Available online at www.sciencedirect.com

ScienceDirect

Journal of Molecular Graphics and Modelling 26 (2008) 1306–1314

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Classification of voltage-gated K^+ ion channels from 3D pseudo-folding graph representation of protein sequences using genetic algorithm-optimized support vector machines

Michael Fernández [a,b,*], Leyden Fernández [a], Jose Ignacio Abreu [a,c], Miguel Garriga [a,d]

[a] Molecular Modeling Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, 44740 Matanzas, Cuba
[b] Department of Bioscience and Bioinformatics, Kyushu Institute of Technology (KIT), 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan
[c] Artificial Intelligence Lab, Faculty of Informatics, University of Matanzas, 44740 Matanzas, Cuba
[d] Plant Biotechnology Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, 44740 Matanzas, Cuba

Received 31 August 2007; received in revised form 3 January 2008; accepted 3 January 2008
Available online 11 January 2008

Abstract

Voltage-gated K^+ ion channels (VKCs) are membrane proteins that regulate the passage of potassium ions through membranes. This work reports a classification scheme of VKCs according to the signs of three electrophysiological variables: activation threshold voltage (V_t), half-activation voltage ($V_{a_{50}}$) and half-inactivation voltage ($V_{h_{50}}$). A novel 3D pseudo-folding graph representation of protein sequences encoded the VKC sequences. Amino acid pseudo-folding 3D distances count (AAp3DC) descriptors, calculated from the Euclidean distances matrices (EDMs) were tested for building the classifiers. Genetic algorithm (GA)-optimized support vector machines (SVMs) with a radial basis function (RBF) kernel well discriminated between VKCs having negative and positive/zero V_t, $V_{a_{50}}$ and $V_{h_{50}}$ values with overall accuracies about 80, 90 and 86%, respectively, in crossvalidation test. We found contributions of the "pseudo-core" and "pseudo-surface" of the 3D pseudo-folded proteins to the discrimination between VKCs according to the three electrophysiological variables.

Keywords: Ion channel; Electrophysiological variables; Kernel-based methods; Graph similarity

1. Introduction

Voltage-gated ion channels (VICs) are intrinsic membrane proteins that respond to changes in the transmembrane electrical field by altering conformation and selectively allowing ions to pass through the membrane [1]. A pore for the passage of ions across the lipid bilayer is formed when a VIC senses a change in the transmembrane electric field. Particularly, voltage-gated K^+ ion channels (VKCs) are membrane proteins that regulate the passage of potassium ions through membranes [1]. The probability that VKCs will open begins to become significant when the voltage difference across a membrane reaches a threshold, then an ion-selective pore in the channel is formed allowing increased potassium ion diffusion through it. Four subunits S1 through S6A, each containing six transmembrane regions, form a functional VKC. The main voltage-sensing domain S4 [1], acts by moving perpendicular to the plane of the membrane upon depolarization [2,3]. A conformational change in the region of the pore opens the "gate" allowing potassium ions to pass through. Several mechanisms have been proposed to explain how the sensor movement changes the channel conformation but this phenomenon remains unclear [4].

Voltage-regulated potassium ion permeability is critical in cellular excitability. Cardiac arrhythmias [5], episodic ataxia [6], and other diseases [7,8] are associated to mutations in VKC genes. Consequently, VKC proteins have been considered good targets for drug design directed at a number of diseases [9,10]. Structure–function relationship of VKCs has been mainly carried

* Corresponding author at: Molecular Modeling Group, Center for Biotechnological Studies, Faculty of Agronomy, University of Matanzas, 44740 Matanzas, Cuba. Tel.: +53 45 26 1251; fax: +53 45 25 3101.
E-mail addresses: michael.fernandez@umcc.cu, michael_llamosa@yahoo.com (M. Fernández).

1093-3263/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.01.001

out by site-directed mutagenesis but the prohibitively time-consuming and costly implementation of this experimental technique has prompted the application of computational modelling tools to VKCs study. Multiple sequence alignment has been usually used to identify conserved regions of VKCs and limit the priority in mutagenesis experiments to evolutionarily conserved residues [11,12]. But understanding by simple inspection of aligned VKC sequences the complex structure–function relationship between individual residues and the electrophysiological properties is very difficult. In this regards, machine learning approaches have been also used to correlate VKC sequences with electrophysiological variables [13].

However, converting sequence information to numerical values in order to applied statistical pattern recognition techniques is essential. Some reports refer the novel extensions of different structure/property relationships approaches to the prediction of protein properties and function from the sequence [14–20]. In such reports, descriptors are calculated over the protein sequences in such a way that several variables are computed considering the protein structure as a simplified molecular pseudo graph of Cα atoms. Specifically, we introduced the amino acid pseudo-folding 3D distances count (AAp3DC) descriptors to the classification of the sign of the change of free energy change upon single mutations of protein mutants [19]. AAp3DC descriptors are an encoding scheme based on a novel 3D pseudo-folding graph representation of protein sequence that enriches the information from the protein primary structure.

In this work, optimum models for the recognition of the signs of three electrophysiological variables: activation threshold voltage (V_t), half-activation voltage ($V_{a_{50}}$) and half-inactivation voltage ($V_{h_{50}}$) of VKC proteins was successfully built from their protein sequences. A total of 35 AAp3DC descriptors were calculated from the amino acids Euclidean distances matrices (EDMs) from the protein 3D graphs. Prediction of the signs of V_t, $V_{a_{50}}$ and $V_{h_{50}}$ was accomplished by support vector machine (SVM) classification optimized by genetic algorithm (GA).

2. Materials and methods

2.1. VKC sequences and electrophysiological variables dataset

Structural and functional VKC data was obtained from the VKCDB database [21], which currently stores 346 voltage-gated potassium channel entries including some ‘‘unknown proteins’’ annotated by automatic genome annotation projects, sharing a high degree of sequence similarity with voltage-gated potassium channels. In the database, authors checked the entries manually for redundancy, sequence conflicts, and isoforms. We collected all the data with reported electrophysiological variables: activation threshold voltage (V_t), half-activation voltage ($V_{a_{50}}$) and half-inactivation voltage ($V_{h_{50}}$) (Table 1SM Supplementary Material). A set of non-redundant VKC sequences with the following electrophysiological classification was collected:

- V_t a total of 143 cases: 60 cases with $V_t < 0$ (class V_t-) and 83 cases with $V_t = 0$ (class V_t+).
- $V_{a_{50}}$ a total of 64 cases: 44 cases with $V_{a_{50}} < 0$ (class V_a-) and 20 cases with $V_{a_{50}} > 0$ (class V_a+).
- $V_{h_{50}}$ a total of 120 cases: 37 cases with $V_{h_{50}} < 0$ (class V_h-) and 83 cases with $V_{h_{50}} = 0$ (class V_h+).

