Predicting Peptide Oligomeric State Through Chemical Artificial Intelligence

Jose Isagani B. Janairo[1]*, Gerardo C. Janairo[2]

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

[2]Chemistry Department, De La Salle University, 2401 Taft Avenue, Manila 0922, Philippines

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

Abstract

Oligomerization plays a crucial role in the structure and function of peptides and proteins, wherein sequence variations can affect the oligomeric stability of the biomolecule. In this study, an artificial neural network classifier that can predict the oligomeric state of peptides is presented, using the p53 tetramerization domain and associated mutants as the model system. The FASGAI vectors were utilized as the peptide descriptors, and the resulting binary classifier exhibits satisfactory predictive ability as demonstrated by a test set accuracy of 86%.

Keywords: artificial neural networks, machine learning, p53 tetramerization domain, tumor suppressor protein

Introduction

Oligomerization is a key characteristic of peptides and proteins that is related to their overall structure and function, such as the binding affinity of metal-binding peptides (Sakaguchi et al. 2017), and peptide antimicrobial activity (Mani et al. 2006). While different factors can affect the oligomeric state of peptides, point mutations or amino acid substitutions can have a tremendous impact. Thus, the application of machine learning (ML) techniques appears to be a promising approach to predict mutation effects on peptide oligomerization. While several ML models are available for biomolecular structural prediction, these models often require protein crystal structures, which may not always be available and applicable for peptides. Moreover, these prediction models may have limited applicability to peptides due to their shorter

length compared to proteins. Thus, there is a compelling need to create structural predictive models that are tailored for peptide oligomerization. Most applications of machine learning algorithms for peptides have mostly been focused on their bio/chemical activity (Janairo 2019; Schaduangrat et al. 2019; Janairo and Sy-Janairo 2020) rather than structure. This study therefore presents the development of a proof-of-concept ML classifier for peptide oligomeric state using the p53 tetramerization domain (p53TD) as a model system.

Tetramerization is a prerequisite for the function of the tumor suppressor p53 protein (Cheá 2001), wherein the p53TD (amino acid 326-356) is responsible for initiating oligomerization upon activation of the protein due to genotoxic stress. The p53TD is one of the five domains of the p53 protein, and in the activated state, the p53TD forms a dimer of dimers from four monomers through hydrophobic interactions leading to the formation of stable tetramers (Clore et al. 1995). The tetramerized p53 protein then carries out its function, such as facilitating the expression of genes involved in apoptosis. Point mutations in the p53TD may destabilize the tetramer, thereby negatively impacting the function of the p53 protein. This may lead to the onset of cancer (Rollenhagen and Chene 1998; DiGiammarino et al. 2002) or aberration in signal transduction pathways, such as ubiquitination (Lang et al. 2014). However, not all point mutations in the p53TD destabilize oligomerization, and have varying effects on the tetramerization (Kamada et al. 2011). This therefore renders the case of the p53TD as an ideal system to train and test ML algorithms for peptide oligomerization prediction.

Methodology

In order to create a peptide oligomerization classifier, 50 p53 TD sequences and their % tetramer values obtained from gel filtration chromatography were taken from the study of Kamada (Kamada et al. 2011). A p53 TD sequence was judged as tetrameric if the reported % tetramer exceeds 75%. For each p53 TD, the sequence was used to calculate peptide descriptors using the Peptides R package 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). These peptide descriptors can be generally classified according to what aspect of the peptide they represent. The Blosum indices are under the similarity measures category; the T-scales and ST-scales are topological descriptors; the FASGAI vectors, ProtFP, VHSE scales, and Z-scales describe the physico-chemical properties of the peptide (Rifaioglu et al. 2019). These descriptors were then used as variables to predict if a given p53 TD sequence is tetrameric or not. Multilayer perceptron artificial neural networks (ANN), classification and regression trees (CART), k-nearest neighbor (KNN), logistic regression (LR), random forest (RF), and support vector machines (SVM) classifiers were formulated for each peptide descriptor using the Caret package (Kuhn et al. 2018). For all created classification algorithms, 70% of the dataset was dedicated for training, followed by a 10-fold cross-validation. The default parameters that yielded the highest accuracy were automatically selected. All computations were conducted in R version 3.5.2 (R Core Team 2018) using a 64 bit Windows machine. The full dataset used in this study is available in the supporting information.

