International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2015): 6.391

Computer Assisted Characterization of Virulence Factors among Helicobacter Pylori Strains

Akoo Benedict[1], Dr. Johnson Kinyua[2], Dr. Daniel Maina[3]

[1]Jomo Kenyatta University of Agriculture & Technology, Nairobi, Kenya
akoobenedict[at]gmail.com
[2]Jomo Kenyatta University of Agriculture & Technology, Nairobi, Kenya
johnsonkinyua[at]jkuat.ac.ke
[3]Aga Khan University Hospital, Nairobi, Kenya
daniel.maina[at]aku.edu

Abstract: ***The global prevalence of Helicobacter pylori infection is more than 50% and varies significantly between developed and developing countries [4, 28, 32]. The fact that the mode of transmission is still not yet clear and the bacterium has been classified as a carcinogen raises concern for a concerted effort in initiating relevant research towards gaining more in-depth knowledge on the bacterium[14]. Application of bioinformatics tools to the sequenced data provides an avenue for more information that could be analyzed to further help in designing precise lab experiments that would direct more knowledge on virulence. Identification of a given strain can help in designing primer specific amplicons to target the virulent strain which can be both strategic in drug development and also designing of diagnostic tool.***

Keywords: Helicobacter pylori, comparative genomics, virulent factors, protein structure prediction, proteomics

1. Introduction

Globally different strains of Helicobacter pylori have been associated with difference in virulence and the resulting interaction with the host's factors [7]. Several related studies have suggested that environmental factors also lead to subsequent difference in the expression of disease [29, 30]. Molecular techniques applied has revealed that independent Helicobacter pylori isolates exhibit extensive genetic diversity which has been predicted to be important in pathogenesis, possibly relating to the wide variation in patient symptomology. Bioinformatics approaches have permitted comparative genomics techniques to identify conserved genes among multiple pathogenic strains or genes that have predicted functions similar to known virulence factors [10, 27]. Interpreting the molecular mechanisms therefore of virulent factors can improve understanding of the cellular and molecular basis of pathogenesis. This can consequently assist in designing precise laboratory experiments to inform new avenues for identifying promising approaches to disease prevention and therapy.

2. Study Methodology

The materials used were all in silico and also included web-based programs. The nucleic acid and amino acid sequences of the reference/template sequence of Helicobacter pylori 26695 were accessed online in the virulence factor database, http:www.mgc.ac.cn/VFs/.The sequences were all downloaded in fasta format.

The query sequences were the complete genomes of Helicobacter pylori with known accession numbers. The accession numbers were used as the query id. They were; Helicobacter pylori B38 accession number NC 000921.1, Helicobacter pylori G27 accession number NC 011333.1 and Helicobacter pylori J99 accession number NC 000921.1. The other online tools and programs that were used included:

- The National Centre for Biotechnology Information database, http:www.ncbi.nlm.nih.gov/.
- Integrated InterPro Consortium (InterPro58.0) database, http://www.ebi.ac.uk/interpro/.
- Integrated protein database, PIR-PSD, http://pir.georgetown.edu/pirwww/.
- The central archive for 3D-structures, www.wwpdb.org.
- Alignment tools, Clustal 2.1, MAFFT, MEGA 7 and Jalview.
- Phylogeny tools used were ClustalW and phylip developed by Felsenstein of the University of Washington.
- Protein comparison tools, Matt and FATCAT, and protein structure prediction I-TASSER, structure visualization program was pymol.
- Reliable internet connection and a lap top were also used.

3. Sequence Retrieval, Sequence Alignment, Similarity Searching and Comparison

3.1 Sequence Retrieval

The reference sequences for virulent factors under study were downloaded directly from the virulence factors database. The selected factors were accessed through the subtitle 'Major virulence factors in the Helicobacter' in the home page of the website. Each gene sequence was retrieved from the major categories of the virulence factors i.e. gluE for endotoxin, futA for molecular mimicry, napA for pro-inflammatory effect, cag1 for secretion system, cagA for type iv secretory protein and vacA for toxin. All were downloaded in fasta format.