The sequences of those VKC proteins were used for calculating AAp3DC descriptors.

2.2. 3D pseudo-folding representations of protein sequences and amino acid pseudo-folding 3D distances count descriptors

Exploitation of protein sequences for prediction and similarity studies have been extended by developing different representations of sequence [9,14–18,20,22–27]. Some of these approaches intended to enrich the primary structure information by mapping sequence residues in a higher-dimensional space. Among these reports, Bai and Wang [26] referred a new approach to represent protein sequences with a 3D Cartesian coordinate system. Protein representations are constructed in the interior of a regular dodecahedron centered at origin of the Cartesian coordinate system and one vertex at the point (0, 0, 1) that is circumscribed inside a unit ‘‘magic sphere’’ (Fig. 1). At the dodecahedron vertexes are positioned 20 amino acids. For calculating 3D position (x_i, y_i, z_i) of amino acid i in a given sequence Bai and Wang [26] multiplied the 3 × 20 matrix of Cartesian coordinates of dodecahedron vertexes by a 1 × 20 vector representing the cumulative occurrence numbers of amino acids from the 1st amino acid to the ith amino acid of the sequence [26].

Differently to Bai and Wang [26], we calculated 3D positions of amino acid residues by using a method similar to the ‘‘moving across the sequence’’ scheme reported by Jeffrey [25] for graphical representation of DNA that was recently applied to the 2D graph representation of protein sequences by Randić et al. [22]. This representation is more suitable for discriminating among highly similar sequences such as it is the case of single point mutants. In this method 3D graphical representation of proteins is obtained by starting in the ‘‘magic

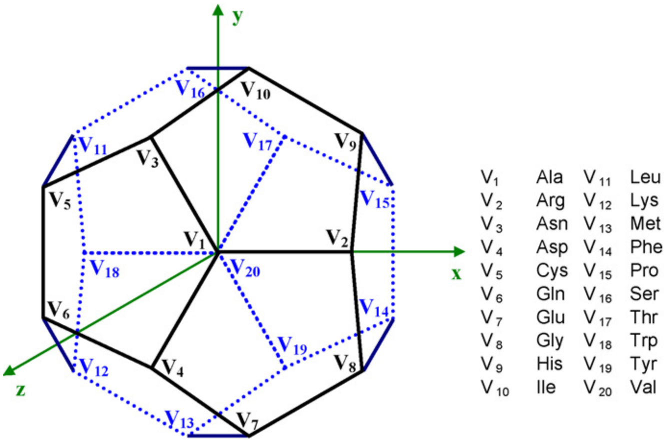

Fig. 1. Graphical representation of the regular dodecahedron proposed by Bai and Wang [26] for the 3D representation of protein sequences.

sphere'' center following amino acid sequence by moving half way (half-step) towards the corresponding amino acid. By moving towards one amino acid at dodecahedron vertexes to another, following the sequence order, protein is pseudo-folded inside the ''magic sphere'' of unit radius. 3D graphical form of depicted protein depends on ordering on amino acids in the sequence, however, the relative variations between sequences remain to a large extent independent of the ordering of amino acids along the dodecahedron vertexes. Amino acids in vertexes from 1 to 20 were assigned in alphabetical order according to three-letter code. As example, Fig. 2A and B depicts graphical representations of two shorter segments of protein of yeast *Saccharomyces cerevisiae* already used by Randić et al. [22] for 2D graph representation of protein sequences:

- Protein I: WTFESRNDPAKDPVILWLNGGPGCSSLTGL.
- Protein II: WFFESRNDPANDPIILWLNGGPGCSSFTGL.

From the protein 3D representations in Fig. 2, the amino acid Euclidean distances matrices of the graphs can be computed. The EDM represents the relative distance among nodes (amino acid residues) in the 3D pseudo-folded representation of the protein sequence. Afterwards, similarity between such graphs can be assessed by performing a Euclidean distances count in such a way that pairs of nodes (amino acid residues) at certain discrete distances are counted. Amino acid pseudo-folding 3D distances count descriptors are then computed using the following equation:

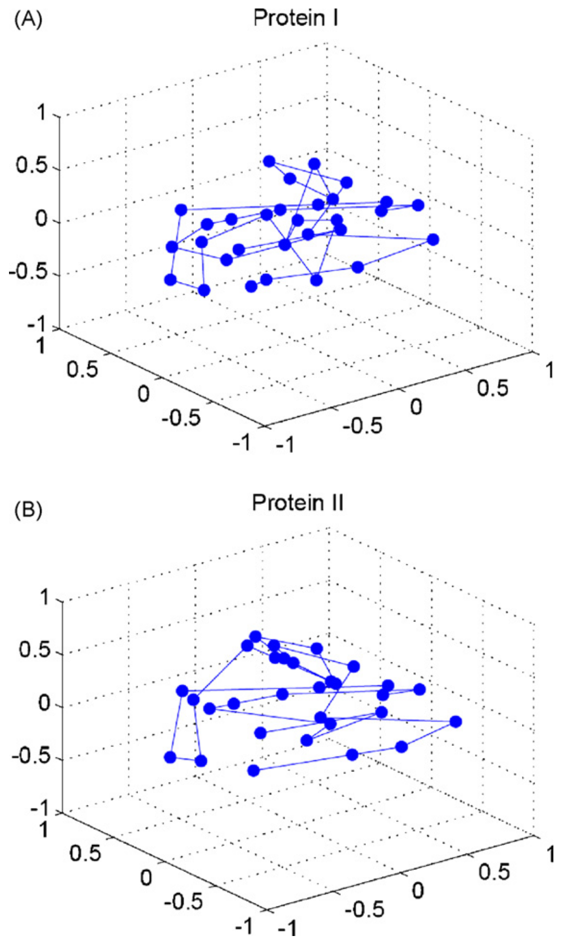

Fig. 2. 3D pseudo-folding graph representations of example proteins according to ''half-step'' method. Protein I: WTFESRNDPAKDPVILWLNGGPGCSSLTGL (A). Protein II: WFFESRNDPANDPIILWLNGGPGCSSFTGL (B).

$$\mathrm{AAp3DC}_l = \frac{1}{L}\sum_i \delta_{ij} \tag{1}$$

where L is the length of the protein sequence used for normalizing according to the size of the sequence and $\delta(l, d_{ij})$ is a Dirac-delta function defined as:

$$\delta(l, s, d_{ij}) = \begin{bmatrix} 1 & \text{if } l - \frac{s}{2} < d_{ij} \leq l + \frac{s}{2}, \\ 0 & \text{otherwise} \end{bmatrix} \tag{2}$$

where the d_{ij} is the Euclidean distance between amino acid residues i and j in the 3D protein graph, l and s are the Euclidean distance and the step used for the distance count, respectively.

