A Screening Algorithm for Gastric Cancer – Binding Peptides

Jose Isagani B. Janairo[1*] Marianne Linley L. Sy-Janairo[2]

[1]Biology Department, De La Salle University, 2401 Taft Avenue, Manila, Philippines

[2]Institute of Digestive and Liver Diseases, St. Luke's Medical Center – Global City, Taguig, Philippines

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

Abstract

Gastric cancer-binding peptides (GCBP) are promising diagnostic and therapeutic agents for gastric cancer management. Their utility lies in their ability to facilitate the early detection of gastric cancer, prevent metastasis, and prevent tumor angiogenesis. In order to promote and accelerate the discovery of more GCBP, this study aims to create a machine-learning classification model that can predict if a given sequence can bind with gastric cancer cells. A systematic literature search was conducted to extract peptides that can and cannot bind with gastric cancer cells. Nine descriptor classes were then calculated for each sequence. The resulting dataset was used to create classifiers using five machine-learning algorithms. Rigorous model optimizations were conducted which included descriptor selection and probability threshold tuning. The combination of the topological descriptor T-scales, and logistic regression were found to satisfactorily predict GCBP class. The optimized classification model exhibited satisfactory accuracy with balanced sensitivity and specificity, and excellent precision. The results brought forward provide the foundation for an alternative screening method for GCBPs. This system is expected to positively contribute in the discovery of new GCBPs, thereby potentially enhancing GC disease diagnostics and management.

Keywords: machine learning, medical bioinformatics, drug design

Introduction

Gastric cancer (GC) is a highly important global health issue, which is the third cause of cancer death worldwide (Herrero et al. 2014). GCs are mostly malignant epithelial neoplasms, wherein chemotherapy is the standard treatment for late-stage GC (Lordick et al. 2014). As with most diseases, early detection of GC allows the prompt adoption of effective therapeutic strategies leading to improved prognosis. However, GC treatment remains challenging due to the possible onset of multidrug resistance by the tumor, which diminishes the efficacy of chemotherapy (Shi and Gao 2016). GC metastatic spread, commonly observed at the liver, peritoneum, lungs, and bones (Sundquist et al. 2016) adds complexity in the overall treatment of GC. Thus, innovative therapeutic advances are needed to further enhance GC treatment and improve prognosis.

An emerging and promising experimental option in GC management are gastric cancer – binding peptides (GCBP). This class of peptides specifically binds with GC cells, and can be used for better visualization of the tumor thereby facilitating early detection (Zhang et al. 2012) (Han et al. 2016). GCBPs can also reverse multidrug resistant GC cells (Kang et al. 2013), prevent metastasis (Hu et al. 2006), and prevent angiogenesis (Chen et al. 2009). The versatility of GCBP makes them attractive therapeutic alternatives worth developing. Currently, the discovery of GCBPs has mainly relied on the combinatorial technique of phage display assay, which involves the following laborious steps: vector system selection, library construction, affinity selection, and peptide characterization (Sidhu et al. 2000). In order to fully realize the potential of GCBPs, a rapid, straightforward, and cost – effective screening method is needed. Such system can accelerate GCBP lead discovery and development.

Classification models built using machine – learning have recently gained traction in boosting the overall workflow of the drug discovery process. Classification models that can determine whether a given compound possesses functional properties can be used to rapidly screen libraries to identify promising lead candidates. This system has been deployed for the screening of functional peptides, such as antihypertensive peptides (Singh Chauhan et al. 2015), cell-penetrating peptides (Sanders et al. 2011),

metal-binding peptides (Janairo 2019), among others. However, screening algorithms for GCBP have yet to be developed. This study therefore aims to construct a screening algorithm for GCBP using machine-learning. Developing this kind of screening technique can potentially lead to the discovery of new GCBPs, and contribute to the better treatment options of GC.

Materials and Methods

A systematic literature search for GCBP was conducted in major databases such as Scopus, Google Scholar, and Pubmed. A peptide sequence was designated as GCBP if the paper compared the binding of the peptide in question with a control, and the relevant peptide exhibited statistically significant difference. Otherwise, the peptide was designated as non-GCBP.

The extracted peptide sequences from the systematic literature search were then subjected to descriptor calculations using the R package Peptides version 2.4 (Osorio et al. 2015). The calculated peptide descriptors were the Blosum indices (Georgiev 2009), Cruciani properties (Cruciani et al. 2004), factor analysis scale of generalized amino acid information (FASGAI) vectors (Liang and Li 2007), Kidera factors (Kidera et al. 1985), ProtFP (van Westen et al. 2013), ST-scales (Yang et al. 2010), T-scales (Tian et al. 2007), VHSE scales (Mei et al. 2005), and Z-scales (Sjöström et al. 2002). The resulting dataset was then used for the formulation of classification models using the following machine-learning algorithms: generalized linear models in the form of logistic regression (LR), k – nearest neighbor (KNN), classification and regression trees (CART), support vector machine (SVM), and random forest (RF). In all classification model formulations, 60% of the dataset was devoted for training, and the remaining 40% served as the test set. A 10 – fold cross-validation was likewise conducted for all models. The “GCBP” status was designated as the positive class for the confusion matrix (y = 1). The default settings of each algorithm which led to the best performance were automatically selected. The R package caret was used for the creation of the classification models (Kuhn et al. 2018). All R packages and their dependents used in the study were executed in R version 3.5.2 (R Core Team 2018) using a Windows 64 bit desktop.