The complete genomes of query sequences were accessed from NCBI database. The Blast program was used to identify probable nucleotide and amino acid sequences for the corresponding genes in selected strains. These were chosen for completeness and taking into account regions of local

similarity determined through identification of significant hits, maximum score, total score and E-value.

However the corresponding amino acids for the each virulent factor in the query genomes were obtained differently. The nucleotide sequence queries, translated in six reading frames and the resulting six-protein sequences were compared each in turn to those of amino acid sequence for each virulence factor as the subject ID. This was achieved through the use of program blastx. The rationale for choosing the program over blastp was because blastx helps to annotate coding regions on nucleotide sequence and is useful in detecting frame shifts in the coding regions. The sequence lengths were also a factor determining probable alignment.

3.2 Sequence Alignment

Sequences of genes and proteins are compared to check the similarities and differences at the level of individual bases or amino acids. The most common comparative method is sequence alignment which provides an explicit mapping between the residues of two or more sequences. The alignment helps to infer structural, functional and evolutionary relationships among the sequences. It also helps to highlight conserved sites/regions, variable sites and uncover changes in gene structure. Comparison of similarities of both the DNA and amino acid sequences was done across genes through the pairwise alignment and multiple alignment.

3.3 Similarity Searching

The amino acid sequence for each deduced gene was compared to a protein database of solved structures to identify similar protein structures. The method used is known as homology modeling based on the premise that two proteins with enough sequence similarity will fold in a similar way and share the same conformation in space. The process through which a tertiary structure is assigned to a given sequence is carried out in three steps, namely: template identification, template alignment and model building.

3.4 Protein Structure Determination

The sequences of which similar structure could not be found within the database; ab initio method (de novo protein structure prediction) was used. This method tries to predict the tertiary structure of protein directly from its sequence properties and the principle idea is that the structure can be determined without any explicit templates but only by means of applying the general principles that govern protein folding and the statistical tendencies of conformational features gathered from structural knowledge. Protein structure models are theoretical models which may contain errors and therefore need to be treated with caution. However, protein structure models can support the design of experiments and also may help in explaining experimental observations. A web based program was used, I-TASSER an online platform which allows academic users to automatically generate high quality predictions of 3 D structure [22, 23, 24] (Figure 1-6). There was no similar structure for the amino acid sequence coding for cag1 gene in strain 26695, G27 and J99 significantly noted was the absence of cagA and cag1 genes respectively in strain B38.

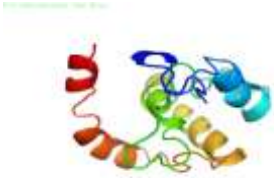

Figure 1: cag1 predicted structure for strain G27

Figure 2: cag1 predicted structure for strain J99

Figure 3: cag1 predicted structure for the reference strain (26695)

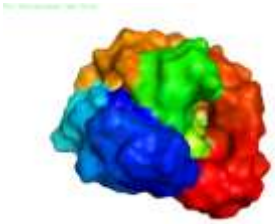

Figure 4: Surface Representation cag1 G27 by pymol

Figure 5: Surface Representation cag1 J99 by pymol

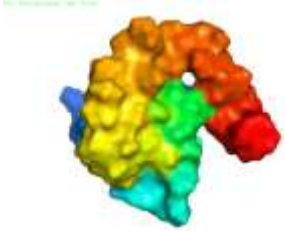

Figure 6: Surface Representation cag1 26695 by pymol

Table 1: cag1 Structure Prediction Results by I-TASSER program

Strain	C-score	Estimated RMSD	Estimated TM Score
26695	-4.25	14.2±3.8Å	0.27±0.08
G27	-4.00	13.5±4.0Å	0.29±0.09
J99	-3.99	13.5±4.0Å	0.29±0.09

