

PROCEEDINGS

Open Access

In silico approach to screen compounds active against parasitic nematodes of major socio-economic importance

Varun Khanna¹, Shoba Ranganathan^{1,2*}

From Asia Pacific Bioinformatics Network (APBioNet) Tenth International Conference on Bioinformatics – First ISCB Asia Joint Conference 2011 (InCoB2011/ISCB-Asia 2011)
Kuala Lumpur, Malaysia. 30 November - 2 December 2011

Abstract

Background: Infections due to parasitic nematodes are common causes of morbidity and fatality around the world especially in developing nations. At present however, there are only three major classes of drugs for treating human nematode infections. Additionally the scientific knowledge on the mechanism of action and the reason for the resistance to these drugs is poorly understood. Commercial incentives to design drugs that are endemic to developing countries are limited therefore, virtual screening in academic settings can play a vital role in discovering novel drugs useful against neglected diseases. In this study we propose to build a robust machine learning model to classify and screen compounds active against parasitic nematodes.

Results: A set of compounds active against parasitic nematodes were collated from various literature sources including PubChem while the inactive set was derived from DrugBank database. The support vector machine (SVM) algorithm was used for model development, and stratified ten-fold cross validation was used to evaluate the performance of each classifier. The best results were obtained using the radial basis function kernel. The SVM method achieved an accuracy of 81.79% on an independent test set. Using the model developed above, we were able to identify novel compounds with potential anthelmintic activity.

Conclusion: In this study, we successfully present the SVM approach for predicting compounds active against parasitic nematodes which suggests the effectiveness of computational approaches for antiparasitic drug discovery. Although, the accuracy obtained is lower than the previously reported in a similar study but we believe that our model is more robust because we intentionally employed stringent criteria to select the inactive dataset thus making it difficult for the model to classify compounds. The method presents an alternative approach to the existing traditional methods and may be useful for predicting hitherto novel anthelmintic compounds.

Background

Besides malaria, infections due to nematodes are the leading cause of ailment to human beings. In particular, parasitic flatworms (cestodes and trematodes) and roundworms (nematodes) are a major cause of considerable suffering, mainly in children. According to a report by the World Health Organization (WHO) it is estimated

that 2.9 billion people are infected with nematodes [1]. Therefore, to search for nematode specific targets is an active area under research. In Table 1, we present the list of successful biochemical targets and corresponding drug classes that are known to be active against those targets in helminths. With the availability of the completely sequenced nematode genomes, currently there is much interest to investigate drugs targeting their gene products.

At present however, only a couple of drugs are being used to control most worm infections in humans and

* Correspondence: shoba.ranganathan@mq.edu.au

¹Dept. of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, Australia

Full list of author information is available at the end of the article

Table 1 List of successful targets in helminths and corresponding drug class known to be active against those target

S.No	Target	Biochemical class	BLAST score	Drug family
1.	Nicotinic acetylcholine receptor beta 1	Ion transport	E: 7e-27 62% identity with human NACh receptor beta 2	Cholinergic Agents
2.	Glutamate-gated chloride channel	Ion transport	E: e-137 54% similarity with human glutamate receptor	Macrolides
3.	Glutathione S-transferase	Transferases transferring alkyl or aryl groups	E: 6e-47 61% identity with human Mu isoform	Isoquinolines
4.	Tubulin beta	–	E: 0 96% similarity to human tubulin beta	Benzimidazoles
5.	Gamma-aminobutyric acid receptor	Chloride channel	–	Piperazines

animals. There are only three major classes of anthelmintic drugs available in the market. Benzimidazoles are broad spectrum anthelmintics and inhibit β -tubulin resulting in impaired microtubule formation during cell division [2]. The benzimidazoles have greater affinity for tubulin in helminth cells than the tubulin found in the cells of mammals as first reported by Friedman and Plazer [3]. They found that fenbendazole was 250 times and mebendazole was 400 times more potent inhibitors of colchicine binding to *A. suum* embryonic tubulin than to mammalian tubulin and concluded that benzimidazoles clearly exhibit higher affinity to helminth tubulins. However, direct binding studies by Kohler and Bachmann [4] failed to find a significant change in benzimidazole affinity using mebendazole and intestinal *A. suum* tubulin. The authors surmised that differential pharmacokinetic behaviour of mebendazole could be responsible for the difference in drug susceptibility between host and parasite. Macrocyclic lactones form the second class of anthelmintics, interacting with a range of ion channels including glutamate-gated [5], γ -aminobutyric acid-gated [6] and acetylcholine-gated [7] chloride channels. Levamisole, pyrantel and morantel belong to the third class and bind to the nicotinic acetylcholine receptors causing muscle paralysis due to extended muscle contraction and spastic paralysis of the parasite [8]. Given the diversity in the chemical structures of these classes, predicting novel anthelmintics is a challenging task.

Nematodes infect the majority of the farm animals, and consequently, present a huge risk to livestock industry and exacerbate global food shortages. It is therefore not surprising that most of the anthelmintic drugs were originally developed to treat animal infections but were subsequently approved for human use with little or no modification. However, due to the disproportionate use of anthelmintics, currently the livestock industry is facing a very serious challenge with drug resistance in farm animals [9,10]. Furthermore, with a limited number of drugs being used, worm strains are able to develop drug resistance easily. In fact, there have also been reports of resistance for the present day anthelmintic drugs in humans [11]. Hence, there is an urgent need to discover novel safe and efficacious classes of anthelmintics with a new mode of action.

Recent efforts in anthelmintic drug discovery

An excellent review on the current anthelmintics and existing research gaps that need to be addressed in order to discover novel anthelmintic drugs are summarized recently by Keiser and Utzinger [12]. Kaminsky *et al.* [13] reported a new class of synthetic anthelmintics, amino-acetonitrile derivatives (AADs) that are active against a variety of livestock pathogenic nematode species. The authors reported that the optimized AADs were able to eliminate fourth larval stages of *H. contortus*,

T. colubriformis in sheep and *Cooperia oncophora*, *Ostertagia ostertagi* in cattle at a single oral dose of 20 mg racemate kg^{-1} . The authors surmised that a unique group of nematode specific nAChR protein from *acr-23* gene is responsible for AAD efficacy. Hu *et al.* [14] have demonstrated that the mechanism of action of a novel anthelmintic drug, tribendimidine, approved recently in P.R. China. They concluded that tribendimidine is an L-subtype nAChR agonist, similar to levamisole pyrantel. The anthelmintic properties of cyclooctadepsipeptides have also been reported recently *in vitro* and *in vivo* [15,16]. Mefloquine is an antimalarial drug and has been used successfully for past four decades to treat prophylaxis of malaria. However, recent research revealed promising antischistosomal properties of mefloquine in *Schistosoma mansoni*- and *Schistosoma japonicum*-infected mouse models [17,18]. Ponce-Marrero *et al.* [19] introduced a novel approach for *in silico* design of new anthelmintic drugs using linear discriminant analysis to obtain a quantitative model that classified anthelmintic drug-like from non-anthelmintic compounds. The developed model correctly classified 88.18% of the compounds in external test set. The model was then used for virtual screening and several compounds from Merck Index and Negwer's handbook were identified by the model as anthelmintic. Subsequently *in vivo* test were carried out to validate the predictions.

