Published in final edited form as:
Comb Chem High Throughput Screen. 2010 January ; 13(1): 54–66.

Applications of Integrated Data Mining Methods to Exploring Natural Product Space for Acetylcholinesterase Inhibitors

Daniela Schuster[1,2], **Lisa Kern**[3], **Dimitar P. Hristozov**[4], **Lothar Terfloth**[4], **Bruno Bienfait**[4], **Christian Laggner**[1], **Johannes Kirchmair**[1,2], **Ulrike Grienke**[3], **Gerhard Wolber**[1,2], **Thierry Langer**[5], **Hermann Stuppner**[3], **Johann Gasteiger**[4], and **Judith M. Rollinger**[*,3]

[1]Department of Pharmaceutical Chemistry, Institute of Pharmacy, University of Innsbruck, Innrain 52c, A-6020 Innsbruck, Austria and Center of Molecular Biosciences Innsbruck (CMBI)

[2]Inte:Ligand Softwareentwicklung und Consulting GmbH, Clemens-Maria-Hofbauer-Gasse 6, A-2344 Maria Enzersdorf, Austria

[3]Department of Pharmacognosy, Institute of Pharmacy, University of Innsbruck, Innrain 52c, A-6020 Innsbruck, Austria and Center of Molecular Biosciences Innsbruck (CMBI)

[4]Molecular Networks GmbH, Henkestr. 91, D-91052 Erlangen, Germany

[5]Prestwick Chemical Inc., Boulevard Gonthier d'Andernach, 67100 Illkirch, France

Abstract

Nature, especially the plant kingdom, is a rich source for novel bioactive compounds that can be used as lead compounds for drug development. In order to exploit this resource, the two neural network-based virtual screening techniques novelty detection with self-organizing maps (SOMs) and counterpropagation neural network were evaluated as tools for efficient lead structure discovery. As application scenario, significant descriptors for acetylcholinesterase (AChE) inhibitors were determined and used for model building, theoretical model validation, and virtual screening. Top-ranked virtual hits from both approaches were docked into the AChE binding site to approve the initial hits. Finally, in vitro testing of selected compounds led to the identification of forsythoside A and (+)-sesamolin as novel AChE inhibitors.

Keywords

Natural products; drug discovery; acetylcholinesterase; virtual screening; counterpropagation network; spinne; novelty detection

INTRODUCTION

The utility of natural products as sources for novel lead structures could be testified by the recently updated analysis of Newman and Cragg [1]. Their work clearly indicates that natural products, their derivatives and mimics play a key role in drug discovery and development process. Only 37% of 974 small molecule new chemical entities have been shown to be truly synthetic in origin. Although rationalized procedures in the search for

[*]Address correspondence to this author at the Department of Pharmacognosy, Institute of Pharmacy, University of Innsbruck, Innrain 52c, A-6020 Innsbruck, Austria; Tel.: +43-512-507-5308; Fax: +43-512-507-2939; Judith.Rollinger@uibk.ac.at.

SUPPLEMENTARY MATERIAL This article also contains supplementary material (Tables and Figures), and it can be viewed online at publisher's website.

bioactive natural products are in great demand to find the needles in a haystack, computational methods have only rarely been applied for natural product research [2-4].

The better understanding of the fundamental principles of protein-ligand-interactions and the steadily growing number of 3D structures of targets and their ligands provide ways toward more rational and more efficient approaches in drug discovery. In recent decades different computational tools for data mining of bioactive molecules from large libraries of 3D structures have contributed substantially to the target-oriented search of new drug leads. The underlying techniques include virtual screening experiments based on docking protocols or pharmacophore models, and different prediction methods based on the analysis of the chemical environment represented by structural 2D or 3D characteristics of molecules, i.e. descriptors, known to act as ligands or non-ligands [5-8]. In the data mining process, the correlation of these structural characteristics with the biological activity of a representative number of defined objects is used to extract knowledge from a large set of data in order to make predictions of new events [9].

