Computers in Biology and Medicine 42 (2012) 504–507

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine

Wei Chen [a,c,*], Hao Lin [b,*]

[a] Department of Physics, College of Sciences, Hebei United University, Tangshan 063000, China
[b] Key Laboratory for Neuroinformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
[c] Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China

ARTICLE INFO

Article history:
Received 31 December 2010
Accepted 12 January 2012

Keywords:
Voltage-gated potassium channel
Subfamily
Feature selection
Support vector machine

ABSTRACT

Proteins belonging to different subfamilies of Voltage-gated K^+ channels (VKC) are functionally divergent. The traditional method to classify ion channels is more time consuming. Thus, it is highly desirable to develop novel computational methods for VKC subfamily classification. In this study, a support vector machine based method was proposed to predict VKC subfamilies using amino acid and dipeptide compositions. In order to remove redundant information, a novel feature selection technique was employed to single out optimized features. In the jackknife cross-validation, the proposed method (VKCPred) achieved an overall accuracy of 93.09% with 93.22% average sensitivity and 98.34% average specificity, which are superior to that of other two state-of-the-art classifiers. These results indicate that VKCPred can be efficiently used to identify and annotate voltage-gated K^+ channels' subfamilies. The VKCPred software and dataset are freely available at http://cobi.uestc.edu.cn/people/hlin/tools/VKCPred/.

1. Introduction

Potassium (K^+) channels are the most widely distributed type of ion channels and are found in virtually all living organisms [1,2]. Voltage-gated K^+ channels (VKC), the largest family of K^+ channels, are specific for K^+ ions and sensitive to voltage changes of cell membrane. VKC response to membrane potential changes by opening and closing an ion-selective permeation pathway of K^+ ions across the cell, and play a crucial role during action potentials in returning the depolarized cell to a resting state. They are also key components in generation and propagation of electrical impulses in nervous system. By contributing to the regulation of the action potential duration in cardiac muscle, malfunction of VKC may cause life-threatening arrhythmias. Moreover, mutations in VKC genes can lead to severe diseases, such as long QT syndrome and epilepsy [3–6]. Thus, VKC have become valuable targets for disease diagnosis and drug design.

In terms of sequence diversity, VKC can be grouped into different subfamilies [7,8] and proteins in these subfamilies are functionally divergent. The sensitivity to the membrane potential and the kinetics of the response to changes in potential vary substantially between different VKC subfamilies, which means that cells expressing different VKC subfamilies repolarize the membrane and shorten the action potential at different rates [9]. Thus, judging the subfamilies of VKC not only provides novel insights into their functions, but also facilitates understanding the intricate pathways regulating cellular processes. Phylogenetic tree [10] is a traditional method and a gold standard for most experimental scholars to classify ion channels. Although this method is not particularly expensive, it is more time consuming than machine learning approaches.

Currently, many computational methods have been developed to study the structure and function of ion channels [11–17]. Some works focused on the prediction of the activity of ion channel proteins using machine learning methods [18,19]. Other works employed computational approaches to find potential drug targets from ion channels [20,21]. A web server VGIchan [22] was developed to predict four types of voltage-gated ion channels using support vector machine (SVM). Lin and Ding [23] developed a soft package IonchanPred to predict ion channels and their types. However, few cases were performed to predict subfamilies of VKC. Recently, Liu et al. [24] predicted five types of VKC on a high similarity benchmark dataset using support vector machine. It has been demonstrated that the predictive accuracy is closely related with sequence identity [25,26] and high sequence similarity can surely lead to the overestimation of predictive performance. Therefore, there is an urgent need to develop efficient computational tools for VKC subfamilies identification.

* Corresponding author at: Department of Physics, College of Sciences, Hebei United University, Tangshan 063000, China.
E-mail addresses: chenwei_imu@yahoo.com.cn (W. Chen), hlin@uestc.edu.cn (H. Lin).

0010-4825/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2012.01.003

Keeping this in mind, we proposed an SVM-based method to predict VKC subfamilies using amino acid compositions and dipeptide compositions. The Correlation-based Feature Subset Selection algorithm [27] was introduced to perform feature selection for improving the predictive accuracy. In the jack knife cross-validation, our method yields an overall accuracy of 93.09% with 93.22% average sensitivity and 98.34% average specificity for VKC subfamilies prediction.

