Int. J. Mol. Sci. **2014**, *15*, 12940-12951; doi:10.3390/ijms150712940

OPEN ACCESS
International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms

Article

Identifying the Subfamilies of Voltage-Gated Potassium Channels Using Feature Selection Technique

Wei-Xin Liu [1], En-Ze Deng [1], Wei Chen [2] and Hao Lin [1,*]

[1] Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China; E-Mails: lweixin316@gmail.com (W.-X.L.); enzeas@gmail.com (E.-Z.D.)

[2] Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China; E-Mail: greatchen@heuu.edu.cn

* Author to whom correspondence should be addressed; E-Mail: hlin@uestc.edu.cn;
Tel.: +86-28-8320-2351; Fax: +86-28-8320-8238.

*Received: 10 June 2014; in revised form: 13 July 2014 / Accepted: 14 July 2014 /
Published: 22 July 2014*

Abstract: Voltage-gated K^+ channel (VKC) plays important roles in biology procession, especially in nervous system. Different subfamilies of VKCs have different biological functions. Thus, knowing VKCs' subfamilies has become a meaningful job because it can guide the direction for the disease diagnosis and drug design. However, the traditional wet-experimental methods were costly and time-consuming. It is highly desirable to develop an effective and powerful computational tool for identifying different subfamilies of VKCs. In this study, a predictor, called iVKC-OTC, has been developed by incorporating the optimized tripeptide composition (OTC) generated by feature selection technique into the general form of pseudo-amino acid composition to identify six subfamilies of VKCs. One of the remarkable advantages of introducing the optimized tripeptide composition is being able to avoid the notorious dimension disaster or over fitting problems in statistical predictions. It was observed on a benchmark dataset, by using a jackknife test, that the overall accuracy achieved by iVKC-OTC reaches to 96.77% in identifying the six subfamilies of VKCs, indicating that the new predictor is promising or at least may become a complementary tool to the existing methods in this area. It has not escaped our notice that the optimized tripeptide composition can also be used to investigate other protein classification problems.

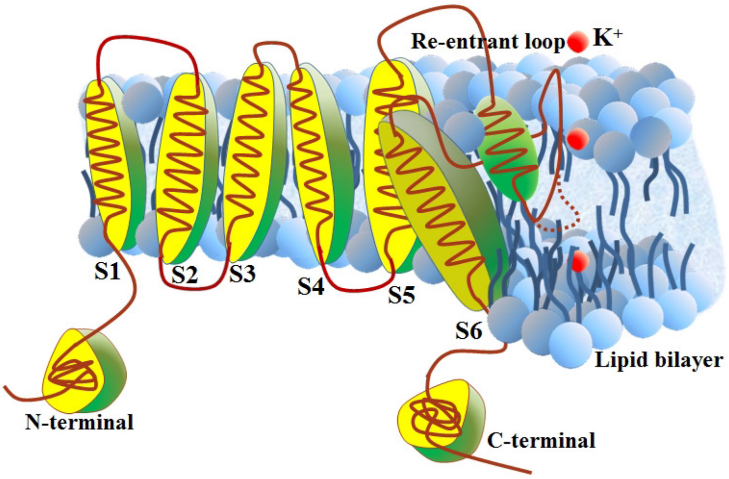
Keywords: voltage-gated potassium channel; subfamily; optimized tripeptide composition; support vector machine; feature selection

1. Introduction

Ion channels located in the surface of cell membrane can maintain the balance of cell microenvironment by selectively penetrating ions and organic molecules in and out of cells. The K^+ channel has been found in all living organisms [1]. The voltage-gated K^+ channel (VKC), which is the largest family of K^+ channels, specifically controls the movement of K^+ under the stimulation of voltage changes in the cell's membrane potential. During action potentials, they play crucial roles in returning the depolarized cell to a resting state [2]. They are also key components in generation and propagation of electrical impulses in nervous system. The mutations in VKC genes can lead to severe diseases, such as long QT syndrome and epilepsy [3]. Thus, VKCs have become valuable targets for disease diagnosis and drug design.

VKCs have four subunits. Each subunit comprises six transmembrane helices. A re-entrant loop forms the ion-selective channel, highly variable *C*- and *N*-terminal domains (Figure 1). According to the *N*- and *C*-terminal domains, VKCs can be grouped into different subfamilies. The proteins in these subfamilies are functionally divergent. Different subfamilies of VKC proteins have different sensitivity to the membrane potential and response to changes in potential [2]. Therefore, recognition of subfamily type of a new VKC is benefit to understand its biological functions. However, the traditional biochemical methods were costly and time-consuming. Thus, it is necessary to develop effective computational methods to identify subfamilies of VKCs.

