A principal component regression model for predicting phytochemical binding to the *H. pylori* CagA protein

Jose Isagani B. Janairo

Biology Department, De La Salle University, 2401 Taft Avenue, Manila 0922, Philippines

Corresponding author email: jose.isagani.janairo@dlsu.edu.ph

Abstract

Helicobater pylori is an important causative factor in the pathogenesis of multiple gastrointestinal diseases. One of the factors responsible for the virulence of *H. pylori* is the cagA protein, which can interfere with a number of cellular signaling processes once this protein is transferred inside the host cell. Thus, inhibiting the interaction of the cagA protein with the host cell membrane using small molecular inhibitors appears to be a promising pharmacological strategy. In this study, a predictive model for the binding free energy of natural compounds towards the cagA protein is presented. The formulated model which is built on principal component – multiple linear regression demonstrates reliable accuracy (r^2_{test} = 0.92, $RMSE_{test}$ = 0.483), while only requiring five independent variables for the prediction. It was further noted that topological descriptors had the greatest influence on the generated principal components which served as the predictors. The created regression model can help promote and accelerate the discovery of natural compounds as cagA binders for the development of anti-*H. pylori* agents.

Keywords: computer-aided molecular design, drug discovery, machine learning, QSAR

1. Introduction

Helicobater pylori is a Gram-negative bacterium that lives in the gastric mucosa causing inflammation, and is implicated in a number of gastrointestinal diseases such as gastric and duodenal ulcers, and gastric cancers (Suerbaum and Michetti 2002). There are two types of *H. pylori* strain, the $cagA^+$ and the $cagA^-$. Evidences suggest that the $cagA^+$ strain is more virulent, since infection of this strain elevates

the risk of the host to gastric and duodenal ulcers (Nomura et al. 2002), gastric carcinoma (Rokkas et al. 1999), colorectal cancer (Shmuely et al. 2001), among others. Upon adhesion of cagA^{+} *H. pylori* to the gastric mucosa, the cagA protein is transferred to the host cell via the type IV secretion system (TFSS) and the cagA protein interacts with the phosphatidylserine (PS) of the cell membrane (Hatakeyama 2017). Once inside the cell, the phosphorylated cagA protein binds with other cellular compounds, resulting to the interference of normal cellular signaling pathways (Handa et al. 2007). Considering that the interaction between the cagA protein and the membrane PS is crucial for the entry of the cagA protein into the host cell, inhibiting the interaction with small molecules appears to be a promising pharmacological strategy.

Phytochemicals have been considered a promising source of therapeutic agents against *H. pylori* (Lawal et al. 2011; Vale and Oleastro 2014). However, only a few compounds have been identified so far as potential binders to cagA. Methylantcinate B which was extracted from *Antrodia camphorata* (Lin et al. 2013), flavonoids from *Syzgium alternifolium* (Chandra Babu et al. 2017), and curcumin (Srivastava et al. 2015) are some of the natural compounds that exhibit promising potential as cagA binders. Racha et al. (2019) previously conducted docking and molecular dynamics simulations between a series of natural compounds and the cagA protein. The study has provided important insights on the nature of ligands that favorably bind with the cagA protein. In order to extend and generalize the findings of this study which are useful for drug design and discovery, the formulation of a quantitative structure-activity relationship model is needed. The development of a regression model that can predict the binding affinity of compounds to cagA can help accelerate the discovery and development of CagA binders. Considering that natural compounds often possess favorable properties, the bioactive compounds that the regression model may discovery can also serve as lead compounds for other pharmacological purposes (Rastelli et al. 2020). In this study, a principal component regression model is presented that can predict the binding free energy (BFE) of phytochemicals to the cagA protein.

2. Methodology

The list of compounds and their corresponding binding free energy towards the CagA protein were obtained from Racha et al. (2019). The 38 compounds were then converted into their SMILES format, thereafter their geometric, constitutional, hybrid, electronic, and topological descriptors were calculated using the rcdk package (Guha 2007). From the resulting dataset composed of the compounds and their chemical descriptors, principal component analysis was conducted using Tibco Statistica version 13.3.