Node pair summations were carried out from an initial distance l of 0.05–1.8 units at distance steps s of 0.05 units, resulting in a total of 35 AAp3DC descriptors computed for discriminating among VKC sequences. Computational codes for protein sequence 3D graph generation and AAp3DC descriptors calculation were written in Matlab environment [28] and M-file is available from the authors upon request. Before SVM optimization by GA, calculated variables were scaled to zero mean and unit variance.

2.3. Support vector machines

The SVM, a new machine learning method, has been used for many kinds of pattern recognition problems. Since excellent introductions to SVM appear in Refs. [29–31] only the main idea of SVM applied to pattern classification problems is stated here. Firstly, the input vectors are mapped into one feature space (possible with a higher dimension). Secondly, a hyperplane which can separate two classes is constructed within this feature space. Only relatively low-dimensional vectors in the input space and dot products in the feature space will involve a mapping function. SVM was designed to minimize structural risk whereas previous techniques were usually based on minimization of empirical risk. It uses a set of models ordered in terms of their complexities. The complexity is generally given by the number of free parameters. Model selection by structural risk minimization then corresponds to finding the model simplest in terms of order and best in terms of empirical error on the data (bias-variance trade-off). So SVM is usually less vulnerable to the overfitting problem, so it can deal with a large number of features.

The mapping into the feature space is performed by a kernel function. There are several parameters in the SVM, including the kernel function and regularization parameter. The kernel function and its specific parameters, together with regularization parameter, cannot be set from the optimization problem but have to be tuned by the user. These can be optimized by the use of Vapnik–Chervonenkis bounds, crossvalidation, an independent optimization set, or Bayesian learning. In this paper, we firstly tried a linear kernel that yields very poor results.

Afterwards, a radial basic function (RBF) kernel was used yielding an improved nonlinear predictor. A crossvalidation was implemented for setting the optimized values of the two parameters: the regularization parameter and the width of the RBF kernel. For implementing the SVM it used the toolbox LIBSVM for Matlab by Chang and Lin [32] that can be downloaded from: http://www.csie.ntu.edu.tw/cjlin/libsvm/.

2.4. Genetic algorithm feature selection and hyperparameter optimization

The use of SVM approach for solving classification and function mapping problems in structure–property/activity relationship studies has been growing very rapidly in the last years [33]. However, choosing the adequate descriptors for predictor training in such studies is difficult because there are no absolute rules that govern this choice. Recently, evolutionary algorithms and specifically genetic algorithms have been used for variable selection problems [16–18,34,35]. Since 35 AAp3DC descriptors were available for classification analysis and only a subset of them is statistically significant in terms of correlation with electrophysiological variables, deriving an optimal classification model through variable selection needs to be addressed.

GAs are governed by biological evolution rules [36]. They are stochastic optimization methods that have been inspired by evolutionary principles. The distinctive aspect of a GA is that it investigates many possible solutions simultaneously, each of which explores different regions in parameter space [37]. In the case where the number of features to be selected can be known beforehand, as in our study, a binary encoding in the prior way due to the much smaller search space of size would be inefficient. Therefore in this case it is reasonable to switch to a decimal encoding indicating the number of the feature which is selected. Of course one has to make sure that each is unique in the code [38]. The first step is to create a population of N individuals. Each individual encodes the same number of randomly chosen descriptors. The fitness of each individual in this generation is determined. In the second step, a fraction of children of the next generation is produced by crossover (crossover children) and the rest by mutation (mutation children) from the parents on the basis of their scaled fitness scores. The new offspring contains characteristics from two or one of its parents.

Misclassification percent of threefold-out (TFO) cross-validation ($\mathrm{MCP_{TFO}}$) was used as individual fitness or cost function. The first step is to create a gene pool (population of SVM classifiers) of N individuals. Each individual encodes the same number of descriptors; the descriptors are randomly chosen from a common data matrix, and in a way such that: (1) no two individuals can have exactly the same set of descriptors and (2) all descriptors in a given individual must be different. The fitness of each individual in this generation is determined by the $\mathrm{MCP_{TFO}}$ of the model and scaled using a scaling function. A top scaling fitness function scaled a top fraction of the individuals in a population equally; these individuals have the same probability to be reproduced while the rest are assigned the value 0.

The next step, a fraction of children of the next generation is produced by crossover (crossover children) and the rest by mutation (mutation children) from the parents. Sexual and asexual reproductions take place so that the new offspring contains characteristics from two or one of its parents. In a sexual reproduction two individuals are selected probabilistically on the basis of their scaled fitness scores and serve as parents. Next, in a crossover each parent contributes a random selection of half of its descriptor set and a child is constructed by combining these two halves of "genetic code". Finally, the rest of the individuals in the new generation are obtained by asexual reproduction when parents selected randomly are subjected to a random mutation in one of its genes, i.e., one descriptor is replaced by another. We also included elitism which protects the fittest individual in any given generation from crossover or mutation during reproduction. The genetic content of this individual simply moves on to the next generation intact. This selection, crossover and mutation process is repeated until all of the N parents in the population are replaced by their children. The fitness score of each member of this new generation is again evaluated, and the reproductive cycle is continued until 90% of the generations showed the same target fitness score [34].

GA was also used for the optimization of kernel regularization parameter C and width of an RBF kernel σ^2 of SVM as Fröhlich et al. [38], suggested. We can simply concatenate a representation of the parameter to our existing chromosome. That means we are trying to select an optimal feature subset and an optimal C at the same time. This is reasonable, because the choice of the parameter is influenced by the feature subset taken into account and vice versa. Usually it is not necessary to consider any arbitrary value of but only certain discrete values with the form: $n \times 10^k$, where $n = 1, \ldots, 9$ and $k = -3, \ldots, 4$. So, this values can be calculated by randomly generating n and k values as integers between $(1, \ldots, 9)$ and $(-3, \ldots, 4)$, respectively. In a similar way we used GA to optimize the width of an RBF kernel but in this case n and k values were integers between $(1, \ldots, 9)$ and $(-2, \ldots, 1)$. Then, our chromosome was concatenate with another gene with discrete values in the interval (0.001–90,000) for encoding the C parameter and similarly the width of the RBF kernel was encoded in a gene containing discrete values ranging in the interval (0.01–90).