Figure 1. Schematic representation of potassium (K^+) channel subunit. The S1, S2, S3, S4, S5, S6 are six transmembrane helices.



In the past decade, some scholars have focused on the identification of VKCs families. Liu *et al.* [4] proposed a dipeptide-based method to predict five subfamilies of VKCs. Subsequently, Chen and Lin [5] developed an SVM-based model to predict six subfamilies of VKCs by using the Correlation-based Feature Subset Selection algorithm (CFSS) to select the optimal features. All these methods could yield quite encouraging results, and each of them did play a role in stimulating the development of this area. However, further work is needed due to the following reasons. (i) The predicted successful rate is still far from satisfaction; (ii) No web-server was provided to most of these methods, and, hence, their usage is quite limited, especially for the majority of experimental scientists.

The present study was initiated in an attempt to improve the prediction of VKC subfamilies from the above two aspects. According to a comprehensive review [6], to establish a really useful statistical predictor for VKC subfamily prediction, an objective benchmark dataset was constructed. Subsequently, a feature selection technique was used to obtain the optimal tripeptides. The support vector machine was used to operate the prediction. The jackknife cross-validation test was utilized to estimate the accuracy of the predictor. Finally, we established a user-friendly web-server for the predictor.

2. Results and Discussion

2.1. Benchmark Dataset

The raw dataset of VKCs were extracted from the updated Voltage-gated K$^+$ Channel Database (VKCDB) [2] and filtered by VKCPred [5]. The following steps were used to construct a reliable benchmark dataset. At first, if the primary structure (amino acid sequence) of a VKC contains ambiguous residues, such as “B”, “X”, and “Z”, the VKC will be removed; Secondly, if the sequence is fragment of other proteins, it will be excluded because its information is redundant and fragmentary; Thirdly, to objectively evaluate the proposed predictor, the CD-HIT software [7] was used to remove highly similar sequences by setting the cutoff of sequence identity to 60%. As a result, we obtained the benchmark dataset S as formulated by:

$$S = S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 \cup S_6 \tag{1}$$

where the subset S_1 contains 82 Kv1 subfamily proteins, S_2 contains 16 Kv2 subfamily proteins, S_3 contains 37 Kv3 subfamily proteins, S_4 contains 32 Kv4 subfamily proteins, S_5 contains 10 Kv6 subfamily proteins and S_6 contains 40 Kv7 subfamily proteins (Table 1) and where U represents the symbol for union in the set theory. For readers’ convenience, the 217 VKCs can be freely downloaded from our webserver.

Table 1. Breakdown of the 217 voltage-gated K$^+$ channels (VKCs) in the benchmark dataset S according to their six subfamilies.

Dataset	Channel Subfamilies	Number of VKC Samples
S_1	Kv1	82
S_2	Kv2	16
S_3	Kv3	37
S_4	Kv4	32
S_5	Kv6	10
S_6	Kv7	40
S	Overall	217

2.2. The Tripeptide Composition

To develop a sequence-based predictor for the prediction of the subfamilies of VKCs, one of the keys is to formulate its sequence with an effective mathematical expression that can truly reflect the intrinsic correlation with the types to be predicted. The most straightforward method to formulate the sample of a VKC protein P with *L* residues is to use its entire amino acid sequence, as can be formulated by:

$$\mathrm{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\ldots\mathrm{R}_L \tag{2}$$

where R_1 represents the 1st residue of the protein P, R_2 represents the 2nd residue of the protein P, and so forth. According to a recent review [8], the general form of PseAAC for a protein P is formulated by:

$$\mathrm{P} = [\Psi_1\ \Psi_2\ \ldots\ \Psi_u\ \ldots\ \Psi_\Omega]^{\mathrm{T}} \tag{3}$$

where the subscript Ω is an integer and its value, as well as the components Ψ_u (u = 1, 2, …, Ω), will depend on how to extract the desired information from the amino acid sequence P (*cf.* Equation (3)).