Multiple linear regression models were then created using R version 3.5.2 (R Core Team 2018) on a 64 bit Windows machine, wherein the binding free energy was used as the dependent variable and the ten principal components served as the independent variables. Sixty percent of the dataset was devoted for training the algorithm, followed by a 10-fold cross-validation. Model performance was evaluated using the correlation efficient between the actual and predicted values (r^2), and the root mean square error (RMSE). Backward elimination of the independent variables based on p-values was conducted in order to optimize and improve the parsimony of the regression model. The multi-collinearity of the optimized model was assessed by determining the variance inflation factor and condition index of the independent variables using the R package "olsrr" (Hebbali 2020). The process on how the regression models were developed and optimized is summarized in figure 1.

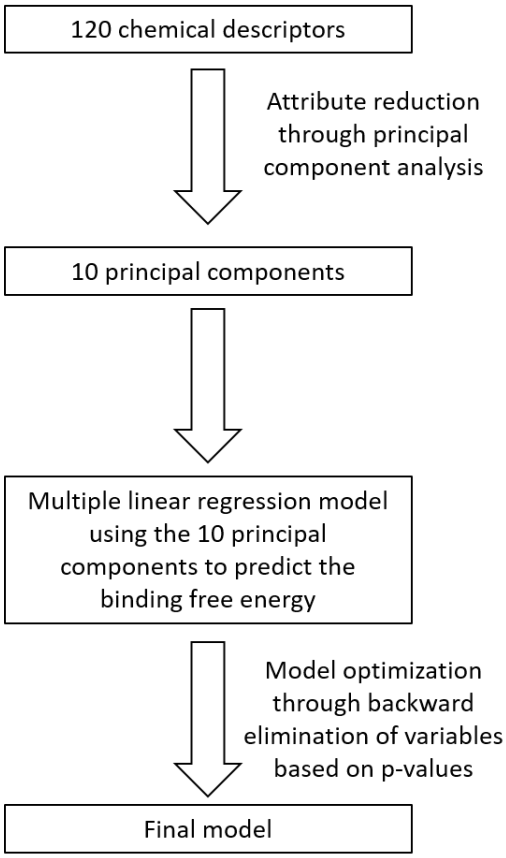

Figure 1. Workflow summarizing the development and optimization of the regression model

3. Results and Discussion

3.1 Narrowing descriptor selection through principal component analysis

One of the difficult and tedious tasks in creating predictive models is the selection of independent variables that can be used to predict the dependent variable. For example, in the present study, 120 descriptors were calculated for each compound (Supporting information). It will be difficult to create predictive models out of the 120 descriptors, or identify which descriptors exhibit correlation with the binding free energy towards the cagA protein. One way to circumvent this problem is to conduct principal component analysis (PCA), which is a form of dimension reduction. By applying PCA, the information from the 120 descriptors is condensed into ten principal components (PC). The ten PCs were then used as the independent variables in order to predict the BFE. As the scree plot shows (figure 2), the ten PCs are

able to explain the variability of the BFE of the dataset. In particular, PC1 can account for over 50% of the variability of BFE of the 38 compounds.