Results and Discussion

The first step in the modelling process involves identifying which among the nine peptide descriptors and six machine learning algorithms are capable of the classification task. The pair-wise performance analysis of the 54 constructed classification models revealed that the combination of ANN and FASGAI vectors yielded the highest classification accuracy (Figure 1).

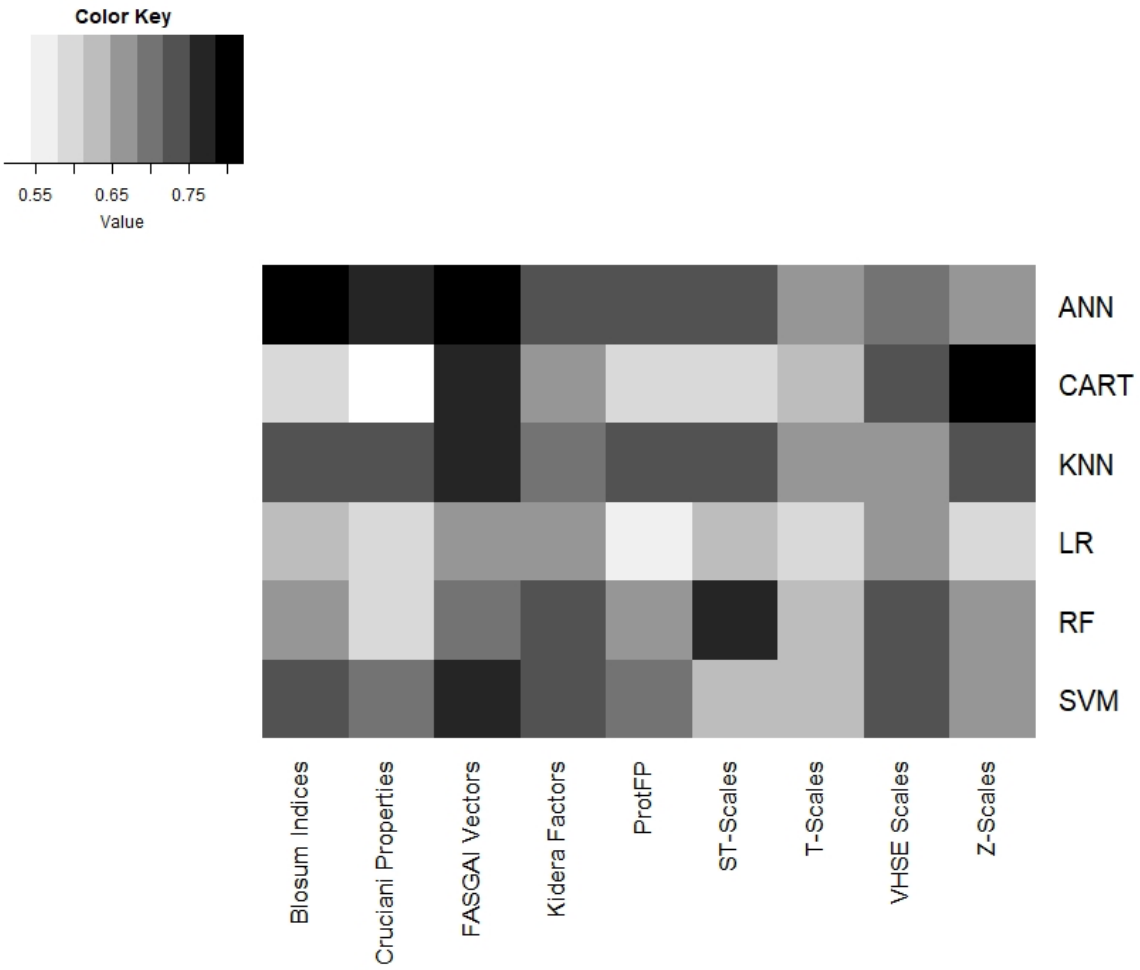

Figure 1. Training accuracy of the descriptor and machine learning algorithm for classifying oligomeric stability of the p53TD mutants. The darker the shade, the higher the training accuracy.

