Comparative Study of Data Generated by Structural Annotation of the Genome for Identifying the Unique Parameter Responsible For Pathogenic Property of *Helicobacter Pylori*

Megha Vaidya* and Hetal Kumar Panchal

Department of Computer Science & Technology (GDCST), Sardar Patel University, Vallabh Vidyanagar – 388120, Gujarat – India

Abstract

Helicobacter pylori are the most common bacterial infection worldwide. They are generally acquired during childhood and the infection can persist in gastric ecosystem throughout the life span of the host, if untreated. There are similarities and the variations in the size of genomes of several strains. The disease causing factor Cag Pathogenic Island is present in almost all the strains but yet the different strains cause different diseases. In present studies, detailed comparative studies of each genomes is carried out to check possibility of prediction of the type of disease it would cause and suggest some target for treatment on the bases of structural annotation. It is found that comparative studies of structural features of genomes give important outcomes but, cannot give detailed idea about the type of disease the strains would cause. Functional analysis of genome or study of gene order and rearrangement of genome may solve the mystery of disease specific genes and give some better target for treatment.

Keywords: *Helicobacter pylori*; Pathogenic island; Structural annotation; Structural features

Introduction

Helicobacter pylori are the most common bacterial infection worldwide. They are generally acquired during childhood and the infection can persist in gastric ecosystem throughout the life span of the host, if untreated [1]. This Gram negative species of ϵ -proteobacteria are reported to cause various diseases like gastritis, gastric (stomach) ulcer, and duodenal ulcer, and is associated with gastric cancer and mucosa-associated lymphoid tissue (MALT) lymphoma. Some clinical work in past at Japan suggests that *H. pylori* eradication reduces the risk of new gastric carcinomas in patients with a history of the disease [2].

Three major virulence factors of *H. pylori* have been described: 1) the cytotoxin-associated gene product (CagA), 2) the vacuolating toxin (VacA) and 3) the adhesion protein BabA2. The cytotoxin-associated gene A (CagA) is a protein with a molecular mass of approximately 125-140 kDa, encoded by the *cagA* gene, [3,4], that is translocated into gastric epithelial cells by a type IV secretion system, encoded by the Cag pathogenicity island (Cag PAI) [5].

The first strain of *H. pylori* to be sequenced in 1997 was strain 26695. The chromosome of strain 26695 is circular and composed of 1667867 base pairs. The average GC content is approximately 39%. In 1999, strain J99 was sequenced which was isolated from an American patient with a duodenal ulcer. Compared to strain 26695, it has a slightly smaller circular chromosome with 1643831 base pair. The overall genomic organization, gene order and predicted proteomes of the sequenced strains were very similar. In 2006, a chronic atrophic gastritis *H. pylori* strain, HPAG1 with 1596366 base pair, was sequenced [6]. Similar to the sequenced strains 26695 and J99, HPAG1 is a type-1 strain that contains CagA and a virulent allele of VacA [7]. The *H. pylori* strain G27 was sequenced recently [8]. It was originally isolated from an Italian patient and has been used widely in *H. pylori* research. The G27 genome has a similar size to the other three sequenced strains. It is 1652983 bp long and has a GC content of 38.9%. In addition, G27 also contains one 10032 bp AT rich (65.2%) plasmid resembling that found in strain HPAG1. The plasmid encodes 11 genes.

Prokaryotic genomes can be annotated based on their structural, operational, and functional properties (structure annotation paper-folter HP). Structural annotation means identification of genomic elements, like ORFs and their localization, gene structure, coding regions, coding density, nucleotide content, location of regulatory motifs etc. Providing biological information to these genomic elements is called as functional annotation.

There are similarities and the variations in the size of genomes several strains. The disease causing factor Pathogenic Island is present in all most all the strains but yet the strains cause different diseases. In present studies detailed comparative studies of each genomes is carried out to check possibility of prediction of the type of disease it would cause and suggest some target for treatment on the bases of structural annotation.

Materials and Methods

From literature study pathological consequences of each strain of *H. pylori* was carried out and grouped on the basis of disease it causes. Mostly *H. Pylori* strains are divided into five major groups, hpEurope, HpAfrica1, hpAfrica2, hp East Asia and hpAsia2 [10]. According to the literature study, currently genome sequences of 36 different *H. pylori* strains are available for study. Past studies have evident total 18 strains contributing to 4 different Gastro intestinal diseases namely, gastric adenocarcinoma, malt lymphoma, peptic ulcer and gastritis. For comparative studies, random selection of strains was done to access

***Corresponding author:** Megha Vaidya, Department of Computer Science & Technology (GDCST), Sardar Patel University, Vallabh Vidyanagar – 388120, Gujarat – India, E-mail: mbv26887@gmail.com, swamihetal@gmail.com

