Peer-reviewed version available at Int. J. Mol. Sci. 2018, 19, 3273; doi:10.3390/ijms19103273

Article

Structural Analysis of Variability and Interaction of the N-terminal of the Oncogenic Effector CagA of *Helicobacter pylori* with Phosphatidylserine

Cindy P. Ulloa-Guerrero [1], Maria del Pilar Delgado [1,*] and Carlos A. Jaramillo [1]

[1] Affiliation 1: Laboratory of Molecular and Bioinformatic Diagnosis, Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia.

* Correspondence: mdelgado@uniandes.edu.co

Version July 5, 2018 submitted to Preprints

Abstract: *Helicobacter pylori* cytotoxin-associated gene A protein (CagA) has been associated with the increase in virulence and risk of cancer. It has been demonstrated that CagA's translocation is dependent on its interaction with phosphatidylserine. We evaluated the variability of the N-terminal CagA in 127 sequences reported in NCBI, by referring to molecular interaction forces with the Phosphatidylserine and the docking of 3 mutations chosen from variations in specific positions. The major sites of conservation of the residues involved in CagA-Phosphatidylserine interaction were 617, 621 and 626 which had no amino acid variation. Position 636 had the lowest conservation score, so mutations in this position were evaluated to observe the differences in intermolecular forces of the CagA-Phosphatidylserine complex. We evaluated the docking of 3 mutations: K636A, K636R and K636N. The models of the crystal and mutations presented a ΔG of -8.919907, -8.665261, -8.701923, -8.515097 Kcal/mol, respectively, while mutations K636A, K636R, K636N and the crystal structure presented 0, 3, 4 and 1 H-bonds, respectively. Likewise, the bulk effect of the ΔG and amount of H-bonds was estimated in all of the docking models. The type of mutation affected both the ΔG ($\chi^2(1) = 93.82$, p-value $< 2.2x10^{-16}$) and the H-bonds ($\chi^2(1) = 91.93$, p-value $< 2.2x10^{-16}$). In all the data, 76.9% of the strains that exhibit the K636N mutation produced a severe pathology. The average H-bond count diminished when comparing the mutations with the crystal structure of all the docking models, which means that other molecular forces are involved in the CagA-Phosphatidylserine complex interaction.

Keywords: CagA; Phosphatidylserine; Phosphatidylserine mutations; Conservation N-terminal CagA; Homology modeling; Molecular docking

1. Introduction

Helicobacter pylori is a bacteria that colonizes and infects the digestive tract and is found in approximately 50% of the population [1]. Among the pathologies it causes are: gastritis, peptic ulcers, adenocarcinomas and mucosa-associated lymphoid tissue (MALT) lymphoma [2]. Nevertheless, only 1-5% of infected individuals present one of these severe gastric diseases [3,4]. Infection due to *H. pylori* has been recognized as an important risk factor for the presence of gastric cancer [5]. According to epidemiologic data, 60-90% of gastric cancer cases can be attributed to the presence of this microorganism [6]. Likewise, the World Health Organization has classified this pathogen as a type I carcinogen [7]. The relative risk of acquiring this pathology increases if patients are infected by positive CagA strains [5].

CagA gene is part of the 40kb pathogenicity island (cag PAI) that codes for a type IV secretion system (T4SS). This system is responsible for the translocation of CagA inside of epithelial cells through its injection [2,5]. Once inside the cell, CagA can modify the signaling pathways according to the presence [5,8–13] or absence [4],[14–20] of phosphorylation [21].

Phosphorylation of CagA is done by kinases Src and c-Abl of the host in the tyrosine residues of EPIYA motifs of the C-terminal region of the protein [3,22,23]. Eventually, all this allows for interactions

with cellular proteins to appear, thus leading to elongation, migration and dispersion of the cells due to the fact that these cells interfere with signaling pathways that control adhesion between cells, cellular growth and motility [3,22,23]. Additionally, independent phosphorylation pathways activate β-catenin, which leads to the disruption of the apical union complexes and the loss of cellular polarity [3,12,22,23]. CagA can also form dimers in cells independent of phosphorylation [24]. Dimerization is mediated by a multimerization sequence (CM), which is essential for the union of CagA with the PAR1 complex (MARK). Dimerization inhibits the kinase activity of PAR1 promoting the loss of cellular polarity [24]. Likewise, CagA can unite with c-Met. This union leads to an increase in the β-catenin regulation and the nuclear factor κB ($NF\kappa B$) that promote proliferation and inflammation, respectively [24]. Crystallographic studies have allowed for the elucidation of only the structure of the N-terminal because the C-terminal corresponds to an intrinsically disordered region possessing versatile folds [16,25]. The N-terminal section possesses three domains: Domain I is composed by 10 α helices, has a small interaction area (374 $Å^2$) and is structurally isolated from the other domains [25]. Domain II corresponds to a singular extended layer of beta sheets and 2 helicoidal subdomains of α helices. Domain III is composed of 4 α helices that form a complex with the C-terminal [25].

The oncogenic effector CagA is an important virulence factor of *H. pylori*. It is strongly associated with the severity of the pathology [26–31]. In several studies, the number and type of EPIYA have been associated with this pathology; nevertheless, in Colombia and elsewhere, a direct association has not been observed [32–34]. Generally, the interactions with motifs located in the C-terminal have been studied. Nonetheless, little attention has been given to the possible role that the N-terminal may play in the observed pathophysiology.

Generally, signal induction of CagA in host cells happens because of its interaction with the motifs located in the C-terminal. However, the N-terminal might also play an important role in this pathophysiology. Bagnolli et al. demonstrated that the N-terminal of CagA helps to direct the proteins to the plasmatic membrane of epithelial cells independent of the C-terminal [22]. Also, Pelz et al. proved that CagA consists of two independent domains [C and N terminals] that interact with membrane structures of host cells [5]. The first 200 amino acids of the N-terminal act as an inhibitory domain to cellular responses evoked by the C-terminal [5]. The N-terminal increases the rate and strength of the newly formed contacts between cells, diminishes cellular elongation and the construction of the apical membrane induced by the C-terminal and reduces the transcription activity of the TCF/ β catenin of the C-terminal. The former mediates the cell to cell adhesion by the E-cadherin-β catenin complex [5,35]. All this suggests that the N-terminal and the C-terminal also interact, and so influence the cellular response presented during the infection. Therefore, the N-terminal is involved in the attachment to the cell membrane. From this we can conclude that it is responsible for the localization of CagA in the phospholipid bilayer.