Overview of the ligand-based virtual screening methods

Antiparasitic drugs historically have been discovered by experimental screening against intact parasites, but due to the enormity of the task and availability of better computational facilities there has been a shift towards computational screening. Computational screening (also known as virtual screening) has inherent advantage over traditional and even experimental high throughput screening (HTS) due to its massive parallel processing ability; millions of compounds per week can be tested. Virtual screening (VS) has been widely used to discover new leads by computationally identifying compounds with higher probability of strong binding affinity to the target protein. Successful studies have led to the identification of molecules either resembling the native ligands of a particular target or novel compounds [20,21]. VS methods can be classified into structure-based and ligand-based approaches based on the amount of structural and bioactivity data available. If the 3D structure of the receptor is known, a structure-based VS methods that can be used is high-throughput docking [22] but where the information on the receptor is scant, ligand-based methods [23] like similarity searching and machine learning techniques are commonly used. Docking involves a complex optimization task of finding the most favourable 3D binding conformation of the ligand to the receptor molecule. Being computationally intensive,

docking is not suitable for very large virtual screening experiments. On the other hand, ligand-based methods are popular because they are computationally inexpensive and easy to use. Furthermore, the assumption that structurally similar molecules exhibit similar biological activity than dissimilar or less similar molecules is generally valid. Thus, ligand-based methods are increasingly playing an important role at the beginning of the drug discovery projects especially where little 3D information is available for the receptor. Particularly interesting are machine learning based approaches such as neural networks, genetic algorithms and support vector machines (SVM). SVM is a powerful classification technique that has found numerous applications in chemistry such as drug design, quantitative structure property prediction and chemical data mining. Many studies in the past have shown SVM to be one of the best methods for correctly classifying molecules [24-26]. Zernov *et al.* [24] used SVM and neural networks to predict the drug-likeness and agrochemical-likeness for large compound collections. They showed that for both kinds of data, SVM outperformed all neural networks under the same training conditions. Warmuth *et al.* [25] investigated a large collection of compounds to find those that bind to the target of interest in as few iterations of biochemical testing as possible. The authors compared various search strategies including maximum margin hyperplane, generated by SVM. They concluded that the strategies based on SVM clearly outperform the simpler ones. Similarly, Burbidge *et al.* [26] carried out a comparative study that involved prediction of the inhibition of dihydrofolate reductase by pyrimidines, using SVM, ANN and decision trees. They found that SVM outperformed the other methods, except in a manually capacity-controlled ANN, which required significantly longer training time. Nonetheless, ligand-based VS still remains an unproven approach in the discovery of antiparasitic medicines [27].

In this investigation, we have developed an *in silico* classification model using SVM to predict potential anthelmintic leads targeted towards parasitic nematodes. Our model has an estimated accuracy of ~82.0% for the test dataset. We have applied this model to a large public database to predict novel anthelmintic compounds and identified a set of 45 compounds, of which six are promising as potential therapeutic agents.

Methods

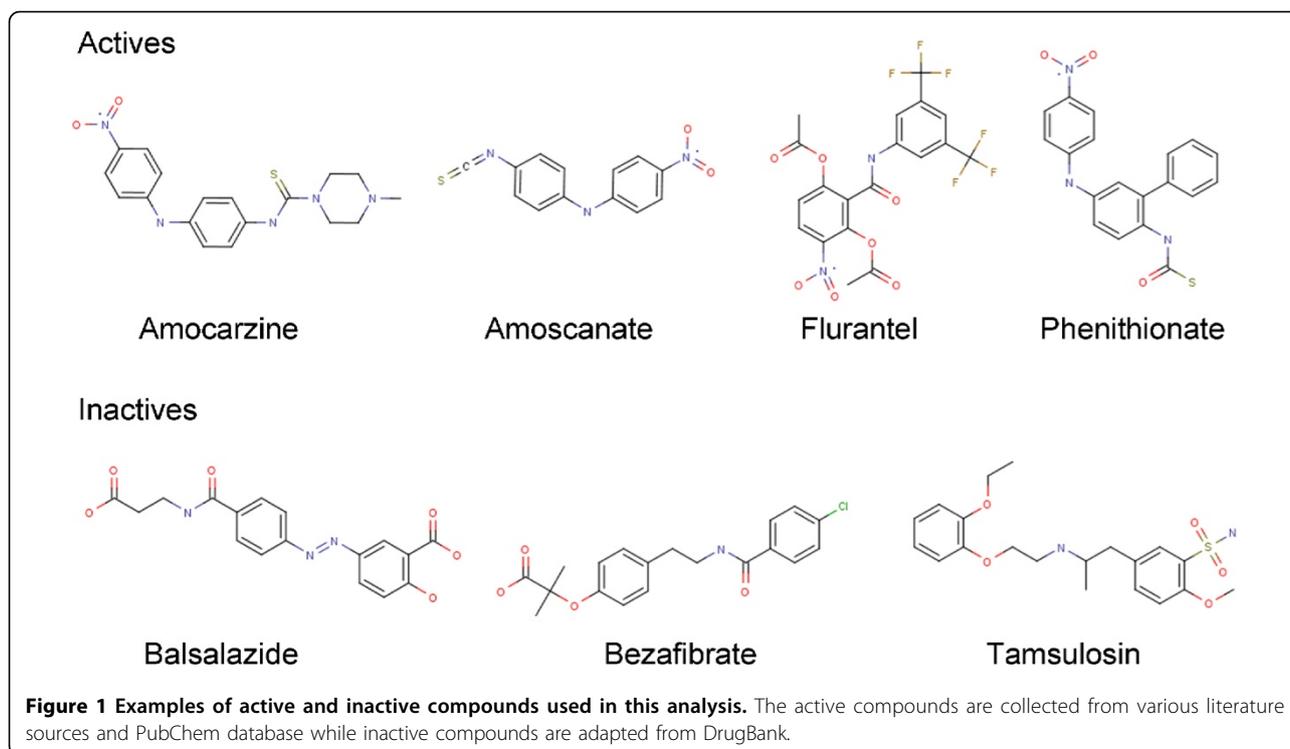
Preparation of the dataset

The quality of the data available largely determines the quality of any machine learning model [28]. Our primary dataset contains 295 unique compounds (148 actives and 147 inactives). The library of active molecules (compounds active against parasitic nematodes) was carefully collated from PubChem [29] and other literature sources [30-33]. For inactive compounds, we searched the