3.5 Protein Structure Alignment and Comparison

The produced models and target structures were subjected to protein structure alignment and comparison. This was achieved through aligning pairs of proteins, using two different programs i.e. fatcat and matt. Most of the available methods for structure alignment start by computing all pair-wise alignments between a set of structures but then use them to generate the optimal consensus alignment between all the structures. The two methods, despite many web-based resources being based on rigid structural alignment algorithms, the two servers developed flexible protein structure comparison algorithm. This is in tandem with the native protein which generally is flexible and undergoes structural rearrangements as part of their function. The fatcat algorithm also has been implemented in a fast and efficient computer program and systematically tested on large alignment benchmarks and has been shown to be unbiased towards introducing twists. The structures for comparison were uploaded in the two servers (i.e. http://fatcat.godziklab.org/ and http://matt.cs.tufts.edu/ respectively) and submitted and the progress of the experiment monitored [25, 26]. The results were integrated from all pair-wise comparisons and analyzed and visualized using pymol. The results were combined and similarity consensus profile produced using different similarity comparison methods as indicated in the result section.

4. Results

There was significant difference in nucleotide and amino acid sequence results obtained after blast experiment. The search for corresponding nucleotide sequence for the query strains for the genes were obtained through blastn. The search results for strain B38 found no significant similarity for cag1 and cagA respectively. However, there was significant single hit for cag1 in each strain G27 and J99 with over 96% similarity and expect value of 4e-164.There was six blast hit each for cagA gene in both strain G27 and J99. All the three query strains obtained similar hits for futA gene, i.e. four blast hits. The was one significant blast hit for all the query strains for both the gluE and napA gene respectively with above 94% similarity in all cases and maximum score equals the total score, an indication of a single alignment for the three strains. There was notable difference in the case of vacA gene, where three blast hits were observed for strain B38 and single blast hit for strain G27 and J99 with maximum score equals total score for the two respectively. The probable corresponding nucleotide sequence selected was based on suggested threshold for expect value (zero) and percent sequence similarity (85%) taking into account the query coverage, maximum score and total score. Based on the aforementioned criteria; two alignments each were selected for futA gene in all the query strains, one alignment automatically selected for cag1, gluE and napA respectively for the respective strains. In case of cagA two alignments each were selected for the respective strains while one alignment each for vacA gene in strain G27 and J99, a deviation otherwise observed in B38 where two alignments were selected. The blastx results obtained realized two sequences each for futA gene, one sequence each for napA and gluE genes respectively. The vacA gene had two sequences for strain B38 while strain G27 and J99 had one each. Notably cagA had two sequences each for strain G27 and J99 while cag1 had one sequence each for the same strains. No results observed for cag1 and cagA for strain B38.Multiple alignment results for cag1 for both nucleotide sequences and amino acid sequences showed a highly conserved sequence especially at protein level although point mutations were noted to be sparsely scattered at nucleotide sequence level, specifically transition mutation. Pairwise alignment between the strains also showed a high percentage similarity of above 96%.The multiple alignment for cagA showed high similarity at specific regions, a block of conserved regions while pairwise alignment obtained percentage similarity range of between 60-80%.The futA showed several regions with conserved sequences with a high percentage similarity for pairwise alignments of between 68-94%.Interesting pattern was observed with gluE gene where multiple alignment results showed several insertions and point mutations at nucleotide level while conservative substitutions were observed at protein level. The percentage similarity was however very high for pairwise alignments with all above 94% similarity at amino acid level but a little low at nucleotide level. Conserved regions were observed at protein level. The same was observed for napA gene with exception of pairwise alignment between strain B38 and G27 which had a 100% similarity along the entire length. There was a notable variation in percentage similarity for vacA gene at both nucleotide sequence level and at protein level with only relatively high percentage identity between the reference sequence (strain 26695) and strain G27.The multiple alignment however observed several deletions and insertions at nucleotide level while at protein level there were insertions and deletions with the sequence less conserved. The structure alignment and comparison results observed were both surprising and interesting. There was no similar structure observed in the protein data bank (pdb) for cag1 hence the predicted structures. C-score, estimated TM-score and Root Mean Square Deviation had only marginal difference (Table 1).There were over three similar structures for cagA but the structure alignment obtained less percentage similarity, 13.7% with only ten core residues. The futA gene had three similar structures in the pdb, while comparison obtained a high percentage similarity, 97% with high z-score for both programs used. The core residues were 338 over alignment length of between 331 and 376.The napA gene had over 56 similar structures in the pdb and structure comparison obtained a 100% similarity and identity, a z-score of 410 with core residues of 143 over alignment length of between 124 and 144.The gluE gene had over 98 similar structures in the pdb while comparison obtained 36.3% identity and 54.3% similarity and core residue of 327 against alignment length of 344.The vacA gene had only one similar structure in the pdb and percentage comparison obtained