Results

The systematic literature search yielded 78 peptide sequences, as summarized in Table 1. The curated peptide sequences were divided into two classes: non-GCBP, and GCBP. The designation of either class was based on the paper from which the peptide was taken.

Table 1. Curated peptide sequences and their designation which were extracted from the systematic literature search.

Name	Class	Sequence	Reference
Kang 2013 GCBP1	GCBP	ETAPLSTMLSPY	(Kang et al. 2013)
Kang 2013 GCBP2	Non-GCBP	TTFSPPSPLSSI	(Kang et al. 2013)
Kang 2013 GCBP3	Non- GCBP	FNNHYPAAATYP	(Kang et al. 2013)
Kang 2013 GCBP4	Non- GCBP	EVFWPLNAPRLL	(Kang et al. 2013)
Kang 2013 GCBP5	Non- GCBP	SWQIPYPISPRS	(Kang et al. 2013)
Kang 2013 GCBP6	GCBP	KTAPTPYALVHD	(Kang et al. 2013)
Kang 2013 GCBP7	Non- GCBP	SNFMHNTRIWSH	(Kang et al. 2013)
Kang 2013 GCBP8	GCBP	SWQITYPISPRS	(Kang et al. 2013)
Kang 2013 GCBP9	GCBP	TDSYHVASARQP	(Kang et al. 2013)
Kang 2013 GCBP10	Non- GCBP	STYSHPLSLRPD	(Kang et al. 2013)
Kang 2013 GCBP11	Non- GCBP	AWTWVLPSSIRA	(Kang et al. 2013)
Kang 2013 GCBP12	Non- GCBP	YTTWPFTSLQLD	(Kang et al. 2013)
Kang 2013 GCBP13	Non- GCBP	AFMETTSQNAWL	(Kang et al. 2013)
Kang 2013 GCBP14	Non- GCBP	ETAPPTPYSVMF	(Kang et al. 2013)
Kang 2013 GCBP15	Non- GCBP	TTFNPLYLRLDT	(Kang et al. 2013)
Kang 2013 GCBP16	Non- GCBP	SSFPQVIPLDYL	(Kang et al. 2013)
Kang 2013 GCBP17	Non- GCBP	APKYSLSDLYLN	(Kang et al. 2013)
Kang 2013 GCBP18	Non- GCBP	QIEKISQHLDMH	(Kang et al. 2013)
Kang 2013 GCBP19	Non- GCBP	TWNQPYIPPLYP	(Kang et al. 2013)
Kang 2013 GCBP20	Non- GCBP	YASPPNPSLRLT	(Kang et al. 2013)
Kang 2013 GCBP21	Non- GCBP	YHGLTPVRYVSV	(Kang et al. 2013)
Kang 2013 GCBP22	Non- GCBP	YTFMPELTTPRT	(Kang et al. 2013)
Kang 2013 GCBP23	GCBP	NSFNYAPLLMPR	(Kang et al. 2013)
Kang 2013 GCBP24	Non- GCBP	TSPYHLLHAHLQ	(Kang et al. 2013)
Kang 2013 GCBP25	Non- GCBP	SSFVALSISPSM	(Kang et al. 2013)
Kang 2013 GCBP26	Non- GCBP	ANLSSHSSPGDS	(Kang et al. 2013)
Kang 2013 GCBP27	Non- GCBP	TTLQFTGQTNKT	(Kang et al. 2013)
Kang 2013 GCBP28	Non- GCBP	TPPPRDASLSRW	(Kang et al. 2013)
Kang 2013 GCBP29	Non- GCBP	HNIGTWGPKSHL	(Kang et al. 2013)
Kang 2013 GCBP30	Non- GCBP	SGFFEPQHSHPL	(Kang et al. 2013)
Kang 2013 GCBP31	Non- GCBP	QIEESFVRGHTT	(Kang et al. 2013)
Kang 2013 GCBP32	Non- GCBP	SPQTDGLVSTPS	(Kang et al. 2013)
Kang 2013 GCBP33	Non- GCBP	YTFDPQIRPAGL	(Kang et al. 2013)
Kang 2013 GCBP34	Non- GCBP	YERSILPFSHVF	(Kang et al. 2013)
Kang 2013 GCBP35	Non- GCBP	YPSVTFPTQTLL	(Kang et al. 2013)