This class of descriptors is composed of six variables, which are related to the amino acid hydrophobicity index (FV1), alpha and turn propensities (FV2), bulky properties (FV3), compositional characteristic index (FV4), local flexibility (FV5), and electronic properties (FV6) (Liang and Li 2007). A stepwise and systematic removal of descriptors was done in order to further improve the performance and parsimony of the classifier (Figure 2). As figure 2 indicates, removing FV5 (model FV5 in Figure 2) led to a more balanced classification performance in terms of accuracy, sensitivity, and specificity. The other FASGAI

models suffer from extremely low specificity indicating that these classifiers cannot discriminate between a tetrameric and non-tetrameric p53TD based on peptide numerical representation using FASGAI vectors. Thus, model FV5 was judged as the best classifier in terms of performance and parsimony.

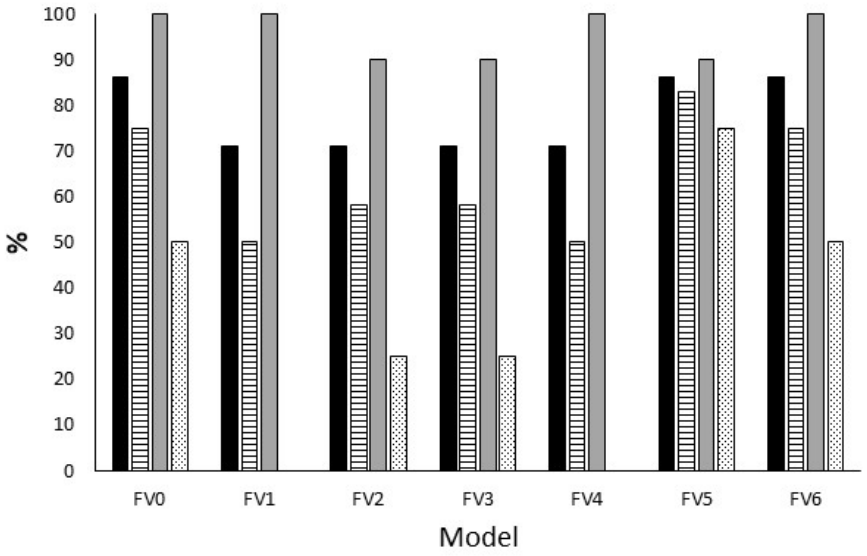

Figure 2. Performance of ANN classifiers using FASGAI Vectors as the descriptor. The model name indicates which FASGAI Vector was removed during optimization. Black fill = test set accuracy, horizontal fill = balanced accuracy, grey fill = sensitivity, dotted fill = specificity

ANN is a powerful ML algorithm that can model both linear and nonlinear relationships between the descriptors and the outcome of interest. Despite the powerful predictive capability, one of the main drawbacks of ANN models is the difficulty in interpreting the relationship of the variables in the model (Zhang et al. 2018). The ANN architecture of model FV5 utilizes four input layers, three hidden layers, two bias nodes, and one output layer (Figure 3). In addition, the optimized ANN model employed entropy fitting with a decay of $1x10^{-4}$. The weights that connect nodes in the model are similar to coefficients in general

linear models, wherein the incoming data is multiplied with the weight. Thus, a greater the magnitude of the weight translates to greater influence on predicting the outcome. The optimization steps taken suggests that the local flexibility of the amino acid composition of the p53TD has the least influence on determining its oligomerization state. On the other hand, FV2, which represents peptide secondary information has the greatest weight in the ANN structure, indicative of its importance in the prediction model. The important residues that contribute in the calculation of FV2 are methionine, glycine, and proline (Liang and Li 2007). The importance of FV2 in predicting the oligomeric state is consistent with current knowledge regarding the structure of the p53TD, wherein secondary structures, in particular alpha helices, play important roles in the formation of the tetrameric structure (Clore et al. 1995). In addition, methionine residues are known to play critical roles in the p53 tetramer stability, wherein oxidation of these residues at the tetramer core can disrupt the oligomer state of the peptide (Nomura et al. 2009). By and large, while the prediction process adopted by the ANN model can be difficult to interpret, the important elements of the ANN match with the current principles of p53 tetramerization.