Tripeptide is a useful and minimal biological recognition signal which can be used for studying molecular modulators of biological function [9] and predicting plausible structures for oligopeptides as well as de novo protein design [8]. Thus, we extract tripeptide composition from the benchmark dataset S to define the components in Equation (3) for the VKC samples concerned in this study. Then a VKC sequence can be formulated by:

$$\mathrm{P} = [f_1, f_2, \cdots, f_i, \cdots, f_{8000}]^{\mathrm{T}} \tag{4}$$

where symbol T denotes the transposition of vector and the f_i is the frequency of the i-th (i = 1, 2, …, 8000) tripeptide in the VKC and expressed as:

$$f_i = n_i / (L - 2) \tag{5}$$

where n_i and L denote the occurrence number of the i-th tripeptide and the length of the VKC sequence, respectively.

2.3. Feature Selection

If all 8000 tripeptides are used for prediction, the predictive result isn't usually satisfactory, such as low generalization ability of prediction model and poor prediction results because irrelevant features and noise is included. On the other hand, it is time-consuming to analyze an 8000 dimensional vector for large amounts of proteins. Using feature selection techniques to optimize feature set can not only gain deeper insight into the intrinsic properties of VKCs, but also improve understandability, scalability, possibility, and accuracy of the proposed models. Moreover, it can also economize the time for model construction and prediction.

Although many dimensionality reduction techniques such as principal component analysis (PCA) [10,11], diffusion Maps [12] and minimal-redundancy-maximal-relevance (mRMR) [13,14] have been proposed to perform feature selection, none of them concerned the statistical significance of the features. According to this, we proposed the binomial distribution to investigate the statistical significance of each tripeptide and the optimal the feature set.

Each of the 8000 tripeptides occurring in one subfamily may be a stochastic event, thus, we must calculate the confidence level (*CL*) of each tripeptide occurring in different VKC subfamilies. For a stochastic event, two possible cases that are occurrence and non-occurrence will happen when one observes the *i*-th tripeptide occurring in the *k*-th VKC subfamily. Each outcome has a fixed probability when benchmark dataset has been fixed. This probability is called prior probability and defined as:

$$p_k = \sum_{i=1}^{8000} n_{ik} \Big/ \sum_{k=1}^{6} \sum_{i=1}^{8000} n_{ik} \tag{6}$$

where $\sum_{k=1}^{6}\sum_{i=1}^{8000} n_{ik}$ denotes the total occurrence number of all tripeptides in the benchmark dataset. $\sum_{i=1}^{8000} n_{ik}$ is the occurrence number of all tripeptides in the *k*-th VKC subfamily. The n_{ik} represents the number of the *i*-th tripeptides occurring in the *k*-th VKC subfamily. Correspondingly, the probability of the non-occurrence in the *k*-th VKC subfamily is defined as $q_k = 1 - p_k$.

Let $N_i = \sum_{k=1}^{6} n_{ik}$ represents the total occurrence number of the *i*-th tripeptide in benchmark dataset. That is to say, under the condition of the prior probability p_k, one performs trial or observation with N_i times. We may calculate the posterior probability P_{ik} of the *i*-th tripeptide occurring n_{ik} or more times in the *k*-th VKC subfamily as following:

$$P_{ik} = 1 - CL_{ik} = \sum_{m=n_{ik}}^{N_i} \frac{N_i}{m!(N_i - m)!} p_k^m (1 - p_k)^{N_i - m} \tag{7}$$

where CL_{ik} is the *CL* of the *i*-th tripeptide in the *k*-th VKC subfamily. Based on small probability event principle, if P_{ik} is a small value, it means the tripeptide *i* appearing in VKC subfamily *k* is not random.

There are six VKC subfamilies in the current study, namely *k* = 1, 2, 3, 4, 5, 6. Hence, for an arbitrary tripeptide *i*, it has six *CLs* corresponding to six VKC subfamilies. Then, we may define the probability of tripeptide *i* in benchmark dataset as:

$$CL_i = \max\left\{CL_{i\,\mathrm{Kv1}}, CL_{i\,\mathrm{Kv2}}, CL_{i\,\mathrm{Kv3}}, CL_{i\,\mathrm{Kv4}}, CL_{i\,\mathrm{Kv6}}, CL_{i\,\mathrm{Kv7}}\right\} \quad (i = 1, 2, \cdots, 8000) \tag{8}$$

