]. Recently, Suhre and Claverie [5] described a database on gene fusion events in bacteria and archaea. Fusion genes gain added advantage by coupling biochemical reactions through tight regulation of fusion partners, compared to individual partners [6]. Yanai *et al.* [7] used gene fusion to establish links between fusion genes and functional networks with their involvement. Gene fusion has also been used to illustrate novel gene function [8], enhanced substrate specificity [9] and multi-functional enzyme specificity [10]. Many databases compile information on gene fusion events integrated with phylogenomic profiling and the identification of conserved chromosomal localization, to propose hypotheses for the characterization of proteins of unknown function and of fusion genes across bacterial and eukaryotic genomes [11-15]. Thus, the formation of complex composite proteins by gene fusion is a dominant process in protein evolution.

H. pylori is one of the most common bacterial pathogens of humans that colonizes the gastric mucous membrane and induces chronic gastric inflammation that can progress to gastric ulcer, peptic ulcer and mucosa-associated lymphoma. Here we report gene fusion analyses for two strains of *H. pylori*, i.e. 26695 [16] and J99 [17]. The two strains of *H. pylori* show highly conserved gene order. For example, 84.7% of the 1479 genes in strain J99 each have the same neighbour on each side in the genomes of both strains. Only 8-10 genome rearrangements consisting of both inversions and translocations have been reported in J99 when compared to 26695 [17]. The absence of extensive gene shuffling between *H. pylori* strains J99 and 26695 is consistent with a low level of evolutionary distance. Studying gene fusion at such close evolutionary distances offers the advantage of “genome conservation” (both the conservation of sequence and gene content between two genomes). It also circumvents the tedious process of generating phylogenetic trees to infer orthology, which is the first step in detecting the fusion or fission events across two or more genomes. The 32 identified cases of gene fusion in the two *H. pylori* genomes and their mechanisms are presented. An analysis on the essentiality of the identified fusion genes revealed most of them to be essential. This observation suggests the potential of the products of fusion genes as putative microbial drug targets.

2. Materials and methods

The genome sequences of the two *H. pylori* strains, i.e. 26695 and J99, were downloaded from the National Center for Biotechnology Information (NCBI) website (<ftp://ftp.ncbi.nlm.nih.gov/genomes/bacteria>). A brief overview of the genomic features of both the genomes is presented in Table 1. The protein sequences of each of the respective genomes were searched against each other using BLASTP at an E value cutoff of $\leq 10^{-10}$. This experiment identified proteins that occur as fusion products in one genome (composites) and as separate gene products in the second genome (components), and vice versa. The coding sequence (CDS) feature annotation was used to extract the component genes showing overlap or separated by intergenic regions [18]. Cases of fusion that cover at least 50% of the length of composite protein were selected. In this analysis, we did not take into consideration component genes that are not neighbours and are involved in gene fusion. Moreover, since there are only two genomes of strains of *H. pylori* for comparison it is noted that fusion cannot be distinguished from fission in our analysis.

Aligning the protein sequences and mapping back to nucleotide sequences led to the identification of the location of components on the composite proteins. The components were classified into three categories, i.e. firstly, those juxtaposed with zero intergenic regions; secondly, those with intergenic regions of length 1 bp or more; and third, components with overlapping reading frames. Mapping of composite and component genes in relation to the recently published experimental data for transposon mutagenesis of *H. pylori* strain G27 was performed to check the biological functions of these genes in *H. pylori* [19].

3. Results and discussion

The genomes of the two strains of *H. pylori* are highly co-linear, and most of the orthologous genes appear in the same order in the genomes of both strains. The overall genomic organization, gene order and predicted proteomes are quite similar [17]. Table 1 summarizes the genome size, number of CDS, number of overlapping genes with directions of overlap and number of fusion cases identified in each of the genomes. Thirty two cases of gene fusion were identified in total. Fifteen of them occur as fusion genes in *H. pylori* J99 but are split in *H. pylori* 26695, whilst 17 cases occur as fusion genes in 26695 but are split in J99 (Tables 1). These genes appear

as juxtaposed component genes (overlapping or non-overlapping) in one strain and are present as composite genes (fused) in the second strain. The observation of fusion genes across microbial genomes raises questions on their origin and the reasons for their fusion. The possible scenarios are elaborated upon.

Table 1. Features of the genomes of *H. pylori* strains J99 and 26695.