DrugBank [34] database for similar molecules to the ones present in the active set with a Tanimoto cut-off range from 0.25 to 0.75. As a result, compounds from various pharmacological uses (anticancer, antibacterial, sedatives, antifungal) were collected into the inactive dataset. Since no true negatives (compounds without any anthelmintic activity) are reported in the literature, inactive compounds used in this study may possess residual anthelmintic activity. In Figure 1, we present representative active and inactive compounds used in this study for developing models. Further, the primary dataset was divided into training (80%) and testing sets (20%). The sampling was carried out at random and compounds in the test set were excluded from model development. In Table 2, we present the composition of the datasets used in this study. The training dataset was used for optimizing and training the SVM classifier [35] in order to predict compounds from an unseen test set. The training dataset contains 240 compounds (126 active and 114 inactive). The test dataset on the other hand was used for evaluating the performance of the SVM method and contains 55 compounds (22 active and 33 inactive). All the training set and test set compounds are available in Additional file 1. Based on our previous study, where we reported that the ChEMBL database [36] is quite diverse, contains many drug-like and interesting compounds, therefore, we used the ChEMBL database compounds for prediction set. Currently, the database holds over 650,000 compounds with calculated physicochemical properties (log P, molecular weight, Lipinski properties) and abstracted bioactivities (binding constant, pharmacology and ADMET data). We downloaded the ChEMBL dataset in SD format. After cleaning the dataset of any inconsistencies and inorganic structures, we removed the compounds with 0.8 or greater Tanimoto similarity to the compounds in primary dataset. Then we clustered the dataset to remove similar structures. Cluster centres were selected from each cluster while singletons were retained as such. For clustering, we employed the functional class substructural fingerprint as implemented in Pipeline Pilot software [37] with the Tanimoto cut-off value 0.7. This reduced our dataset to around 300,000 compounds. Finally, we randomly selected 10,000 compounds from ChEMBL dataset for descriptor calculation and further analysis.

Defining scaffolds

In order to study the patterns in chemical compounds, it is important to decompose the molecules into fragments. There are a number of ways to fragment molecules as discussed elsewhere [38]. We describe below the specific method used in this study to obtain molecular scaffolds, where the term scaffold describes the core structure of the molecule (carbon skeleton). To obtain the carbon



skeleton of the molecule, all the heavy atoms are represented as carbon and all bonds are converted to single bonds as shown in Figure 2.

Descriptor calculation and selection

The determination of relevant features is an important step in any machine learning process [39]. Moreover, with hundreds of descriptors available it is essential to choose the best subset of descriptors because many of the descriptors are noisy and some are irrelevant to the target activity. Feature selection is the effective way to remove noisy or irrelevant descriptors and reduce the dimensionality of the feature space to avoid overfitting. This leads to simple and robust computational models with improved prediction accuracy.

There are two main approaches for feature selection in a supervised learning context. The first one is the filter approach [40]. It consists of selecting the best subset of features in an independent way, with *ad hoc* criteria. Filter methods are fast and can be easily implemented;

however, there is no guarantee that the best subset of descriptors has been selected. The second method is the wrapper approach [41] which evaluates the performance of a predetermined learning algorithm and uses it as an evaluation criterion to select the optimum subset of features.

The Molecular Operating Environment (MOE) [42] software was used for descriptor calculation. It calculates 333 descriptors, which are classified as one-dimensional (physicochemical properties), two-dimensional (topological) and three-dimensional (volume and surface area) descriptors. In Figure 3 we show the overall methodology adopted for descriptor calculation and selection. Due to the large number of descriptors available, we first filtered out constant and near constant descriptors (descriptors with <0.3 standard deviation). This resulted in the removal of 81 descriptors. Following this, we removed descriptors with a correlation coefficient greater than or equal to 0.8. The removal of correlated descriptors resulted in a set of 113 descriptors. Before performing univariate analysis, we normalized the dataset using the z-transformation. We then performed the normality test and those descriptors that passed the normality test were retained while the others were rejected. This reduced our previous set of 113 descriptors to 34 descriptors. For further selection of descriptors, we used the Stepwise Discriminant Analysis (SDA) [43] using a free data mining tool Tanagra [44]. SDA is often associated with

Table 2 Composition of the datasets used in this study

Dataset	Training set	Testing set	Total
Active	126	22	148
Inactive	114	33	147
Total	240	55	295
Prediction set (from ChEMBL)	–	–	10,000

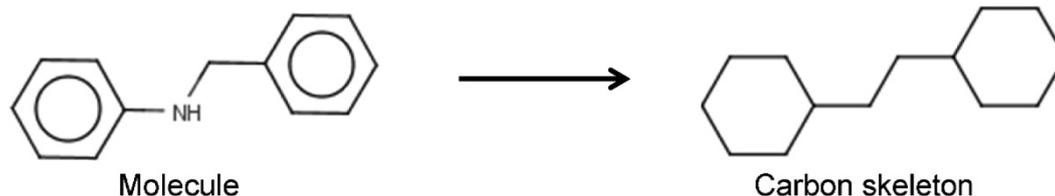


Figure 2 Definition of the scaffold used in this study. The scaffold is obtained by iteratively removing side chains and converting all the bonds to single bonds

discriminant analysis but in fact, it can be applied to various linear models such as linear SVM and logistic regression. However, it is not suitable for non-linear models such as multi-layer neural networks and nearest neighbours. We implemented SDA with both forward and backward elimination strategies. In the forward approach, at each step, all the variables are evaluated to determine which variables contribute maximum to the discrimination between the groups. Variables with significant contributions are included and the process starts again till there is no attribute to add to the model. In the backward approach, all the descriptors are included in the model and then, at each step, the descriptor that contributes least to the discrimination is eliminated, terminating when there is no descriptor to remove. For our

problem, we found that the forward approach performs better than the backward elimination strategy. We used F statistics as the termination criterion, with a predefined threshold value of 3.44, where the F value for a descriptor indicates its statistical significance to discriminate between the positive and negative data groups. This resulted in the selection of final 14 descriptors out of 34. In Table 3 we present the final 14 descriptors used in this study.