It should be noted that the larger the CL_i is, the more likely this feature has a better discriminative capability. Therefore, we ranked the tripeptides according to their CL_i. Based on the ranked tripeptides, we used the Incremental Feature Selection (IFS) strategy to find an optimal subset of features that gives the highest overall accuracy. During the IFS procedure, the feature subset started with one feature with the largest *CL*. A new feature subset was composed when one feature with the second largest *CL* had been added. By adding features one by one from larger to smaller rank, this process repeated 8000 times until all the features were evaluated. Thus, the 8000 feature sets thus formed would be composed of 8000 ranked features. The *τ*-th feature set can be formulated as:

$$\mathrm{S}_\tau = \left\{f_1, f_2, \cdots f_i, \cdots, f_\tau\right\} \qquad \left(1 \le \tau \le 8000\right) \tag{9}$$

where f_i has been defined by Equation (5). For each of the feature sets, the cross-validation test was used to investigate the accuracy by using proposed predictive algorithm. Through the method referred above, we got an IFS curve in Descartes Curvilinear Coordinate System, which used τ as *X* axis, *CL* as *Y* axis and overall accuracy as *Z* axis. The optimal feature set is expressed as:

$$\mathrm{S}_\Theta = \left\{f_1, f_2, \cdots f_i, \cdots, f_\Theta\right\} \tag{10}$$

with which the IFS curve reaches its peak. In other words, in the 3D Cartesian coordinate system, when $X = \Theta$, the value of overall accuracy is the maximum. Thus, we used the Θ features to build the final predictor.

2.4. Support Vector Machine

Support vector machine has been widely applied in bioinformatics [15–20]. The basic idea of applying SVM to pattern classification is to map samples with low dimensional feature space into a high dimensional space, and then seek an optimal separating hyperplane with the maximal margin in this space by using the decision function:

$$f(\vec{X}) = sgn(\sum_{i=1}^{N} y_i \alpha_i \cdot K(\vec{X}, \vec{X}_i) + b) \tag{11}$$

where $\vec{X}_i$ is the *i*-th training vector. The y_i represents the type of the *i*-th training vector. α_i is coefficient which can be solved by quadratic programming. The *b* is the intercept parameter. $K(\vec{X}, \vec{X}_i)$ is a kernel function which defines an inner product in a high dimensional feature space. Because of its effectiveness and speed in nonlinear classification process, the radial basis kernel function (RBF) $K(\vec{X}_i, \vec{X}_j) = \exp(-\gamma \| \vec{X}_i - \vec{X}_j \|^2)$ was used to in this work.

The traditional SVM was designed for two-class problems. For handling a multi-class problem, "one-versus-one (OVO)" and "one-versus-rest (OVR)" are often applied to extend the traditional SVM. The present study adopted OVO strategy for multi-class prediction. The software toolbox used to implement SVM is LibSVM [21]. A grid search method was used to optimize the regularization parameter *c* and kernel parameter *γ* by using cross-validation test. The search spaces for *c* and *γ* are (2^{15}, 2^{-5}) and (2^{-5}, 2^{-15}) with steps being 2^{-1} and 2, respectively.

2.5. Prediction Assessment

The predictive capability and reliable of the method is estimated by the four parameters: the sensitivity (*Sn*), specificity (*Sp*), Matthew's correlation coefficient (*MCC*) and overall accuracy (*OA*), which were employed to measure the performance of the method and can be defined as follows:

$$Sn_k = \frac{TP_k}{TP_k + FN_k} \tag{12}$$

$$Sp_k = \frac{TN_k}{TN_k + FP_k} \tag{13}$$

$$MCC_k = \frac{TP_k \times TN_k - FP_k \times FN_k}{\sqrt{(TP_k + FN_k)(TN_k + FP_k)(TP_k + FP_k)(TN_k + FN_k)}} \tag{14}$$

$$OA = \frac{1}{N} \sum_{k=1}^{6} TP_k \tag{15}$$

where *k* is the *k*-th VKC subfamily, *N* is the total sequence number of benchmark dataset. TP_k, TN_k, FP_k and FN_k represent true positive, true negative, false positive and false negative of the *k*-th VKC subfamily, respectively.

3. Experimental

In statistical prediction, the following four cross-validation test methods were often used to build a predictor for its effectiveness in practical application: self-consistency test, independent dataset test, *n*-fold cross-validation and jackknife cross-validation. Among them, the jackknife test method makes best use of the data, involves no random sub-sampling and achieves unique results [6,22]. It has been widely and increasingly adopted in bioinformatics [5,12–14,23–25]. Therefore, the jackknife cross-validation was used in all procession of feature selection and parameter optimization of SVM.

