A method for predicting natriuretic peptides in congenital heart disease using support vector machine

Atul Tyagi[1], Sudeep Roy[1,] Ivo Provaznik[2]
[1]BUT Brno, [2]Brno University of Technology

Abstract

Congenital heart disease (CHD) includes structural abnormalities of the heart that occur before birth. Congenital heart defects happen during the first eight weeks of the fetus development. CHD affected 0.8% of live births in the past few decades. The discovery of cardiac natriuretic peptides (NPs) such as ANP, BNP, CNP and research on their function and regulation in health and disease has led to breakthroughs in more profound understanding and clinical management of heart failure. Among NP properties are inhibition of cardiac remodeling. In cardiology, NPs are a valuable biomarker of heart failure. Studies investigating NP levels in the fetuses are quite limited. However, recent findings suggest that elevated NP levels are mainly attributed to increased central venous pressure secondary to arrhythmia caused by CHD. The features of ANP, BNP, and their related peptides in the umbilical cord blood and amniotic fluid provide a potential basis for their use as biomarkers.

In our recent study we analyzed 182 natriuretic peptides obtained from the UniProt database to predict and classify these peptides using Support Vector Machine (SVM). The di-peptide amino acid composition model achieved an accuracy of 92.86%, with Matthews correlation coefficient of 0.86.

1. Introduction

Congenital heart disease (CHD) is the most common type of genetic disability. The heart's structure and function problem that is presented at birth is called Congenital heart disease. It affects blood flow through the heart and to other parts of the body.1 in 100 children has a heart defect due to genetic or chromosomal abnormalities such as Down syndrome[1].The natriuretic peptides (NPs) are the efficient hormones: that includes three family’s atrial natriuretic peptide (ANP), brain natriuretic peptide (BNP), and C-type natriuretic peptide. Atrial natriuretic peptides and BNP are mainly synthesized in the heart and other organs. At the same time, CNP is produced primarily by endothelial cells[2]. These peptides are cyclic and have natriuretic, diuretic and vasodilator properties. These natriuretic peptides release by the heart is stimulated by atrial and ventricular distension and neurohormonal stimuli, which are usually in response to heart failure. In this study, we predict natriuretic peptide characteristics containing natural residues. We designed support vector machine (SVM) based models that include various features such as amino acid composition (AAC), dipeptide composition (DPC), tri-peptide composition (TPC) binary patterns. These models were used in SVM classifiers predicting natriuretic peptides. The study revealed that Ala, Asp, Glu, Gly, Leu, Trp, Lys, Pro and Val dominate various natriuretic peptides positions. Therefore, these computational models used for peptide prediction and therapeutic peptides may promote further drug development and congenital heart disease management on a global scale.

2. Methods

2.1. Datasets

We extracted 182 natriuretic peptides from the UniProt database to understand amino acid residues[3].
All these peptides were unique and considered positive examples. Since there are very few experimentally proved non- natriuretic peptides, we derived 182 random peptides from SwissProt proteins. In this study, we assign these random peptides as non-NPs (negative dataset).

2.2. Support vector machine

This study developed models for discriminating natriuretic and non-natriuretic peptides using a highly successful machine learning technique, support vector machine (SVM). This study successfully used machine learning techniques to develop a model for distinguishing natriuretic and non-natriuretic peptides. We developed the SVM model using the SVMlight package. Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression challenges.

2.3. Performance measures

The performance of models were evaluated using threshold-dependent and threshold-independent parameters. Sensitivity (Sn), specificity (Sp), accuracy (Ac) and Matthew's correlation coefficient (MCC) were used as threshold-dependent parameters as previously described. For threshold-independent parameters, ROC (Receiver Operating Characteristic) for all of the models were created to evaluate the performance of models.

2.4. Cross-validation technique

The five-fold cross-validation technique was used to evaluate the performance of various SVM models. In this technique, sequences are randomly divided into five sets, of which four sets are used for training and the remaining fifth set for testing. The process is repeated five times in such a way that each set is used once for testing. The final performance is obtained by averaging the performance of all five sets.

3. Results

3.1 Analysis of the natriuretic peptides

The comparison of the average whole amino acid composition of natriuretic and non-natriuretic peptides is suggesting that Ala, Asp, Glu, Gly, Leu, Lys, Pro, ser and Val are dominated at various positions in natriuretic peptides compared to non-natriuretic peptides (Figure 1). In the compositional analysis of NPs, it was noticed that specific residues are more dominant. Model-based observations of amino acid-based composition (AAC), dipeptide composition (DPC), and tripeptide composition (TPC) were developed using SVM models on the main data set (Table 1.1). The model developed on the NP main dataset (DPC-based model) achieved maximum accuracy of 99.27% with an MCC and AUC (area under the curve) of 0.99 and 0.9944, respectively (Table 1.1).

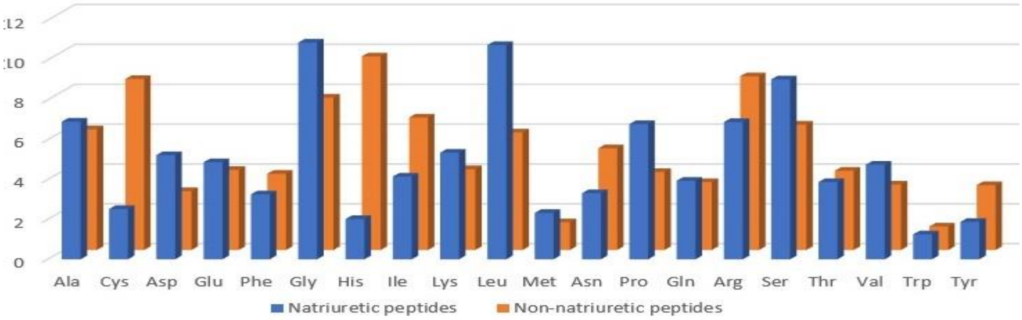

Figure 1. Comparison of the average whole amino acid composition of natriuretic and non-natriuretic peptides