	<i>H. pylori</i> J99	<i>H. pylori</i> 26695
Genome size (bp)	1,643,831	1,667,867
Number of genes (CDS)	1,491	1,651
Overlapping genes (++/--)	156 / 141	179 / 177
Overlapping genes (+/-/+)	14 / 1	18 / 6
Number of fusion genes	15	17

3.1 Juxtaposed genes, intergenic DNA and gene fusion:

The mechanism of gene fusion in cases where the component genes are non-overlapping, juxtaposed and separated by intergenic DNA was investigated. It was found that fusion genes are formed by loss of the stop codon in one of the component genes. The absence of the stop codon results in extension of the 3' end of the gene's coding region. This loss of stop codon arises from a change in reading frame by insertion or deletion or point mutation in the stop codon. It is interesting to note that any intergenic regions between the components are absorbed into the coding regions in the composite proteins. This addition of extra DNA segments may be important in contributing to the expression of accreted functions in fusion genes as these intergenic regions have in some cases been reported to encode protein motifs [20]. Additional motifs in composite proteins may add new or enhanced functionality to the fusion gene product by contributing incremental structural architectures and functional capabilities that are reported as characteristic features of fusion gene products. The formation of a fusion gene product will also result in a domain-domain interaction rather than a subunit-subunit interaction if the two proteins are interacting. The transition may be thermodynamically favourable as fusion proteins acquire reduced entropy compared to their physically separated fusion partners.

Even though only about 15% of the typical bacterial genome is non-coding, the function of non-coding DNA remains poorly understood. These cases of gene fusion may be a step towards understanding the evolving roles of intergenic DNA in microbial genome evolution. The results hint at the plasticity of bacterial genomes.

3.2 Overlapping genes and gene fusion

Our results show that, as previously reported, many cases of gene fusion involve overlapping component genes [21]. All of the fusions that involve overlapping components have a single representation of the overlapped segment in the fusion gene product (composite protein), i.e. overlapping fragments in components meld into one composite gene. The frequent occurrence of the unidirectional overlapping structure probably reflects the most common orientation of adjacent genes in the chromosomes, as prokaryotic genes are often organized into operons i.e. clusters of genes that are transcribed together. Since all genes in an operon must be transcribed in the same direction, this organization will be reflected by a tendency for nearby genes to adopt the same orientation. Our results demonstrate that overlapping genes evolve as a result of the extension of an open reading frame caused by a switch to an upstream initiation codon, substitutions in initiation or termination codons and deletions. In all the three cases reported, it is observed that insertions or deletions (non multiples of 3) cause frameshifts that eliminate termination codons and cause readthrough transcription. Our results suggest the important role of overlapping genes in gene fusion. Thus, gene fusion of overlapping genes represents a potential source of new domains and domain accretion.

3.3. Fusion genes as potential microbial drug targets

Exploiting genome sequence data for the identification of new antimicrobial targets has received considerable attention from the pharmaceutical and biotechnological communities. Galperin and Koonin [22] alluded to the demonstration of essentiality of a particular gene as the first step towards using it as a possible drug target, and suggested the selection of virulence genes, membrane proteins, uncharacterized essential genes and species-specific genes as drug targets. The identification of two virulence proteins and two membrane proteins as fusion genes further motivated us to explore the essentiality of the remaining hypothetical proteins identified as fusion genes. Recently, Salama et al. [19] performed global transposon mutagenesis in *H. pylori* G27 and analyzed essential genes. Being indispensable, essential genes are more evolutionarily conserved than non-essential genes [Jordan et al. 2002]. This is because negative selection acting on essential genes is expected to be more stringent than on non-essential genes. We performed a match of the 32 identified cases of gene fusion to the essential gene list reported by Salama et al. [19]. This mapping of the composite proteins onto the transposon mutagenesis data for *H. pylori* G27 identified all but four genes as essential genes. Two of the four apparently inessential genes are annotated as FrpB and are iron-regulated outer membrane proteins. The remaining two are hypothetical proteins, thus implying that they are species-specific and no function could be assigned based on sequence homology. Since almost all of the identified fusion genes are reported as either essential genes [22] or membrane proteins or species-specific, which are proposed categories of putative microbial drug targets, the results suggest that gene fusion may serve as a marker for putative microbial drug target identification. These findings thus suggest that gene fusion can be exploited as a strategy to identify essential genes. Further filtration of the identified essential genes for the detection of non-homologous human genes represents a promising means of identifying novel drug targets and facilitates the search for new antibiotics. Concordance analyses on these essential gene products that show sets of proteins conserved across one set of user-specified genomes, but are absent in another set of user-specified genomes may help to address the “complexities and conundra” in identification of drug targets and pathogen-specific drugs.

4. Conclusions

Whole genome sequencing of microorganisms is providing an opportunity for computer-based genetic analyses that can highlight interesting features such as fusion genes in the genomes. Our analyses revealed that fusion genes have incremental structural and functional architectures. It also appears that intergenic DNA may have a significant role in these enhanced attributes. All of the identified fusion genes (components) are reported as essential genes in *H. pylori* strain G27 and their utility as potential drug targets is suggested.