Kang 2013 GCBP36	Non- GCBP	ASTNVFARPMYL	(Kang et al. 2013)
Kang 2013 GCBP37	Non- GCBP	SPWYMTPSPNTA	(Kang et al. 2013)
Kang 2013 GCBP38	Non- GCBP	HISVINYTTKIS	(Kang et al. 2013)
Kang 2013 GCBP39	Non- GCBP	MNVTVSGRLSGP	(Kang et al. 2013)
Kang 2013 GCBP40	Non- GCBP	TIPYPFSLLNNP	(Kang et al. 2013)
Kang 2013 GCBP41	Non- GCBP	PFLYSQVAWRS	(Kang et al. 2013)
Kang 2013 GCBP42	GCBP	TALPNHWSDASP	(Kang et al. 2013)
Kang 2013 GCBP43	Non- GCBP	SVSVGMKPSPRP	(Kang et al. 2013)
Kang 2013 GCBP44	Non- GCBP	YQEETPASSFSR	(Kang et al. 2013)
Kang 2013 GCBP45	Non- GCBP	NSSQLAPYTTHR	(Kang et al. 2013)
Kang 2013 GCBP46	GCBP	TLHPSVLSYVLK	(Kang et al. 2013)
Kang 2013 GCBP47	Non- GCBP	HNGLPNFFQTRL	(Kang et al. 2013)
Kang 2013 GCBP48	Non- GCBP	EAAPNFYPPLTF	(Kang et al. 2013)
Kang 2013 GCBP49	Non- GCBP	DLFQFAFPLNTI	(Kang et al. 2013)
Kang 2013 GCBP50	Non- GCBP	FTFSYAESVSYF	(Kang et al. 2013)
Hui 2008 GX1	GCBP	CGNSNPKSC	(Hui et al. 2008)
Hu 2006 GCBP1	Non- GCBP	SMSIASPQIPWS	(Hu et al. 2006)
Hu 2006 GCBP2	Non- GCBP	TPRNLRTSNTHR	(Hu et al. 2006)
Hu 2006 GCBP3	GCBP	GRRIAGPYIALE	(Hu et al. 2006)
Hu 2006 GCBP4	Non- GCBP	SMPINSPYIPWS	(Hu et al. 2006)
Hu 2006 GCBP5	GCBP	GRRPMKLNKTP	(Hu et al. 2006)
Hu 2006 GCBP6	GCBP	GRRINRLILPRN	(Hu et al. 2006)
Hu 2006 GCBP7	GCBP	GRRTRSRRLRRS	(Hu et al. 2006)
Hu 2006 GCBP8	GCBP	GRRTRSSRLRNS	(Hu et al. 2006)
Zhang 2012 AAD	GCBP	AADNAKTKSFPV	(Zhang et al. 2012)
Liang 2006 GEBP 11	GCBP	CTKNSYLMC	(Liang et al. 2006)
Liang 2006 GEBP 9	GCBP	CKNSLTMAC	(Liang et al. 2006)
Liang 2006 GEBP4	Non- GCBP	CSSTTPNAC	(Liang et al. 2006)
Liang 2006 GEBP 15	Non- GCBP	CNSTKYNQC	(Liang et al. 2006)
Liang 2006 GEBP 12	GCBP	CTNTMLPQC	(Liang et al. 2006)
Liang 2006 GEBP 13	Non- GCBP	CTTTFDLRAC	(Liang et al. 2006)
Liang 2006 GEBP 1	Non- GCBP	CPVSLQALC	(Liang et al. 2006)
Liang 2006 GEBP 2	Non- GCBP	CDRHNLTFC	(Liang et al. 2006)
Liang 2006 GEBP 3	GCBP	CPPSKMSQC	(Liang et al. 2006)
Liang 2006 GEBP 5	Non- GCBP	CQPALQMKC	(Liang et al. 2006)
Liang 2006 GEBP 6	GCBP	CLSSSLSDC	(Liang et al. 2006)
Liang 2006 GEBP 14	Non- GCBP	CIPTHPRLC	(Liang et al. 2006)
Zhi 2004	GCBP	CGNSNPKSC	(Zhi et al. 2004)
Wang 2014 GP1	Non- GCBP	YTHNEKPSSDTH	(Wang et al. 2014)
Wang 2014 GP2	Non- GCBP	YTVPDNHKYSAH	(Wang et al. 2014)
Wang 2014 GP3	Non- GCBP	THPWQVSTINFK	(Wang et al. 2014)
Wang 2014 GP4	Non- GCBP	DVFPPRSHADEL	(Wang et al. 2014)
Wang 2014 GP5	GCBP	IHKDKNAPSLVP	(Wang et al. 2014)

The next step involves identifying which among the 9 peptide descriptor class and 5 machine-learning models are suitable for GCBP classification. Thus, a pairwise performance analysis was executed to

identify the optimum peptide descriptor and machine-learning model pair. The performance analysis of the 45 constructed classification models revealed that the optimum peptide descriptor is the T-scales, and the best machine-learning model for the classification task is the LR model. This pair demonstrated the highest training accuracy (Figure 1) and kappa score (Figure 2).