The medium domain presents a union site for the $\alpha 5\beta 1$ integrin necessary for the delivery of CagA in epithelial gastric cells [16]. Additionally, it possesses a positively charged region that is important in the union with the membrane, especially with phosphatidylserine (PS) [17,25]. Murata-Kamiya et al. [17] have reported that this union generates a rapid and transitory externalization of the PS, independent of apoptosis, in the bacterial union site [16]. The union of CagA with PS does not happen with domains that unite phospholipids but with the Lys-Xn-Arg-X-Arg [K-Xn-R-X-R] motif encountered in the N-terminal region of CagA [16]. The electrostatic interaction between the negative charge of the PS and the positive charge of the lysine and arginine residues of the K-Xn-R-X-R motif are highly conserved on the binding region of acidic phospholipids like PS [36]. Specifically, two arginine residues (R619 and R621) are conserved in Western strains of *H. pylori* [26695, G27, J99] and the F75 strain from East Asia [16]. Through the preparation of CagA mutants, the test for the union to lipids in vitro and through the transitory expression in MDCK cells, it was found that the residues K613, K614, K617, K621, R624, R626, K631, K635 and K636 are involved in the CagA-PS interaction [25]. In this interaction, CagA uses different association mechanisms depending on the polarity status of the epithelial cell. In a polar epithelial cell, CagA is distributed selectively to the inner face of the plasmatic

Peer-reviewed version available at Int. J. Mol. Sci. 2018, 19, 3273; doi:10.3390/ijms19103273

membrane, initiating the disruption of stretch unions, causing epithelial apico-basal polarity losses and inhibiting the kinase activity of PAR1 through the physical formation of a complex [16,22]. Also, in non-polarized cells, CagA is located in the membrane through a mechanism dependent on the EPIYA motif of the C-terminal [10].

CagA uses PS as a receptor that allows it to enter the cell [17]. The CagA-PS interaction plays an important role in mediating the delivery, intracellular location and pathophysiologic action of CagA where this protein could finally cause deregulation of several pathways that might eventually lead to cancer generation [17].

In this study, the variability of the amino terminal of the oncogenic effector CagA was determined by referring to molecular interaction forces with the PS. This was done through i) evaluating of the variability of CagA in the amino terminal region, ii) determine the possible amino acid variation shown in the CagA specific positions that interact with the PS, iii) assessing 3 mutations (K636A, K636R, and K636N) that were chosen from variations in specific positions shown by the MSA and iv) calculating the intermolecular interaction strength (free energy and Hydrogen bonds) from the chosen mutations. Finally, an association between the amino acid variability in position 636 and the severity of the pathology was found.

2. Results

2.1. Sequence Selection and Multiple Sequence Alignment (MSA)

Out of the 191 different sequences found in the search from NCBI, only 127 met all the criteria (pathology, isolation region and type of EPIYA). These sequences where translated to amino acids and aligned using different methods. From the different alignments obtained, the best alignment was selected using AQUA's NorMD score, with an acceptance score above 0.6 [37]. Consequently, the best alignment corresponded to the highest NorMD score, which was 1 for the selected MSA. Two sequences (88 and 94) that produced some gaps in the alignment were removed, generating lower total scores in AQUA. For this reason, the MSA used for the Consurf analysis included all of the 127 sequences.

2.2. Consurf analysis

The major sites of conservation found in CagA basically correspond to alpha helices of all the three domains. In these areas the conservation scores varied from 5 to 9. In Domain I, these corresponded to helices α1, α2 and α7 in the approximate positions 29-38, 45-121 and 157-172, respectively; in Domain II, they corresponded to the middle helices α13, α14 and α18; and in Domain III, they corresponded to initial helices α19 and α23 (Fig 1).

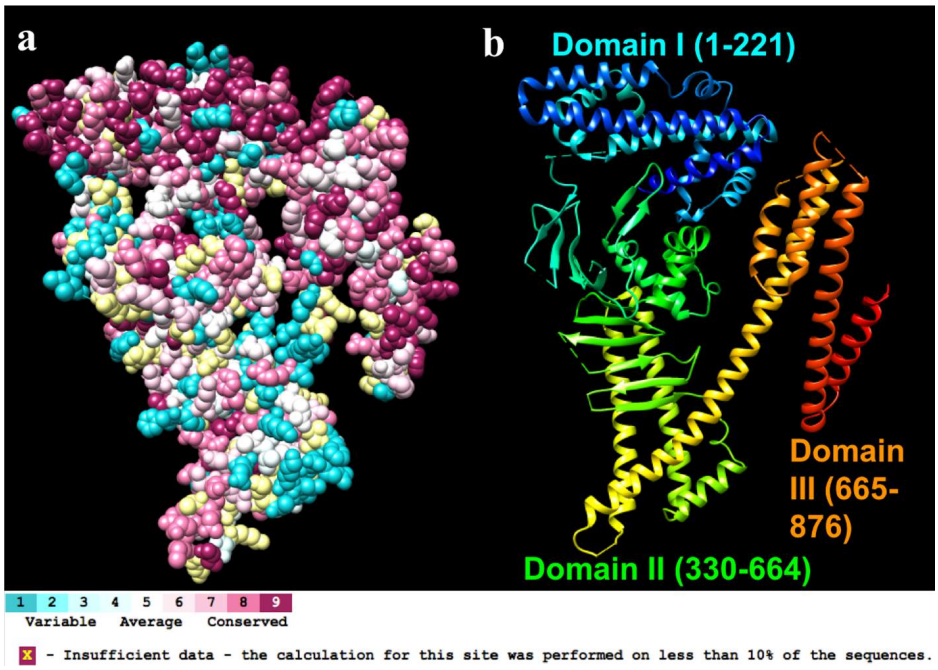

Figure 1. Consurf Results obtained from AQUA MSA. (A) Amino acids with conservations scores of CagA. (B) Ribbon view of each domain is specified.

Helix α18 in Domain II is the specific region of interaction between CagA and PS. There is a positive patch that attaches in a Velcro-like form to the negatively charged PS found in the plasmatic membrane [25]. Most of the residues involved in this interaction are highly conserved with scores from 7-9 (Fig 2). In this region, the most conserved positions were 617, 621, 626 with a score of 9 and a lysine/arginine amino acid, which is consistent with the findings by *Roujeinikova* [14].

Nevertheless, there were only two positions in which the score was below 5: in 619 and 636. These suggest possible positions of considerable variability where an amino acid change could alter the interaction forces of the complex CagA-PS. The amino acid in position 619 had an unreliable result due to the fact that it had excessively large confidence intervals [38]. Therefore, position 636 was used to create mutations in the crystal.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0085.v1

Peer-reviewed version available at Int. J. Mol. Sci. 2018, 19, 3273; doi:10.3390/ijms19103273