Based on Equations (4)–(5), we may define the 8000 tripeptide composition as the original feature set. Generally, the larger the feature set is, the more information the representation bears. However, the tripeptides with low *CL* (or large posterior probability) maybe randomly appear in six VKC subfamilies. Including these tripeptides into feature set will add redundant information or reduce the cluster-tolerant capacity so as to lower down the cross-validation accuracy. For example, 8000 tripeptides can only produce the overall accuracy of 92.17% for predicting different VKC subfamilies. In contrast, the tripeptides with larger *CL* (or small posterior probability) give more reliable information for classification. The occurrence of these tripeptides prefers to different VKC subfamilies. However, if the number of tripeptide in feature set is very small, they are still not the optimized features for prediction because they cannot reflect real characteristics of VKCs and afford enough information, which deduces the poor predictive accuracy. For instance, by selecting 29 tripeptides with *CL*~100% (p value = 10^{-7}), we can only achieve 81.10%.

Therefore, it is a key step to obtain the best feature set which can product the maximum overall accuracy. According to the equation from Equation (6) to (9), we calculated the cross-validated accuracy of all 8000 feature sets using SVM and plotted a three-dimensional curve for *CL*, feature dimension and OA in Figure 2. As we can see from Figure 2, the overall accuracy reaches its maximum of 96.77% when the *CL* is selected as 99.99%. The optimized feature set contains 648 tripeptides. Results in Table 2 show that the average *Sn* and average *Sp* are 93.92% and 99.20%, respectively, indicating that the proposed method is indeed very powerful in identifying proteins which belongs to different subfamilies of VKCs.

Recently, the optimized dipeptide composition (DPC) and amino acid composition (AAC) selected by Correlation-based Feature Subset Selection (CFSS) algorithm were used to predict six VKC subfamilies by Chen and Lin [5]. In jackknife cross-validation, the overall accuracies of 93.09%, 85.71% and 82.03% were obtained by SVM, Naïve Bayes and Random Forest, respectively. The comparative results in Table 2 demonstrate that the method proposed in this paper is superior to the published methods [5].

Figure 2. The IFS curve (red) in a 3D Cartesian coordinate system for predicting six subfamilies of VKCs. The blue, green and yellow lines are the projections of the IFS curve on the Overall accuracy/Confidence level plane, the Overall accuracy/Feature dimension plane, the Feature dimension/Confidence level plane, respectively.

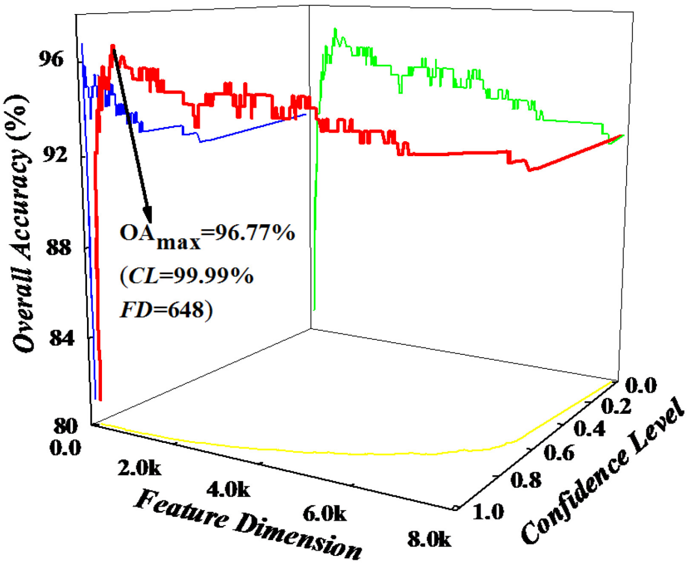


Table 2. Comparison with other published methods.