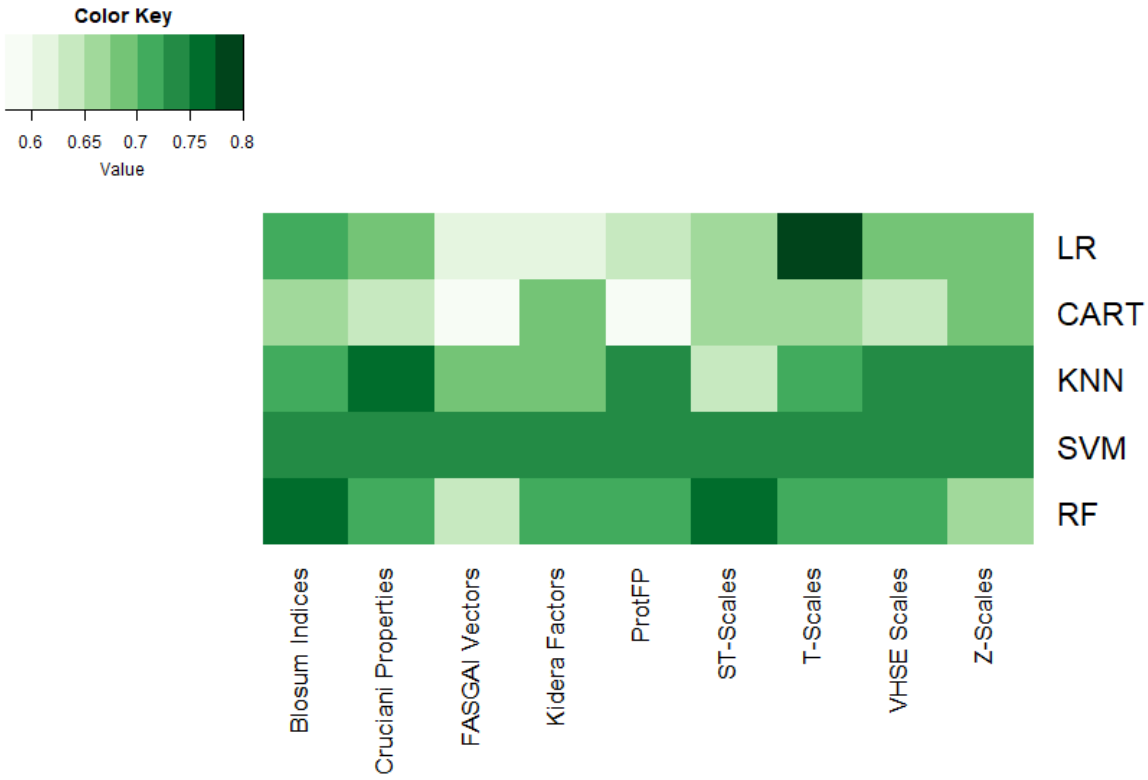

Fig. 1 Training accuracy of the descriptor and machine-learning algorithm for the classification of GCBP. The darker the shade, the higher the training accuracy.