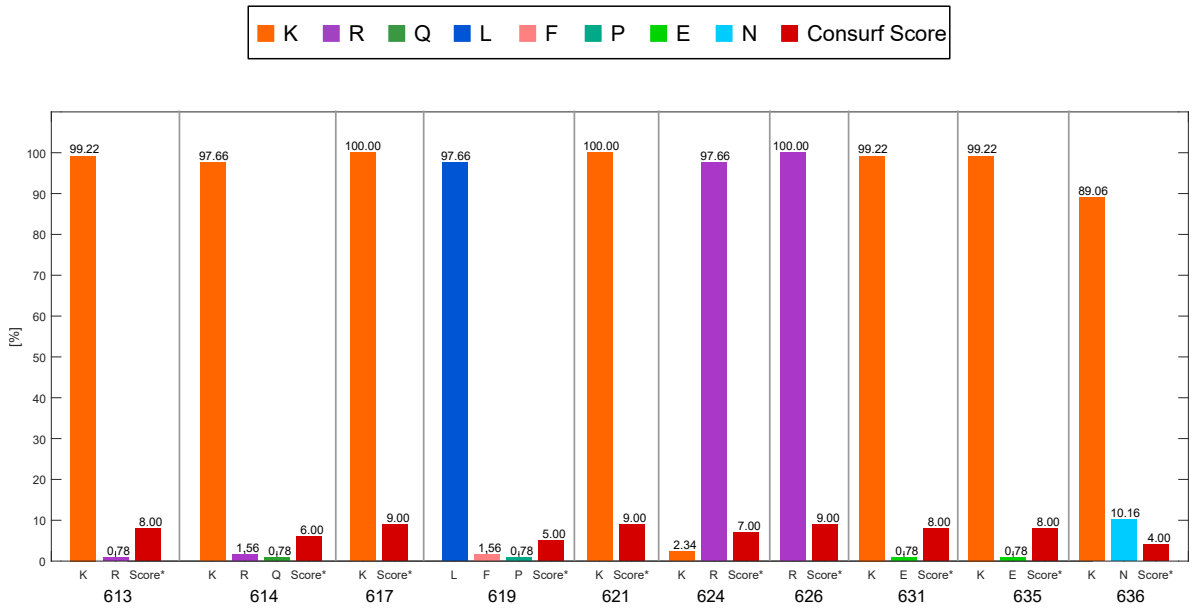

Figure 2. Percentage of Amino Acid Variation in CagA.

This results were obtained from the Consurf plataform. Amino acids in position 617, 621 and 626 had the highest conservation score, while the amino acid 636 had the lowest conservation score.

We evaluated 3 different mutations of the crystal (4DVY): K636A, K636N and K636R. The natural variation of lysine to asparagine was assessed in one of the mutations (K636N). The other two mutations were chosen in order to evaluate the reliability of the data obtained from the different dockings. For each of the mutations, 256 different binding models were acquired. Nonetheless, only the models that interacted with the α18 were selected. For all models, the free energy (ΔG) and hydrogen bonds were obtained. The best model from each mutation was determined using the following criteria: the highest ΔG and number of hydrogen bonds.

2.3. Docking of all Mutations

The ΔG for all the mutations varied within a very small range (Table 1). The highest value obtained was from the crystal (-8.919907 Kcal/mol), as was expected. The lowest free energy value was from the mutation K636N (-8.515097 Kcal/mol). On the other hand, the number of hydrogen bonds had the inverse effect of ΔG. While mutations K636A, K636R, K636N presented 0 (Fig 4), 3 (Fig 5), and 4 (Fig 6) hydrogen bonds, respectively, the crystal structure presented one type of this bond (Fig 3). Moreover, in each of the different interactions evaluated there was a hydrophobic interaction pocket between the CagA and the PS surrounded by highly polar interactions where the formation of hydrogen bonds occurred (Fig 3-6).

Table 1. Results for the best model from each mutation. The ΔG obtained from the molecular docking in Swiss-Dock and the hydrogen count obtained from Chimera.

Mutation	Model Number	Cluster	ΔG (Kcal/mol)	H-bonds
Crystal	1.81	10	-8.919907	1
K636A	1.69	8	-8.665261	0
K636R	1.74	8	-8.701923	3
K636N	1.60	0	-8.515097	4

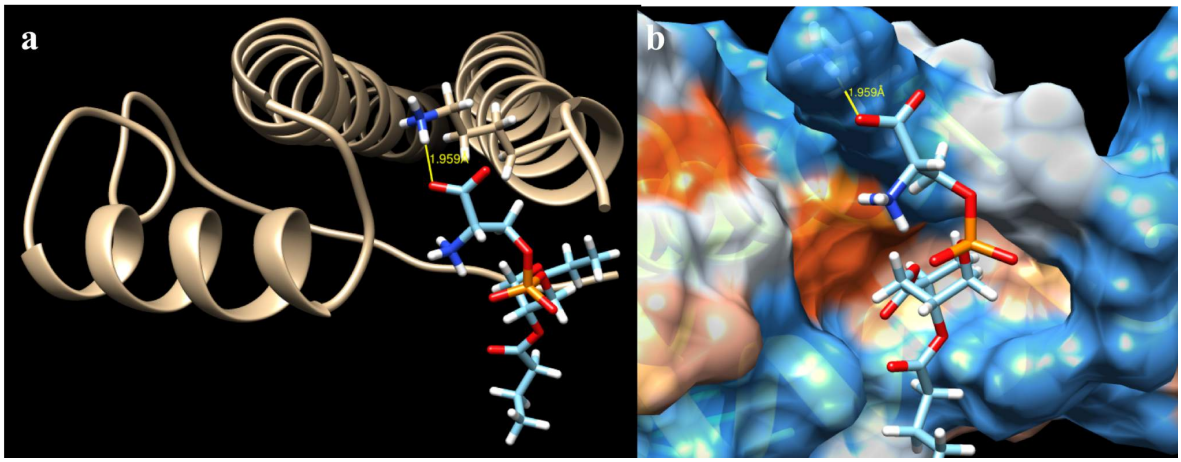

Figure 3. Interaction CagA-PS of the 4DVY crystal obtained from Chimera. (A) Hydrogen bonds and their distance (yellow lines) of interacting atoms of both molecules. (B) Surface hydrophobic view of CagA, color coded from dodger blue for the most hydrophilic, to white, to orange red for the most hydrophobic.

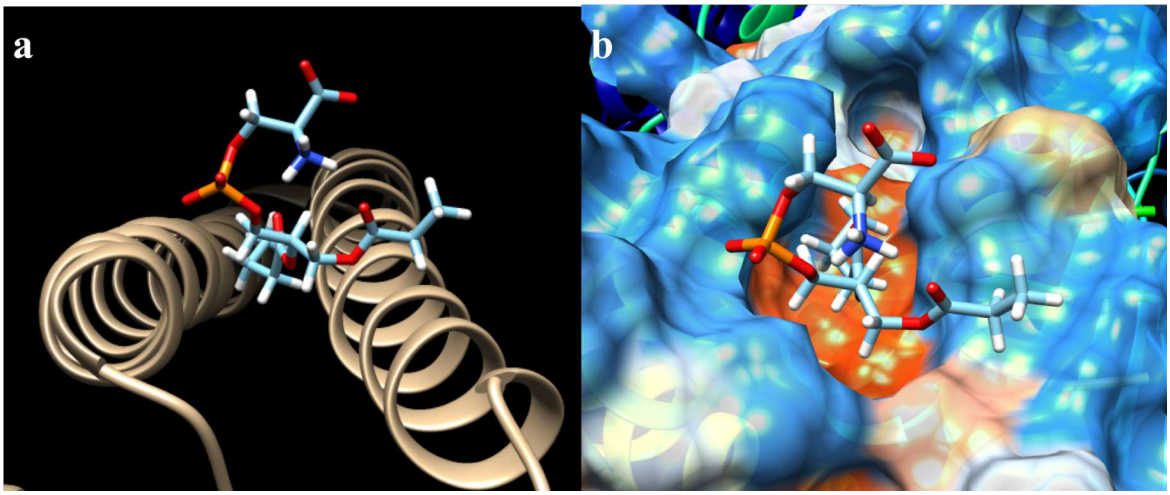

Figure 4. Interaction CagA-PS of the mutation K636A obtained from Chimera. (A) Hydrogen bonds and their distance (yellow lines) of interacting atoms of both molecules. (B) Surface hydrophobic view of CagA, color coded from dodger blue for the most hydrophilic, to white, to orange red for the most hydrophobic.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2018 doi:10.20944/preprints201807.0085.v1