Family	This Paper			SVM [5]			Naïve Bayes [5]			Random Forest [5]		
	Sn (%)	*Sp* (%)	*MCC*	*Sn* (%)	*Sp* (%)	*MCC*	*Sn* (%)	*Sp* (%)	*MCC*	*Sn* (%)	*Sp* (%)	*MCC*
Kv1	100.00	96.30	0.95	93.90	93.98	0.86	93.90	83.85	0.76	97.56	78.51	0.76
Kv2	93.75	100.00	0.96	87.50	98.95	0.86	81.25	100.00	0.89	75.00	98.78	0.82
Kv3	97.30	98.89	0.95	89.19	97.69	0.87	81.08	95.12	0.75	59.45	97.44	0.67
Kv4	100.00	100.00	1.00	93.75	100.00	0.96	87.50	100.00	0.92	65.38	98.73	0.75
Kv6	80.00	100.00	0.89	100.00	100.00	1.00	40.00	100.00	0.62	80.00	98.82	0.87
Kv7	92.50	100.00	0.95	95.00	99.39	0.95	85.00	98.70	0.87	85.00	99.29	0.89
Average *Sn* (%)	93.92			93.22			78.12			77.07		
Average *Sp* (%)	99.20			98.34			96.28			95.26		
OA (%)	96.77			93.09			85.71			82.03		

For verifying the advantage of optimized tripeptide composition, it is necessary to investigate the performance of other parameters. Hence, we estimated the accuracies of traditional pseudo amino acid composition (PseAAC) [6], optimal tripeptides combined with PseAAC and optimal tripeptides combined with dipeptides on six subfamilies of voltage-gated ion channels. Results were recorded in Table 3. It is obviously that the optimized tripeptide composition is superior to other parameters. It should be noted that the two mixture features can only achieve the overall accuracies of 96.31% and 95.39% which are lower than that (96.77%) of our optimal tripeptides, suggesting that information redundancy or noise were included in mixture feature sets.

For testifying the capability of the proposed feature selection technique, a powerful feature selection technique, namely SVM-RFE [26,27], was introduced to optimize the tripeptides. Subsequently, the

IFS strategy was used to find an optimal subset of features that gives the highest overall accuracy. The maximum accuracy was recorded in Table 3. Comparison demonstrated that our feature selection technique is more powerful.

Table 3. Comparison with different methods on training set.

Method	Sn (%)						OA (%)
	Kv1	Kv2	Kv3	Kv4	Kv6	Kv7	
Optimal tripeptides (Our method)	100.00	93.75	97.30	100.00	80.00	92.50	96.77
Optimal tripeptides (SVM-RFE)	100.00	81.25	91.67	96.88	80.00	87.55	93.09
Traditional PseAAC	82.93	81.25	72.97	78.13	80.00	87.50	81.11
Optimal tripeptides (Our method) + PseAAC	100.00	87.50	97.30	100.00	80.00	92.50	96.31
Optimal tripeptides (Our method) + Dipeptides	100.00	81.25	94.59	100.00	80.00	92.50	95.39

4. Conclusions

In this work, we developed a promising feature selection technique to optimize feature set and applied these selected features to identify six VKC subfamilies. An overall accuracy of 96.77% was achieved, demonstrating that the proposed model is a powerful tool for the study of VKC subfamilies prediction. For the convenience of experimental scientists, a free web server iVKC-OTC was built to implement the prediction. A friendly guide was given to describe the way to use the iVKC-OTC web server. We believe that the predictor will be helpful for wet lab scientists who focus on VKC research. We hope the predictor will pave the way for the future research of VKC.

5. Web-Server and User Guide

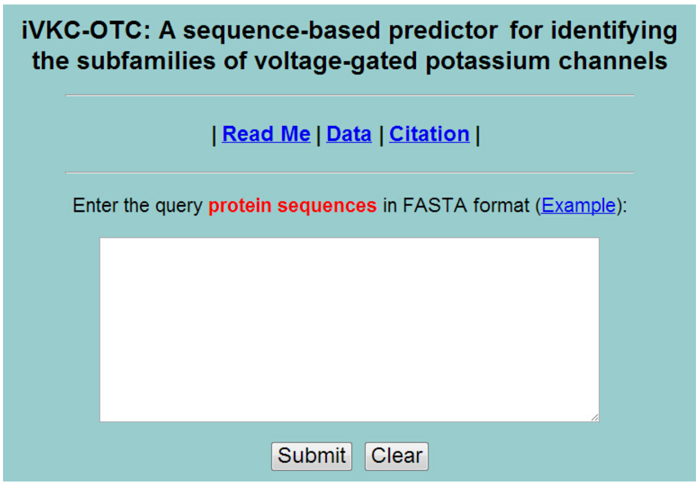
Establishing a user-friendly web-server will improve the efficiency and avoid repeating a complicated mathematics and program for studying VKC. The predictor established via aforementioned procedures is called iVKC-OTC, where "i" stands for "identify", "VKC" for "Voltage-gated K^+ channel" and "OTC" for "optimized tripeptide composition".