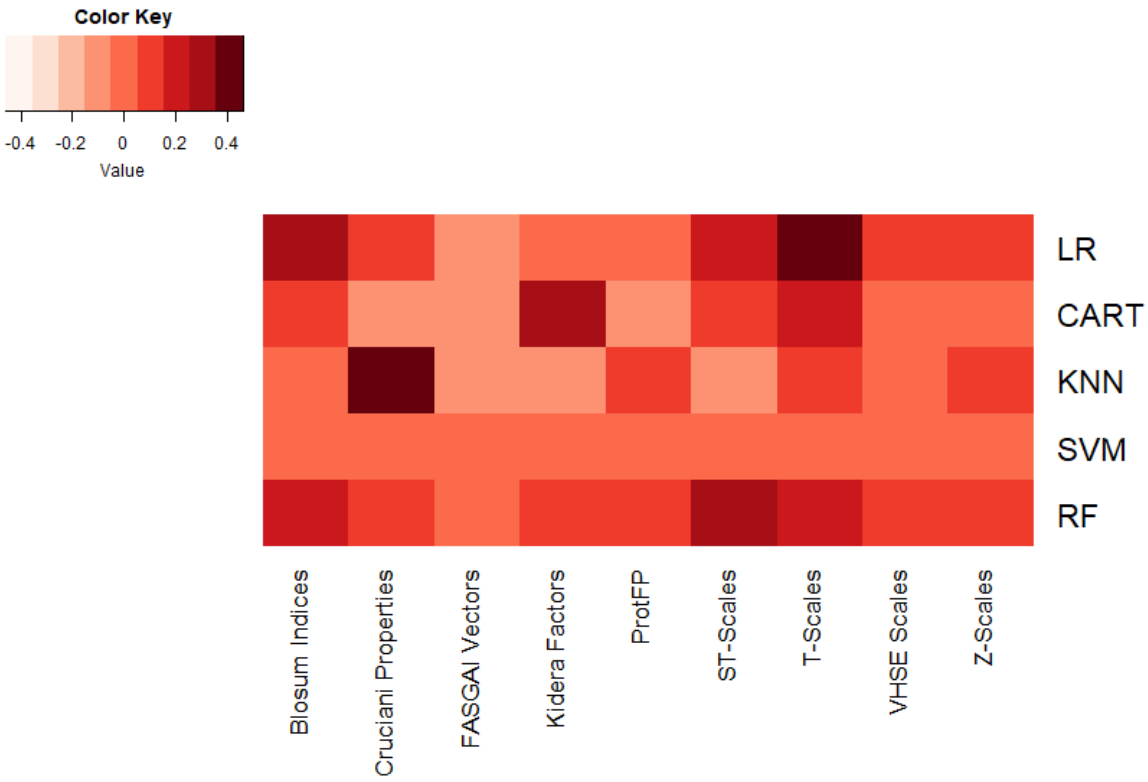

Fig. 2 Training accuracy in terms of the kappa scores of the descriptor and machine-learning algorithm for the classification of GCBP. The darker the shade, the higher the kappa scores.

The T-scales descriptor class is composed of 5 descriptors, and the second optimization phase is intended to identify if the 5 T-scales descriptors can be reduced without compromising classification accuracy. A step-wise back elimination of the 5 T-scales descriptors was therefore conducted, and the model performance is based on the diagnostic metrics of the test set (Table 2). Model A is the full model set, which means the 5 T-scales descriptors were used for the GCBP classification. It is worth mentioning that the full model exhibited similar classification accuracy with training set, an indication that the T-scales descriptor – LR model did not exhibit over fitting. Model B only included 4 T-scales descriptors, wherein the T-scales descriptor T4 was removed. The T4 descriptor had the highest p-value in the full model, indicating that this descriptor had little predictive value. This elimination process was continued, wherein Table 2 summarizes the descriptors removed for each model, and their corresponding p-values.

Table 2. Results of stepwise back elimination of T-scales descriptors for each model, and their corresponding p-values

Descriptor	**Model A (Full Model)**	**Model B**	**Model C**	**Model D**
Intercept	0.135	0.0233*	0.0236*	0.00172**
T1	0.507	0.3825	0.455	-
T2	0.624	0.6166	-	-
T3	0.113	0.097	0.0267*	0.00492**
T4	0.933	-	-	-
T5	0.525	0.4969	0.3133	0.0086**

*Significant at 5% level; **Significant at 1% level

Despite Model D had all significant descriptors (p-value < 0.05), it however performed poorly in the classification task using a probability threshold of 0.5 for non-GCBP classification (Table 3). It appears that Model B presents the best compromise between parsimony and performance of the model, since this model exhibited the best performance while only requiring 4 descriptors. The optimum classification model for classifying non-GCBP from GCBP therefore assumes the form of

$$P(GCBP) = \frac{e^{(-5.136-0.6844T1+0.5583T2-3.9506T3+2.3354T5)}}{1 + e^{(-5.136-0.6844T1+0.5583T2-3.9506T3+2.3354T5)}}$$

Table 3. Performance of optimized classification models on the test set. Characters in parenthesis under the model title correspond to the T-scale descriptors used.

Metrics	**Model A**	**Model B**	**Model C**	**Model D**	**Model B1**
Overall Accuracy	0.80	0.83	0.73	0.77	0.76
Balanced Accuracy	0.70	0.77	0.62	0.64	0.76
Kappa Score	0.44	0.56	0.26	0.32	0.44
Sensitivity	0.50	0.63	0.38	0.38	0.70
Specificity	0.91	0.91	0.86	0.91	0.82
Positive Predictive Value	0.67	0.71	0.50	0.60	0.65
Negative Predictive Value	0.83	0.87	0.79	0.80	0.90

While Model B exhibited high overall accuracy in classifying non-GCBP from GCBPs, discrepancy in the sensitivity and specificity can be observed. This can be attributed in the imbalanced nature of the dataset. One way to address this is to tune the probability threshold of the classification model. Currently, when the calculated probability is greater than 0.5, the classifier designates the case as GCBP, and non-GCBP for less than 0.5. Figure 3 shows the probability threshold – tuning as a function of sensitivity, specificity, and balanced accuracy of model B.