Peer-reviewed version available at Int. J. Mol. Sci. 2018, 19, 3273; doi:10.3390/ijms19103273

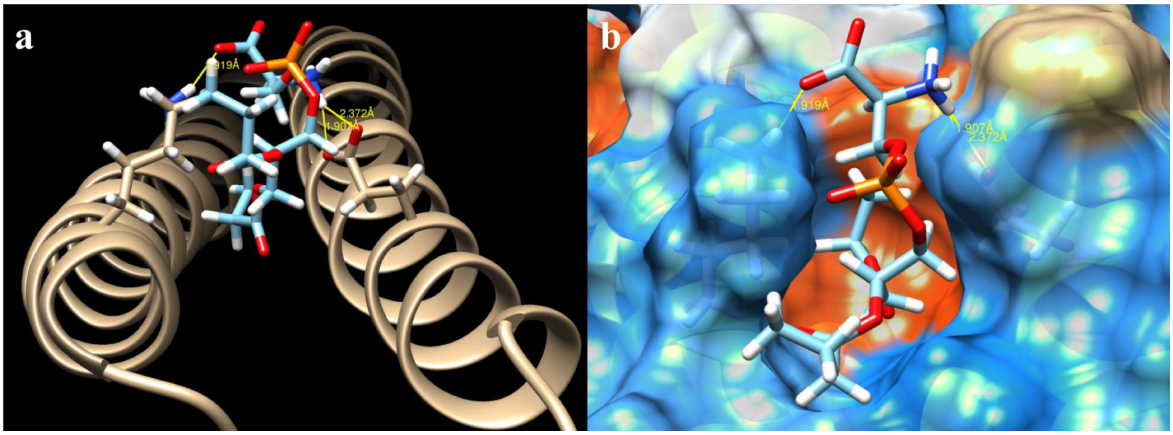

Figure 5. Interaction CagA-PS of the mutation K636R obtained from Chimera. (A) Hydrogen bonds and their distance (yellow lines) of interacting atoms of both molecules. (B) Surface hydrophobic view of CagA, color coded from dodger blue for the most hydrophilic, to white, to orange red for the most hydrophobic.

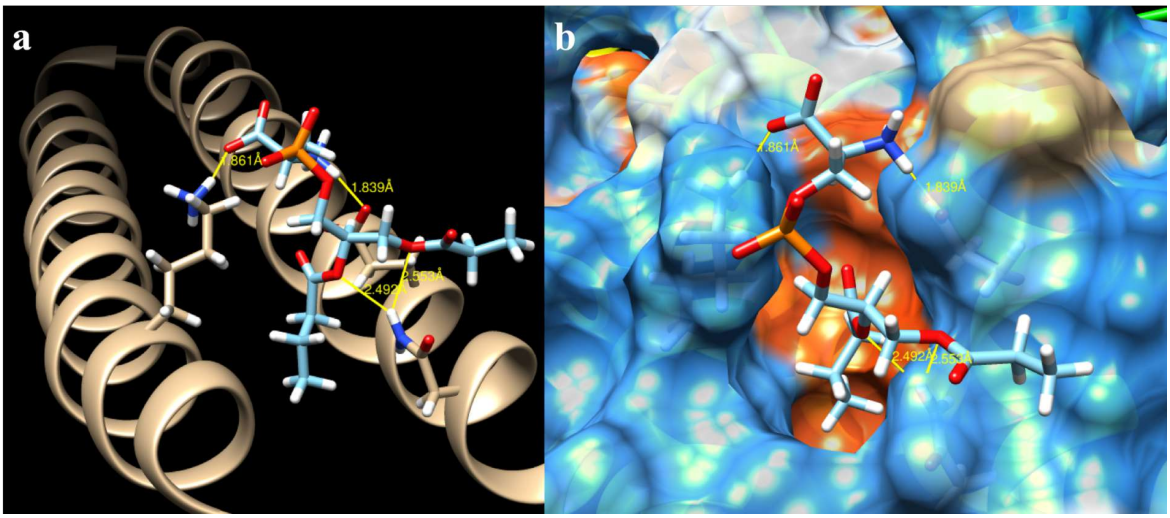

Figure 6. Interaction CagA-PS of the mutation K636N obtained from Chimera.(A) Hydrogen bonds and their distance (yellow lines) of interacting atoms of both molecules. (B) Surface hydrophobic view of CagA, color coded from dodger blue for the most hydrophilic, to white, to orange red for the most hydrophobic.

The change of a lysine to asparagine in position 636 generated two different hydrogen bridges between CagA and PS. This suggests an increase of the interactive force in the presence of this specific mutation (Fig 6). The asparagine is a non-polar amino acid that provides the same hydrogen bond donor counts and acceptor count as the lysine [39], but it is a smaller residue that allows a more favorable interaction with the negatively charged membrane surface.

Thus we proceeded to evaluate the sequences with this specific mutation in the entire database. We found that 13 of the 127 individuals (10.24%) had the K636N mutation, and 10 of those 13 sequences presented a severe pathology (76.9%) (Table 2). All of the above suggests that the increase in interaction may cause a higher degree of pathology. To evaluate the bulk effect of the ΔG and the amount of hydrogen bonds on the entire docking model for each mutation (Fig 7), a likelihood ratio test comparing the LMM of these two response variables with a null model was performed. The type of mutation affected both the free energy ($\chi^2(1) = 93.82$, p-value $< 2.2x10^{-16}$) and the hydrogen bonds ($\chi^2(1) =$

91.93, p-value $< 2.2x10^{-16}$). The mutation K636A diminished the ΔG by 1.105 ± 0.20 Kcal/mol and the quantity of hydrogen bonds by 3.32 ± 0.30. This matches the results obtained by *Hayashi et al.* in which there was a decrease in the interaction of co-immunoprecipitation assays comparing CagA-PS and CagA K636A mutation-PS [25]. Likewise, mutation K636R reduced the ΔG by 0.5729 ± 0.16 Kcal/mol and the quantity of hydrogen bonds by 2.90 ± 0.26 as was expected since the change of lysine to arginine, which are both positively charged amino acids, would have a smaller effect on the affinity of the interaction. However, the K636N mutation increased the ΔG by 0.6262 ± 0.23 Kcal/mol and diminished the quantity of hydrogen bonds by 3.32 ± 0.30. This means that the interaction increased the affinity of the molecule; however, the hydrogen bonding alone does not explain this conclusion. It may be that other types of interactions are playing an important role in the increase of free binding energy.

Table 2. Sequences with the K636N mutation obtained from NCBI.