Continuing our efforts to discover new bioactive compounds from the plant kingdom, we focused on acetylcholinesterase (AChE) inhibitors. This enzyme (EC 3.1.1.7) is a key component of cholinergic brain synapses and neuromuscular junctions. The major biological role of AChE is the termination of impulse transmissions within the nervous system by rapid hydrolysis of the neurotransmitter, acetylcholine. The inhibition of the enzyme’s hydrolytic activity has been exploited in medicine for treating disease states associated with inadequate levels of acetylcholine, like neurodegenerative impairments [10, 11]. Besides its classical role in terminating synaptic transmission, AChE was found to be involved in a number of other functions, e.g., in neurite growth and in accelerating the assembly of β-amyloid into amyloid fibrils which are characteristically found in the brain cells of Alzheimer’s patients [12].

Today’s structural knowledge on AChE is mostly based on work carried out on *Tc*AChE from the electric eel *Torpedo californica* showing high similarity to the human AChE [13]. X-ray crystallographic studies of AChE/ligand complexes show an almost identical 3D structure of the active site (only Phe330 is replaced by Tyr337 in *Tc*AChE) in which the active center is located nearly at the bottom of a 20 Å deep and narrow gorge [14]. Different positions of known inhibitors in the binding pocket suggest that more than one clearly defined binding site exists [15]: In the active site, the esteratic subsite contains the catalytic triad, whereas the anionic subsite is largely composed of aromatic residues which are believed to stabilize the ligand/substrate by π-cation interactions. A further initial binding is thought to occur at the outer rim of the gorge, at the peripheral anionic site (PAS [16]). Ligands of the PAS as well as compounds with gorge-spanning binding modes, reaching from the PAS to the active site, have been reported [17-19].

Our aim is to detect descriptors relevant for compounds with inhibiting effect on AChE. Based on a validated descriptor set, we employed different chemoinformatics methods to mine a natural product 3D-database. By simultaneous presentation of this large dataset and the training set, it is expected that compounds co-localizing with highly active compounds from the training set also show high AChE inhibiting activity. The combination of these ligand-based approaches with structure-based virtual screening (i.e. docking) should confirm the virtual hits and help prioritizing candidates for biological testing.

Scrutinizing the resulting clusters of the applied methods, this approach was intended as rationale for the discovery of potential new AChE inhibitors from nature without being restricted to only one binding mode.

RESULTS AND DISCUSSION

Workflow Overview

First, a compound set of 313 AChE inhibitors was collected from literature. For the selection of descriptors used in this study, the compound set was split manually into a training set and a test set (1:3, respectively. Supplementary Material S-1 and S-2). Subsequently, the whole dataset together with the previously determined descriptors was used to train models for activity prediction (counterpropagation neural networks and novelty detection method). After model validation, an external dataset - the natural products database DIOS [20] – was screened. In the resulting hitlists, basic, neutral, and acidic compounds were present. However, due to the aromatic environment of the AChE inhibitor binding site, binding of neutral or basic compounds is more favorable than binding of acidic ones. Virtual hits were filtered in order to exclude acidic compounds and then docked to confirm their potential to fit into the AChE active site. From the resulting hit lists, promising natural compounds were selected for biological testing. A graphical workflow is given in Fig. (1).

Datasets

A dataset composed of 313 compounds with known activity on AChE was collected from literature (Supplementary Material Tables S-1 and S-2). Compound names and activity data extracted from literature were assembled and evaluated for their use in this study. As it would be arbitrary to define single cut-off values for neighboring activity classes, we decided to form activity ranges that are separated from each other. The dataset was split manually into a training set and test set accounting to a comparable chemical diversity and equal distribution of activity clusters. The training set consisting of 80 compounds with a broad activity spectrum was grouped into three classes: (i) non-inhibitors (i.e. class 0, 36 compounds, AChE inactive compounds or IC_{50} higher than 600 μM), (ii) medium inhibitors (class 1, 21 compounds, IC_{50} between 50 and 100 μM), and (iii) strong inhibitors (class 2, 23 compounds, IC_{50} lower than 0.1 μM) with all classes containing about the same range of entities (Supplementary Material Table 1). As our overall data pool also included compounds with activities between these activity classes,the test set compounds (n = 233) had to be divided into the following three groups (Supplementary Material Table 2): Non-inhibitors (IC_{50} > 200 μM, 92 compounds), medium inhibitors (IC_{50} between 10 and 200 μM, 57 compounds), and strong inhibitors (IC_{50} < 10 μM, 84 compounds). CAS numbers, the assigned activity classes and the corresponding references of the test set compounds are provided in the Supplementary Material (Supplementary Material Table S-2).