100% similarity and identity and core residues of 447 against alignment length of between 777 and 1290.

5. Discussions

The genes under study with their function and product were cag1, product- type iv secretion system, function- acts like a channel that helps the secretory protein to cross insert into the host's membrane.cagA, product- cytotoxicity-associated immunodominant antigen- function, interaction with the host proteins and trigger the development of inflammatory response. futA, product, -fucosyltransferase, -function, play a role in molecular antigenic mimicry suppressing immune response against the bacteria.gluE, product-UDP-glucose 4-epimerase, function-promotes disruption of epithelial cell basement contributing to the disruption of gastric mucosal integrity. napA, product-neutrophil activating protein-function, promotes adhesion of human neutrophils to endothelial cells and production of reactive oxygen radicals inducing inflammation and the last gene vacA, product-vacuolating cytotoxin and function is formation of ion channels in membranes and induces apoptosis.[7, 16, 17, 28]. The phenotypes for these genes include but are not limited to duodenal ulcer, duodenitis, stomach ulcer, gastric cancer, MALT lymphoma, acute/chronic gastritis and gastric atrophy. The rationale for picking on each gene in the categories of major virulence factor was informed by the diversity of the mode of action on the host while the exclusion criteria for some was because some of the virulent factors in the same category have already been widely studied and well characterized by other researchers; this category include the urease for enzyme category, flagella for motility category, LPS for the endotoxin category and OipA for the proinflammatory. Four different multiple sequence alignment programs were used for comparison i.e. Clustal, MAFFT, MEGA 7 and Jalview. This was for comparison purposes and also each program has got both merits and demerits. The MAFFT and Clustal methods highlights identical, similar residues and similar substitution by color sheds, positive sign, dot, a star or colon respectively. MEGA 7 output can be easily be exported to excel for farther analysis and also provides a variety of statistical results such as nucleotide and amino acid frequencies, codon bias results among others. Jalview method is so pronounced in color coding and easily identifies conserved regions, it also provides a consensus sequence that can be helpful in primer design and other experimental design such as typing of biological markers.

The reference sequence proteins had two genes, cagA and napA as experimented evidence at protein level, two genes, gluE and vacA protein inferred from homology although several similar structures of the proteins in the database were experimentally determined by x-ray diffraction. The remaining genes cag1 and futA were protein predicted. The futA gene had several similar structures in the database experimentally determined by x-ray diffraction while cag1 had no similar structure in the database. The gluE and napA genes respectively had similar structures in the database from different bacterial species and hence orthologous. The results of futA indicate that the gene is duplicated and occupy different positions in the same genome hence paralogous. The other results for the remaining genes that multiple sequence was observed only pointed to a small piece of a much larger gene and therefore just a fragment of the respective genes noted.

Protein structure comparison is essential in almost every aspect of modern structural biology for example high resolution models with an RMSD of 1 to 1.5Å are useful for almost any application, including studying catalytic mechanism of enzymes and also a variety of structure-based protein engineering, such as drug design [34].The cag1 structure predicted in this study had RMSD of over 10Å therefore the prediction is significantly wrong. This is supported by both the approximated TM-score and C-score.TM-score of less than 0.17 means a random similarity, while C-score is the confidence score calculated based on the significance of threading template alignment and the convergence parameters of the structure assembly simulations. The C-score is in the range of negative five and two and a C-score of higher value signifies a model with a high confidence [22, 24].In this case the confidence is low.