Sequence Number	Accession Number	Region*	Pathology
39	22335784	Eastern	Severe
104	259123360	Western	Severe
106	259123364	Eastern	Severe
112	307135434	Eastern	Severe
115	307135440	Eastern	Severe
116	307135442	Eastern	Severe
117	307135444	Eastern	Mild
119	307135448	Eastern	Severe
121	307135452	Eastern	Mild
125	307135460	Eastern	Severe
127	307135464	Eastern	Severe
135	335335488	Western	Severe
150	345421953	Eastern	Mild

*Region of origin of the sample

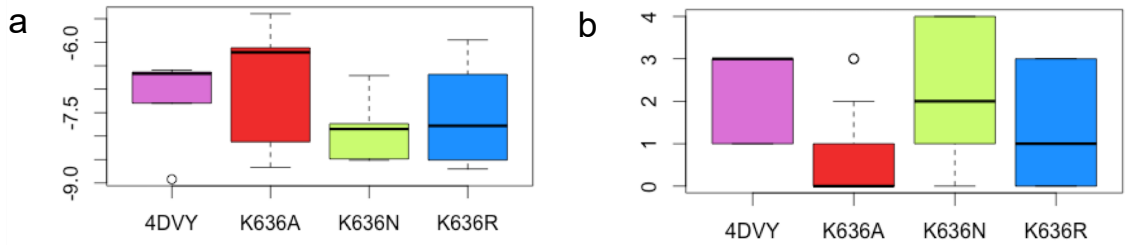

Figure 7. Comparison of the Docking results.(A) Delta G values for each mutation. (B) Hydrogen Count for each mutation.

3. Discussion

Bacterial-borne effector protein CagA plays an essential role in pathogenic activity due to its tethering to the plasmatic membrane [25]. The translocation of the protein is dependent on the interaction interface of several regions with the phospholipid membrane [5,16,17,22,35]. PS is one of the phospholipids that composes the eukaryotic membrane and is characterized by having a negatively charged head group [40]. As PS is involved in a number of cellular signaling pathways where several molecules like kinases, small GTPases and fusogenic proteins depend on this phospholipid to carry out their normal function, its disruption would trigger an homeostatic imbalance and apoptotic interference [40].

Signaling is mediated by PS functions in two ways: either via domains that stereospecifically recognize the head group, or by electrostatic interactions with a negatively charged surface of

membranes with rich PS and positively charged groups. In CagA, there is a positively charged helix α18 (residues 610-639) that has an exposed cluster of lysine/arginine residues at positions 613, 614, 617, 621, 624, 626, 631, 635 and 636 [17,25]. It is known that most of these residues of the positive patch of CagA-PS interaction are highly conserved between some *H. pylori* strains (26695, G27, J99, F75) [16]. Recently, the similarity between the membrane tethering helices of CagA and eukaryotic F-BAR domains was revealed [14]. Since these domains are involved in the interaction with lipids, this would explain the interaction of CagA with lipid membranes of human epithelial cells. As was observed by their MSA, the positively charged patch of residues found on the lipid binding face in F-BAR domains are also present in CagA [14]. Therefore, mutations in these interacting areas change the pathogenesis of *H. pylori*, explicitly in the degree of the hummingbird phenotype observed in MDK cells [17]. Hence, it may be possible to assume that if we take all the possible sequences of CagA into account and find the specific variations in the positively charged patch, we could find that variability in certain positions could explain the degree of pathology presented by a patient. This finding is important because of the lack of correlation between the number and type of EPIYA with the pathology found in some studies [32–34]. Thus, if we consider an additional factor, like the variation and force of interaction in the N-terminal, we could predict the prognosis of a patient more accurately.

In this study, we found that major sites of conservation in CagA mainly correspond to alpha helices of all three domains. Residues in positions 617, 621 and 626 are highly conserved and have no amino acid variation, which is consistent with the results from Roujeinikova [14]. On the other hand, the position with the highest variability was 636; therefore, different mutations were performed in this position to evaluate how the amino acid change could alter the interaction forces of the complex CagA-PS. To test this hypothesis, the free energy and amount of hydrogen bonds was determined. We found that these values are comparable and have the same order of magnitude as in other studies where computational and experimentally acquired ΔG and hydrogen bond count from proteins with small ligands [41–45] used, thus validating our data.

Additionally, we showed that more than half of the sequences that exhibit the K636N mutation correspond to a severe pathology (76.9%). This may be due to the fact that there is an increase in the interacting forces shown by the generation of a higher, bulk-free binding energy and more hydrogen bonds in this specific model. However, when taking into account the entire model for each mutation, the average hydrogen count diminished as compared to the mutations with the crystal structure. This may be due to the increased variation of the different interacting clusters and the fact that there are other possible interactions that were not taken into account. The latter could explain the increases in free energy. For example, the 636 residue in the CagA has a role in the electrostatic effect because the positively charged basic patch influences the strength of CagA binding to PS. So, if we could calculate the degree of electrostatic interaction, this may account for the discrepancies found with hydrogen bonds.

Moreover, studies have also suggested additional roles of the N-terminal CagA in the regulation and function of the entire protein. It is known that the N-terminal has a binding segment with the C-terminal that serves as a regulatory element. The interaction of the N-terminal and C-terminal enhances the localization of CagA via the positive patch, and strengthens the pathogenic scaffold/hub function of the protein [25]. This characteristic also accounts for the promiscuity of CagA as it promotes interaction with several host proteins [21,25,46]. The interaction of both segments allows for a determined, folded state of the protein that eventually leads to its oncogenic action.

In addition, epithelial cells are not all the same; they display a different polarity status. CagA has different mechanisms by which it can enter these cells depending on the degree of epithelial polarity. Polarized epithelial cells are rich in PS, so the CagA contains a binding motif to PS that initiates a disruption of tight junctions and causes loss of epithelial, apico-basal polarity by inhibiting kinase activity of PAR1 through physical complex formation [17,22]. In non-polarized epithelial cells, CagA is located in the plasma membrane through a C-terminal EPIYA motif in CagA [11]. These EPIYA motifs allow CagA to bind to several host cell proteins, a process which generates cellular elongation,

Peer-reviewed version available at Int. J. Mol. Sci. 2018, 19, 3273; doi:10.3390/ijms19103273

migration and dispersion since these motifs interfere with signaling pathways involved in cellular adhesion, growth and motility [3,12,22,23].

Consequently, as was mentioned earlier, CagA may include both pathways of N-terminal PS binding and EPIYA motif binding proteins via independent and dependent phosphorylation on polar and non-polar epithelial cells, respectively. Cellular disruptions caused by CagA, give rise to the first steps in the transformation to neoplastic tissue. This tissue, in its transformed state, eventually causes carcinogenic gastric epithelial cells. Our study underscores the importance of considering the molecular forces for the interactions of CagA-PS as there are implications of this on the pathogenic development and the consideration of several variables that influence the interaction of CagA with host cells. All of this could possibly have important therapeutic implications on how *H. pylori* infections are handled in the future.

The study of molecular forces involved in the interaction of CagA binding with proteins in host cells helps us to have a better understanding of how it could cause different degrees of pathology. However, the development of different pathologies requires a multiple step process that involves several different variables of the N-terminal as well as the C-terminal. The interaction of CagA with PS requires a set of positively charged residues that is highly conserved among the sequences analyzed. The most variable position naturally found was the K636N mutation, which generated a higher free energy change and a lower hydrogen count with respect to the crystal structure 4DVY when considering all docking models. Nevertheless, the amount of hydrogen bonds increased when comparing the best model of all the mutations; specifically, when there was a lysine to asparagine change, it generated two additional hydrogen bonds. This mutation was also associated with a severe pathology, which means that there could be other molecular forces involved in the CagA-PS complex interaction that were not taken into account. For future studies, it is important to include other types of intermolecular interactions to evaluate the bulk effect of the affinity between CagA-PS. Similarly, to have a complete model of how CagA affects the final pathology, both the N-terminal and C-terminal interactions must be considered. In this study, only the N-terminal was assessed due to limitations of the existing crystallographic structure on which all the results evaluated depended.