The GA implemented in this paper is a version of a previous report of our group [35] but incorporating SVM hyperparameter optimization. GlibSVM toolbox for Matlab [39] was programmed within the Matlab environment using genetic algorithm [40] and libSVM [32] toolboxes.

2.5. Model's validation

TFO crossvalidation accounted for model predicted power. Three data subsets were created, in the crossvalidation process two subsets are used for training the models and the rest subset is then predicted. This process is repeated until all the subsets have been predicted. The efficiency of the SVM predictor for VKC classification was evaluated by the set of statistics listed below.

The overall accuracy is:

$$Q^2 = \frac{p}{N} \tag{3}$$

where p is the total number of correct predicted mutations and N is the total number of mutations.

The correlation coefficient C is defined as follows:

$$\mathrm{Co}(s) = \frac{[p(s)n(s) - u(s)o(s)]}{D} \tag{4}$$

where D is the normalization factor:

$$D = [(p(s) + u(s))(p(s) + o(s))(n(s) + u(s))(n(s) + o(s))]^{1/2} \tag{5}$$

for each class s (+ and − for positive (or zero) and negative signs of V_t, $V_{a_{50}}$ and $V_{h_{50}}$ values, respectively); $p(s)$ and $n(s)$ are the number of correct predictions and correctly rejected assignments, respectively, and $u(s)$ and $o(s)$ are the number or under- and over-predictions.

The coverage for each discriminant structure s is evaluated as:

$$Q_S = \frac{p(s)}{p(s) + u(s)} \tag{6}$$

The accuracy for s is computed as:

$$P_S = \frac{p(s)}{p(s) + o(s)} \tag{7}$$

where $p(s)$ and $u(s)$ are the same as in Eq. (4).

3. Results and discussion

3D graph representation of proteins attempts to discriminate among protein sequences according to a differentiated 3D spatial distribution of residues in a tri-dimensional map. Thus, the one-dimensional space of protein primary structure is then translated to a tri-dimensional one. Consequently, sequence information is converted into a more easily readable format in order to compute some descriptors for statistical pattern recognition studies. After 3D pseudo-folding graph representation of VKC full sequences, we applied graph similarity measurements for obtaining a feature data matrix for SVM training. Hyperparameters and optimum training AAp3DC descriptors for SVMs were optimized by GA. The procedure yielded optimum subsets of AAp3DC descriptors calculated over the 3D graph representations of the VKC full sequences (see Section 2.2) for each electrophysiological variable.

Action potentials are triggered when an initial depolarization reaches the threshold. VKCs are activated as a result of the transmembrane movement of charge carriers located in the voltage-sensing domain. This 'gating current' corresponds to the translocation of the equivalent of ~13 elementary charges per channel across the membrane. The S4 helices, bearing a regular array of positively charged amino acids, are the principal structural elements responsible for voltage sensing. It is generally accepted that the first four arginines of S4 (R117–R126 in S6) account for the gating current, moving toward the extracellular solvent upon channel activation in response to membrane depolarization [40].

Due to complexity of the electrophysiological variables, we aimed to predict the signs instead of its actual values for achieving reliable results. A prior attempt to model the actual values yielded very poor results. However, by characterizing the signs of the electrophysiological variables studied a useful general description of the action potential landscape of the VKC can be depicted. In this sense, electrophysiological variables were classified into class: (−) having negative and (+) having positive or zero values, respectively. The classification schemes differentiate VKC sequences according to the sign of the values of three electrophysiological variables V_t, $V_{a_{50}}$ and $V_{h_{50}}$. Since the classification training dataset is unbalanced (see Section 2.1), higher penalties for the misclassification of less represented classes were established inside the SVM framework in order to avoid classifiers biased to the most represented classes.

GA-optimizing SVMs were implemented. In a first attempt, we implemented the simpler linear kernel but the highest crossvalidation accuracy was not higher than 70% for the classification of VKCs according to any electrophysiological variable. Afterwards, nonlinearity was accessed by using a RBF kernel inside the SVM framework. As was point out in Section 2.4, the GA algorithm was also used for optimizing the hyperparameters, the kernel regularization parameter C and the width of an RBF kernel σ^2. Nonlinear subspace in the dataset was searched varying problem dimension from 2 to 6. Fig. 3 depicts the behaviours of the minimum $\mathrm{MCP_{FFO}}$ values yielded throughout the GA search versus the number of SVM inputs.

Modelling of V_t yielded an optimum five-input SVM. Nonlinear subspace in the dataset was searched varying number of variables from 2 to 6. Fig. 3A depicts behaviour of $\mathrm{MCP_{TFO}}$ versus number of inputs in the SVM model for the classification of VKCs according to the electrophysiological variable V_t. The predictive powers of the models increase with the increment of inputs until a minimum $\mathrm{MCP_{TFO}} = 19.6\%$ is reached for five inputs in the classifier. Table 1 shows input AAp3DC descriptors, hyperparameters and crossvalidation statistics of the optimum AAp3DC-SVM model for the activation electrophysiological variable V_t (model AAp3DC-SVM 1). $\mathrm{AAp3DC_{0.25}}$, $\mathrm{AAp3DC_{0.45}}$, $\mathrm{AAp3DC_{1.2}}$, $\mathrm{AAp3DC_{1.3}}$ and $\mathrm{AAp3DC_{1.8}}$ are sequence-length normalized counts of amino acids at distance ranges 0.225–0.275, 0.425–0.475, 1.175–1.225, 1.275–1.325 and 1.775–1.825 in the pseudo-folded proteins at the "magic sphere". The overall accuracy of the model was about 80% but the VKC sequences with $V_t < 0$ (class V_t−) were recognized with lower accuracy about 73% in comparison to 86% of the VKC sequences with $V_t = 0$ (class V_t+). Despite we used a higher penalty for class V_t− misclassification, the nature of the data of 143 VKC with reported V_t values provoked that predictor over-classified class V_t+ in detriment of class V_t−. However, the prediction accuracies yielded by the model over 85 and 70% for both negative and zero V_t values are adequate. Recognition of VKC proteins belonging to V_t− and V_t+ classes discriminated between ion channel proteins that need a negative voltage range