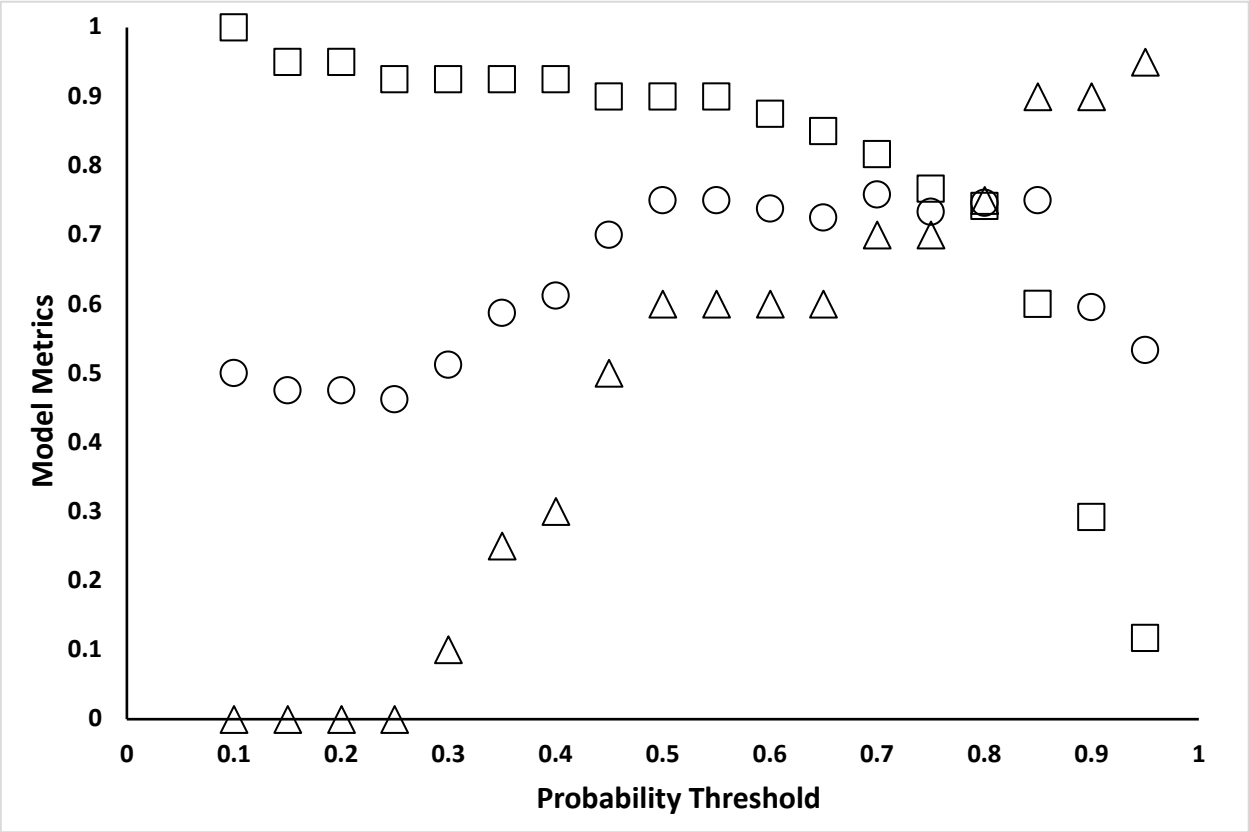

Fig. 3 Effects of varying the probability threshold of the logistic regression model on the specificity (square), sensitivity (triangle), and balanced accuracy (circle) of model B.

Figure 3 suggests that changing the probability threshold to 0.7 for GCBP classification creates a more robust and balanced model (Model B1). On the other hand, changing the probability threshold also impaired some aspects of the classifier. Thus, the decision on which probability threshold to utilize depends on which aspect of the classifier is prioritized.

In order to further validate the GCBP classification model, external validation was carried out. In this validation phase, peptide sequences which were not part of the model training and testing were used to assess model performance. Using model B, the classification model had an accuracy rating of 70% (Table 4). This result further affirms the reliability and potential immediate practical applicability of the classifier to screen potential GCBPs.

Table 4. External validation of the GCBP classification model using peptide sequences that were not part of the model training and testing.

Peptide Sequence (Reference)	Actual Class	Predicted Class
KLP (Akita et al. 2006)	GCBP	GCBP
AADNAKTKSFPV (Zhang et al. 2012)	GCBP	Non- GCBP
YIGSR (Matsuoka et al. 1998)	GCBP	GCBP
RGD (Matsuoka et al. 1998)	GCBP	GCBP
SWKLPPS(Akita et al. 2006)	GCBP	Non- GCBP
IFLLWQR (Hatakeyama et al. 2011)	Non- GCBP	Non- GCBP
CTTHWGFTLC (Koivunen et al. 1999)	Non- GCBP	Non- GCBP
CDTRL (Hoffman et al. 2003)	Non- GCBP	Non- GCBP
CTPSPFSHC (Li et al. 2010)	Non- GCBP	Non- GCBP
CRGRRST (Joyce et al. 2003)	Non- GCBP	GCBP

Discussion

One of the challenges in creating predictive models is identifying which variables and algorithm can account and predict the variability of the cases. This challenge is amplified for peptide modeling since numerous peptide descriptors are available, making the permutation with machine-learning models highly diverse. Thus, the first optimization phase of this study sought to determine which descriptor class and machine-learning model should be used in creating GCBP classifiers. The peptide descriptors were limited to 9 classes, which are easily calculated by the Peptides R Package. These descriptor classes are mostly principal components derived from the properties of the amino acids that makeup the peptide under investigation. The selected descriptors can be generally subdivided depending on what particular aspect of the peptide they describe. The Blosum indices are under the similarity measures of descriptor category; the T-scales and ST-scales are topological descriptors, FASGAI vectors, ProtFP, VHSE scales,

and Z-scales are variables that describe the physico-chemical properties of the peptide (Atas et al. 2018), and the Cruciani properties and Kidera factors are combinations of different descriptor classes. On the other hand, the machine-learning models selected for optimization are the logistic regression, a form of generalized linear model; CART and KNN are types of nonlinear models, while SVM and RF are examples of complex nonlinear models. Thus, the selected algorithms for optimization are good representation of the general algorithm classes.