4. Material and Methods

4.1. Database

NCBI was consulted for all the DNA sequences of the complete segment of cagA under "*Helicobacter pylori* CagA complete cds" with data up to January 2014 as the criteria. Sequences that presented information about the pathology, the region from which the sample was taken and EPIYA type were selected for the data base construction.

4.2. Translation of Gene Bank DNA sequences

The translation from the DNA sequences to amino acids was performed by the AASA (Amino acid Sequence Analyzer) program [34] using an open reading frame that codified for the complete protein. The C-terminal was eliminated from each sequence after the 877^{th} amino acid, due to the size of the 4DVY crystal of 876 amino acids [25].

4.3. Multiple Sequence Alignment (MSA) of Translated Sequences and Quality Assessment

All sequences were aligned with the 4DVY crystal with MUSCLE [47], T-COFFEE [48] and MAFT [49]. Residues from the sequences that presented amino acids before the initiation residue (methionine) were eliminated.

RASCAL [37] was used to determine which sequences should be eliminated and to refine and improve the multiple alignments obtained with MUSCLE, T-COFFEE and MAFT.

All of these programs were used within Automated Quality Improvement for Multiple Sequence alignments (AQUA) [50].

4.4. Conservation level and variability of the interaction residues with PS

Using the purified MSA, the Consurf [38] platform was used to determine the conservation level of the protein. From the alignment results, we observed the residues that are directly involved in the interaction with the PS (K613, K614, K617, K621, R624, R626, K631, K635 and K636) and determined which amino acid variations were present between each of these positions. Then, the one that presented the most variability was chosen (K636).

A mutation was generated in a Swiss PDB viewer [51] based on different variations that were found in the most variable residue (K636A, K636N, K636R).

4.5. Determination of interaction forces between CagA N-terminal and PS

From the PDBs generated from the respective mutations, SwissDock was used to dock with PS [52]. Putative complexes found in the CagA-PS interaction region were taken. Using Chimera [53], we found the amount of hydrogen bonds present in each interaction. Finally, using the results obtained from the docking, we obtained different changes in the Gibbs free energy for all the models of each mutation.

4.6. Statistical analysis

For the statistical analysis of the data, we redefined certain groups: for the region, data was grouped in eastern and western; for the pathology, we took into account the mild and severe disease criteria. A slight gastric disease included patients with erythematous and/or chronic nodular gastritis. On the other hand, severe disease included patients that presented erosive gastric disease and an acute gastric ulcer, dysplasia/metaplasia and gastric cancer. For the analysis of the delta G and the hydrogen bond count of all the docking models, a linear mixed model (LMM) with random effects was calculated for each of the two response parameters using R programming [54]. This LMM was constructed considering the clustering in each mutation as a random effect in the two models. Followed by a likelihood ratio test using the ANOVA function comparing the LMM the effect of the free energy and hydrogen bond with a null model.

Author Contributions: C.P.U contributed to the study design, collected the data, interpreted and analysis of the results, wrote and prepared the final version of the manuscript; C.A.J. conceived and supervised the whole project, interpreted the results and prepared the final version of the manuscript; M.P.D. supervised the whole project and prepared the final version of the manuscript. All authors read and approved the final manuscript for publication.

Funding: This research was funded by Vicerrectoria de Investigaciones of Universidad de Los Andes, Bogotá (Colombia) grant number RV-UA04022013.

Acknowledgments: We would like to thank Juan Manuel Cordovéz, Claudia Chica, Adolfo Amézquita and Andres Gonzalez for scientific assistance, for assessing the analysis of the MSA results, statistical tests and docking results, respectively. We also thank Sergio Vásquez for translating and Nicole Bruskewitz for proofreading the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CagA	Cytotoxin-associated gen A
Cag PAI	CagA pathogenic island
c-Abl	Mammalian Abelson murine leukemia viral oncogene
CM	Multimerization sequence
LMM	Linear Mixed Model
MALT	Mucosa-associated lymphoid tissue
MARK	Microtubule affinity-regulating kinase
MSA	Multiple Sequence Alignment
NFκB	Nuclear factor κB
PAR1	Protease-activated receptor 1
PS	Phosphatidylserine
Src	Proto-oncogene tyrosine-protein kinase
TCF	Transcription factor
T4SS	Type IV secretion system

References

1. Covacci, A.; Telford, J.L.; Del Giudice, G.; Parsonnet, J.; Rappuoli, R. Helicobacter pylori virulence and genetic geography. *Science (New York, N.Y.)* **1999**, *284*, 1328–1333. doi:10.1126/science.284.5418.1328.
2. Woon, A.P.; Tohidpour, A.; Alonso, H.; Saijo-Hamano, Y.; Kwok, T.; Roujeinikova, A. Conformational analysis of isolated domains of Helicobacter pylori CagA. *PLoS ONE* **2013**, *8*, 1–11. doi:10.1371/journal.pone.0079367.
3. Wroblewski, L.E.; Peek, R.M.; Wilson, K.T. Helicobacter pylori and gastric cancer: Factors that modulate disease risk. *Clinical Microbiology Reviews* **2010**, *23*, 713–739. doi:10.1128/CMR.00011-10.
4. Tsang, Y.H.; Lamb, A.; Romero-Gallo, J.; Huang, B.; Ito, K.; Peek, R.M.J.; Ito, Y.; Chen, L.F. Helicobacter pylori CagA targets gastric tumor suppressor RUNX3 for proteasome-mediated degradation. *Oncogene* **2010**, *29*, 5643–5650. doi:10.1038/onc.2010.304.
5. Pelz, C.; Steininger, S.; Weiss, C.; Coscia, F.; Vogelmann, R. A novel inhibitory domain of Helicobacter pylori protein CagA reduces CagA effects on host cell biology. *Journal of Biological Chemistry* **2011**, *286*, 8999–9008. doi:10.1074/jbc.M110.166504.
6. Malfertheiner, P.; Sipponen, P.; Naumann, M.; Moayyedi, P.; Mégraud, F.; Xiao, S.D.; Sugano, K.; Nyrén, O. Helicobacter pylori eradication has the potential to prevent gastric cancer: A state-of-the-art critique. *American Journal of Gastroenterology* **2005**, *100*, 2100–2115. doi:10.1111/j.1572-0241.2005.41688.x.
7. Vogiatzi, P.; Cassone, M.; Luzzi, I.; Lucchetti, C.; Otvos, L.; Giordano, A. Helicobacter pylori as a class I carcinogen: Physiopathology and management strategies. *Journal of Cellular Biochemistry* **2007**, *102*, 264–273. doi:10.1002/jcb.21375.
8. Baek, H.Y.; Lim, J.W.; Kim, H. Interaction between the Helicobacter pylori CagA and alpha-Pix in gastric epithelial AGS cells. *Annals of the New York Academy of Sciences* **2007**, *1096*, 18–23. doi:10.1196/annals.1397.065.
9. Churin, Y.; Al-Ghoul, L.; Kepp, O.; Meyer, T.F.; Birchmeier, W.; Naumann, M. Helicobacter pylori CagA protein targets the c-Met receptor and enhances the motogenic response. *Journal of Cell Biology* **2003**, *161*, 249–255. doi:10.1083/jcb.200208039.
10. Suzuki, M.; Mimuro, H.; Suzuki, T.; Park, M.; Yamamoto, T.; Sasakawa, C. Interaction of CagA with Crk plays an important role in Helicobacter pylori-induced loss of gastric epithelial cell adhesion. *The Journal of experimental medicine* **2005**, *202*, 1235–1247. doi:10.1084/jem.20051027.
11. Tsutsumi, R.; Higashi, H.; Higuchi, M.; Okada, M.; Hatakeyama, M. Attenuation of Helicobacter pylori CagA·SHP-2 signaling by interaction between CagA and C-terminal Src kinase. *Journal of Biological Chemistry* **2003**, *278*, 3664–3670. doi:10.1074/jbc.M208155200.
12. Kurashima, Y.; Murata-Kamiya, N.; Kikuchi, K.; Higashi, H.; Azuma, T.; Kondo, S.; Hatakeyama, M. Deregulation of β-catenin signal by Helicobacter pylori CagA requires the CagA-multimerization sequence. *International Journal of Cancer* **2008**, *122*, 823–831. doi:10.1002/ijc.23190.
13. Krueger, S.; Hundertmark, T.; Kuester, D.; Kalinski, T.; Peitz, U.; Roessner, A. Helicobacter pylori alters the distribution of ZO-1 and p120ctn in primary human gastric epithelial cells. *Pathology Research and Practice* **2007**, *203*, 433–444. doi:10.1016/j.prp.2007.04.003.