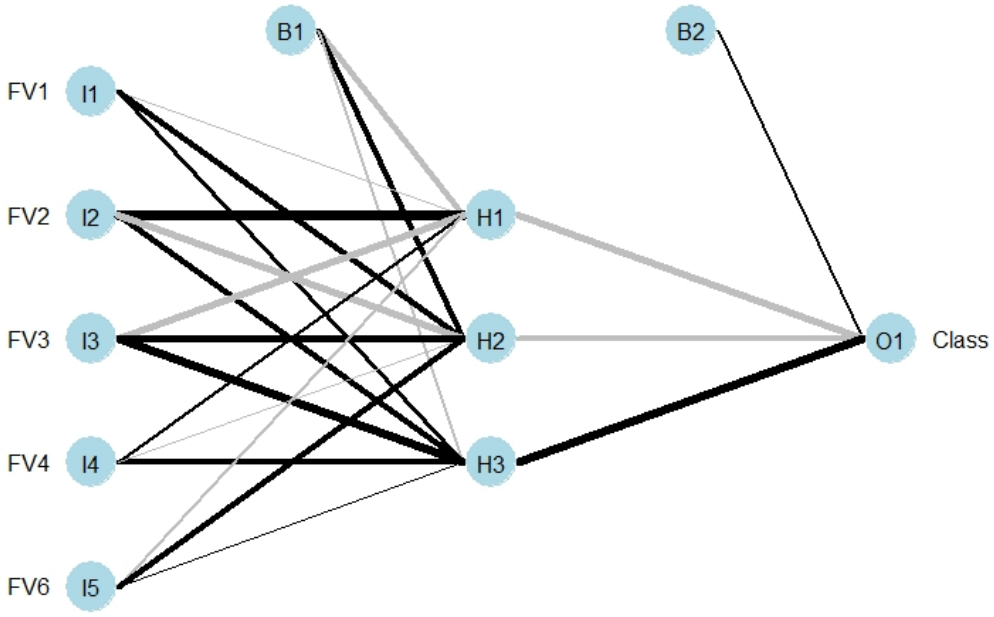

Figure 3. ANN structure for Model FV5. The darker and thicker the lines connecting the nodes, the greater the magnitude of the weight. The ANN classifier has five input nodes (I1 – I5), three hidden nodes (H1 – H3), two bias nodes (B1-B2), and one output node (O1).

Table 1. Details of the network structure, such as weight values and assignments.

B1→H1: -31.80	I1→H1: -13.65	I2→H1: 56.91
I3→H1: -42.44	I4→H1: 17.21	I5→H1: -21.88
B1→H2: 32.78	I1→H2: 31.55	I2→H2: -44.98
I3→H2: 45.67	I4→H2: -5.14	I5→H2: 33.53
B1→H3: -14.54	I1→H3: 26.27	I2→H3: 31.63
I3→H3: 50.84	I4→H3: 39.24	I5→H3: 6.24
B2→O: 18.03	H1→O: -40.59	H2→O: -37.97
H3→O: 49.44		

The presented classification algorithm offers both fundamental and practical contributions to peptide structural analysis. From a fundamental perspective, the algorithm has identified peptide data patterns, in the form of the FASGAI vectors, which can be used to describe and study the oligomerization state of the

p53TD. The presented method is more straightforward in analyzing the structure of the p53 tetramer as opposed to previous computer-aided studies that examined the structure of the p53 tetramer using molecular dynamics simulations (Rohani et al. 2016). In terms of practicality, the presented proof-of-concept ML model has promising potential to supplement routine experimental methods for peptide structural characterization such as gel filtration chromatography, circular dichroism spectroscopy, among others.