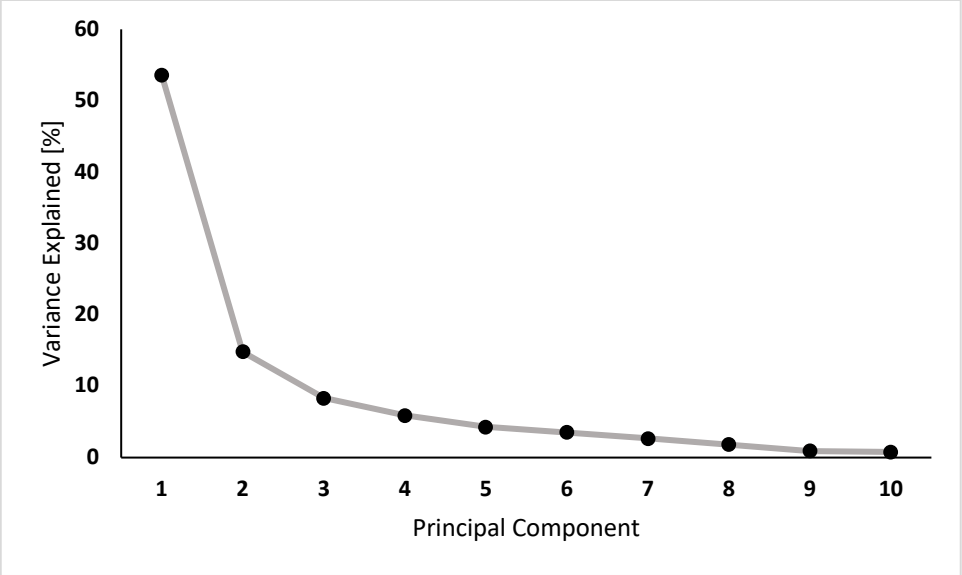

Figure 2. Scree plot showing the percentage of variance explained for each principal component.

The top chemical descriptors that bear weight on the PCs are listed in table 1. From the list, it can be observed that most of the descriptors that influence the PCs are from the topological descriptor class. This class of descriptors describe the arrangement and connectivity of atoms using group theory, wherein the heavy atoms are viewed as vertices and the bonds as the edges (Gozalbes et al. 2002). Various topological descriptors are available, and have been thoroughly reviewed by Dearden (2017).

Table 1. The top ten most important descriptors in the formulation of the principal components. All the listed descriptors have a power of 1.

Descriptor	Class
nB (number of bonds)	Constitutional
Zagreb index	Topological

SP.0 (Chi path)	Topological
WPOL (Wiener polarity number)	Topological
nAtom (number of atoms)	Constitutional
SP.4 (Chi path)	Topological
ATSp2 (autocorrelation polarizability)	Topological
VP.3 (Chi path)	Topological
ATSp1 (autocorrelation polarizability)	Topological
ATSm2 (autocorrelation mass)	Topological

3.2. Predicting the Binding Free Energy (BFE) through multiple linear regression models

The loadings of the 10 PCs for each of the 38 compounds were then used to predict the BFE using multiple linear regression (MLR). The full dataset is shown in table 2. MLR was prioritized in creating the predictive model due to its simplicity and ease of interpreting the model.

Table 2. Full dataset that was used to create predictive regression models. The 10 PCs were used as the independent variable to predict the binding free energy.

Compound	Binding free energy (kcal/mol)	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Spinasterol	-9.24	-10.3501	6.31136	2.53511	-1.52038	-5.04	3.98937	2.43779	3.36055	1.33677	0.75527
Luteoxanthin	-9.21	-19.9122	9.62335	-4.92736	1.66749	1.74492	-1.8074	0.5075	-0.87971	0.20964	0.05465
3'-prenylrubranine	-8.68	-10.8072	-1.60042	1.59595	-2.39447	-1.41341	2.80782	1.49481	-1.3314	-0.75305	0.95169
Neoxanthin	-8.23	-18.5786	9.08455	-6.73488	1.23641	1.33694	-1.48123	-0.52946	-1.17131	-0.0502	0.57758
Berberine	-7.47	-2.1773	-3.25789	3.78545	1.08582	-1.81439	-2.64641	3.52215	-1.43278	1.34186	1.56233