For the convenience of the vast majority of experimental scientists, we provided a guide on how to use the web-server to get the desired results.

Step 1. Open the web server and you will see the top page of iVKC-OTC [28] on your computer screen, as shown in Figure 3 Click on the Read Me button to see a brief introduction about the predictor and the caveat when using it.

Step 2. Either type or copy/paste the query peptide sequences into the input box at the center of Figure 3 The input sequence should be in the FASTA format. A sequence in FASTA format consists of a single initial line beginning with a greater-than symbol (">") in the first column, followed by lines of sequence data. The words right after the ">" symbol in the single initial line are optional and only used for the purpose of identification and description. The sequence ends if another line starting with a ">" appears; this indicates the start of another sequence. Example sequences in FASTA format can be seen by clicking on the Example button right above the input box.

Figure 3. A semi-screenshot for the top page of the iVKC-OTC.



iVKC-OTC: A sequence-based predictor for identifying the subfamilies of voltage-gated potassium channels

| Read Me | Data | Citation |

Enter the query protein sequences in FASTA format (Example):

Step 3. Click on the Submit button to see the predicted result. After clicking the Submit button, you will see the following shown on the screen of your computer: the outcome for the 1st query example is "Kv1 subfamily protein"; the outcome for the 2nd query sample is "Kv2 subfamily protein"; the outcome for the 3rd query sample is "Kv3 subfamily protein"; the outcome for the 4th query sample is "Kv4 subfamily protein"; the outcome for the 5th query sample is "Kv6 subfamily protein" and the outcome for the 6th query sample is "Kv7 subfamily protein". All these results are fully consistent with the experimental observations. It takes about few seconds for the above computation before the predicted result appears on your computer screen; the more number of query sequences and longer of each sequence, the more time it is usually needed.

Step 4. Click on the Data button to download the benchmark datasets used to train and test the iVKC-OTC predictor.

Step 5. Click on the Citation button to find the relevant papers that document the detailed development and algorithm of iVKC-OTC.

Caveats. Each of the input query sequences cannot any illegal character: such as "B", "X", "U", "Z".

Acknowledgments

This work was supported by the National Nature Scientific Foundation of China (Nos. 61202256, 61301260 and 61100092), the Nature Scientific Foundation of Hebei Province (No. C2013209105), and the Fundamental Research Funds for the Central Universities (Nos. ZYGX2012J113, ZYGX2013J102).