2. Materials and methods

2.1. Dataset

VKC belonging to 6 phylogenetic subfamilies (Kv1, Kv2, Kv3, Kv4, Kv6 and Kv7) were derived from the updated Voltage-gated K^+ Channel DataBase (VKCDB, http://vkcdb.biology.ualberta.ca) [9]. To prepare a high quality dataset, the following procedures were performed: (1) Protein sequences containing ambiguous residues ("B", "X" and "Z") were removed. (2) Sequences which are not full-length were excluded. (3) Highly similar sequences were excluded. For balancing the number of samples and providing a significant statistics, sequences which have $>60\%$ sequence similarity were removed from the dataset using CD-HIT program [28]. If the sequence identity cutoff is set to a stringent threshold of 25%, the results will be more objective and reliable. However, in this study we did not use such a stringent criterion because the currently available data do not allow us to do so. Otherwise, the number of proteins for some subfamilies would be too few to have statistical significance. Finally, 217VKC were retained in the final dataset, composed of 82 Kv1, 16 Kv2, 37 Kv3, 32 Kv4, 10 Kv6 and 40 Kv7, respectively.

2.2. Features

Each sequence in the dataset is encoded by 420 features, including 20 amino acid compositions (AAC) and 400 dipeptide compositions (DPC) defined as the following equations:

$$AAC(i)=\frac{x(i)}{\sum_{i=1}^{20}x(i)} \tag{1}$$

$$DPC(j)=\frac{y(j)}{\sum_{j=1}^{400}y(j)} \tag{2}$$

where i can be any one of the 20 amino acids and j represents one out of the 400 dipeptides. The $x(i)$ and $y(j)$ are their numbers in each sequence, respectively.

2.3. Feature selection

Inclusion of irrelevant, redundant and noisy attributes in the model building process can result in poor predictive performance and increased computation. To remove redundant features and identify features that could distinguish subfamilies of VKC from the original feature set, we performed feature selection using the Correlation-based Feature Subset Selection algorithm that couples the evaluation formula with an appropriate correlation measure and a heuristic search strategy [27]. This procedure was implemented using Weka [29] by ten-fold cross-validation on the final dataset.

2.4. Support vector machine

Support vector machine (SVM) is an effective method for supervised pattern recognition [13], which is well founded theoretically [30] and has been widely used in the field of bioinformatics, such as subcellular localization prediction [31–36] and membrane protein identification [37–39]. The basic idea of SVM is to transform the data into a high dimensional feature space, and then determine the optimal separating hyperplane. For handling a multi-class problem, "one-versus-one (OVO)" and "one-versus-rest (OVR)" are generally applied to extend the traditional SVM. Since this work deals with proteins from 6 families, OVO strategy was employed for making prediction using radial basis function (RBF), which outperforms the other three kinds of kernel functions, namely, linear function, polynomial function and sigmoid function, in empirical studies. The implementation of SVM is based on LibSVM 2.84 written by Chang and Lin [40]. The grid search method is applied to tune the regularization parameter C and the kernel width parameter γ.

2.5. Prediction assessment

The performance of the method was measured in terms of sensitivity (Sn), specificity (Sp), Matthew's correlation coefficient (MCC) [41] and overall accuracy (OA) defined as follows:

$$Sn(i)=\frac{TP(i)}{TP(i)+FN(i)} \tag{3}$$

$$Sp(i)=\frac{TN(i)}{TN(i)+FP(i)} \tag{4}$$

$$MCC(i)=\frac{TP(i)\times TN(i)-FP(i)\times FN(i)}{\sqrt{(TP(i)+FN(i))\times(TN(i)+FP(i))\times(TP(i)+FP(i))\times(TN(i)+FN(i))}} \tag{5}$$

$$OA=\frac{1}{N}\sum_{i=1}^{k}TP(i) \tag{6}$$

here k ($k=6$) is the number of families, N is the total number of sequences. $TP(i)$, $TN(i)$, $FP(i)$, and $FN(i)$ represent true positive, true negative, false positive and false negative of family i, respectively.

3. Results

3.1. Prediction of VKC subfamilies

We trained our SVM classifier (VKCPred) on the dataset containing 217 sequences. Three cross-validation methods, i.e., subsampling test, independent dataset test and jack knife test are often employed to evaluate the predictive capability of a predictor. Among the three methods, the jack knife test is deemed the most objective and rigorous one [42] that can always yield a unique outcome as demonstrated by a penetrating analysis in a recent comprehensive review [43] and has been widely and increasingly adopted [44–47]. Accordingly, the jack knife test was employed to examine the performance of VKCPred. For the jackknife cross-validation, each sample in the dataset is in turn singled out as an independent test sample and all the rule parameters are calculated based on the remaining samples without including the one being identified. As shown in Table 1, an overall accuracy of 87.39% with an average sensitivity of 84.54% and an average specificity of 96.43% was obtained using all 420 features.