14. Roujeinikova, A. Phospholipid binding residues of eukaryotic membrane-remodelling F-BAR domain proteins are conserved in Helicobacter pylori CagA. *BMC Research Notes* **2014**, *7*, 1–7. doi:10.1186/1756-0500-7-525.
15. Nešić, D.; Buti, L.; Lu, X.; Stebbins, C.E. Structure of the Helicobacter pylori CagA oncoprotein bound to the human tumor suppressor ASPP2. *Proceedings of the National Academy of Sciences of the United States of America* **2014**, *111*, 1562–7. doi:10.1073/pnas.1320631111.
16. Kaplan-Turkoz, B.; Jimenez-Soto, L.F.; Dian, C.; Ertl, C.; Remaut, H.; Louche, a.; Tosi, T.; Haas, R.; Terradot, L. Structural insights into Helicobacter pylori oncoprotein CagA interaction with 1 integrin. *Proceedings of the National Academy of Sciences* **2012**, *109*, 14640–14645. doi:10.1073/pnas.1206098109.
17. Murata-Kamiya, N.; Kikuchi, K.; Hayashi, T.; Higashi, H.; Hatakeyama, M. Helicobacter pylori exploits host membrane phosphatidylserine for delivery, localization, and pathophysiological action of the CagA oncoprotein. *Cell Host and Microbe* **2010**, *7*, 399–411. doi:10.1016/j.chom.2010.04.005.
18. Sokolova, O.; Maubach, G.; Naumann, M. MEKK3 and TAK1 synergize to activate IKK complex in Helicobacter pylori infection. *Biochimica et biophysica acta* **2014**, *1843*, 715–724. doi:10.1016/j.bbamcr.2014.01.006.
19. Lamb, A.; Yang, X.D.; Tsang, Y.H.N.; Li, J.D.; Higashi, H.; Hatakeyama, M.; Peek, R.M.; Blanke, S.R.; Chen, L.F. Helicobacter pylori CagA activates NF-kappaB by targeting TAK1 for TRAF6-mediated Lys 63 ubiquitination. *EMBO reports* **2009**, *10*, 1242–1249. doi:10.1038/embor.2009.210.
20. Coombs, N.; Sompallae, R.; Olbermann, P.; Gastaldello, S.; Göppel, D.; Masucci, M.G.; Josenhans, C. Helicobacter pylori affects the cellular deubiquitinase USP7 and ubiquitin-regulated components TRAF6 and the tumour suppressor p53. *International Journal of Medical Microbiology* **2011**, *301*, 213–224. doi:10.1016/j.ijmm.2010.09.004.
21. Backert, S.; Tegtmeyer, N.; Selbach, M. The versatility of helicobacter pylori caga effector protein functions: The master key hypothesis. *Helicobacter* **2010**, *15*, 163–176. doi:10.1111/j.1523-5378.2010.00759.x.
22. Bagnoli, F.; Buti, L.; Tompkins, L.; Covacci, A.; Amieva, M.R. Helicobacter pylori CagA induces a transition from polarized to invasive phenotypes in MDCK cells. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 16339–16344. doi:10.1073/pnas.0502598102.
23. Amieva, M.R.; Vogelmann, R.; Covacci, A.; Tompkins, L.S.; Nelson, W.J.; Falkow, S. Disruption of the epithelial apical-junctional complex by Helicobacter pylori CagA. *Science (New York, N.Y.)* **2003**, *300*, 1430–1434. doi:10.1126/science.1081919.
24. Yamaoka, Y. Mechanisms of disease: Helicobacter pylori virulence factors. *Nature Reviews Gastroenterology and Hepatology* **2010**, *7*, 629–641, [NIHMS150003]. doi:10.1038/nrgastro.2010.154.
25. Hayashi, T.; Senda, M.; Morohashi, H.; Higashi, H.; Horio, M.; Kashiba, Y.; Nagase, L.; Sasaya, D.; Shimizu, T.; Venugopalan, N.; Kumeta, H.; Noda, N.N.; Inagaki, F.; Senda, T.; Hatakeyama, M. Tertiary Structure-Function Analysis Reveals the Pathogenic Signaling Potentiation Mechanism of Helicobacter pylori Oncogenic Effector CagA. *Cell Host & Microbe* **2012**, *12*, 20–33. doi:10.1016/j.chom.2012.05.010.
26. Ferreira, R.M.; Machado, J.C.; Leite, M.; Carneiro, F.; Figueiredo, C. The number of Helicobacter pylori CagA EPIYA C tyrosine phosphorylation motifs influences the pattern of gastritis and the development of gastric carcinoma. *Histopathology* **2012**, *60*, 992–998. doi:10.1111/j.1365-2559.2012.04190.x.
27. Beltrán-Anaya, F.O.; Poblete, T.M.T.M.; Román-Román, A.; Reyes, S.S.; de Sampedro, J.J.; Peralta-Zaragoza, O.; Rodríguez, M.Á.; del Moral-Hernández, O.; Illades-Aguiar, B.; Fernández-Tilapa, G. The EPIYA-ABCC motif pattern in CagA of Helicobacter pylori is associated with peptic ulcer and gastric cancer in Mexican population. *BMC Gastroenterology* **2014**, *14*, 1–11. doi:10.1186/s12876-014-0223-9.
28. Batista, S.a.; Rocha, G.a.; Rocha, A.M.C.; Saraiva, I.E.B.; Cabral, M.M.D.a.; Oliveira, R.C.; Queiroz, D.M.M. Higher number of Helicobacter pylori CagA EPIYA C phosphorylation sites increases the risk of gastric cancer, but not duodenal ulcer. *BMC microbiology* **2011**, *11*, 61. doi:10.1186/1471-2180-11-61.
29. Ferreira, R.M.; Machado, J.C.; Figueiredo, C. Clinical relevance of Helicobacter pylori vacA and cagA genotypes in gastric carcinoma. *Best Practice & Research Clinical Gastroenterology* **2014**, *28*, 1003–1015. doi:10.1016/j.bpg.2014.09.004.
30. Sicinschi, L.a.; Correa, P.; Peek, R.M.; Camargo, M.C.; Piazuelo, M.B.; Romero-Gallo, J.; Hobbs, S.S.; Krishna, U.; Delgado, a.; Mera, R.; Bravo, L.E.; Schneider, B.G. CagA C-terminal variations in Helicobacter pylori strains from Colombian patients with gastric precancerous lesions. *Clinical Microbiology and Infection* **2010**, *16*, 369–378. doi:10.1111/j.1469-0691.2009.02811.x.