Rottlerin	-7.07	-10.2964	-7.82969	-1.33266	-0.88563	3.55631	3.78278	-2.50853	-1.49377	0.28575	1.4065
3-Farnesyl-2-hydroxybenzoic	-7.1	1.1883	-1.11104	-3.40584	-0.90125	-1.63096	-0.39737	-1.16374	0.99862	-2.4664	2.71994
Tatridin-A	-6.73	3.3621	0.19417	2.27872	0.2063	2.62114	-0.30553	-1.50656	2.29887	-0.43595	0.7907
(2S)-4′,7-dihydroxy-8-methylflavan	-6.47	2.8386	-2.0288	1.77726	-1.2643	-2.02833	-1.49344	-1.32836	-1.48683	1.92394	0.25501
Psoracorylifols D	-6.46	-1.83	4.1415	3.20391	-1.48793	-2.7013	0.65136	-0.78367	-1.68638	0.21687	-1.06532
Altissin	-6.45	0.3176	1.92441	4.29941	0.41638	2.11589	1.54443	-1.25257	1.60965	0.33983	-0.25295
Protopine	-6.42	-1.755	-3.20999	3.20047	1.97267	0.93905	-2.58285	3.86166	-0.57323	0.35977	0.77495
β-Hydrastine	-6.24	-3.8727	-3.57587	3.94007	2.82111	1.94996	-2.29451	3.95345	0.49509	-1.1581	-0.13365
Psoracorylifols B	-6.24	-1.0369	2.39258	2.62929	-1.0041	-1.7743	-0.17025	-0.66433	-1.02876	-1.56044	-1.05483
Psoracorylifols E	-6.2	-1.83	4.1415	3.20391	-1.48793	-2.7013	0.65136	-0.78367	-1.68638	0.21687	-1.06532
Psoracorylifols C	-6.16	-1.0369	2.39258	2.62929	-1.0041	-1.7743	-0.17025	-0.66433	-1.02876	-1.56044	-1.05483
Desacetyl-β-cyclopyrethrosin	-6.03	0.3757	1.97175	4.33991	0.41394	2.20395	1.64514	-1.02965	1.59165	0.33498	-0.20856
Speciformin	-5.99	0.6951	4.4328	1.79513	1.20493	5.34746	-4.37627	-1.35975	1.10634	1.35044	-0.91194
Resveratrol	-5.99	5.3149	-2.13985	0.06719	-1.47254	-2.08626	-2.20604	-2.60248	-2.63658	2.13518	1.0674
Dehydroleucodine	-5.94	2.3642	1.7177	4.28644	-0.21894	1.50756	-0.97663	0.38252	1.00624	-1.47883	0.88471
Sivasinolide	-5.87	0.3176	1.92441	4.29941	0.41638	2.11589	1.54443	-1.25257	1.60965	0.33983	-0.25295
6-gingerol	-5.81	4.03	-1.4613	-2.61536	-0.31236	-1.56906	-0.94712	-1.42093	1.819	0.8064	-0.07441
Boropinic acid	-5.65	5.1309	-1.16137	-0.90098	-0.18532	-0.73572	-2.28434	-1.23582	0.06475	-2.46958	0.90341
8-Gingerol	-5.52	3.0263	-1.64115	-3.51165	-0.23097	-2.01939	-0.87532	-0.99992	2.52771	0.79929	0.08722
Psoracorylifols A	-5.44	1.0045	0.73204	0.88732	-1.09277	-2.23212	-0.26309	-2.04427	-0.33639	-1.52473	-1.23994
Cochinchinenins B	-5.22	-9.3407	-8.39865	-2.60562	-0.01188	-1.05646	-0.7368	1.43509	0.78782	-0.25723	-1.89411
Curcumin	-5.1	0.4953	-4.55447	-2.45109	-0.23931	-0.35308	-2.19352	-1.29022	0.28515	0.04897	-1.02829
MeONQ	-5.07	6.7547	-0.64618	2.32172	-1.26614	-0.00901	-2.35807	-1.57589	-1.14588	0.08824	0.63409
Cochinchinenins C	-4.8	-9.2295	-8.40059	-2.61808	0.08639	-1.05824	-0.8433	1.33726	0.86709	-0.19065	-1.89534
Isorottlerin	-4.6	-11.4367	-7.90263	0.47003	-0.92666	3.61403	3.96078	-1.50399	-0.94637	0.47947	-0.11399
10-Gingerol	-4.47	2.0161	-1.82391	-4.43415	-0.14936	-2.46763	-0.80619	-0.58309	3.22464	0.78783	0.25296
L-Sulforaphene	-4.37	11.293	1.3952	-2.62614	-1.322	1.50336	1.06571	1.41687	-1.0166	0.04564	-0.25996
Erysolin	-4.14	10.3967	1.63332	-3.05625	-0.89921	2.24374	1.91761	1.41943	-0.95947	-0.33419	-0.79028
Alyssin	-3.95	10.6489	1.50646	-3.12699	-1.31959	1.29241	1.42479	1.87137	-0.13497	0.11861	-0.43853
L-Sulforaphane	-3.9	11.218	1.62788	-2.78066	-1.17984	1.46744	1.51617	1.72783	-0.5572	0.12162	-0.54004
Iberin	-3.7	8.0573	-0.22696	-0.94925	14.39912	-2.48934	3.06874	-1.22083	-1.13478	-0.01137	-0.02386
Berteroin	-3.63	11.0388	1.85136	-2.91432	-1.6428	0.606	1.27366	2.04945	-0.27176	0.27941	0.35962
Erucin	-3.4	11.6055	1.97182	-2.55471	-1.50717	0.78854	1.37179	1.88746	-0.71353	0.28397	0.26108