Received May 24, 2013; **Accepted** June 26, 2013; **Published** June 28, 2013

Citation: Vaidya M, Panchal HK (2013) Comparative Study of Data Generated by Structural Annotation of the Genome for Identifying the Unique Parameter Responsible For Pathogenic Property of *Helicobacter Pylori*. Pharm Anal Acta 4: 249. doi:10.4172/2153-2435.1000249

Copyright: © 2013 Vaidya M, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

the relation between genomic features of strains and there pathogenic ability. For each disease except Malt lymphoma, four strains were selected for analysis. It is to be noted that only one strain is confirmed and is reported which causes Malt lymphoma. Genomic data of strain F32, F57, HPAG1 and PaCan4 is taken for gastric Adenocarcinoma; strain B38 for Malt lymphoma; strain F16, 26695, SJM180 and V225d for Gastritis and strain J99, B8, G17and F30 for Peptic ulcer were obtained from GENBANK genomic data repository at NCBI.

Structural annotation of many of above mentioned strains is available but comparative study of them was not available. So we have majorly focused on the comparative studied of structural genomic features. The detailed study of all the above 13 genomes was carried out by Artemis software. Artemis was preferred for analysis because it is a free genome browser and annotation tool that allows visualization of sequence features, next generation data and the results of analyses within the context of the sequence, and also its six-frame translation. Artemis is written in Java, and is available for UNIX, Macintosh and Windows systems. It can read EMBL and GENBANK database entries or sequence in FASTA, indexed FASTA or raw format. Other sequence features can be in EMBL, GENBANK or GFF format, which easily reads FASTA, EMBL or GENBANK format sequences and feature tables, and also can work on sequences of any size.

Results and Discussion

By analyzing genome of 13 strains we found that the genome size of *Helicobacter pylori* B38 strain is comparatively less than other 12 genomes. We have also noticed that Cag Pathogenic island is absent in it but still it causes Malt Lymphoma. So we may say that pathogenic nature of *h.pylori* cannot be removed by removal of the Cag pathogenic island form the organism. Other remarkable observation is that *Helicobacter pylori* B38 strain do not content genes like babB, babC, sabB, and homB which are also contributing to virulent activity. Thus, we can say that there are certain genes other than Cag, babB, babC, sabB, and homB, which are pathogenic but yet undiscovered.

Comparative structural genomic analysis of selected 13 strains was carried out using Artemis software for number of bases in forward and reverse strand, Number of features in the active entries Genes (CDS features without a /pseudo qualifier), gene density (genes per kb), average gene length ,average gene length (including introns), coding percentage, coding percentage (including introns), gene sequence composition (A,C,G,T content) GC percentage, Overall sequence composition (A,C,G,T content) (Table 1).

Results of comparative analysis show that there are no unique features shared by strains responsible to cause same disease. The Analysis of each of all the 13 genome of *H.Pylori* showed that we cannot predict about the disease caused by the strains form its structural features (Table 2).

It is reported that transmission of bacteria to a new host may affect the bacterial genome contents and this can be a one of the factor contributing for inter-strain genomic diversity. Apparently alteration in bacterial genome according to host environment may cause either deletion or insertion of some coding or non coding segment which shall affect the coding percentage of genome. But according to present comparative studied of 13 different genomes none of strain showed dramatic change in coding percentage (Figure 1). We also found that genome of all the strains have same nucleic acid base content even though there is variation in size of genome which concludes that there is equal distribution of each four nucleotide throughout the genome (Figure 2).

Similarly the analysis of gene density feature (Figure 3) had shown that over all there is minor difference except for genomes of V225d and G27 genomes. It is evident from the chart that none of the strains F16, 26695 or SJM180 had shown high gene density like V225d even it falls in same group of diseases, Gastritis. Likewise Peptic ulcer causing strains j99, B8 and F30 had not shown high gene density like G27. Both V225d and G27 have close values but they fall in different diseases category.

Conclusion

From comparative studies of structural features like total Bases, gene (predicted ORF's), gene density, average gene length, coding percentage, GC % content and total base count of each nucleic acid base, we found that there are no specific patterns with regards to structural features in strains responsible for particular disease. It was observed that, content of each of the four nucleotides were evenly distributed throughout the genome. Further we can conclude *H.pylori* strains with absence of Cag-PAI are also likely to cause GI tract disease specifically MALT Lymphoma. It should be noted that there are fair chances of getting the pathogenic genes other than so far reported genes, whose virulence is yet undiscovered. Thus, comparative studies of structural genomic features of genomes give important outcomes but, cannot give detailed idea about the kind of disease the strains