31. Breurec, S.; Michel, R.; Seck, a.; Brisse, S.; Côme, D.; Dieye, F.B.; Garin, B.; Huerre, M.; Mbengue, M.; Fall, C.; Sgouras, D.N.; Thiberge, J.M.; Dia, D.; Raymond, J. Clinical relevance of cagA and vacA gene polymorphisms in Helicobacter pylori isolates from Senegalese patients. *Clinical Microbiology and Infection* **2012**, *18*, 153–159. doi:10.1111/j.1469-0691.2011.03524.x.
32. Fajardo, C.A.; Quiroga, A.J.; Coronado, A.; Labrador, K.; Acosta, N.; Delgado, P.; Jaramillo, C.; Bravo, M.M. CagA EPIYA polymorphisms in Colombian Helicobacter pylori strains and their influence on disease-associated cellular responses. *World journal of gastrointestinal oncology* **2013**, *5*, 50–9. doi:10.4251/wjgo.v5.i3.50.
33. Püls, J.; Fischer, W.; Haas, R. Activation of Helicobacter pylori CagA by tyrosine phosphorylation is essential for dephosphorylation of host cell proteins in gastric epithelial cells. *Molecular Microbiology* **2002**, *43*, 961–969. doi:10.1046/j.1365-2958.2002.02780.x.
34. Acosta, N.; Quiroga, A.A.; Delgado, P.; Bravo, M.M.M.M.; Jaramillo, C. Helicobacter pylori CagA protein polymorphisms and their lack of association with pathogenesis. *World journal of gastroenterology* **2010**, *16*, 3936–3943. doi:10.3748/wjg.v16.i31.3936.
35. Steininger, S.; Pelz, C.; Vogelmann, R. Purpose of recently detected inhibitory domain of the Helicobacter pylori protein CagA. *Gut Microbes* **2011**, *2*, 37–41. doi:10.4161/gmic.2.3.15872.
36. Lemmon, M.A. Membrane recognition by phospholipid-binding domains. *Nat Rev Mol Cell Biol* **2008**, *9*, 99–111.
37. Thompson, J.D.; Thierry, J.C.; Poch, O. RASCAL: Rapid scanning and correction of multiple sequence alignments. *Bioinformatics* **2003**, *19*, 1155–1161. doi:10.1093/bioinformatics/btg133.
38. Landau, M.; Mayrose, I.; Rosenberg, Y.; Glaser, F.; Martz, E.; Pupko, T.; Ben-Tal, N. {ConSurf} 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Research* **2005**, *33*, W299—-W302. doi:10.1093/nar/gki370.
39. Barnes, M.R.; Gray, I.C. Bioinformatics for Geneticists; John Wiley & Sons, Ltd, 2003; Vol. 4, chapter Chapter 14, pp. 289–314. doi:10.1002/0470867302.
40. Kay, J.G.; Grinstein, S. Lipid-mediated Protein Signaling; Springer: Blacksburg,VA,USA, 2013; chapter Chapter 10, pp. 117–193. doi:10.1007/978-94-007-6331-9.
41. Wilchek, M.; Bayer, E.a.; Livnah, O. Essentials of biorecognition: The (strept)avidin-biotin system as a model for protein-protein and protein-ligand interaction. *Immunology Letters* **2006**, *103*, 27–32. doi:10.1016/j.imlet.2005.10.022.
42. Friesner, R.a.; Murphy, R.B.; Repasky, M.P.; Frye, L.L.; Greenwood, J.R.; Halgren, T.a.; Sanschagrin, P.C.; Mainz, D.T. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry* **2006**, *49*, 6177–6196. doi:10.1021/jm051256o.
43. Zoete, V.; Michielin, O. EADock : Docking of Small Molecules Into Protein Active Sites With a Multiobjective Evolutionary Optimization **2007**. *1025*, 1010–1025. doi:10.1002/prot.
44. Matias, P.M.; Donner, P.; Coelho, R.; Thomaz, M.; Peixoto, C.; Macedo, S.; Otto, N.; Joschko, S.; Scholz, P.; Wegg, A.; Basler, S.; Schafer, M.; Egner, U.; Carrondo, M.A. Structural evidence for ligand specificity in the binding domain of the human androgen receptor. Implications for pathogenic gene mutations. *The Journal of biological chemistry* **2000**, *275*, 26164–26171. doi:10.1074/jbc.M004571200.
45. Looger, L.L.; Dwyer, M.A.; Smith, J.J.; Hellinga, H.W. Computational design of receptor and sensor proteins with novel functions. *Nature* **2003**, *423*, 185–190. doi:10.1038/nature01578.1.
46. Hatakeyama, M. Linking epithelial polarity and carcinogenesis by multitasking Helicobacter pylori virulence factor CagA. *Oncogene* **2008**, *27*, 7047–54. doi:10.1038/onc.2008.353.
47. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **2004**, *32*, 1792–1797. doi:10.1093/nar/gkh340.
48. Di Tommaso, P.; Moretti, S.; Xenarios, I.; Orobitg, M.; Montanyola, A.; Chang, J.M.; Taly, J.F.; Notredame, C. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Research* **2011**, *39*, W13–W17. doi:10.1093/nar/gkr245.
49. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **2013**, *30*, 772–780. doi:10.1093/molbev/mst010.
50. Muller, J.; Creevey, C.J.; Thompson, J.D.; Arendt, D.; Bork, P. AQUA: Automated quality improvement for multiple sequence alignments. *Bioinformatics* **2010**, *26*, 263–265. doi:10.1093/bioinformatics/btp651.

460 51. Biasini, M.; Bienert, S.; Waterhouse, A.; Arnold, K.; Studer, G.; Schmidt, T.; Kiefer, F.; Gallo Cassarino, T.;
461 Bertoni, M.; Bordoli, L.; Schwede, T. SWISS-MODEL: Modelling protein tertiary and quaternary structure
462 using evolutionary information. *Nucleic Acids Research* **2014**, *42*, 252–258. doi:10.1093/nar/gku340.
463 52. Grosdidier, A.; Zoete, V.; Michielin, O. EADock: docking of small molecules into protein active sites with a
464 multiobjective evolutionary optimization. *Proteins* **2007**, *67*, 1010–1025. doi:10.1002/prot.21367.
465 53. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF
466 Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*
467 **2004**, *25*, 1605–1612. doi:10.1002/jcc.20084.
468 54. Team, R.D.C. R: A language and environment for statistical computing, 2012. doi:3-900051-07-0.