Chapter IV

Bioinformatic study of Pathogenicity related genes of three species of Helicobacter

4.1: Introduction

Isolated first by Marshall and Warren in 1982, from the gastric mucosa of a patient infected with gastritis and peptic ulceration, *Helicobacter pylori* represents a genus of gram negative, helix shaped, flagellated and motile, microaerophilic bacteria that persistently colonizes the human stomach (Warren & Marshall 1983). *Helicobacter pylori* have been linked with chronic antral gastritis and diseases like MALT lymphoma, peptic ulcer and gastric adenocarcinoma (Covacci *et al.,* 1999: Cover & Blaser 1999). More than 50% of the world's population harbor *H. pylori* in their upper gastrointestinal tract but merely 10% suffer from overt disease (Telford et al., 1994) Gastric cancer caused by H. pylori is a major cause of death worldwide particularly in the developing countries. The estimated current incidence of gastric cancer is approximately 16.2/100,000 persons per year (world standardized rate), with highest incidences in Eastern Asia, Eastern Europe, and South America (Ferlay *et al.,* 2004. The mechanisms of *H. pylori* transmission are incompletely characterized. Person-to-person transmission is most commonly implicated with fecal/oral, oral/oral, or gastric/oral pathways. The prevalence of H. pylori infection varies widely by geographic area, age, race, ethnicity, and SES. (Brown 2000)

The mechanisms by which bacterial infection leads to gastric mucosal damage include the direct effects of virulence factors produced by *H. pylori* like as *cag*A, *vac* A, or urease; the propagation and perpetuation of inflammation; oxidative stress; and the induction of apoptosis in infected gastric epithelial cells (Park & Hahm 2007). Virulent H. pylori strain genomes are characterized by the presence of a 40kb stretch or block of DNA containing nearly 31 open reading frames known as the cag pathogenicity island or cag PAI. The cag PAI has approximately 40 genes part of which codes for a complex type IV secretion system, which is able to translocate virulence proteins into the host cells. The most virulent gene, the cag A gene located within the PAI codes for a 1186 amino acid containing protein which confers H. pylori strains with the ability to cause ulcers.

Like other branches of biology, the study of pathogenic microorganisms has undergone a paradigm shift. The incredible deluge of information from genome sequencing projects is revolutionizing the science of bacterial pathogenicity. The accessibility of the complete genome sequences of *Helicobacter* provides a scope to undertake bioinformatics based approaches focusing on synonymous codon usage and investigating the gene expression profile of the organism.

The non-random usages of synonymous codons are well accredited (Ikemura 1985). Synonymous codon usage is species specific and differs appreciably between the genes in the same organism (Banerjee *et al.,* 2004). Unlike patterns of codon usage crop up from diverse factors. Mutational pressure and natural selection operating at the level of translation are the primary reasons behind codon usage variation among the genes in different organisms (Sen *et al.,* 2008). Codon bias is quite high in the highly expressed genes compared to lowly expressed ones inside a genome(Sharp & Li 1986; Sharp & Li 1987; Lafay *et al.,* 2000; Dos Reis 2003). The bias of highly expressed genes is influenced by translational selection in contrast to lowly expressed genes, which is governed by mutational bias (Banerjee *et al.,* 2004). In order to inspect the patterns and cause of codon usage, many indices have been projected to assess the degree and direction of codon bias Amongst them, the codon adaptation index (CAI) was proposed as an estimate of codon usage within a gene relative to a reference set of genes (by and large ribosomal protein genes) (Sharp & Li 1987). This index has been revealed to relate better with mRNA expression levels (Ikemura 1981). Over and above codon adaptation index, the effective number of codons (Nc) (Wright 1990), which is described as the amount of equal codons producing the same codon usage bias as observed, and the incidence of optimal codons

(Fop)(Sen *et al.,* 2008), defined as the proportion of synonymous codons that are optimal codons, are also used.