5. Acknowledgements

The authors gratefully acknowledge Dr. N. Salama and Prof. S. Falkow (Fred Hutchinson Cancer Research Center, Seattle, WA, USA) for providing the transposon mutagenesis data for *H. pylori*. MKS acknowledges A*STAR-BMRC, Singapore Grant #03/1/22/19/242.

6. References

- [1] Yourno J.D. 1972. Gene fusion. Brookhaven Symp. Biol. 23: 95-120.
- [2] Isono K. and Yourno J. 1973. Mutation leading to gene fusion in the histidine operon of *Salmonella typhimurium*. J. Mol. Biol. 76: 455-461.
- [3] Sali A. 1999. Functional links between proteins. Nature 402: 23-26.
- [4] Galperin M.Y. and Koonin E.V. 2000. Who's your neighbor? New computational approaches for functional genomics. Nat. Biotechnol. 18: 609-613.
- [5] Suhre K. and Claverie J.M. 2004. FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. Nucleic Acids Res. 32: D273-D276.
- [6] Tsoka S. and Ouzounis C.A. 2001. Functional versatility and molecular diversity of the metabolic map of *Escherichia coli*. Genome Res. 11: 1503-1510.
- [7] Yanai I., Derti A. and DeLisi C. 2001. Genes linked by fusion events are generally of the same functional category: A systematic analysis of 30 microbial genomes. Proc. Natl. Acad. Sci. USA 98: 7940-7945.
- [8] Long M. 2000. A new function evolved from gene fusion. Genome Res. 10: 1655-1657.

- [9] Katzen F.M., Deshmukh F.D., Daldal F. and Beckwith J. 2002. Evolutionary domain fusion expanded the substrate specificity of the transmembrane electron transporter DsbD. *EMBO J.* 21: 3960-3969.
- [10] Berthonneau E. and Mirande M. 2000. A gene fusion event in the evolution of aminoacyl-tRNA synthetases. *FEBS Lett.* 470: 300-304.
- [11] Marcotte E.M. 2000. Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.* 10: 359-365.
- [12] Marcotte E.M., Pellegrini M., Ng H.L., Rice D.W., Yeates T.O. and Eisenberg D. 1999a. Detecting protein function and protein-protein interactions from genome sequences. *Science* 285: 751-753.
- [13] Enright A.J. and Ouzounis C.A. 2001. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol.* 2: RESEARCH0034.
- [14] Mellor J.C., Yanai I., Clodfelter K.H., Mintseris J. and DeLisi C. 2002. Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.* 30: 306-309.
- [15] von Mering C., Huynen M., Jaeggi D., Schmidt S., Bork P. and Snel B. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 31: 258-261.
- [16] Tomb J.F., White O., Kerlavage A.R., Clayton R.A., Sutton G.G., Fleischmann R.D., Ketchum K.A., Klenk H.P., Gill S., Dougherty B.A., Nelson K., Quackenbush J., Zhou L., Kirkness E.F., Peterson S., Loftus B., Richardson D., Dodson R., Khalak H.G., Glodek A., McKenney K., Fitzgerald L.M., Lee N., Adams M.D., Hickey E.K., Berg D.E., Gocayne J.D., Utterback T.R., Peterson J.D., Kelley J.M., Cotton M.D., Weidman J.M., Fujii C., Bowman C., Watthey L., Wallin E., Hayes W.S., Borodovsky M., Karp P.D., Smith H.O., Fraser C.M. and Venter J.C. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388: 539-547.
- [17] Alm R.A., Ling L.S., Moir D.T., King B.L., Brown E.D., Doig P.C., Smith D.R., Noonan B., Guild B.C., deJonge B.L., Carmel G., Tummino P.J., Caruso A., Uria-Nickelsen M., Mills D.M., Ives C., Gibson R., Merberg D., Mills S.D., Jiang Q., Taylor D.E., Vovis G.F. and Trust T.J. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* 397: 176-180.
- [18] Sakharkar M.K., Passetti F., de Souza J.E., Long M. and de Souza S.J. 2002. ExInt: an Exon Intron Database. *Nucleic Acids Res.* 30: 191-194.
- [19] Salama N.R., Shepherd B. and Falkow S. 2004. Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *J. Bacteriol.* 186: 7926-7935.
- [20] Zhang Z.L., Harrison P.M. and Gerstein M. 2002. Digging deep for ancient relics: a survey of protein motifs in the intergenic sequences of four eukaryotic genomes. *J. Mol. Biol.* 323: 811-822.
- [21] Fukuda Y., Nakayama Y. and Tomita M. 2003. On dynamics of overlapping genes in bacterial genomes. *Gene* 323: 181-187.
- [22] Galperin M.Y. and Koonin E.V. 1999. Searching for drug targets in microbial genomes. *Curr. Opin. Biotechnol.* 10: 571-578.