In order to identify prominent features that distinguish proteins from different VKC subfamilies, we applied Correlation-based Feature Subset Selection algorithm to eliminate redundant features. This algorithm evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while

Table 1
Performance of VKCPred for VKC subfamily classification with different features.

Family	Best 24 features			Best 50 features			Best 118 features			420 features		
	Sn(%)	*Sp*(%)	*MCC*	*Sn*(%)	*Sp*(%)	*MCC*	*Sn*(%)	*Sp*(%)	*MCC*	*Sn*(%)	*Sp*(%)	*MCC*
Kv1	90.24	87.77	0.77	92.68	90.98	0.83	**93.90**	**93.98**	**0.86**	93.90	89.31	0.82
Kv2	75.00	99.46	0.82	75.00	98.93	0.79	**87.50**	**98.95**	**0.86**	81.25	100.00	0.89
Kv3	75.68	97.11	0.76	83.78	95.95	0.79	**89.19**	**97.69**	**0.87**	75.68	96.51	0.75
Kv4	100.00	99.35	0.98	87.50	100.00	0.92	**93.75**	**100.00**	**0.96**	87.50	97.08	0.84
Kv6	90.00	100.00	0.95	100.00	100.00	1.00	**100.00**	**100.00**	**1.00**	80.00	100.00	0.89
Kv7	91.11	98.73	0.91	88.89	97.52	0.87	**95.00**	**99.39**	**0.95**	88.89	95.65	0.83
Average *Sn*(%)	87.01			87.98			**93.22**			84.54		
Average *Sp*(%)	97.07			97.23			**98.34**			96.43		
OA (%)	88.29			88.74			**93.09**			87.39		

Table 2
Comparison of SVM with other machine learning methods.

Family	SVM			Naïve Bayes			RF		
	Sn(%)	*Sp*(%)	*MCC*	*Sn*(%)	*Sp*(%)	*MCC*	*Sn*(%)	*Sp*(%)	*MCC*
Kv1	**93.90**	**93.98**	**0.86**	93.90	83.85	0.76	97.56	78.51	0.76
Kv2	**87.50**	**98.95**	**0.86**	81.25	100.00	0.89	75.00	98.78	0.82
Kv3	**89.19**	**97.69**	**0.87**	81.08	95.12	0.75	59.46	97.44	0.67
Kv4	**93.75**	**100.00**	**0.96**	87.50	100.00	0.92	65.38	98.73	0.75
Kv6	**100.00**	**100.00**	**1.00**	40.00	100.00	0.62	80.00	98.82	0.87
Kv7	**95.00**	**99.39**	**0.95**	85.00	98.70	0.87	85.00	99.29	0.89
Average *Sn*(%)	**93.22**			78.12			77.07		
Average *Sp*(%)	**98.34**			96.28			95.26		
OA (%)	**93.09**			85.71			82.03		

having low inter-correlation are preferred. The prediction rate is improved in each feature selection step until the number of features increased to 118. To demonstrate this point, we listed the representative results (subsets with best 24, 50 and 118 best features) in Table 1. The highest overall accuracy of 93.09% with 93.22% average sensitivity and 98.34% average specificity was obtained by using 118 features.

3.2. Comparison with other methods

It is objective to compare the proposed methods with previously published classifiers using an independent dataset. However, the currently available data do not allow us to do so because of lacking enough samples. We can only perform a rough comparison with previously published classifiers. Liu et al. [24] have predicted five types of VKC on a high similarity benchmark dataset using 2000 features. Using 118 optimized features, our proposed model classified six types of VKC on a lower similarity benchmark dataset and achieved a comparable accuracy with that of Liu et al. [24].

In addition, our proposed SVM model was also compared with two state-of-the-art classifiers, i.e. Random Forest (RF) and Naïve Bayes (NB) implemented in WEKA. We also optimized the features for both RF and NB algorithms and enumerated their best jack knife cross-validated results in Table 2. The prediction accuracy of SVM is approximately 8% and 11% higher than Naïve Bayes and Random Forest classifiers, respectively. These results indicate that it is creditable to use VKCPred to annotate voltage-gated K^+ channels' subfamilies.