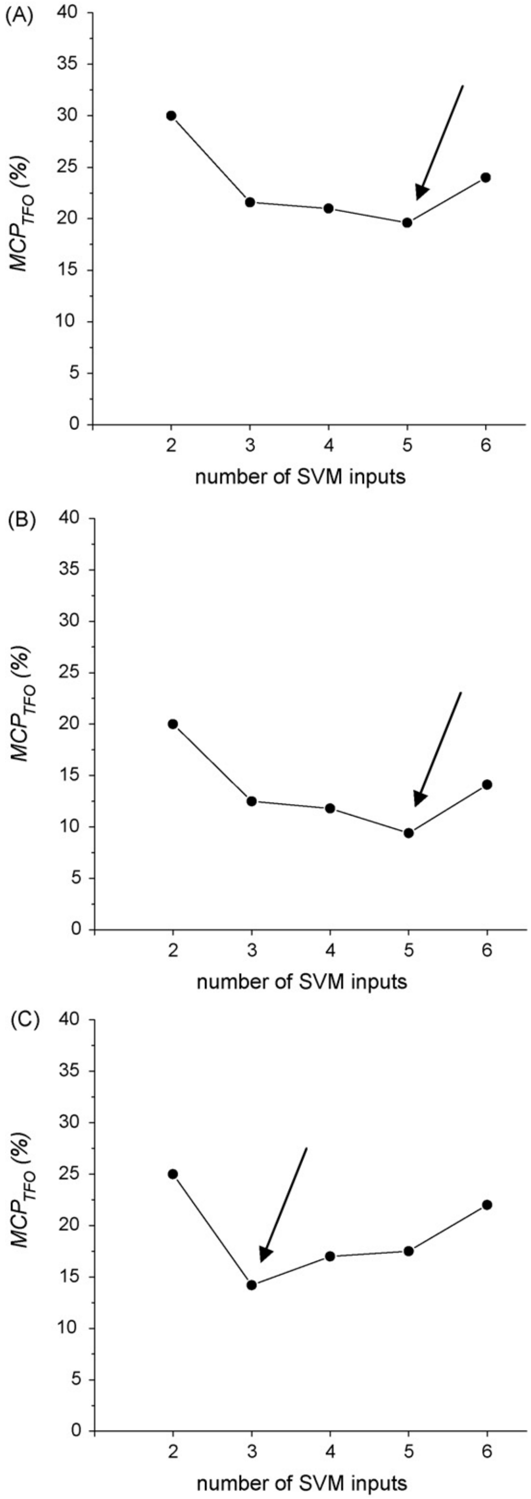

Fig. 3. Plots of $\mathrm{MCP_{TFO}}$ throughout the GA–SVM search vs. the number of inputs in the models for VKC discrimination according to three electrophysiological classification schemes: V_{t} (A), $V_{\mathrm{a}_{50}}$ (B) and $V_{\mathrm{h}_{50}}$ (C) values. Arrows point out the optimum number of inputs.

for initiation of the ion current through the ion pore and proteins which ion current passing through the pore at null voltage value.

By inspection of the optimum variables in Table 1 it can be observed that they gather information from the residue distributions at short, middle and large distances inside the ‘‘magic sphere’’ in which VKC proteins are 3D pseudo-folded for descriptor calculation. This fact suggested that optimum SVM recognized pseudo-folding features of the sequences corresponding to residues abundance at the ‘‘pseudo-core’’ and ‘‘pseudo-surface’’ of the 3D pseudo-folded proteins that are relevant for discriminating between VKC proteins with negative and zero V_{t} values.

Another electrophysiological variable related to the activation phase of VKC was $V_{\mathrm{a}_{50}}$. Similarly to V_{t} variable, two VKC classes to model were defined according to $V_{\mathrm{a}_{50}}$ values, but $V_{\mathrm{a}}+$ corresponds to VKC proteins with $V_{\mathrm{a}_{50}} > 0$. SVM models were found varying the number of inputs from 2 to 6. Fig. 3B depicts plot of $\mathrm{MCP_{TFO}}$ versus number of inputs in SVM models for the classification of VKC sequence according to sign of $V_{\mathrm{a}_{50}}$ values. The accuracy of prediction was increased with the increment of SVM inputs till a minimum $\mathrm{MCP_{TFO}} = 9.41\%$ was achieved by a model with five AAp3DC descriptors. Table 2 shows optimum inputs, hyperparameters and training and crossvalidation statistics of optimum AAp3DC-SVM predictor for $V_{\mathrm{a}_{50}}$ signs classification (model AAp3DC-SVM 2). $\mathrm{AAp3DC_{0.40}}$, $\mathrm{AAp3DC_{0.75}}$, $\mathrm{AAp3DC_{1.05}}$, $\mathrm{AAp3DC_{1.75}}$ and $\mathrm{AAp3DC_{1.8}}$ are sequence-length normalized counts of amino acids at distance ranges 0.375–0.425, 0.725–0.775, 1.025–1.075, 1.7250–1.775 and 1.775–1.825 in the pseudo-folded proteins at the ‘‘magic sphere’’.

As can be observed in Table 2, high overall prediction accuracy about 90% was achieved in the TFO crossvalidation test. However, in this case, the classifier was more accurate predicting the class $V_{\mathrm{a}}-$ with $Q(-) = 0.955$, which is higher than $Q(+) = 0.800$, the accuracy for predicting class $V_{\mathrm{a}}+$. Despite this accuracy differences for each VKC class, more than 80% of both classes are correctly recognized in the crossvalidation experiments that means a very robust behaviour of this SVM classifier. Similarly to classifier AAp3DC-SVM 1, the optimum model for $V_{\mathrm{a}_{50}}$ classification included contributions of AAp3DC descriptors calculated at short, middle and distance ranges at the ‘‘magic sphere’’. In this sense, we found that the discrimination between VKC proteins according to the signs of the $V_{\mathrm{a}_{50}}$ values also has contributions of the residues distributions at the ‘‘pseudo-core’’ and ‘‘pseudo-surface’’ of the pseudo-folded protein.