Identification of structurally conserved active sites is possible in this case for genes with larger core residues (common core), common core being a set of residues that can be simultaneously superimposed with small structural variation [13].In this study therefore gluE, napA, futA and vacA are the right candidates. Further research is however required both in silico and experimental for cag1 and cagA genes respectively to exhaustively determine their structures because from this study there is the likelihood that they may be part of a bigger gene family acting like an operon.

References

[1] **Barh, D., Tiwari, S., Jain, N., Ali, A., Santos, A. R., Misra, A. N., & Kumar, A. (2011).** In silico subtractive genomics for target identification in human bacterial pathogens. Drug Development Research, **72**(2), 162-177.

[2] **Bernarde, C., Lehours, P., Lasserre, J. P., Castroviejo, M., Bonneu, M., Mégraud, F., & Ménard, A. (2010).** A complexomic study of two Helicobacter pylori strains of two pathological origins: potential targets for vaccine development and new insight in bacteria metabolism. Molecular & Cellular Proteomics, mcp-M110.

[3] **Brinkman, F. S., & Leipe, D. D. (2001).** Phylogenetic analysis. Bioinformatics: a practical guide to the analysis of genes and proteins, **2,** 349.

[4] **Brown, L. M. (2000).** Helicobacter pylori: epidemiology and routes of transmission. Epidemiologic reviews, **22** (2), 283-297.

[5] **Huson, D. H., & Bryant, D. (2006).** Application of phylogenetic networks in evolutionary studies. Molecular biology and evolution, **23**(2), 254-267.

[6] **Moreira, E. D., et al**. "Risk factors for Helicobacter pylori infection in children: is education a main determinant?" Epidemiology and infection **132.02 (2004):** 327-335.

[7] **Olbermann, P., Josenhans, C., Moodley, Y., Uhr, M., Stamer, C., Vauterin, M., & Linz, B. (2010).** A global overview of the genetic and functional diversity in the Helicobacter pylori cag pathogenicity island. PLoS Genet, **6** (8), e1001069.

[8] **Owen, R. J. (1998).** Helicobacter-species classification and identification. British medical bulletin, **54** (1), 17-30.

[9] **Salih, B. A. (2009).** Helicobacter pylori infection in developing countries: the burden for how long? Saudi Journal of Gastroenterology, **15** (3), 201.

[10] **Schell, M. A., Lipscomb, L., & DeShazer, D. (2008).** Comparative genomics and an insect model rapidly identify novel virulence genes of Burkholderia mallei. Journal of bacteriology, **190**(7), 2306-2313.

[11] **Wu, H. J., Wang, A. H., & Jennings, M. P.** (2008). Discovery of virulence factors of pathogenic bacteria. Current opinion in chemical biology, **12** (1), 93-101.

[12] **Xiaoyun "Sean" Liao (2008).** .Orthologs, Paralogs and Evolutionary Genomics. Paper presented on 20th November.

[13] **M.Menke, Berger, L. Cowen,** "Matt: Local Flexibility Aids Protein Multiple Structure Alignment", PLO Computational Biology, Vol.**4**, No1. **2008**.

[14] Tanih, N. F., et al. "Helicobacter pylori infection in Africa: Pathology and microbiological diagnosis." African Journal of Biotechnology 7.25 **(2008).**

[15] **Parsonnet, J. (1995).** Bacterial infection as a cause of cancer. Environmental health perspectives, **103**(Suppl 8), 263.

[16] **Montecucco, C., & de Bernard, M. (2003).** Molecular and cellular mechanisms of action of the vacuolating cytotoxin (VacA) and neutrophil-activating protein (HP-NAP) virulence factors of Helicobacter pylori. Microbes and infection, **5**(8), 715-721.

[17] **Naito, Y., & Yoshikawa, T. (2002).** Molecular and cellular mechanisms involved in Helicobacter pylori-induced inflammation and oxidative stress 1, 2. Free Radical Biology and Medicine, **33**(3), 323-336.