SVM algorithm

The SVM algorithm was developed by Vapnik [45]. Recently, SVM has been applied to chemoinformatics, due to its robustness and ability to classify objects into two classes as a function of their features [46,47]. Many

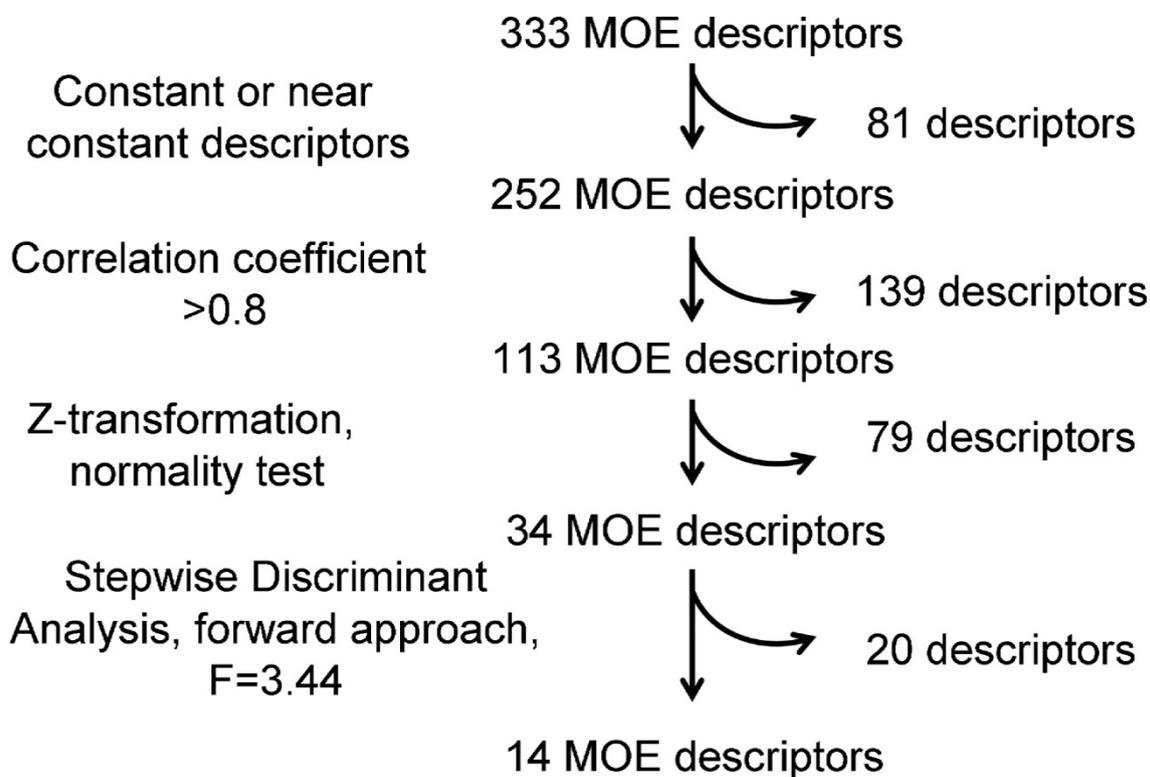


Figure 3 Overall methodology adopted for descriptor selection. Out of the total 333 MOE descriptors only 14 are used in this analysis.

Table 3 List of final 14 descriptors used in this analysis

S.No.	Descriptor	Description
1.	AM1_HF	The heat of formation (kcal/mol)
2.	AM1_HOMO	The energy (eV) of the Highest Occupied Molecular Orbital
3.	ASA+	Water accessible surface area of all atoms with positive partial charge
4.	ASA-	Water accessible surface area of all atoms with negative partial charge
5.	ASA_P	Water accessible surface area of all polar
6.	E_ele	Electrostatic component of the potential energy.
7.	KierFlex	Kier molecular flexibility index
8.	LogS	Log of the aqueous solubility (mol/L).
9.	Std_dim3	The square root of the third largest eigenvalue of the covariance matrix of the atomic coordinates.
10.	Vsurf_CP	
11.	Vsurf_CW2	Capacity factor
12.	Vsurf_D8	Hydrophobic volume
13.	Vsurf_EWmin	Lowest hydrophilic energy
14.	Vsurf_HB1	H-bond donor capacity

All the descriptors are derived from MOE software.

studies in the past have shown SVM to be one of the best methods for correctly classifying molecules [25,48]. A standard application of SVM involves defining two classes of objects, determining the set of features that distinguish these objects and use the trained SVM model to predict the classes of unknown data. Detailed accounts of the SVM methodology are present in literature [35,49]. Briefly, SVM is a new algorithm and is based on structural risk minimization principle from statistical learning theory. Each molecule to be classified by SVM is represented by a feature vector x_i ($i=1,2,\dots,N$) of M real numbers (descriptors) with the corresponding label $y_i \in \{+1,-1\}$, where $y_i = -1$ means inactive and $y_i = +1$ means active. To classify the data, the SVM attempts to find the optimal hyperplane $\{x \in R^m: w \cdot x + b = 0\}$ that best separates the input data into two classes in M dimensional space. The optimal hyperplane is defined in such a way that margin of separation between positive $\{x \in R^m: w \cdot x + b \geq 0\}$ and negative $\{x \in R^m: w \cdot x + b \leq 0\}$ examples is maximized with minimal error; where w is the normal vector of the hyperplane and b is the scalar. In other words, the optimal hyperplane passes through the "midpoint" between these sets. The decision function for new predictions on unseen examples is given in equation 1:

$$f(x) = \text{sign} \left(\sum_{i=1}^N \alpha_i y_i K(x_i, x_j) + b \right) \quad (1)$$

where $K(x_i, x_j)$ is the kernel function and the parameters are determined by maximizing the following equation 2:

$$\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (2)$$

under the conditions (equation 3):

$$\sum_{i=1}^N \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C \quad (3)$$

The penalty constant C serves as a regularization parameter and represents the trade-off between minimizing the training set error and maximizing the margin. Higher number of support vectors is due to a small C and *vice versa*. If we use a very small C value, then almost all the samples would influence the model equally to build a decision boundary regardless of their position. As a result, virtually all the samples become support vectors. On the other hand, if we use a large C it may cause overfitting.

Since there are different types of kernels present (linear, polynomial, radial basis function, sigmoid) we explored various kernels for the efficacy of SVM prediction. From our analysis we note that radial basis function (RBF) kernel (equation 4) was found to be most effective (data not shown) therefore we have chosen the RBF kernel for further analysis.

$$K(x_i, x_j) = \exp \left(- \frac{\|x_i - x_j\|^2}{2\gamma^2} \right) \quad (4)$$

Two parameters *viz.*, γ which determines the capacity of the RBF kernel and the regularization parameter, C are required for optimization of SVM classifiers. To optimize the SVM parameters, C and γ , we carried out an extensive grid search to build accurate models. The

resulting optimized parameters were $C = 1.4$ and $\gamma = 0.43$.