Diseases	CAG-PAI Island	Strain
Gastric Adeno Carcinomal	Present	Heloco bacter Pylori F32
		Heloco bacter Pylori F57
		Heloco bacter Pylori HPAG1
Gastritis	Present	Heloco bacter Pylori PeCan4
		Heloco bacter Pylori F16
		Heloco bacter Pylori 26695
Peptic Ulcer	Present	Heloco bacter Pylori SJM180
		Heloco bacter Pylori v225d
		Heloco bacter Pylori P 12
		Heloco bacter Pylori F 30
		Heloco bacter Pylori 2017
		Heloco bacter Pylori 2018
		Heloco bacter Pylori 51
		Heloco bacter Pylori 908
		Heloco bacter Pylori B8
		Heloco bacter Pylori G27
		Heloco bacter Pylori J99
		Heloco bacter Pylori B38
Malt lymphoma	Absent	Heloco bacter Pylori
Unknown	Present	Heloco bacter Pylori Cuz20
		Heloco bacter Pylori Sat464
		Heloco bacter Pylori 35A
		Heloco bacter Pylori 83
		Heloco bacter Pylori ELS 37
		Heloco bacter Pylori Gambia94/24
		Heloco bacter Pylori HUP-B14
		Heloco bacter Pylori Lithuania75
		Heloco bacter Pylori PeCan18
		Heloco bacter Pylori Puno120
		Heloco bacter Pylori Puno135
		Heloco bacter Pylori SNT49
		Heloco bacter Pylori Shi112
		Heloco bacter Pylori Shi169
		Heloco bacter Pylori Shi417
		Heloco bacter Pylori South Africa7
		Heloco bacter Pylori XZ274

Table 1: List of strains with presence / absence of CAG-PAI with respect to various diseases.

Strands	Features	Gastric edinocarcinoma				Malt Lymphoma	Gastritis				Peptic Ulcers			
		F32	F57	HPAG1	PaCan4	B38	F16	25595	SJM180	V225sd	J99	B8	G27	F30
Forward	No.of Bases present in CDS	798816	845766	839527	793743	755493	827554	830877	845328	830522	855024	918432	833424	800451
	Gene (ORF's)	934	979	954	921	931	983	959	958	939	988	994	980	925
	Gene Density (per Kb)	0.591	0.608	0.597	0.565	0.59	0.623	0.574	0.577	0.591	0.601	0.593	0.592	0.588
	Average Gene length	855	863	880	851	821	841	855	882	884	865	923	850	865
	Coding percentage	50.5	52.5	52.6	48.7	47.9	52.5	49.8	50.9	52.2	52	54.8	50.4	50.9
	GC percentage content	39.51	39.46	39.91	39.71	39.88	39.52	39.77	39.69	39.85	39.9	39.52	39.65	39.51
	Adenine content (%)	31.3	31.53	31.31	31.34	31.27	31.53	30.3	31.55	31.28	31.34	31.86	31.47	31.57
	Cytosine content (%)	18.5	18.42	18.68	18.61	18.85	18.48	19.61	18.62	18.48	18.77	18.26	18.62	18.39
	Guanine content (%)	21	21.02	21.22	21.09	21.03	21.03	19.26	21.08	21.38	21.11	12.25	21.02	21.11
	Thymine content (%)	29.18	29.01	28.77	28.94	28.84	28.94	30.82	28.75	28.86	28.76	28.61	28.87	28.91

Strands	Features	Gastric edinocarcinoma				Malt Lymphoma	Gastritis				Peptic Ulcers			
		F32	F57	HPAG1	PaCan4	B38	F16	25595	SJM180	V225sd	J99	B8	G27	F30
Forward	No.of Bases present in CDS	843291	821607	827070	893274	883416	805396	897264	872028	834081	857232	818520	891756	828855
	Gene (ORF's)	939	906	946	974	935	878	1007	971	966	943	964	997	927
	Gene Density (per Kb)	0.594	0.563	0.592	0.5	0.592	0.557	0.603	0.585	0.608	0.573	0.575	0.603	0.59
	Average Gene length	898	906	874	913	944	917	891	898	863	909	849	894	894
	Coding percentage	53.4	51	51.8	54.8	56	51.1	53.7	52.5	52.5	52.1	48.8	53.9	52.7
	GC percentage content	39.66	39.46	39.8	39.68	39.92	39.67	39.54	39.67	39.46	40.03	39.6	39.65	39.58
	Adenine content (%)	31.85	31.85	31.48	31.81	31.44	31.58	31.9	31.65	31.8	31.49	31.53	31.75	31.58
	Cytosine content (%)	19.27	18.27	18.57	18.33	18.39	18.35	18.29	18.36	18.39	18.57	18.63	18.38	18.37
	Guanine content (%)	19.58	21.18	21.23	21.33	21.52	21.31	21.24	21.3	21.07	21.45	20.96	21.25	21.19
	Thymine content (%)	30.27	28.26	28.71	28.51	28.62	28.74	28.55	28.66	28.72	28.48	28.86	28.59	28.83