4.2: Materials and Methods

The complete genome sequences for three *Helicobacter* strains [*Helicobacter pylori* J99, *Helicobacter pylori* G27 & *Helicobacter pylori* P12 (hence forth, these strains will be referred to as HPJ, HPG & HP respectively) were obtained from the IMG website (www.img.jgi.doe.gov) (Markowitz 2006) Table 1 shows some of general features of those genomes. Table 2 shows Pathogenicity Related gene (PRG) in detail.

Table 1: Salient features of *Helicobacter* genomes analyzed in this study.

Organism	HPJ	HPG	HP
DNA, total number of bases	1643831	1663013	1684038
DNA G+C content (%)	39.19%	38.87%	38.79
Genes total number	1541	1547	1620
Protein coding genes	1494	1504	1578
RNA genes	47	43	42
rRNA genes	6	7	6
tRNA genes	36	36	36
Protein coding genes with function	897	1008	1128
Pseudogenes	3	0	0
Protein coding genes with enzymes	516	515	516
Protein coding genes with COGs	1087	1080	1082

HPJ: *Helicobacter pylori J99*, HPG: *Helicobacter pylori G27*, HP: *Helicobacter pylori P12*

Table-2: Pathogenicity Related gene (PRG) detail

Gene Name	HPJ	HPG	HP
Cag	15	20	22
adhesin	2	2	3
catalase	1	2	2
glr	1	1	1
trpA	1		1
trpB	1		1
urease	6	8	8
Virulence		1	2

All of the protein coding genes, genes associated with pathogenicity and ribosomal protein genes were examined using Codon W software (http://bioweb2.pasteur.fr) (sen *et al.,* 2008) and E-CAI server(http://genomes.urv.es/CAIcal/E-CAI.) (Puigbò et al., 2008).

The software Codon W (Sen *et al.,* 2008) was employed to inspect G or C in the third position of codons (GC3s), effective number of codons (Nc) (Wright 1990) and the frequency of optimal codons (Fop) (Sen *et al.,* 2008). The effective number of codons (Nc) is a straightforward measure of codon bias (Wu *et al.,* 2005). It ranges from 20 (when merely one codon is used per amino acid) to 61(when each and every codon is used in equal likelihood). Fop (Sen *et al.,* 2008) determines the section of synonymous codons that are optimal codons. Its value varies form 0 (meaning a gene has no optimal codons) to 1.0 (when a gene is exclusively comprised of optimal codons).

The 'codon adaptation index' (CAI) (Sen *et al.,* 2008) values were computed using a web-based application: 'E-CAI server' (http://genomes.urv.es/CAIcal/E-CAI.) (Puigbò et al., 2008) taking the codon usage table as a reference. It quantifies the relative adaptiveness of a gene's codon usage towards the codon usage of highly expressed genes. The relative adaptiveness of each codon is the quantity of the usage of each codon, to that of the most plentiful codon inside the same synonymous family (Sen *et al.*, 2008). The CAI value varies from 0 to 1.0 with higher CAI values signifying that the gene of concern has a codon usage pattern resembling that in the reference genes.

Using the Integrated Microbial Genomes database (www.img.jgi.doe.gov) (Markowitz *et al.*, 2006), the sorted pathogenicity related genes for each strain were subjected to IMG Genome BLAST against the studied strains to find out the sequence homologs. The minimum percent identity was set at 90% and the maximum E value 1e-2.

Correspondence analysis(COA) was performed usingCodonW (http://bioweb2.pasteur.fr)

(Sen *et al.*, 2008). This method explores the major trends in codon and amino acid variations among the genes

4.3: Results

The initial step in our work was to discriminate the three genomes of *Helicobacter* on the basis of their codon usage configuration. The Nc index of Wright (Wright 1990) is an appropriate tool to study codon degeneracy. It is a simple measure of overall codon bias and ranges from twenty to sixty one where 20 is the value obtained when only one codon is used for each amino acid (i.e., the codon bias is maximum) and 61 is the value obtained when all synonymous codon for each amino acid are equally used (i.e., no codon bias). The whole genome of H. pylori J99, H. pylori G27 and H. pylori P12 contains 1541, 1547 and 1620 genes respectively. The Nc score of all the genes in the genome of J99, G27 and P12 ranges between 22.56-61.0, 23.05-61.0 and 21.5-61.0 with a mean of 46.67, 46.27 and 46.33 respectively. The percentage of guanine and cytosine at the third position of a triplet, known as GC3 was also calculated alongside Nc to determine codon heterogeneity among the three species of H. pylori. Table 3 shows the mean values of different indices used to study codon usage patterns

Table 3: Mean values of Nc, GC, GC3, CAI , of the genes in three *Helicobacter* strains.