Conclusion

In summary, a binary classifier built using artificial neural networks and FASGAI vectors was formulated that can predict the oligomeric state of peptides, as demonstrated in the case of p53 tetramerization domain and associated mutants. The optimized ANN classifier has reliable predictive ability as displayed by an accuracy rating of 86%, 83% balanced accuracy, 90% sensitivity, and 75% specificity. The preliminary model provides encouraging results that rationalize the continuous development of the algorithm to accommodate other peptide sequences. Moving forward, it is envisioned that the classifier be expanded as more data becomes accessible in order to further improve the accuracy of the model, and make it available to more peptide classes.

Compliance with Ethical Standards

Conflict of Interest. None

Informed Consent. Not applicable

Ethical Approval. Not applicable

References

Cheá P (2001) The role of tetramerization in p53 function. Oncogene 20:2611–2617

Clore GM, Ernst J, Clubb R, et al (1995) Refined solution structure of the oligomerization domain of the

tumour suppressor p53. Nat Struct Mol Biol 2:321–333

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. https://doi.org/10.1002/cem.856

DiGiammarino EL, Lee AS, Cadwell C, et al (2002) A novel mechanism of tumorigenesis involving pH-dependent destabilization of a mutant p53 tetramer. Nat Struct Biol 9:12–16. https://doi.org/10.1038/nsb730

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

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

Janairo JIB, Sy-Janairo MLL (2020) A Screening Algorithm for Gastric Cancer-Binding Peptides. Int J Pept Res Ther 26:667–674. https://doi.org/10.1007/s10989-019-09874-8

Kamada R, Nomura T, Anderson CW, Sakaguchi K (2011) Cancer-associated p53 tetramerization domain mutants: Quantitative analysis reveals a low threshold for tumor suppressor inactivation. J Biol Chem 286:252–258. https://doi.org/10.1074/jbc.M110.174698

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. https://doi.org/10.1007/BF01025492

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

Lang V, Pallara C, Zabala A, et al (2014) Tetramerization-defects of p53 result in aberrant ubiquitylation and transcriptional activity. Mol Oncol 8:1026–1042. https://doi.org/10.1016/j.molonc.2014.04.002

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. https://doi.org/10.1002/qsar.200630145

Mani R, Cady SD, Tang M, et al (2006) Membrane-dependent oligomeric structure and pore formation of a-hairpin antimicrobial peptide in lipid bilayers from solid-state NMR. Proc Natl Acad Sci 103:16242–16247

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. https://doi.org/10.1002/bip.20296

Nomura T, Kamada R, Ito I, et al (2009) Oxidation of methionine residue at hydrophobic core destabilitizes p53 tetrameric structure. Biopolymers 91:78–84. https://doi.org/10.1002/bip.21084

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

Rifaioglu AS, Atas H, Martin MJ, et al (2019) Recent applications of deep learning and machine intelligence on in silico drug discovery: Methods, tools and databases. Brief. Bioinform. 20:1878–1912

Rohani L, Morton DJ, Wang X-Q, Chaudhary J (2016) Relative Stability of Wild-Type and Mutant p53 Core Domain: A Molecular Dynamic Study. J Comput Biol 23:80–89. https://doi.org/10.1089/cmb.2015.0163

Rollenhagen C, Chene P (1998) Characterization of p53 mutants identified in human tumors with a missense mutation in the tetramerization domain. Int J Cancer 78:372–376

Sakaguchi T, Janairo JIB, Lussier-Price M, et al (2017) Oligomerization enhances the binding affinity of a silver biomineralization peptide and catalyzes nanostructure formation. Sci Rep 7:1400. https://doi.org/10.1038/s41598-017-01442-8

Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombuatong W (2019) ACPred: A computational tool for the prediction and analysis of anticancer peptides. Molecules 24:1973. https://doi.org/10.3390/molecules24101973

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. https://doi.org/10.1021/jm9700575

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. https://doi.org/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. https://doi.org/10.1186/1758-2946-5-41

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. https://doi.org/10.1007/s00726-009-0287-y

Zhang Z, Beck MW, Winkler DA, et al (2018) Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. Ann Transl Med 6:216–216. https://doi.org/10.21037/atm.2018.05.32