Author Contributions

Conceived and designed the experiments: Hao Lin, Wei Chen. Performed the experiments: Wei-Xin Liu, En-Ze Deng. Analyzed the data: Hao Lin, Wei-Xin Liu. Contributed reagents/materials/analysis tools: Wei-Xin Liu, En-Ze Deng, Hao Lin. Wrote the paper: Wei-Xin Liu, Hao Lin, Wei Chen.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Littleton, J.T.; Ganetzky, B. Ion channels and synaptic organization: Analysis of the Drosophila genome. *Neuron* **2000**, *26*, 35–43.
2. Gallin, W.J.; Boutet, P.A. VKCDB: Voltage-gated K$^+$ channel database updated and upgraded. *Nucleic Acids Res*. **2011**, *39*, D362–D366.
3. Lehmann-Horn, F.; Jurkat-Rott, K. Voltage-gated ion channels and hereditary disease. *Physiol. Rev.* **1999**, *79*, 1317–1372.
4. Liu, L.X.; Li, M.L.; Tan, F.Y.; Lu, M.C.; Wang, K.L.; Guo, Y.Z.; Wen, Z.N.; Jiang, L. Local sequence information-based support vector machine to classify voltage-gated potassium channels. *Acta Biochim. Biophys. Sin.* **2006**, *38*, 363–371.
5. Chen, W.; Lin, H. Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine. *Comput. Biol. Med.* **2012**, *42*, 504–507.
6. Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247.
7. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659.
8. Anishetty, S.; Pennathur, G.; Anishetty, R. Tripeptide analysis of protein structures. *BMC Struct. Biol.* **2002**, doi:10.1186/1472-6807-2-9.
9. Ung, P.; Winkler, D.A. Tripeptide motifs in biology: Targets for peptidomimetic design. *J. Med. Chem.* **2011**, *54*, 1111–1125.
10. Ma, J.; Gu, H. A novel method for predicting protein subcellular localization based on pseudo amino acid composition. *BMB Rep.* **2010**, *43*, 670–676.
11. Olivier, I.; Loots du, T. A metabolomics approach to characterise and identify various Mycobacterium species. *J. Microbiol. Methods* **2012**, *88*, 419–426.
12. Yin, J.B.; Fan, Y.X.; Shen, H.B. Conotoxin superfamily prediction using diffusion maps dimensionality reduction and subspace classifier. *Curr. Protein Pept. Sci.* **2011**, *12*, 580–588.
13. Jia, P.; Qian, Z.; Feng, K.; Lu, W.; Li, Y.; Cai, Y. Prediction of membrane protein types in a hybrid space. *J. Proteome Res*. **2008**, *7*, 1131–1137.
14. Huang, T.; Xu, Z.; Chen, L.; Cai, Y.D.; Kong, X. Computational analysis of HIV-1 resistance based on gene expression profiles and the virus-host interaction network. *PLoS One* **2011**, *6*, e17291.
15. Rashid, M.; Saha, S.; Raghava, G.P. Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. *BMC Bioinform.* **2007**, doi:10.1186/1471-2105-8-337.
16. Liu, B.; Xu, J.; Zou, Q.; Xu, R.; Wang, X.; Chen, Q. Using distances between top-n-gram and residue pairs for protein remote homology detection. *BMC Bioinform.* **2014**, doi:10.1186/1471-2105-15-S2-S3.
17. Liu, B.; Wang, X.; Lin, L.; Tang, B.; Dong, Q.; Wang, X. Prediction of protein binding sites in protein structures using hidden Markov support vector machine. *BMC Bioinform.* **2009**, doi:10.1186/1471-2105-10-381.

18. Liu, B.; Wang, X.; Lin, L.; Dong, Q.; Wang, X. A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis. *BMC bioinform.* **2008**, doi:10.1186/1471-2105-9-510.
19. Liu, B.; Zhang, D.; Xu, R.; Xu, J.; Wang, X.; Chen, Q.; Dong, Q.; Chou, K.C. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* **2013**, *30*, 472–479.
20. Liu, B.; Xu, J.; Fan, S.; Xu, R.; Zhou, J.; Wang, X. Protein remote homology detection by combining Chou's pseudo amino acid composition and profile—Based protein representation. *Mol. Inform*. **2013**, *32*, 775–782.
21. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol*. **2011**, *2*, 1–27.
22. Chou, K.C.; Zhang, C.T. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275–349.
23. Lin, H.; Chen, W.; Yuan, L.F.; Li, Z.Q.; Ding, H. Using over-represented tetrapeptides to predict protein submitochondria locations. *Acta Biotheor.* **2013**, *61*, 259–268.
24. Fan, G.L.; Liu, Y.L.; Zuo, Y.C.; Mei, H.X.; Rang, Y.; Hou, B.Y.; Zhao, Y. acACS: Improving the prediction accuracy of protein subcellular locations and protein classification by incorporating the average chemical shifts composition. *Sci. World J.* **2014**, doi:org/10.1155/2014/864135.
25. Lin, H.; Ding, H.; Guo, F.B.; Huang, J. Prediction of subcellular location of mycobacterial protein using feature selection techniques. *Mol. Divers*. **2010**, *14*, 667–671.
26. Li, L.; Yu, S.; Xiao, W.; Li, Y.; Li, M.; Huang, L.; Zheng, X.; Zhou, S.; Yang, H. Prediction of bacterial protein subcellular localization by incorporating various features into Chou's PseAAC and a backward feature selection approach. *Biochimie* **2014**, doi:10.1016/j.biochi.2014.06.001.
27. Li, L.; Cui, X.; Yu, S.; Zhang, Y.; Luo, Z.; Yang, H.; Zhou, Y.; Zheng, X. PSSP-RFE: Accurate prediction of protein structural class by recursive feature extraction from psi-blast profile, physical-chemical property and functional annotations. *PLoS One* **2014**, *9*, e92863.
28. The Webserver iVKC-OTC. Available online: http://lin.uestc.edu.cn/server/iVKC-OTC (accessed on 14 July 2014).