The full regression model, composed of the 10 PC yielded a satisfactory prediction accuracy as demonstrated by r^2_{train} = 0.92. However, when the full regression model was applied for the test set, the accuracy substantially decreased. This result suggests the occurrence of overfitting, wherein the created algorithm was too tailored for the training set, leading to poor performance in the test set. In order to optimize the performance of the model, backward elimination was conducted wherein non-significant descriptors were removed on the basis of their p-values. From the full regression model, it was found that only PC1, PC2, PC3, PC5, and PC10 were significant independent variables (Table 3).

Table 3. Structure of the full multiple linear regression model using the 10 principal components as the independent variables. P-values with asterisk are less than 0.05.

Variable	Coefficient	p-value
Intercept	-5.811	<0.001*
PC1	0.105	0.00142*
PC2	-0.0759	0.0178*
PC3	-0.0966	0.0317*
PC4	-0.447	0.157
PC5	0.326	0.00811*
PC6	-0.139	0.329
PC7	0.0843	0.325
PC8	0.117	0.307
PC9	0.0733	0.536
PC10	-0.610	<0.001*

Thus, an optimized model (Model A) was formulated using these principal components. The training and test accuracies of the optimized model are excellent, with a significant decrease in the RMSE. Model A was further optimized, wherein PC5 was removed since it had a p-value greater than 0.05 (Table 4).

Table 4. Structure of the multiple linear regression model A using 5 principal components as the independent variables. P-values with asterisk are less than 0.05.

Variable	Coefficient	p-value
Intercept	-5.813	<0.001*
PC1	0.133	<0.001*
PC2	-0.113	0.00247*
PC3	-0.132	0.00831*
PC5	0.110	0.190
PC10	-0.606	0.00378*

The resulting regression model, Model B (Table 5), which is composed of only 4 PC, exhibited satisfactory prediction accuracy, as summarized in Table 6. However, the prediction performance of Model A is judged to be the best since it exhibited the highest r^2 and lowest RMSE in test sets among all of the created regression models.

Table 5. Structure of the multiple linear regression model B using 4 principal components as the independent variables. P-values with asterisk are less than 0.05.

Variable	Coefficient	p-value
Intercept	-5.823	<0.001*
PC1	0.121	<0.001*

PC2	-0.117	0.002*
PC3	-0.124	0.123*
PC10	-0.513	0.008*

Table 6. Summary of the predictive performance of the full and optimized multiple linear regression models

Model	**r^2**	**RMSE**
Full model (PC1 – PC10)	Train = 0.92 Test = 0.80	Train = 1.926 Test = 0.756
Model A (PC1, PC2, PC3, PC5, PC10)	Train = 0.87 Test = 0.92	Train = 0.686 Test = 0.483
Model B (PC1, PC2, PC3, PC10)	Train = 0.89 Test = 0.90	Train = 0.674 Test = 0.569

Thus, Model A was assessed for multi-collinearity through determining the VIF and condition indices of the independent variables. A VIF of less than 4, and a condition index of less than 30 indicate that the variables are not correlated with each other (Hebbali 2020). As table 7 demonstrates, Model A does not suffer from multi-collinearity and this assumption in which linear regression models are built is met and satisfied.