Finally our approach was used to discriminate between VKC proteins according to the signs of $V_{\mathrm{h}_{50}}$, an inactivation electrophysiological variable. For this electrophysiological variable, classes $V_{\mathrm{h}}-$ and $V_{\mathrm{h}}+$ corresponded to negative and zero $V_{\mathrm{h}_{50}}$ values. SVM models were optimized varying the number of inputs from 2 to 6. Fig. 3C depicts the variation of $\mathrm{MCP_{TFO}}$ versus number of AAp3DC descriptors for the electrophysiological variable $V_{\mathrm{h}_{50}}$. There is a minimum value of $\mathrm{MCP_{TFO}} = 14.2\%$ for a models with three inputs. Differently to the other two electrophysiological variables, this optimum classifier AAp3DC-SVM 3 was found with three variables which appear in Table 3 with the optimum hyperparameters and the training and crossvalidation statistics. Variables $\mathrm{AAp3DC_{0.40}}$, $\mathrm{AAp3DC_{1.25}}$ and $\mathrm{AAp3DC_{1.7}}$ are sequence-length normalized counts of amino acids at distance ranges

Table 1
Hyperparameters and statistics of training and crossvalidation of optimum model AAp3DC-SVM 1 for the classification of VKCs according to electrophysiological variable V_t

SVM inputs: $AAp3DC_{0.25}$, $AAp3DC_{0.45}$, $AAp3DC_{1.2}$, $AAp3DC_{1.3}$, $AAp3DC_{1.8}$

	C	σ^2	Q^2	$P(+)$	$P(-)$	$Q(+)$	$Q(-)$	Co
Training	50	0.167	0.979	0.988	0.967	0.976	0.983	0.957
Crossvalidation			0.804	0.818	0.782	0.857	0.729	0.593

+ and −: the indexes account for classes V_t+ and V_t-.

Table 2
Hyperparameters and statistics of training and crossvalidation of optimum model AAp3DC-SVM 2 for the classification of VKCs according to electrophysiological variable $V_{a_{50}}$

SVM inputs: $AAp3DC_{0.40}$, $AAp3DC_{0.75}$, $AAp3DC_{1.05}$, $AAp3DC_{1.75}$, $AAp3DC_{1.8}$

	C	σ^2	Q^2	$P(+)$	$P(-)$	$Q(+)$	$Q(-)$	Co
Training	4000	0.5	1.00	1.00	1.00	1.00	1.00	1.00
Crossvalidation			0.906	0.889	0.913	0.800	0.955	0.778

+ and −: the indexes account for classes V_a+ and V_a-.

Table 3
Hyperparameters and statistics of training and crossvalidation of optimum model AAp3DC-SVM 3 for the classification of VKCs according to electrophysiological variable $V_{h_{50}}$

SVM inputs: $AAp3DC_{0.40}$, $AAp3DC_{0.75}$, $AAp3DC_{1.05}$, $AAp3DC_{1.75}$, $AAp3DC_{1.8}$

	C	σ^2	Q^2	$P(+)$	$P(-)$	$Q(+)$	$Q(-)$	Co
Training	6	0.025	0.967	0.965	0.971	0.988	0.917	0.920
Crossvalidation			0.858	0.868	0.828	0.941	0.667	0.778

+ and −: the indexes account for classes V_h+ and V_h-.

0.375–0.425, 1.225–1.275 and 1.675–1.725 in the pseudo-folded proteins at the "magic sphere". A maximum overall crossvalidation accuracy about 86% was achieved, but in this cases the accuracies for each class V_h+ and V_h- significantly differed. About a 94% of the VKC proteins belonging to the most represented class V_h+ were correctly recognized in crossvalidation but only a 67% of the VKC sequences included in class V_h- were correctly classified in the same experiment. Despite we used different penalty combinations for SVM training, the optimum predictor AAp3DC-SVM 3 has about 30% higher probability to recognize any VKC protein as an ion channel with $V_{h_{50}} = 0$.

In order to compare the performance of the GA–SVM methodology with other classifiers, a comparative analysis of the classification results for the VKC protein according to the three classification schemes was carried out using the Weka software [42]. A wrapper algorithm with GA variable selection was used for model optimization. Results of the crossvalidation results according to nine different classifiers appear in Table 4. As it can be observed all the classifiers tested underperformed the GA–SVM models reported for the three electrophysiological variables. The maximum accuracies were 71% (Bayes Net and RBF Network), 69% (Conjunctive Rule, ADTree and Naïve Bayes) and 77% (Random Tree) for V_t, $V_{a_{50}}$ and $V_{h_{50}}$ classification schemes, respectively. These values are lower than the optimum accuracies of 80, 91 and 86% yielded by the GA–SVM approach for each classification scheme, respectively.

The electrophysiological variable $V_{a_{50}}$ of VKCs had been previously modelled using machine learning techniques [13]. $V_{a_{50}}$ were predicted based on its amino acid sequence by

Table 4
Overall accuracies of crossvalidation of different classifiers from Weka software [42] for the classification of VKCs according to the three electrophysiological variables V_t, $V_{a_{50}}$, $V_{h_{50}}$

Classifier	Q^2		
	V_t	$V_{a_{50}}$	$V_{h_{50}}$
PART	0.64	0.54	0.73
Conjunctive Rule	0.66	0.69	0.70
ADTree	0.70	0.69	0.78
J48	0.69	0.63	0.72
Random Tree	0.70	0.63	0.77
AdaBoostM1	0.68	0.57	0.76
Bayes Net	0.71	0.59	0.72
Naïve Bayes	0.68	0.69	0.63
RBF Network	0.71	0.64	0.72

different learning algorithms, combined in various implementations. Predictors were explicitly trained with sequence residues and the best result was obtained with a *k*-nearest neighbour classifier combined with a wrapper algorithm for feature selection. The method identified some residues that are suggested to be involved in the voltage sensitive conformation changes and therefore are good target candidates for mutagenesis analysis [41].

Although our work does not provide information about relevant residue sites or conformations for the electrophysiological function of VKCs, it predicts a triple characterization of VKC proteins according to relevant electrophysiological variables. 2D and 3D graph depicting of protein sequences has been successfully used for protein similarity and classification studies. In this regard, similarities and dissimilarities for the protein sequences of nine nerve genes were reported based on the Euclidean distances of the nine different nerve gene calculated on 3D protein representation [26]. Recently, our group successfully reported SVM classifiers for the conformational stability of protein mutants trained with descriptors calculated from 2D to 3D representation of protein sequences [19,20]. Despite the artificial and artefact-nature of the 3D coordinates values of protein residues obtained from the 3D pseudo-folding representations of protein sequence, the modelling results for the three electrophysiological classification schemes can be concluded that contributions of the residues distributions at the ‘‘pseudo-core’’ and ‘‘pseudo-surface’’ of the pseudo-folded protein to VKCs classification is already a general finding.