Organism	Genes	Mean Nc	Mean GC%	Mean GC3%	Mean CAI
HPJ	RPG	47.10926	41.65556	43.48333	0.736722
	PRG	49.25185	38.996	40.4963	0.711074
HPG	RPG	47.32857	41.783673	43.54693878	0.751734694
	PRG	48.01765	37.92941	38.44706	0.7485588
HP	RPG	46.792	41.712	43.172	0.7548
	PRG	47.595	37.7725	38.3075	0.748775

RPG = ribosomal protein genes; PRG = pathogenicity related gene;

Special emphasis was given to those genes of *H. pylori* which are an integral part of its genome and empowers *H. pylori* with pathogenic abilities. These genes include the cytotoxin associated pathogenic island genes commonly known as cag PAI (cagA, cagB, cagC, cagD, cagE, cagF, cagG, cagH, cagI, cagL, cagM, cagN, cagP, cagQ, cagS, cagT, cagU, cagV, cagW, cagX, cagy and cagZ) along with genes coding for a variety of adhesins, catalase, glutamate racemase, tryptophan synthase, urease and virulence factors. The cag pathogenicity island is represented by nearly 30 open reading frames and many of the genes are homologous to virulence associated genes of other bacterial pathogens. The cag PAI genes are also linked with amplified severity of disease in the human host (Eaton *et al.,* 2001). The mean Nc score of the genes responsible for pathogenicity in J99, G27 and P12 is 49.25, 48.01 and 47.59 respectively. Codon Adaptation Index or CAI was calculated for the entire genome of three strains of Helicobacter pylori. Codon Adaptation Index is a well-accepted parameter for studying the expressivity of a gene and assesses the extent to which the selection has been effective in moulding the pattern of codon usage. The Codon Adaptation Index ranges from 0 to 1.0, with higher CAI values signifying that the gene of concern has a higher degree of expressivity (Sharp & Li 1987). The CAI values ranges between 0.077 to 0.519 for J99, 0.075 to 0.419 for G27 and 0.077 to 0.454 for P12 respectively. Using ribosomal proteins of each genome of *Helicobacter* included in this study as standard for measuring gene expression levels it was observed that genes that may be considered as potentially highly expressive in all the three strains should have CAI values ranging from 0.65 to 0.8. Correspondence analyses of codon count of the protein coding genes, ribosomal protein genes and pathogenicity-related genes for the *Helicobacter* strains were performed and shown in Figure 1 which reveals the positions of the genes on the planes defined by the first and second principal axes generated by COA of codon count for the protein coding genes, pathogenicity-related genes, and ribosomal protein genes.

A multiple alignment of the pathogenicity related genes of the three species of *H. pylori* reveals that most of the *H. pylori* G27 pathogenicity genes such as cagA, cagD, cagF, cagG, cag M and others has significant identity with *H. pylori* P12 genes. Pair wise comparison of the pathogenicity genes of the three species based on the identity scores shows that the identity value mostly above 95%, the only exception being cagA gene. The identity score of *H pylori* P12 cagA gene with both *H. pylori* G27 and *H. pylori* J99 is significantly less at 87.