Table 7. Model diagnostics for assessing multi-collinearity for the optimized multiple linear regression model A.

Principal Component	**Variance Inflation Factor**	**Condition Index**

	(VIF)	
PC1	1.283	1.087
PC2	1.050	1.159
PC3	1.099	1.373
PC5	1.384	1.422
PC10	1.163	1.854

Therefore, the predictive model assumes the form of:

$$\widehat{\boldsymbol{BFE}} = \boldsymbol{-5.813 + 0.133(PC1) - 0.113(PC2) - 0.132(PC3) + 0.110(PC5) - 0.606(PC10)}$$

The formulated regression model provides valuable support in the development of phytochemicals as CagA binders. As mentioned earlier, only a few compounds have been identified so far as promising CagA binders. However, the discovery of alternative treatment options against *H. pylori* infection is of paramount importance since new therapeutic strategies are needed to fight antibiotic resistance (Gerrits et al. 2006). Currently, the first line of treatment for *H. pylori* infection is clarithromycin triple therapy, which involves the administration of a proton pump inhibitor (PPI), clarithromycin, and amoxicillin (Chey et al. 2017). Other clinical treatments for *H. pylori* infection involve a PPI and a cocktail of antibiotics. The PPI plays an important role in enhancing the efficacy of the antibiotics since the PPI increases the pH of the gastric environment, increasing *H. pylori* susceptibility to antibiotics. Thus, studies have often focused on the discovery and development of compounds that can increase the pH of the gastric environment, such as urease inhibitors (Ul-Haq et al. 2016; Zhou et al. 2017; Kataria and Khatkar 2019). The present work is therefore a positive contribution towards the identification of promising CagA protein binders, which may later be developed as anti – *H. pylori* agents.

Conclusion

An accurate and parsimonious principal component - multiple linear model for the prediction of the binding free energy of phytochemicals to the CagA protein has been formulated. The application of principal component analysis led to the successful and efficient reduction of the variables to make the regression modelling straightforward. The model was created after a series of optimization steps based on backward elimination wherein the final model exhibits excellent accuracy (r^2 = 0.92, RMSE = 0.483). It was found that topological chemical descriptors had the greatest influence in the principal components which were used for predicting the binding free energy of the compounds to the CagA protein. The formulated regression models are expected to aid in the discovery and development of natural products-based therapeutic agents against *H. pylori*.

Compliance with ethical standards

Conflict of interest: None

Research involving human participants and / or animals: Not applicable

Informed consent: Not applicable

Reference

Chandra Babu TM, Rajesh SS, Bhaskar BV, et al (2017) Molecular docking, molecular dynamics simulation, biological evaluation and 2D QSAR analysis of flavonoids from Syzygium alternifolium as potent anti-Helicobacter pylori agents. RSC Adv 7:18277–18292. doi: 10.1039/c6ra27872h

Chey WD, Leontiadis GI, Howden CW, Moss SF (2017) ACG Clinical Guideline: Treatment of Helicobacter pylori Infection. Am J Gastroenterol 112:212–238. doi: 10.1038/ajg.2016.563

Dearden JC (2017) The Use of Topological Indices in QSAR and QSPR Modeling. In: Roy K (ed) Advances in QSAR Modeling, Challenges and Advances in Computational Chemistry and Physics. Springer International Publishing AG, Cham, pp 57–88

Gerrits MM, Van Vliet AHM, Kuipers EJ, Kusters JG (2006) Helicobacter pylori and antimicrobial resistance: molecular mechanisms and clinical implications. Lancet Infect Dis 6:699–709