4. Conclusions

Structure–function relationship of VKCs has been mainly carried out by site-directed mutagenesis but the prohibitively time-consuming and costly implementation of this experimental technique has prompted the application of computational modelling tools to VKCs study. Protein primary structure-based methods are less computational intense and do not require X-ray crystal structure of proteins for implementation. However, encoding protein sequences are needed in order to apply statistical pattern recognition techniques for protein prediction studies. A novel 3D pseudo-folding graph representation of protein sequence allows calculating AAp3DC descriptors, novel protein indexes. SVMs trained with subsets of AAp3DC descriptors discriminated between VKC sequences according to three classification schemes based on the signs of three electrophysiological variables. In crossvalidation test, the best models optimized by GA well recognized about 80, 85 and 90% of the binary VKC classes depending of the signs of V_{t}, $V_{\mathrm{a}_{50}}$ and $V_{\mathrm{h}_{50}}$ values, respectively.

3D pseudo-folding graph representation of protein sequence in combination with SVMs could be very useful in protein prediction studies. Despite the disadvantage of requiring some previous experimental data for generating a training set, our prediction technique is an alternative approach for proteins with known sequences but unsolved X-ray structures.

Acknowledgement

Financial supports of this research by Cuban Ministerio de Ciencia, Tecnología y Medio Ambiente (CITMA) through a grant to M. Fernandez (Grant no. 20104102).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2008.01.001.

References

[1] G. Yellen, The voltage-gated potassium channels and their relatives, Nature 419 (2002) 35–42.

[2] Y. Jiang, A. Lee, J. Chen, V. Ruta, M. Cadene, B.T. Chait, R. MacKinnon, X-ray structure of a voltage-dependent K^+ channel, Nature 423 (2003) 33–41.

[3] H.P. Larsson, O.S. Baker, D.S. Dhillon, E.Y. Isacoff, Transmembrane movement of the shaker K^+ channel S4, Neuron 16 (1996) 387–397.

[4] Z. Sands, A. Grottesi, M.S. Sansom, Voltage-gated ion channels, Curr. Biol. 15 (2005) 44–47.

[5] T.J. Jentsch, Neuronal KCNQ potassium channels: physiology and role in disease, Nat. Rev. Neurosci. 1 (2000) 21–30.

[6] S. Comu, M. Giuliani, V. Narayanan, Episodic ataxia and myokymia syndrome: a new mutation of potassium channel gene Kv1.1, Ann. Neurol. 40 (1996) 684–687.

[7] M. Abdul, N. Hoosein, Voltage-gated potassium ion channels in colon cancer, Oncol. Rep. 9 (2002) 961–964.

[8] P.A. Koni, R. Khanna, M.C. Chang, M.D. Tang, L.K. Kaczmarek, L.C. Schlichter, R.A. Flavella, Compensatory anion currents in Kv1.3 channel-deficient thymocytes, J. Biol. Chem. 278 (2003) 39443–39451.

[9] E.C. Cooper, Potassium channels: how genetic studies of epileptic syndromes open paths to new therapeutic targets and drugs, Epilepsia 42 (2001) 49–54.

[10] J.W. Ford, E.B. Stevens, J.M. Treherne, J. Packer, M. Bushfield, Potassium channels: gene family, therapeutic relevance, high-throughput screening technologies and drug discovery, Prog. Drug Res. 58 (2002) 133–168.

[11] L. Heginbotham, T. Abramson, R. MacKinnon, A functional connection between the pores of distantly related ion channels as revealed by mutant K^+ channels, Science 258 (1992) 1152–1155.

[12] C. Miller, 1990: annus mirabilis of potassium channels, Science 252 (1991) 1092–1096.

[13] B. Li, W.J. Gallin, Computational identification of residues that modulate voltage sensitivity of voltage-gated potassium channels, BMC Struct. Biol. 5 (5) (2005) 16.

[14] R. Ramos de Armas, H. González-Díaz, R. Molina, E. Uriarte, Markovian backbone negentropies: molecular descriptors for protein research. I. Predicting protein stability in arc repressor mutants, Proteins 56 (2004) 715–723.

[15] Y. Marrero-Ponce, R. Medina-Marrero, J.A. Castillo-Garit, V. Romero-Zaldivar, F. Torrens, E.A. Castro, Protein linear indices of the ‘Macromolecular Pseudograph α-Carbon Atom Adjacency Matrix’ in bioinformatics. Part 1. Prediction of protein stability effects of a complete set of alanine substitutions in arc repressor, Bioorg. Med. Chem. 13 (2005) 3003–3015.

[16] J. Caballero, L. Fernández, J.I. Abreu, M. Fernández, Amino acid sequence autocorrelation vectors and ensembles of Bayesian-regularized genetic neural networks for prediction of conformational stability of human lysozyme mutants, J. Chem. Inform. Model 46 (2006) 1255–1268.

[17] L. Fernández, J. Caballero, J.I. Abreu, M. Fernández, Amino acid sequence autocorrelation vectors and Bayesian-regularized genetic neural networks for modeling protein conformational stability: gene v protein mutants, Proteins 67 (2007) 834–852.

[18] J. Caballero, L. Fernández, M. Gariga, J.I. Abreu, S. Collina, M. Fernández, Proteometric study of ghrelin receptor function variations upon

mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines, J. Mol. Graph. Model. 26 (2007) 166–178.

[19] M. Fernández, J. Caballero, L. Fernández, J.I. Abreu, G. Acosta, Classification of conformational stability of protein mutants from 3D pseudo-folding graph representation of protein sequences using support vector machines, Proteins 70 (2008) 167–175.

[20] M. Fernández, J. Caballero, L. Fernández, J.I. Abreu, G. Acosta, Classification of conformational stability of protein mutants from 2D graph representation of protein sequences using support vector machines, Mol. Simulat. 33 (2007) 889–896.

[21] B. Li, W.J. Gallin, VKCDB: voltage-gated potassium channel database, BMC Struct. Biol. 5 (2004) 3.

[22] M. Randić, D. Butina, J. Zupan, Novel 2-D graphical representation of proteins, Chem. Phys. Lett. 419 (2006) 528–532.

[23] C. Raychaudhury, A. Nandy, Indexing scheme and similarity measures for macromolecular sequences, J. Chem. Inform. Comput. Sci. 39 (1999) 243–247.

[24] G. Agüero-Chapin, H. González-Díaz, R. Molina, J. Varona-Santos, E. Uriarte, Y. González-Díaz, Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L, FEBS Lett. 580 (2006) 723–730.