Model validation

The prediction accuracy of the models developed was tested using ten-fold cross-validation technique. In a ten-fold cross-validation, the dataset was split into ten subsets of equal proportions. One of the subsets was used as the test set while the rest were used for training the classifier. The trained classifier was tested using the test set. This was repeated ten times using a different subset for testing and thus ensuring that every compound was used in prediction once.

Performance measure

The prediction results from SVM were evaluated for the test dataset using the following statistical measures.

- TP, true positive – the number of correctly classified active compounds.
- TN, true negative – the number of correctly classified non-active compounds.
- FP, false positives – the number of incorrectly classified non-active compounds.
- FN, false negative – the number of incorrectly classified active compounds.

Using the variables above, a series of metrics were computed sensitivity (SN), specificity (SP), balanced accuracy (BA), F-measure and Matthews correlation coefficient (MCC).

The recall rate for the members of positive class (actives) is given by sensitivity, equation 5:

$$\text{sensitivity} = \frac{TP}{TP + FN} * 100 \quad (5)$$

Similarly, the recall rate for the members of the negative class (inactives) is given by the specificity, equation 6:

$$\text{specificity} = \frac{TN}{TN + FP} * 100 \quad (6)$$

Accuracy measures the ratio of correct predictions to the total number of classes evaluated. We calculated balanced accuracy which is given by the equation 7:

$$\text{balanced accuracy} = \frac{\text{specificity} + \text{sensitivity}}{2} \quad (7)$$

Further, we calculated the F-measure, which is given by equation 8:

$$F - \text{measure} = \frac{2 TP}{2TP + FN + FP} \quad (8)$$

Finally we calculated MCC from equation 9; the coefficient returns a value between +1 and -1. The higher the value of MCC, the better the classification result.

$$\text{matthews correlation coefficient} = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (9)$$

Results and discussion

The main aim of this study was to classify and predict novel compounds active against parasitic nematodes. The various molecular descriptors (333 in total) were calculated initially, using MOE [42]. After removing insignificant attributes (standard deviation ≤ 0.3) and applying a correlation test with a cutoff value of 0.8 we were able to reduce the total number of attributes to 113. Subsequently the SDA algorithm was applied and finally a set of 14 descriptors was selected for the development of classification model (details in Methods section).

The obtained model correctly classified 87.56% of the active compounds and 85.30% of the inactive compounds with the overall accuracy of 86.43% in the training set while 81.82% in the test set. The F-measure of the training and test sets are 86.52% and 79.17% respectively. Table 4 shows the result of the classification for the training and testing sets. All the predicted compounds can be found in Additional File 2.

The machine learning systems such as this could clearly reduce the cost involved in experimental methods involved in drug discovery pipeline. As the SVM algorithm has been effectively applied in various classification problems, we investigated the utility of SVM approach for the prediction of potential anthelmintic lead compounds. The accuracy of the model on the training dataset may indicate the effectiveness of a prediction model however; it may not be able to accurately show how the model will perform on novel compounds. Therefore, it is critical to test the model on an independent dataset, not used in training. In our case we trained and optimized the SVM classifier separately using the entire training set and evaluated the model on the test set. As shown in Table 4, the SVM model obtained an accuracy of 81.79% for the test set. On careful examination of our prediction result, we find that structural similarity of many false

Table 4 Performance measure of SVM classifier in training and test dataset

Dataset	SN (%)	SP (%)	BA (%)	F-measure (%)	MCC
Training set	87.56	85.38	86.43	86.52	0.75
Test set	83.82	79.76	81.79	79.17	0.63

SN: sensitivity, SP: specificity, BA: balanced accuracy, MCC: Matthews correlation coefficient

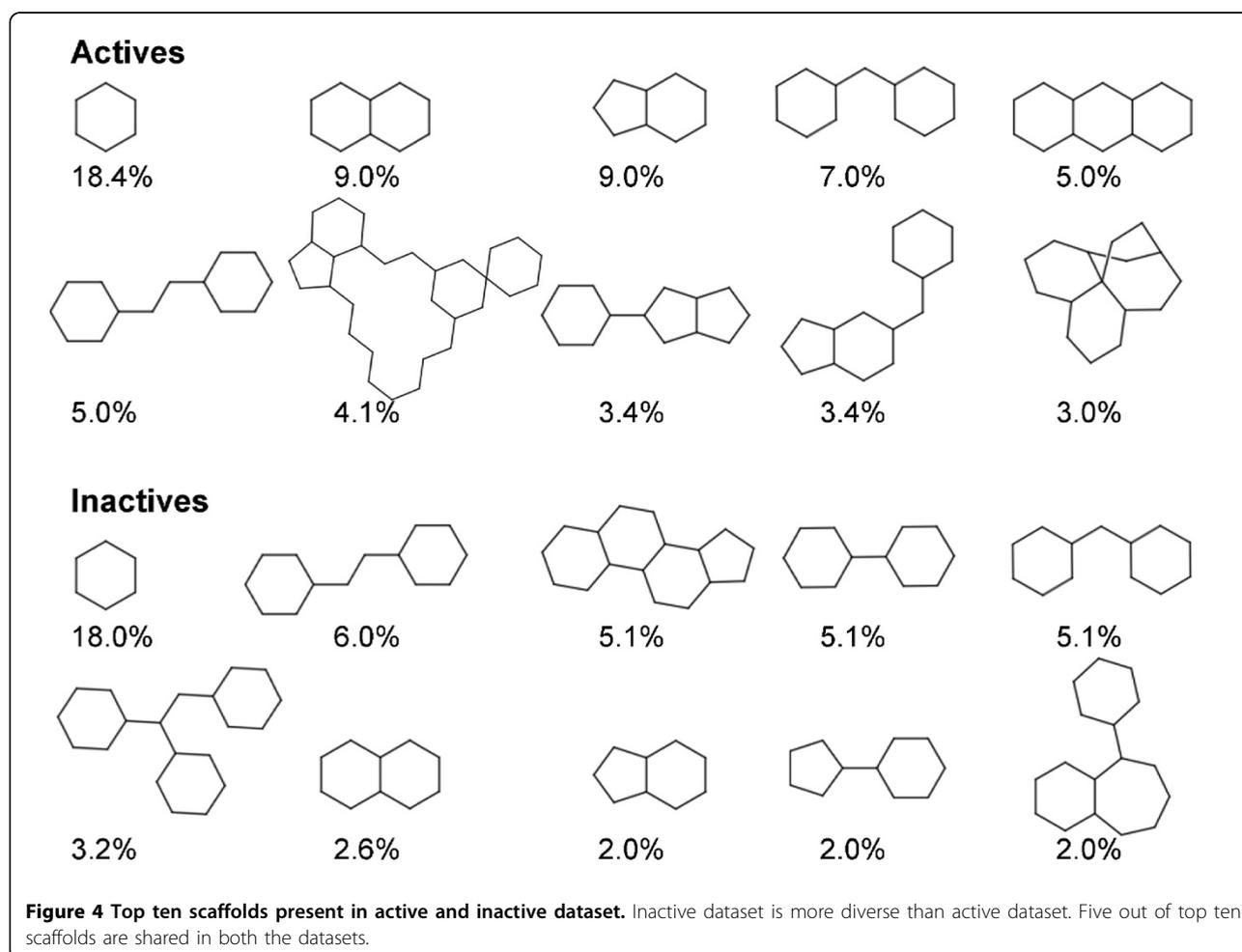
Table 5 The number of unique scaffolds found in active and inactive sets along with the percentage relative to the dataset size