The T-scales topological descriptor generally correlated well with GCBP classification, for all the machine-learning algorithms tested (Figures 1 and 2). T-scales are topological descriptors that are derived from the principal components of 67 structural and topological variables of 135 amino acids (Tian et al. 2007). Topological descriptors refer to the numerical representation of compounds based on graph theory, wherein the atoms are treated as vertices and the bonds that connect them are the edges (Dearden 2017). Topological descriptors therefore relate to the arrangement of atoms in two dimensions. The good agreement between GCBP classification with T-scales suggests that the topological properties of the GCBP correlate well with their function. In particular, the T-scale T3 has the greatest weight in the classification model. Generally, tumor-binding peptides bind with overexpressed or distinct receptors of tumors, such as integrins and CD13 (Li and Cho 2012). Considering that there are possible multiple binding targets of GCBP, the observed relationship between peptide classification and topological descriptors may describe the general desirable attributes of peptides that can bind with gastric cancer.

Conclusion

A reliable classification model that can determine if a given peptide sequence can bind to GC cells was constructed. The study identified that the topological descriptors, T-scales, and logistic regression were the optimum pair for GCBP classification. The resulting classification model demonstrated a satisfactory classification accuracy, which was built after rigorous optimizations and external validation. The results brought forward provide the foundation for an alternative screening

method for GCBPs. This system is expected to positively contribute in the discovery of new GCBPs, thereby potentially enhancing GC disease diagnostics and management.

Conflict of interest: The authors declare no conflict of interest.

Acknowledgement

The authors are grateful to Mr. Frumencio Co for his initial review on the manuscript.

References

Akita N, Maruta F, Seymour LW, et al (2006) Identification of oligopeptides binding to peritoneal tumors of gastric cancer. Cancer Sci 97:1075–1081. doi: 10.1111/j.1349-7006.2006.00291.x

Atas H, Rifaioglu AS, Cetin-Atalay R, et al (2018) Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. Brief Bioinform. doi: 10.1093/bib/bby061

Chen B, Cao S, Zhang Y, et al (2009) A novel peptide (GX1) homing to gastric cancer vasculature inhibits angiogenesis and cooperates with TNF alpha in anti-tumor therapy. BMC Cell Biol 10:63. doi: 10.1186/1471-2121-10-63

Cruciani G, Baroni M, Carosati E, et al (2004) Peptide studies by means of principal properties of amino acids derived from MIF descriptors. J Chemom 18:146–155. doi: 10.1002/cem.856

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

Georgiev AG (2009) Interpretable Numerical Descriptors of Amino Acid Space. J Comput Biol 16:703–723. doi: 10.1089/cmb.2008.0173

Han J, Gao X, Duan W, et al (2016) The further characterization of the peptide specifically binding to gastric cancer. Mol Cell Probes 30:125–130. doi: 10.1016/j.mcp.2016.01.007

Hatakeyama S, Sugihara K, Shibata TK, et al (2011) Targeted drug delivery to tumor vasculature by a carbohydrate mimetic peptide. Proc Natl Acad Sci 108:19587–19592. doi: 10.1073/pnas.1105057108

Herrero R, Park JY, Forman D (2014) The fight against gastric cancer - The IARC Working Group report. Best Pract Res Clin Gastroenterol 28:1107–114. doi: 10.1016/j.bpg.2014.10.003

Hoffman JA, Giraudo E, Singh M, et al (2003) Progressive vascular changes in a transgenic mouse model of squamous cell carcinoma. Cancer Cell 4:383–391. doi: 10.1016/S1535-6108(03)00273-3

Hu S, Guo X, Xie H, et al (2006) Phage display selection of peptides that inhibit metastasis ability of gastric cancer cells with high liver-metastatic potential. Biochem Biophys Res Commun 341:964–972. doi: 10.1016/j.bbrc.2006.01.047

Hui X, Han Y, Liang S, et al (2008) Specific targeting of the vasculature of gastric cancer by a new tumor-homing peptide CGNSNPKSC. J Control Release 131:86–93. doi:

10.1016/j.jconrel.2008.07.024

Janairo JIB (2019) Predictive Analytics for Biomineralization Peptide Binding Affinity. Bionanoscience 9:74–78. doi: 10.1007/s12668-018-0578-4