[25] H.I. Jeffrey, Chaos game representation of gene structure, Nucleic Acid Res. 18 (1990) 2163–2170.

[26] F. Bai, T. Wang, On graphical and numerical representation of protein sequences, J. Biomol. Struct. Dyn. 23 (2006) 537–545.

[27] (a) H. González-Díaz, S. Vilar, L. Santana, E. Uriarte, Medicinal chemistry and bioinformatics—current trends in drugs discovery with networks topological indices, Curr. Top. Med. Chem. 7 (2007) 1025–1039;
(b) K.C. Chou, Prediction of protein subcellular locations by incorporating quasi-sequence-order effect, Biochem. Biophys. Res. Commun. 278 (2000) 477–483;
(c) H. Lin, Q.Z. Li, Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components, J. Comput. Chem. 28 (2007) 1463–1466;
(d) R. Agarwala, S. Batzoglou, V. Dancik, S.E. Decatur, S. Hannenhalli, M. Farach, S. Muthukrishnan, S. Skiena, Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model, J. Comput. Biol. 4 (1997) 275–296;
(e) B. Berger, T. Leighton, Protein folding in the hydrophobic-hydrophilic (HP) model is NP complete, J. Comput. Biol. 5 (1998) 27–40;
(f) M. Chen, W.Q. Huang, A branch and bound algorithm for the protein folding problem in the HP lattice model, Geno. Prot. Bioinfo. 3 (2005) 225–230;
(g) A. Gupta, J. Manuch, L. Stacho, Structure-approximating inverse protein folding problem in the 2D HP model, J. Comput. Biol. 12 (2005) 1328–1345;
(h) H. González-Díaz, Y. Pérez-Castillo, G. Podda, E. Uriarte, Comparison of stable/nonstable protein mutants classification models based on 3D and topological indices, J. Comput. Chem. 28 (2007) 1990–1995;
(i) H. González-Díaz, L. Saiz-Urra, R. Molina, L. Santana, E. Uriarte, A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions, J. Proteome Res. 6 (2007) 904–908;
(j) H. González-Díaz, L. Saiz-Urra, R. Molina, Y. Gonzalez-Diaz, A. Sanchez-Gonzalez, Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments, J. Comput. Chem. 28 (2007) 1042–1048;
(k) K.-C. Chou, H.B. Shen, Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites., J. Proteome Res. 6 (2007) 1728–1734;
(l) K.C. Chou, Y.D. Cai, Predicting protein-protein interactions from sequences in a hybridization space, J. Proteome Res. 5 (2006) 316–322;
(m) Z. Bajzer, M. Randic, D. Plavšic, S.C. Basak, Novel map descriptors for characterization of toxic effects in proteomics maps, J. Mol. Graph. Model. 22 (2003) 1–9;
(n) M. Randić, N. Lers, D. Vukicević, D. Plavsić, S.C. Basak, B.D. Gute, Canonical labeling of proteome maps, J. Proteome Res. 4 (2005) 1347–1352;
(o) P. Ping, T.M. Vondriska, C.J. Creighton, T.K. Gandhi, Z. Yang, R. Menon, M.-S. Kwon, G. Drwal, M. Kellman, S. Peri, S. Suresh, M. Gronborg, H. Molina, R. Chaerkady, B. Rekha, B. Muthusamy, A.S. Shet, R.E. Gerszten, H. Wu, M. Raftery, V. Wasinger, P. Schulze-Knappe, S.M. Hanash, Y.-K. Paik, W.S. Hancock, D.J. States, G.S. Omenn, A. Pandey, A functional annotation of subproteomes in human plasma, Proteomics 5 (2005) 3506–3519;
(p) M. Randić, J. Zupan, D. Vikić-Topić, On representation of proteins by star-like graphs, J. Mol. Graph. Model. 26 (2007) 290–305;
(q) M. Randic, G. Krilov, Characterization of 3D sequences of proteins, Chem. Phys. Lett. 272 (1997) 115–119.

[28] MATLAB 7.0. program, available from The Mathworks Inc., Natick, MA, http://www.mathworks.com.

[29] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.

[30] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (1998) 1–47.

[31] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[32] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[33] (a) W. Lua, N. Donga, G. Náray-Szabó, Predicting anti-HIV-1 activities of hept-analog compounds by using support vector classification, QSAR Comb. Sci. 24 (2005) 1021–1025;
(b) X. Yao, H. Liu, R. Zhang, M. Liu, Z. Hu, A. Panaye, J.P. Doucet, B. Fan, QSAR and classification study of 1,4-dihydropyridine calcium channel antagonists based on least squares support vector machines, Mol. Pharm. 2 (2005) 348–356;
(c) H. Fröhlich, J.K. Wegner, A. Zell, Towards optimal descriptor subset selection with support vector machines in classification and regression, QSAR Comb. Sci. 23 (2004) 311–318.

[34] B. Hemmateenejad, M.A. Safarpour, R. Miri, N. Nesari, Toward an optimal procedure for PC-ANN model building: prediction of the carcinogenic activity of a large set of drugs, J. Chem. Inform. Model. 45 (2005) 190–199.

[35] (a) J. Caballero, M. Garriga, M. Fernándeza, Genetic neural network modeling of the selective inhibition of the intermediate-conductance Ca^{2+}-activated K^+ channel by some triarylmethanes using topological charge indexes descriptors, J. Comput. Aided Mol. Des. 19 (2005) 771–789;
(b) J. Caballero, M. Fernández, Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks, J. Mol. Model. 12 (2006) 168–181.

[36] H. Holland, Adaption in Natural and Artificial Systems, The University of Michigan Press, Ann Arbor, MI, 1975.

[37] H.M. Cartwright, Applications of Artificial Intelligence in Chemistry, Oxford University Press, Oxford, 1993.

[38] H. Fröhlich, O. Chapelle, B. Schölkopf, Feature selection for support vector machines by means of genetic algorithms, in: Proceedings of the 15th IEEE International Conference on Tools with AI, 2003, pp. 142–148.

[39] GlibSVM Toolbox for Matlab Version 1.0, Molecular Modeling Group, University of Matanzas, 2007.

[40] The MathWorks Inc., Genetic Algorithm and Direct Search Toolbox User's Guide for use with MATLAB, The Mathworks Inc., MA, 2004.

[41] Z. Sands, A. Grottesi, M.S. Sansom, Voltage-gated ion channels, Curr. Biol. 15 (44) (2005) 7.

[42] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition, Morgan Kaufmann, San Francisco, 2005.