Table 1.1 Performances of SVM models developed using mono, di, and tri-peptide acid composition of peptides using fivefold cross-validation

Main dataset						
Input vector	**SVM parameters**	**Sensitivity**	**Specificity**	**Accuracy**	**MCC**	**ROC**
Mono-peptide	g:0.005 c:1 j:2	97.25	99.45	98.35	0.97	0.99598
Di-peptide	g:0.001 c:1 j:1	98.53	100	99.27	0.99	0.99444
Tri-peptide	g:0.0005 c:3 j:1	97.8	97.8	97.8	0.96	0.98575

In the subsets, NT5, CT5, NTCT5, NT10, CT10, and NTCT10 and performances of these models are summarized in (Table 1.2). For example, based on five-fold cross-validation, a model developed with the NTCT10 dataset achieved a higher accuracy of 91.71% with MCC 0.84 and AUC 0.96, respectively (Table 1.2). Using binary patterns as input features, SVM models have been generated for window sizes of 5 and 10 residues. To this end, 5 and 10 residues have been extracted from the N-terminal and C-terminal of each data set, respectively. We achieved a maximum accuracy of 92.76% in NTCT5-BIN binary pattern in the five-fold cross-validation method with MCC 0.86 and ROC 0.96701, respectively (Table 1.3).

4. Discussion

To test the performance of the models, a five-fold cross-validation technique is used. Four sets were used for training and the remaining one is used for testing. The whole process is repeated five times. Threshold dependent parameter based on accuracy and the MCC measurements revealed Di- and Tri-peptide SVM models were selected as best models.

Table 1.2 Performance of SVM models developed using split acid composition using five-fold cross-validation.

Main dataset						
Input vector	**SVM parameters**	**Sensitivity**	**Specificity**	**Accuracy**	**MCC**	**ROC**
NT5	g:0.0005 c:7 j:4	80.34	93.37	86.91	0.74	0.91398
CT5	g:0.001 c:3 j:1	78.02	92.31	85.16	0.71	0.9152
NT10	g:0.005 c:2 j:1	84.62	88.2	86.39	0.73	0.93049
CT10	g:0.005 c:1 j:2	82.42	93.89	88.12	0.77	0.93547
NTCT5	g:0.0005 c:1 j:1	80.34	96.13	88.3	0.78	0.9472
NTCT10	g:0.001 c:1 j:2	86.81	93.82	90.28	0.81	0.962

Table 1.3 Performances of SVM models developed using a binary profile-based method using five-fold cross-validation

Main dataset						
Input vector	SVM parameters	Sensitivity	Specificity	Accuracy	MCC	ROC
NT5-BIN	g:0.05 c:8 j:1	84.83	97.24	91.09	0.83	0.93367
CT5-BIN	g:0.5 c:1 j:2	79.12	95.6	87.36	0.76	0.92093
NT10-BIN	g:0.05 c:3 j:2	89.01	88.2	88.61	0.77	0.94598
CT10-BIN	g:0.1 c:2 j:1	80.22	94.44	87.29	0.75	0.94081
NTCT5-BIN	g:0.5 c:1 j:2	88.2	97.24	92.76	0.86	0.96701
NTCT10-BIN	g:0.1 c:1 j:2	87.36	96.07	91.67	0.84	0.97284

A large number of support vector machine-based studies PhytoAFP, CellPPD, TumorHPD, and AntiCP have been developed to predict and design plant antifungal peptides, cell-penetrating, tumor homing, anticancer, respectively[4-7].The present study demonstrates that features like amino acid composition, di, split acid composition and hybrid approach can be used to train an SVM classifier that can predict natriuretic peptides with higher accuracy. The DPC-based model achieved maximum accuracy described in this study. Based on the above research, the di-peptide-based amino acid composition model (DPC) of natriuretic peptides can help biologists and users easily predict the natriuretic peptides.Users can download the natriuretic peptides-related dataset and other essential files from our GitHub account https://github.com/ks26atul/Tyagi.

Acknowledgments

The authors gratefully acknowledge the support provided by the Brno University of Technology, Brno, Czech Republic.

References

1. Sun R, Liu M, Lu L, Zheng Y, Zhang P. Congenital heart disease: causes, diagnosis, symptoms, and treatments. Cell biochemistry and biophysics. 2015;72(3):857-60.
2. Volpe M, Rubattu S, Burnett Jr J. Natriuretic peptides in cardiovascular diseases: current use and perspectives. European heart journal. 2014;35(7):419-25.
3. Consortium U. UniProt: a worldwide hub of protein knowledge. Nucleic acids research. 2019;47(D1):D506-D15.
4. Tyagi A, Roy S, Singh S, Semwal M, Shasany AK, Sharma A, et al. PhytoAFP: In Silico Approaches for Designing Plant-Derived Antifungal Peptides. Antibiotics. 2021;10(7):815.
5. Gautam A, Chaudhary K, Kumar R, Sharma A, Kapoor P, Tyagi A, et al. In silico approaches for designing highly effective cell penetrating peptides. Journal of translational medicine. 2013;11(1):1-12.
6. Sharma A, Kapoor P, Gautam A, Chaudhary K, Kumar R, Chauhan JS, et al. Computational approach for designing tumor homing peptides. Scientific reports. 2013;3(1):1-7.
7. Tyagi A, Kapoor P, Kumar R, Chaudhary K, Gautam A, Raghava G. In silico models for designing and discovering novel anticancer peptides. Scientific reports. 2013;3(1):1-8.