[18] **Moran, A. P. (1996).** The role of lipopolysaccharide in Helicobacter pylori pathogenesis. Alimentary pharmacology & therapeutics, **10**(Sup1), 39-50.

[19] **Stothard P (2000).** The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques **28**:1102-1104.

[20] **Institute of Pathogen Biology, CAMS & PUMC, Bejing, China. (2003-2017).** Virulence Factors of Pathogenic Bacteria. Retrieved May 25, 2017, from Virulence Factors of Pathogenic Bacteria: http://www.mgc.ac.cn/VFs/

[21] **Yuzhen, Ye., & Adam, G.** "Flexible structure alignment by chaining aligned fragment pairs allowing twists" Bioinformatics Journal 19 (Supp2) (**2003).**

[22] Y Zhang. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics, 9: 40 **(2008).**

[23] **A Roy, A Kucukural, Y Zhang**. I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols, 5: 725-738 **(2010).**

[24] **J Yang, R Yan, A Roy, D Xu, J Poisson, Y Zhang.** The I-TASSER Suite: Protein structure and function prediction. Nature Methods, 12: 7-8 **(2015).**

[25] **Zhanwen Li, Yuzhen Ye and Adam Godzik**. "Flexible Structural Neighborhood - a database of protein structural similarities and alignments." Nucleic Acids Res., 34(Database issue):**D277-80, 2006.**

[26] **M. Menke, B. Berger, L. Cowen**, "Matt: Local Flexibility Aids Protein Multiple Structure Alignment", PLOS Computational Biology, Vol. 4, No 1. **2008.**

[27] **Fournier, P. E., Dubourg, G., & Raoult, D. (2014).** Clinical detection and characterization of bacterial pathogens in the genomics era. Genome medicine, 6(11), **114**.

[28] **Xiang, Z., Censini, S., Bayeli, P. F., Telford, J. L., Figura, N., Rappuoli, R., & Covacci, A. (1995).** Analysis of expression of CagA and VacA virulence factors in 43 strains of Helicobacter pylori reveals that clinical isolates can be divided into two major types and that CagA is not necessary for expression of the vacuolating cytotoxin. Infection and immunity, 63(1), **94-98.**

[29] **Kersulyte, D., Mukhopadhyay, A. K., Velapatiño, B., Su, W., Pan, Z., Garcia, C., ... & Yuan, Y. (2000).** Differences in genotypes of Helicobacter pylori from different human populations. Journal of bacteriology, 182(11), 3210-3218.

[30] **Brown, L. M. (2000).** Helicobacter pylori: epidemiology and routes of transmission. Epidemiologic reviews, 22(2), 283-297.

[31] **Atherton, J. C. (1998).** H. pylori virulence factors. British Medical Bulletin, 54(1), 105-120.

[32] **Hunt, R. H., Xiao, S. D., Megraud, F., Leon-Barua, R., Bazzoli, F., Van der Merwe, S., ... & Malfertheiner, P. (2011).** World Gastroenterology Organisation Global Guidelines Helicobacter pylori in developing countries August 2010: WGO global guidelines. South African Gastroenterology Review, 9(3), 16-22.

[33] **Xu, J., & Zhang, Y. (2010).** How significant is a protein structure similarity with TM-score= 0.5? Bioinformatics, 26(7), 889-895.

[34] **Kihara, D., Chen, H., & Yang, Y. D. (2009).** Quality assessment of protein structure models. Current Protein and Peptide Science, 10(3), 216-228.

Author Profile

Akoo Benedict received B.Sc. degree major in Biochemistry and Chemistry from Egerton University (Kenya) in 2005.He is currently a post-graduate student at Jomo Kenyatta University of Agriculture and Technology (JKUAT), (Kenya) studying Bioinformatics and Molecular Biology. During 2006-2017, he worked in the Information Communication and Technology Division, Coding and Transcription Department, as Coding Team Leader at a University Hospital in Kenya, the Aga Khan University Hospital, Nairobi. He is now a full time student at JKUAT.