Datasets	Size of the dataset	Non-redundant scaffolds	Percentage (relative to dataset size)
Actives	148	48	32.43%
Inactives	147	80	54.42%

positives to the compounds in the active set is quite high, which may suggest a lower accuracy figure for the test set, due to our stringent threshold values. Further, we also note that a few false negatives are at the borderline and are thus classified as inactive by our model. To best of our knowledge, there are not many reported studies on the prediction of anthelmintic compounds therefore we were able to compare our results with only one study. We find that our results are comparable to that study. Marrero-Ponce *et al.*[19] used linear discriminant analysis to classify anthelmintic drug-like from non-anthelmintic compounds. The authors reported the accuracy of around 90.4 % in the training set while 88.2% in the test set which is slightly higher than ours. However, we believe our model is more robust because our selection

criterion to pick inactive compounds was quite stringent. We selected molecules within the Tanimoto range of 0.25 to 0.75 of the compounds present in the active set which would make it relatively difficult to classify than if chosen randomly. The idea was to build a robust model that can classify compounds into separate groups even with structural similarity. Further, we surmise that since DrugBank covers most of the FDA approved drugs, the inclusion of DrugBank compounds in our inactive dataset would allow us to navigate to the unexplored regions of drug-like chemical space.

The results obtained are particularly interesting from a clinical perspective. From our scaffold analysis we note that even though the size of both the dataset (active and inactive) is approximately same, the number of unique



scaffolds found in the inactive set is almost twice the number of unique scaffolds found in active set. This clearly indicates that the inactive set is more diverse than the active set. The number of unique scaffolds, along with the relative percentage according to the total number of molecules present in the dataset is reported in Table 5. In Figure 4, we report the top ten molecular scaffolds in both the datasets. We note that, over 70.0% of the active compounds are represented by the top 10 scaffolds whereas only 51.1% of the inactive compounds are represented by the same number of scaffolds. This again suggests high scaffold diversity in inactive dataset. It should also be noted that five of the top ten scaffolds shown in Figure 4 are shared by both datasets.

In the 45 predicted compounds, we note that piperazine-like substructures appear frequently suggesting that the nitrogen atom in the piperazine ring might be involved in binding to the receptor. Figure 5 shows an example set of predicted active compounds. Also, we note that many predicted compounds either contain benzimidazole scaffold or are derived from it *e.g.* in Figure 5,

six compounds out of twelve are a derivative product of the benzimidazole scaffold. This shows the validity of the above method since the benzimidazole class of compounds are well recognized for anthelmintic activity [2]. Further, we searched the ChEMBL database for the binding affinity, assay type and target information of the identified compounds. We note that many predicted compounds bind to targets of interest in model organisms but experimental validation in the case of nematodes needs to be further carried out. Out of the total 45 predicted compounds six compounds are particularly interesting. Compound 3 with antiviral activity, compound 10 with inhibitory activity against *Ancylostoma ceylanicum* (a nematode), compound 12, compound 37 with antimicrobial activity against *Staphylococcus aureus*, compound 26 with activity to inhibit SARS-CoV 3CL protease enzyme and compound 40 with activity against Rhinovirus. In addition, there are compounds that bind to nicotinic acetylcholine receptor and tubulin β -1 chain in rats or humans. Since these two receptors are successful targets in nematodes, predicted compounds that bind

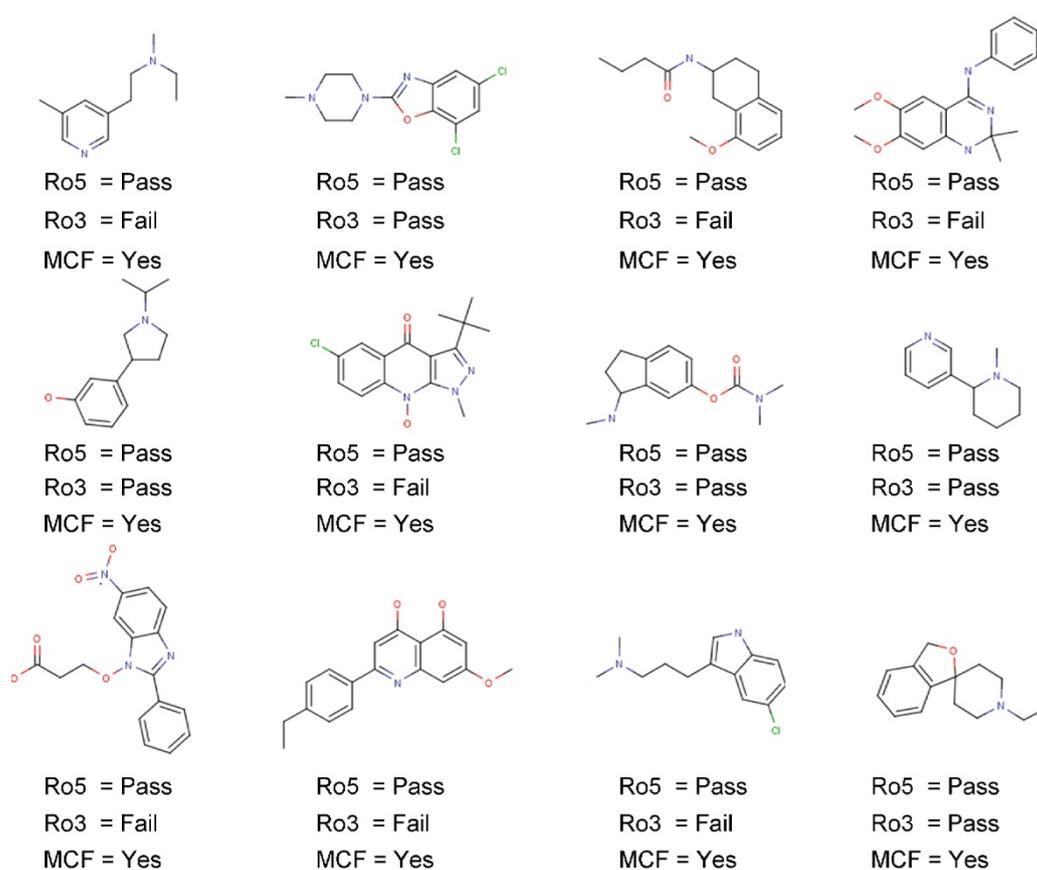


Figure 5 Examples of the actives predicted in the prediction set derived from ChEMBL database. All the molecules shown in the figure pass "rule of five" (Ro5) test and are medicinal chemist friendly (MCF). Further a few of them also pass lead-likeness "rule of three" (Ro3) test.