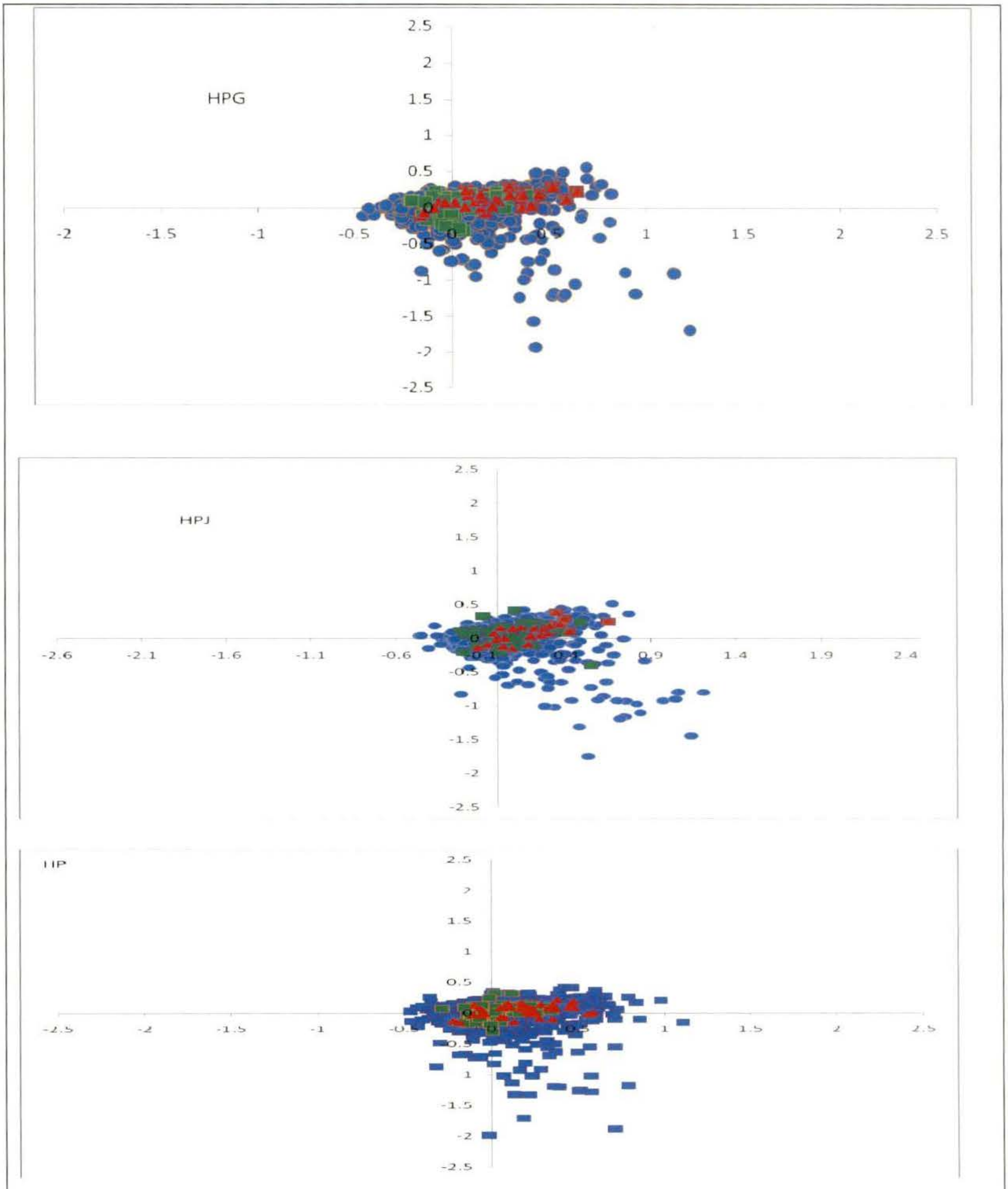

Figure 1: Correspondence analysis of codon usage patterns on codon count for *Helicobacter pylori* J99(HPJ), *Helicobacter pylori* G27(HPG) & *Helicobacter pylori* P12(HP). In all the plots *X* and *Y* axis correspond to axis 1 and 2 of the analysis. Protein coding genes = "blue" PRG (pathogenicity related gene) = "Red" Ribosomal protein genes = "Green"

4.4: Discussion

Detailed codon usage analysis of the cag PAI genes clearly demonstrates that cagD genes have a relatively higher Nc compared to the other cag PAI genes. The protein encoded by cagD gene is a covalent dimer which is an important component of the type IV secretion system (T4SS) that plays a crucial role in CagA translocation (Cendron *et al.,* 2008).