Joyce JA, Laakkonen P, Bernasconi M, et al (2003) Stage-specific vascular markers revealed by phage display in a mouse model of pancreatic islet tumorigenesis. Cancer Cell 4:393–403. doi: 10.1016/S1535-6108(03)00271-X

Kang J, Zhao G, Lin T, et al (2013) A peptide derived from phage display library exhibits anti-tumor activity by targeting GRP78 in gastric cancer multidrug resistance cells. Cancer Lett 339:247–259. doi: 10.1016/j.canlet.2013.06.016

Kidera A, Konish Y, Oka M, et al (1985) Statistical Analysis of the Physical Properties of the 20 Naturally Occurring Amino Acids. J Protein Chem 4:23–55. doi: 10.1007/BF01025492

Koivunen E, Arap W, Valtanen H, et al (1999) Tumor targeting with a selective gelatinase inhibitor. Nat Biotechnol 17:768–774

Kuhn M, Wing J, Weston S, et al (2018) caret: Classification and Regression Training

Li ZJ, Cho CH (2012) Peptides as targeting probes against tumor vasculature for diagnosis and drug delivery. J Transl Med 10:S1. doi: 10.1186/1479-5876-10-s1-s1

Li ZJ, Wu WKK, Ng SSM, et al (2010) A novel peptide specifically targeting the vasculature of orthotopic colorectal cancer for imaging detection and drug delivery. J Control Release 148:292–302. doi: 10.1016/j.jconrel.2010.09.015

Liang G, Li Z (2007) Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides. QSAR Comb Sci 26:754–763. doi: 10.1002/qsar.200630145

Liang S, Lin T, Ding J, et al (2006) Screening and identification of vascular-endothelial-cell-specific binding peptide in gastric cancer. J Mol Med 84:764–773. doi: 10.1007/s00109-006-0064-2

Lordick F, Allum W, Carneiro F, et al (2014) Unmet needs and challenges in gastric cancer: The way forward. Cancer Treat Rev 40:692–700. doi: 10.1016/j.ctrv.2014.03.002

Matsuoka T, Hirakawa K, Chung Y, et al (1998) Adhesion polypeptides are useful for the prevention of peritoneal dissemination of gastric cancer. Clin Exp Metastasis 16:381–388

Mei H, Liao ZH, Zhou Y, Li SZ (2005) A new set of amino acid descriptors and its application in peptide QSARs. Biopolym - Pept Sci Sect 80:775–786. doi: 10.1002/bip.20296

Osorio D, Rondon-Villarreal P, Torres R (2015) Peptides: A Package for Data Mining of Antimicrobial Peptides. R J 7:4–14

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

Sanders WS, Johnston CI, Bridges SM, et al (2011) Prediction of Cell Penetrating Peptides by Support Vector Machines. PLoS Comput Biol 7:e1002101. doi: 10.1371/journal.pcbi.1002101

Shi W-J, Gao J-B (2016) Molecular mechanisms of chemoresistance in gastric cancer. World J

Gastrointest Oncol 15:673–681. doi: 10.4251/wjgo.v8.i9.673

Sidhu SS, Lowman HB, Cunningham BC, Wells JA (2000) [21] Phage display for selection of novel binding peptides. In: Thorner J, Emr SD, Abelson JN (eds) Applications of Chimeric Genes and Hybrid Proteins - Part C: Protein-Protein Interactions and Genomics. Academic Press, London

Singh Chauhan J, Kumar R, Nagpal G, et al (2015) An in silico platform for predicting, screening and designing of antihypertensive peptides. Sci Rep 5:12512. doi: 10.1038/srep12512

Sjöström M, Sandberg M, Wold S, et al (2002) New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. J Med Chem 41:2481–2491. doi: 10.1021/jm9700575

Sundquist K, Riihimäki M, Hemminki A, et al (2016) Metastatic spread in patients with gastric cancer. Oncotarget 7:52307–52316. doi: 10.18632/oncotarget.10740

Tian F, Zhou P, Li Z (2007) T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. J Mol Struct 830:106–115. doi: 10.1016/j.molstruc.2006.07.004

van Westen GJ, Bender A, Swier RF, et al (2013) Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. J Cheminform 5:41. doi: 10.1186/1758-2946-5-41

Wang H, Li R, Ma C, et al (2014) Selection and characterization of a peptide specifically targeting to gastric cancer cell line SGC-7901 using phage display. Int J Pept Res Ther 20:87–94. doi: 10.1007/s10989-013-9367-7

Yang L, Shu M, Ma K, et al (2010) ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. Amino Acids 38:805–816. doi: 10.1007/s00726-009-0287-y

Zhang WJ, Sui YX, Budha A, et al (2012) Affinity peptide developed by phage display selection for targeting gastric cancer. World J Gastroenterol 18:2053–2060. doi: 10.3748/wjg.v18.i17.2053

Zhi M, Wu KC, Dong L, et al (2004) Characterization of a specific phage-displayed peptide binding to vasculature of human gastric cancer. Cancer Biol Ther 3:1232–1235. doi: 10.4161/cbt.3.12.1223