to these targets can be used as leads to design novel compounds with high binding affinity to nematodes nicotinic acetylcholine and tubulin β -1 chain receptor.

Conclusions

We were able to compile an extensive dataset of anthelmintic compounds for the development and validation of support vector machine model. We thoroughly tested the SVM approach for identifying the potential compounds with anthelmintic activity. From our results we conclude that SVM method is well suited for the prediction of anthelmintic (or antiparasitic) compounds. We were also able to identify a number of interesting compounds with potential activity against parasitic nematodes however; experimental validation of the predicted compounds is needed.

Additional material

Additional file 1: Table S1 Dataset used for training, testing and validation of the model.

Additional file 2: Table S2 Predicted compounds with AlogP, molecular weight and SMILES information.

Acknowledgements

We thank Dr. Dominique Gorse for useful discussions during this study. VK is grateful to Macquarie University for the award of MQRES research scholarship.

This article has been published as part of *BMC Bioinformatics* Volume 12 Supplement 13, 2011: Tenth International Conference on Bioinformatics – First ISCB Asia Joint Conference 2011 (InCoB/ISCB-Asia 2011): Bioinformatics. The full contents of the supplement are available online at <http://www.biomedcentral.com/1471-2105/12?issue=S13>.

Author details

¹Dept. of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, Australia. ²Dept. of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.

Authors' contributions

VK curated the datasets and conducted the analysis work, SR directed the study and both the authors prepared the manuscript.

Conflicts of interests

None declared.

Published: 30 November 2011

References

- Ranganathan S, Menon R, Gasser RB: **Advanced in silico analysis of expressed sequence tag (EST) data for parasitic nematodes of major socio-economic importance-fundamental insights toward biotechnological outcomes.** *Biotechnol Adv* 2009, **27**(4):439-448.
- Lacey E: **Mode of action of benzimidazoles.** *Parasitol Today* 1990, **6**(4):112-115.
- Friedman PA, Platzer EG: **Interaction of anthelmintic benzimidazoles with *Ascaris suum* embryonic tubulin.** *Biochim Biophys Acta* 1980, **630**(2):271-278.
- Kohler P, Bachmann R: **Intestinal tubulin as possible target for the chemotherapeutic action of mebendazole in parasitic nematodes.** *Mol Biochem Parasitol* 1981, **4**(5-6):325-336.
- Cully DF, Vassiliatis DK, Liu KK, Paresse PS, Van der Ploeg LH, Schaeffer JM, Arena JP: **Cloning of an avermectin-sensitive glutamate-gated chloride channel from *Caenorhabditis elegans*.** *Nature* 1994, **371**(6499):707-711.
- Holden-Dye L, Walker RJ: **Avermectin and avermectin derivatives are antagonists at the 4-aminobutyric acid (GABA) receptor on the somatic muscle cells of *Ascaris*; is this the site of anthelmintic action?** *Parasitology* 1990, **101**(Pt 2):265-271.
- Bokisch AJ, Walker RJ: **The action of Avermectin (MK 936) on identified central neurones from *Helix* and its interaction with acetylcholine and gamma-aminobutyric acid (GABA) responses.** *Comp Biochem Physiol C* 1986, **84**(1):119-125.
- Martin RJ, Verma S, Levandoski M, Clark CL, Qian H, Stewart M, Robertson AP: **Drug resistance and neurotransmitter receptors of nematodes: recent studies on the mode of action of levamisole.** *Parasitology* 2005, **131**(Suppl):S71-84.
- Sutherland IA, Leathwick DM: **Anthelmintic resistance in nematode parasites of cattle: a global issue?** *Trends Parasitol* 2011, **27**(4):176-181.
- James CE, Hudson AL, Davey MW: **Drug resistance mechanisms in helminths: is it survival of the fittest?** *Trends in Parasitology* 2009, **25**(7):328-335.
- Geerts S, Gryseels B: **Drug resistance in human helminths: current situation and lessons from livestock.** *Clin Microbiol Rev* 2000, **13**(2):207-222.
- Keiser J, Utzinger J, Xiao-Nong Zhou RBRO, Jürg U: **The drugs we have and the drugs we need against major helminth infections.** In *Advances in Parasitology. Volume 73.* Academic Press; 2010:197-230.
- Kaminsky R, Ducray P, Jung M, Clover R, Rufener L, Bouvier J, Weber SS, Wenger A, Wieland-Berghausen S, Goebel T, et al: **A new class of anthelmintics effective against drug-resistant nematodes.** *Nature* 2008, **452**(7184):176-180.
- Hu Y, Xiao SH, Aroian RV: **The new anthelmintic tribendimidine is an L-type (levamisole and pyrantel) nicotinic acetylcholine receptor agonist.** *PLoS Negl Trop Dis* 2009, **3**(8):e499.
- Harder A, von Samson-Himmelstjerna G: **Cyclooctadepsipeptides—a new class of anthelmintically active compounds.** *Parasitol Res* 2002, **88**(6):481-488.
- Harder A, Schmitt-Wrede HP, Krucken J, Marinovski P, Wunderlich F, Willson J, Amliwala K, Holden-Dye L, Walker R: **Cyclooctadepsipeptides—an anthelmintically active class of compounds exhibiting a novel mode of action.** *Int J Antimicrob Agents* 2003, **22**(3):318-331.
- Keiser J, Chollet J, Xiao S-H, Mei J-Y, Jiao P-Y, Utzinger Jr, Tanner M: **Mefloquine—an aminoalcohol with promising antischistosomal properties in mice.** *PLoS Negl Trop Dis* 2009, **3**(1):e350.
- Xiao S-H, Keiser J, Chen M-G, Tanner M, Utzinger J, Xiao-Nong Zhou RBRO, Jürg U: **Research and development of antischistosomal drugs in the people's republic of China: a 60-year review.** In *Advances in Parasitology. Volume 73.* Academic Press; 2010:231-295.
- Marrero-Ponce Y, Castillo-Garit JA, Olazabal E, Serrano HS, Morales A, Castanedo N, Ibarra-Velarde F, Huesca-Guillen A, Jorge E, del Valle A, et al: **TOMOCOMD-CARDD, a novel approach for computer-aided 'rational' drug design: I. Theoretical and experimental assessment of a promising method for computational screening and in silico design of new anthelmintic compounds.** *J Comput Aided Mol Des* 2004, **18**(10):615-634.
- Reddy S, Pati P, Kumar P, Pradeep HN, Sastry N: **Virtual screening in drug discovery – a computational perspective.** *Current protein & peptide science* 2007, **8**(4):329-351.
- Freitas RF, Oprea TI, Montanari CA: **2D QSAR and similarity studies on cruzain inhibitors aimed at improving selectivity over cathepsin L.** *Bioorganic & Medicinal Chemistry* 2008, **16**(2):838-853.
- Sousa Sr, Fernandes P, Ramos M: **Protein-ligand docking: Current status and future challenges.** *Proteins* 2006, **65**(1):15-26.
- Geppert H, Vogt M, Bajorath J: **Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation.** *Journal of Chemical Information and Modeling* 2010, **50**(2):205-216.
- Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV: **Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions.** *Journal of Chemical Information and Computer Sciences* 2003, **43**(6):2048-2056.