The most unpredictable observation was the apparent lack of genes with similar CAI values in the three *Helicobacter* strains. The cag PAI and other pathogenicity related genes of the three strains *of H. pylori* did not show a significant amount of anti-correlation of Nc with CAI as expected. But, it was quite remarkable to observe the fact that though the Nc values indicated lack of codon biasness, the CAI values pointed towards higher expression levels. This observation is quite significant taking into account the fact that codon biasness is one of the most, if not the most important factor in achieving higher gene expression levels.

The correspondence analysis of the three species of *Helicobacter pylori* indicates that in the case of *H. pylori* J99 the expression level of the cag and pathogenicity related genes is substantially similar to that of the ribosomal protein which clearly indicates that the pathogenicity related proteins are potentially highly expressed. Some sort of similar picture is also observed in the case of the other two species included in our study but in this case there are probably five pathogenicity related genes whose probable expression level does tally with that of any other ribosomal protein genes.

There is a high level of identity among the cagG, cagH and cage genes of the three species of *H. pylori.* The conservedness is consistent with their high CAI value, i.e., expression level and these genes have a crucial role stomach disease.

4.5: References

Banerjee T, Basak S, Gupta SK & Ghosh TC (2004) Evolutionary forces in shaping the codon and amino acid usages in *Blochmannia floridanus*. *J. Biomol. Struct. Dyn.*, **22**:13-23.

Brown LM (2000) Helicobacter pylori: epidemiology and routes of transmission. Epidemiol Rev.**22**:283-97

Cendron L, Couturier M, Angelini A, Barison N, Stein M & Zanotti G (2009) The *Helicobacter pylori* CagD (HP0545, Cag24) Protein Is Essential for CagA Translocation and Maximal Induction of Interleukin-8 Secretion Journal of Molecular Biology Volume 386:204-217

Covacci A, Telford JL, Giudice GD, Parsonnet J, & Rappuoli R (1999). *Helicobacter pylori* virulence and genetic geography. Science **284:**1328– 1333.

Cover TL & Blaser MJ (1999) *Helicobacter pylori* factors associated with disease. Gastroenterology **117:**257–261.

Dos Reis M, Wernisch L & Savva R (2003) Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucl. Acids Res.*, **31**:6976-85.

Eaton KA, Kersulyte D, Mefford M, Danon SJ, Krakowka S & Douglas E. Berg DE (2001) Role of *Helicobacter pylori cag* Region Genes in Colonization and Gastritis in Two Animal Models *Infection and Immunity*, 69:2902-2908

Ferlay J, Bray F, Pisani P & Parkin DM (2004) *GLOBOCAN 2002: Cancer Incidence, Mortality and Prevalence Worldwide.* IARC CancerBase No. 5, version 2 0, Lyon, France: IARC Press

Ikemura T (1981) Correlation between abundance of *E. coli* tRNAs and their occurrence of the respective codons in protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* system. *J. Mol. Biol.,* **146**:1-21.

Ikemura T (1985) Codon usage and transfer-RNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.,* **2**:13-34.

Lafay B, Atherton JC & Sharp PM (2000) Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori. Microbiology,* **146**:851-60.

Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, Anderson I, Mavromatis K, Kunin V, Garcia Martin H, Dubchak I, Hugenholtz P & Rocha EP (2004) Codon usage bias from tRNA.s point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.***14**:2279-86.

Park SW & Hahm KB(2007) Pathogenesis of Helicobacter pylori-induced Gastric Cell Injury *Hanyang Med Rev.* 27(3):34-40

Telford JL, Covacci A, Ghiara P, Montecucco C & R. Rappuoli R(1994) Unravelling the pathogenic role of *Helicobacter pylori* in peptic ulcer: potential new therapies and vaccines. Trends Biotechnol. **12:**420– 426.

Warren JR & Marshall B (1983) Unidentified curved bacilli on gastric epithelium in active chronic gastritis. Lancet. i:1273–1275.

Wright F (1990) The effective number of codons'' used in a gene. *Gene,* **87**:23-9

Wu G, Culley DE & Zhang W (2005) Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. *Microbiology,***151**: 2175-87.