Strands	Features	Gastric edinocarcinoma				Malt Lymphoma	Gastritis				Peptic Ulcers			
		F32	F57	HPAG1	PaCan4	B38	F16	25595	SJM180	V225sd	J99	B8	G27	F30
Over all	Total Bases	1578824	1509005	1595355	1529557	1576758	1575399	1557867	1558051	1588278	1543831	1573997	1552982	1570564
	Gene (ORF's)	1873	1885	1900	1899	1866	1966	1905	1929	1905	1931	1958	1977	1852
	Gene Density (per Kb)	1,186	1,171	1,19	1,165	1,183	1,181	1,178	1,163	1,199	1,174	1,169	1,196	1,179
	Average Gene length	876	884	877	888	878	877	879	890	873	886	887	872	879
	Coding percentage	104.7	103.6	104.4	103.5	104	103.6	103.6	103.5	104.8	104.1	103.7	104.3	103.7
	GC percentage content	38.86	38.73	39.08	38.49	39.16	38.88	38.87	38.9	38.97	39.19	39.56	38.89	38.83
	Adenine content (%)	31.3	31.53	31.31	31.34	31.27	31.53	30.3	31.55	31.28	31.34	31.86	31.47	30.53
	Cytosine content (%)	19.58	19.4	19.57	19.65	19.66	19.5	19.61	19.56	19.47	19.69	18.44	19.58	19.45
	Guanine content (%)	19.27	19.32	19.5	19.29	19.29	19.37	19.26	19.33	19.51	19.49	21.11	19.3	19.37
	Thymine content (%)	30.86	30.75	30.49	30.8	30.62	30.82	30.82	30.62	30.67	30.48	28.72	30.74	30.63

Table 2: Structural features of 13 *H.pylori* strains.

would it cause. Functional analysis of genome or study of gene order and rearrangement of genome may solve the mystery of disease specific genes and give some better target for treatment.

Acknowledgment

We are heartily thankful to Prof. (Dr.) P.V. Virparia Director, GDCST, S.P. University, V.V. Nagar, for providing us facilities for the research work. We are also thankful to DST-PURSE program and Center for Interdisciplinary Studies in Science and Technology (CISST), S.P. University, V.V. Nagar, Gujarat (India) for providing financial assistance in form of fellowship.

References

- Everhart JE (2000) Recent developments in the epidemiology of *Helicobacter pylori*. *Gastroenterol Clin North Am* 29: 559-578.
- Kawai M, Furuta Y, Yahara K, Tsuru T, Oshima K, et al. (2011) Evolution in an oncogenic bacterial species with extreme genome plasticity: *Helicobacter pylori* East Asian genomes. *BMC Microbiol* 11:104.
- Covacci A, Censini S, Bugnoli M, Petraccia R, Burroni D, et al. (1993) Molecular characterization of the 128-kDa immunodominant antigen of *Helicobacter pylori* associated with cytotoxicity and duodenal ulcer. *Proc Natl Acad Sci USA* 90: 5791-5795.
- Tummuru MK, Cover TL, Blaser MJ (1993) Cloning and expression of a high-molecular-mass major antigen of *Helicobacter pylori*: evidence of linkage to cytotoxin production. *Infect Immun* 61:1799-1809.
- Odenbreit S, Puls J, Sedlmaier B, Gerland E, Fischer W, et al. (2000) Translocation of *Helicobacter pylori* CagA into gastric epithelial cells by type IV secretion. *Science* 287: 1497-1500.

6. Oh JD, Kling-Backhed H, Giannakis M, Xu J, Fulton RS, et al. (2006) The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: evolution during disease progression. *ProcNatlAcadSci USA* 103: 9999-10004.
7. Enroth H, Kraaz W, Engstrand L, Nyren O, Rohan T (2000) *Helicobacter pylori* strain types and risk of gastric cancer: a case-control study. *CancerEpidemiol Biomarkers Prev* 9: 981-985.
8. Baltrus DA, Amieva MR, Covacci A, Lowe TM, Merrell DS, et al. (2009) The complete genome sequence of *Helicobacter pylori* strain G27. *J Bacteriol* 191: 447-448.
9. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, et al. (2003) Traces of human migrations in *Helicobacter pylori* populations. *Science* 299: 1582-1585.
10. Linz B, Balloux F, Moodley Y, Manica A, Liu H, et al. (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* 445: 915-918.