25. Warmuth MK, Liao J, Ratsch G, Mathieson M, Putta S, Lemmen C: **Active learning with support vector machines in the drug discovery process.** *Journal of Chemical Information and Computer Sciences* 2003, **43**(2):667-673.
26. Burbidge R, Trotter M, Buxton B, Holden S: **Drug design by machine learning: support vector machines for pharmaceutical data analysis.** *Computers & Chemistry* 2001, **26**(1):5-14.
27. Woods D, Williams T: **The challenges of developing novel antiparasitic drugs.** *Invertebrate Neuroscience* 2007, **7**(4):245-250.
28. Tropsha A: **Best practices for QSAR model development, validation, and exploitation.** *Molecular Informatics* 2010, **29**(6-7):476-488.
29. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH: **PubChem: a public information system for analyzing bioactivities of small molecules.** *Nucleic Acids Research* 2009, **37**(suppl 2):W623-W633.
30. Holden-Dye L, Walker RJ: **Anthelmintic drugs.** *WormBook* 2007, 1-13.
31. Mayer AM, Hamann MT: **Marine pharmacology in 2001-2002: marine compounds with anthelmintic, antibacterial, anticoagulant, antidiabetic, antifungal, anti-inflammatory, antimalarial, antiplatelet, antiprotozoal, antituberculosis, and antiviral activities; affecting the cardiovascular, immune and nervous systems and other miscellaneous mechanisms of action.** *Comp Biochem Physiol C Toxicol Pharmacol* 2005, **140**(3-4):265-286.
32. Mayer AM, Rodriguez AD, Berlinck RG, Hamann MT: **Marine pharmacology in 2003-4: marine compounds with anthelmintic antibacterial, anticoagulant, antifungal, anti-inflammatory, antimalarial, antiplatelet, antiprotozoal, antituberculosis, and antiviral activities; affecting the cardiovascular, immune and nervous systems, and other miscellaneous mechanisms of action.** *Comp Biochem Physiol C Toxicol Pharmacol* 2007, **145**(4):553-581.
33. Mayer AM, Rodriguez AD, Berlinck RG, Hamann MT: **Marine pharmacology in 2005-6: Marine compounds with anthelmintic, antibacterial, anticoagulant, antifungal, anti-inflammatory, antimalarial, antiprotozoal, antituberculosis, and antiviral activities; affecting the cardiovascular, immune and nervous systems, and other miscellaneous mechanisms of action.** *Biochim Biophys Acta* 2009, **1790**(5):283-308.
34. Wishart D, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets.** *Nucleic Acids Research* 2008, **36**(Database issue):D901-906.
35. Trotter MWB, Holden SB: **Support vector machines for ADME property classification.** *Qsar & Combinatorial Science* 2003, **22**(5):533-548.
36. Overington J: **ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr.** *J Comput Aided Mol Des* 2009, **23**(4):195-198.
37. **Pipeline Pilot.** [http://accelrys.com/].
38. Bemis GW, Murcko MA: **The properties of known drugs. 1. Molecular frameworks.** *Journal of Medicinal Chemistry* 1996, **39**(15):2887-2893.
39. Dutta D, Guha R, Wild D, Chen T: **Ensemble feature selection: consistent descriptor subsets for multiple QSAR models.** *Journal of Chemical Information and Modeling* 2007, **47**(3):989-997.
40. Duch W: **Filter Methods.** In *Feature Extraction: Foundations and Applications. Volume 207.* Berlin, Germany: Springer; Guyon I, Gunn S, Nikravesh M, Zadeh L 2006.
41. Marchiori E, Moore J, Soto A, Cecchini R, Vazquez G, Ponzoni I: **A wrapper-based feature selection method for ADMET prediction using evolutionary computing.** In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. Volume 4973.* Springer Berlin / Heidelberg; 2008:188-199.
42. **MOE(2009.10).** [http://www.chemcomp.com/].
43. Jennrich RI: **Stepwise discriminant analysis.** In *Statistical methods for digital computers. Volume 3.* New York: Wiley; Enslin K, Ralston A, Wilf HS 1977:76-96.
44. **Tanagra: free data mining software.** [http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html].
45. Cortes C, Vapnik V: **Support-vector networks.** *Machine Learning* 1995, **20**(3):273-297.
46. Jorissen RN, Gilson MK: **Virtual screening of molecular databases using a support vector machine.** *Journal of Chemical Information and Modeling* 2005, **45**(3):549-561.
47. Liew CY, Ma XH, Liu X, Yap CW: **SVM model for virtual screening of Lck inhibitors.** *Journal of Chemical Information and Modeling* 2009, **49**(4):877-885.
48. Byvatov E, Fechner U, Sadowski J, Schneider G: **Comparison of support vector machine and artificial neural network systems for drug/nondrug classification.** *J Chem Inf Comput Sci* 2003, **43**(6):1882-1889.
49. Ivanciuc O: **Applications of support vector machines in chemistry.** *Reviews in Computational Chemistry* 2007, **23**:291-400.

doi:10.1186/1471-2105-12-S13-S25

Cite this article as: Khanna and Ranganathan: *In silico* approach to screen compounds active against parasitic nematodes of major socio-economic importance. *BMC Bioinformatics* 2011 **12**(Suppl 13):S25.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