Gozalbes R, Doucet JP, Derouin F (2002) Application of Topological Descriptors in QSAR and Drug Design: History and New Trends. Curr Drug Targets-Infectious Disord 2:93–102

Guha R (2007) Chemical Informatics Functionality in R. J Stat Softw 18:1–16. doi: 10.18637/jss.v018.i05

Handa O, Naito Y, Yoshikawa T (2007) CagA protein of Helicobacter pylori: A hijacker of gastric epithelial cell signaling. Biochem Pharmacol 73:1697–1702. doi: 10.1016/j.bcp.2006.10.022

Hatakeyama M (2017) Structure and function of helicobacter pylori caga, the first-identified bacterial protein involved in human cancer. Proc. Japan Acad. Ser. B Phys. Biol. Sci. 93:196–219

Hebbali A (2020) olsrr: Tools for Building OLS Regression Models. https://cran.r-project.org/package=olsrr

Kataria R, Khatkar A (2019) Molecular docking, synthesis, kinetics study, structure–activity relationship and ADMET analysis of morin analogous as Helicobacter pylori urease inhibitors. BMC Chem 13:. doi: 10.1186/s13065-019-0562-2

Lawal TO, Soni KK, Saxena RC, et al (2011) Anti-Helicobacter Pylori Activities of Compounds of Natural Origin. In: Brahmachari G (ed) Bioactive Natural Products. World Scientific , Singapore, pp 475–497

Lin CJ, Rao YK, Hung CL, et al (2013) Inhibition of helicobacter pylori CagA-induced pathogenesis by methylantcinate B from Antrodia camphorata. Evidence-based Complement Altern Med 682418. doi: 10.1155/2013/682418

Nomura AMY, Pérez-Pérez GI, Lee J, et al (2002) Relation between Helicobacter pylori cagA Status and Risk of Peptic Ulcer Disease. Am J Epidemiol 155:1054–1059

R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Racha S, Wongrattanakamon P, Raiwa A, Jiranusornkul S (2019) Discovery of Novel Potent Small Natural Molecules Able to Enhance Attenuation of the Pathobiology of Gastric Cancer-Associated Helicobacter pylori by Molecular Modeling. Int J Pept Res Ther 25:881–896. doi: 10.1007/s10989-018-9737-2

Rastelli G, Pellati F, Pinzi L, Gamberini MC (2020) Repositioning natural products in drug discovery. Molecules 25:1154. doi: 10.3390/molecules25051154

Rokkas T, Ladas S, Liatos C, et al (1999) Relationship of Helicobacter pylori CagA status to gastric cell proliferation and apoptosis. Dig Dis Sci 44:487–493

Shmuely H, Passaro D, Figer A, et al (2001) Relationship between Helicobacter pylori CagA status and colorectal cancer. Am J Gastroenterol 96:3406–3410

Srivastava AK, Tewari M, Shukla HS, Roy BK (2015) In Silico Profiling of the Potentiality of Curcumin and Conventional Drugs for CagA Oncoprotein Inactivation. Arch Pharm (Weinheim) 348:548–555. doi: 10.1002/ardp.201400438

Suerbaum S, Michetti P (2002) Helicobacter pylori infection. N Engl J Med 347:1175–1186

Ul-Haq Z, Ashraf S, Al-Majid AM, Barakat A (2016) 3D-QSAR studies on barbituric acid derivatives as urease inhibitors and the effect of charges on the quality of a model. Int J Mol Sci 17:. doi: 10.3390/ijms17050657

Vale FF, Oleastro M (2014) Overview of the phytomedicine approaches against Helicobacter pylori. World J Gastroenterol 20:5594–5609. doi: 10.3748/wjg.v20.i19.5594

Zhou JT, Li CL, Tan LH, et al (2017) Inhibition of Helicobacter pylori and its associated urease by Palmatine: Investigation on the potential mechanism. PLoS One 12:e0168944. doi: 10.1371/journal.pone.0168944