4. Discussion

Proteins from different VKC subfamilies are functionally divergent. Identification of VKC subfamily is an essential and difficult task. We implemented an SVM based approach to predict VKC subfamilies using primary sequence information. Using Correlation-based Feature Subset Selection algorithm, we identified the 118 prominent features that could improve predictive accuracies remarkably. However, the detailed analysis of the selected features are required to provide more information about their roles in biological activity. High accuracies indicate that the proposed method is an effective tool for VKC subfamily identification. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors [48], we have provided a predictive tool of the proposed method at http://cobi.uestc.edu.cn/people/hlin/tools/VKCPred/ and hope that this method will pave the way for the further research on VKC.

Conflict of interest statement

The author has no conflict of interests concerning this work.

Acknowledgments

The authors would like to thank the anonymous referees for constructive comments on the manuscript. This work was supported by the National Natural Science Foundation of China (no. 61100092), The Scientific and Technological Department Foundation of Hebei Province (no. 11275532) and The Doctoral Scientific Research Start-up Foundation of Hebei United University (no. 10101115) to CW, The Fundamental Research Funds for the Central Universities (ZYGX2009J081) and The Scientific Research Foundation of Sichuan Province (2009JY0013) to LH.

References

[1] T.M. Perney, L.K. Kaczmarek, The molecular biology of K channels, Curr. Opinion Cell Biol. 3 (1991) 663–670.

[2] J.T. Littleton, B. Ganetzky, Ion channels and synaptic organization: analysis of the Drosophila genome, Neuron 26 (2000) 35–43.
[3] F. Lehmann-Horn, K. Jurkat-Rott, Voltage-gated ion channels and hereditary disease, Physiol. Rev. 79 (1999) 1317–1372.
[4] T.J. Jentsch, Neuronal KCNQ potassium channels: Physiology and role in disease, Nat. Rev. Neurosci. 1 (2000) 21–30.
[5] G.J. Kaczorowski, M.L. Garcia, Pharmacology of voltage-gated and calcium-activated potassium channels, Curr. Opinion Chem. Biol. 3 (1999) 448–458.
[6] O. Pongs, Voltage-gated potassium channels: from hyperexcitability to excitement, FEBS Lett. 452 (1999) 31–35.
[7] O. Pongs, Molecular biology of voltage-dependent potassium channels, Physiol. Rev. 72 (1992) S69–S87.
[8] W.A. Coetzee, Y. Amarillo, J. Chiu, A. Chow, D. Lau, T. McCormack, H. Moreno, M.S. Nadal, A. Ozaita, D. Pountney, M. Saganich, E. Vega-Saenz de Miera, B. Rudy, Molecular diversity of K+ channels, Ann. NY Acad. Sci. 868 (1999) 233–285.
[9] W.J. Gallin, P.A. Boutet, VKCDB: voltage-gated K+ channel database updated and upgraded, Nucleic Acids Res. 39 (2011) D362–D366.
[10] E.Z. Mangubat, T.T. Tseng, E. Jakobsson, Phylogenetic analyses of potassium channel auxiliary subunits, J. Mol. Microbiol. Biotechnol. 5 (2003) 216–224.
[11] K.C. Chou, Insights from modelling three-dimensional structures of the human potassium and sodium channels, J. Proteome Res 3 (2004) 856–861.
[12] V.S. Satuluri, J. Seelam, S.P. Gupta, A quantitative structure-activity relationship study on some series of potassium channel blockers, Med. Chem. 5 (2009) 87–92.
[13] J.F. Wang, K.C. Chou, Insights from studying the mutation-induced allostery in the M2 proton channel by molecular dynamics, Protein Eng. Des. Sel. 23 (2010) 663–666.
[14] J.R. Schnell, J.J. Chou, Structure and mechanism of the M2 proton channel of influenza A virus, Nature 451 (2008) 591–595.
[15] H. Wei, C.H. Wang, Q.S. Du, J. Meng, K.C. Chou, Investigation into adamantane-based M2 inhibitors with FB-QSAR, Med. Chem. 5 (2009) 305–317.
[16] Q.S. Du, R.B. Huang, C.H. Wang, X.M. Li, K.C. Chou, Energetic analysis of the two controversial drug binding sites of the M2 proton channel in influenza A virus, J. Theor. Biol. 259 (2009) 159–164.
[17] Q.S. Du, R.B. Huang, S.Q. Wang, K.C. Chou, Designing inhibitors of M2 proton channel against H1N1 swine influenza virus, PLoS ONE 5 (2010) e9388.
[18] W. Peter, W. David, Prediction of ion channel activity using binary kernel discrimination, Chem. Inf. Model 47 (2007) 1961–1966.
[19] E. Pourbasheer, S. Riahi, M.R. Ganjali, P. Norouzi, Application of genetic algorithm-support vector machine (GA-SVM) for prediction of BK-channels activity, Eur. J. Med. Chem. 44 (2009) 5023–5028.
[20] X. Guo, J.F. Wang, Y. Zhu, D.Q. Wei, Recent progress on computer-aided inhibitor design of H5N1 influenza a virus, Curr. Comput. Aided Drug Des. Apr (2010) 6. [Epub ahead of print].
[21] Z. He, J. Zhang, X.H. Shi, L.L. Hu, X. Kong, Y.D. Cai, K.C. Chou, Predicting drug-target interaction networks based on functional groups and biological features, PLoS ONE 5 (2010) e9603.
[22] S. Saha, J. Zack, B. Singh, G.P. Raghava, VGIchan: prediction and classification of voltage-gated ion channels, Genomics Proteomics Bioinformatics 4 (2006) 253–258.
[23] H. Lin, H. Ding, Prediction of ion channels and their types by the dipeptide mode of pseudo amino acid composition, J. Theor. Biol. 269 (2011) 64–69.
[24] L.X. Liu, M.L. Li, F.Y. Tan, M.C. Lu, K.L. Wang, Y.Z. Guo, Z.N. Wen, L. Jiang, Local sequence information-based support vector machine to classify voltage gated potassium channels, Acta Biochim. Biophys. Sin 38 (2006) 363–371.
[25] R. Nair, B. Rost, Sequence conserved for subcellular localization, Protein Sci. 11 (2002) 2836–2847.
[26] C.S. Yu, Y.C. Chen, C.H. Lu, J.K. Hwang, Prediction of protein subcellular localization, Proteins 64 (2006) 643–651.
[27] M.A. Hall, Correlation-based feature selection for machine learning: Data Mining, Inference and Prediction, Second ed, Springer-Verlag, 2008.
[28] W. Li, A. Godzik, Cd-hit:a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.
[29] E. Frank, M. Hall, L. Trigg, G. Holmes, I.H. Witten, Data mining in bioinformatics using Weka, Bioinformatics 20 (2004) 2479–2481.
[30] V.N. Vapnik, The Nature of Statistical Learning Theory, Wiley-Interscience, 1998.
[31] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect, J. Cell Biochem. 84 (2002) 343–348.
[32] K.C. Chou, Y.D. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, J. Biol. Chem. 277 (2002) 45765–45769.
[33] S.J. Hua, Z.R. Sun, Support vector machine approach for protein subcellular localization prediction, Bioinformatics 17 (2001) 721–728.
[34] H. Lin, H. Ding, F.B. Guo, J. Huang, Prediction of subcellular location of mycobacterial protein using feature selection techniques, Mol. Divers. 14 (2010) 667–671.
[35] H. Lin, H. Wang, H. Ding, Y.L. Chen, Q.Z. Li, Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition, Acta Biotheor. 57 (2009) 321–330.
[36] W. Chen, H. Lin, Prediction of midbody, centrosome and kinetochore proteins using gene ontology, Biochem. Biophys. Res. Commun. 401 (2010) 382–384.
[37] Y.D. Cai, P.W. Ricardo, C.H. Jen, K.C. Chou, Application of SVM to predict membrane protein types, J. Theor. Biol. 226 (2004) 373–376.
[38] Y.D. Cai, G.P. Zhou, K.C. Chou, Support vector machines for predicting membrane protein types by using functional domain composition, Biophys. J. 84 (2003) 3257–3263.
[39] K.J. Park, M.M. Gromiha, P. Horton, M. Suwa, Discrimination of outer membrane proteins using support vector machines, Bioinformatics 21 (2005) 4223–4229.
[40] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines. Software available at 〈http://www.csie.ntu.edu.tw/_cjlin/libsvm〉.
[41] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta 405 (1975) 442–451.
[42] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.
[43] K.C. Chou, H.B. Shen, Review: recent progress in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.
[44] Y.D. Cai, K.C. Chou, Predicting subcellular localization of proteins in a hybridization space, Bioinformatics 20 (2004) 1151–1156.
[45] K.C. Chou, Y.D. Cai, Prediction of membrane protein types by incorporating amphipathic effects, J. Chem. Inf. Model. 45 (2005) 407–413.
[46] K.C. Chou, D.W. Elrod, Prediction of enzyme family classes, J. Proteome Res. 2 (2003) 183–190.
[47] K.C. Chou, H.B. Shen, ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information, Crit. Rev. Biochem. Mol. Biol. 376 (2008) 321–325.
[48] K.C. Chou, H.B. Shen, Recent advances in developing web-servers for predicting protein attributes, Natural Science 2 (2009) 63–92.