We employed an external dataset – the natural products database DIOS [20] – for testing of our models. It consists of 9,676 (6,702 unique, uncharged) natural products from medicinal plants described in the ancient ethnopharmacological source *De materia medica* by Pedanios Dioscorides (1st cent AD). The prediction of this external dataset was the basis for our selection of compounds for *in vitro* testing.

Selection of Suitable Descriptors for AChE Inhibitors

Computation and Selection of Descriptors—As a biological property cannot be calculated directly from a chemical structure, meaningful descriptors have to be computed in order to handle calculations. These descriptors may be derived from the constitution of a molecule, its 3D structure or molecular surface properties [21]. As starting point for these calculations, a 3D structure of the molecule is needed. In this study, the 3D geometry of all compounds was generated using CORINA [22]. Subsequently, all descriptors available in ADRIANA.Code 2.0 [23] were calculated for the dataset. The descriptors were evaluated according to their ability to cluster compounds from equal activity classes in one area using self-organizing maps (SOMs) [24]. The aim of a SOM is to create a low-dimensional map of a high-dimensional landscape. The created map is usually two-dimensional and allows the exploration of relationships among the data by various methods, including simple visual inspection, which is not feasible within a high-dimensional space. During the projection, the topology of the input space is preserved, which means that objects adjacent in the high-dimensional space are also neighbors in the low-dimensional target space. Each compound is assigned to one specific neuron. Neurons can be occupied by none or also by several compounds. The quality of a map can be analyzed by investigating conflict neurons (neurons that are occupied with compounds from different activity classes). In this study, SONNIA [25] was used to generate SOMs. For SOM training, only the descriptors computed by ADRIANA.Code were submitted to the program (in a first step, each descriptor by its own) which is referred to as unsupervised learning. Subsequently, the training set compounds together with the information on their activity class were laid onto the generated map. We then investigated whether the used descriptor(s) led to neuron occupancy with compounds of the same activity classes and whether the activity neurons formed clusters in the SOM. The molecular descriptors that led to the SOMs with the least conflict and empty neurons were then combined in groups of two or three neurons to lower the signal-to-noise ratio. However, this systematic approach of descriptor combination did not yield satisfying separation of highly active from inactive compounds.

In a second attempt, we decided to use our knowledge on AChE-inhibitor interactions derived from X-ray crystallography for the combination of suitable descriptors. The PDB entry 1w6r [26] (*Tc* AChE in complex with a highly active galanthamine derivative, IC_{50} = 702 nM [26]) was submitted to automated chemical interaction determination using LigandScout 2.0 [27] as illustrated by Fig. (2).

The ligand is anchored in the active site of AChE by hydrogen bonding to His440, Ser200, Gly118, and Glu199. Cation-π interactions at the entrance of the active site additionally stabilize the orientation of the ligand. Furthermore, hydrophobic interactions with aromatic residues of the binding pocket contribute to the stabilization of the complex. Based on these observations, descriptors related to electrostatic interactions, hydrophobicity, and the overall shape of the compounds were selected for further evaluation. Descriptors available in ADRIANA.Code accounting for electrostatic interactions are the molecular dipole moment, topological or 3D autocorrelation vectors for σ-, π-, and in particular total charges as well as 3D surface autocorrelation of the electrostatic potential. Hydrophobicity-related descriptors within ADRIANA.Code are XlogP, 3D surface autocorrelation of the hydrophobicity potential, and topological or 3D autocorrelation vectors for the effective atomic polarizability and π-charges. The overall shape of the compounds is reflected by 3D autocorrelation vectors and the radial distribution function (RDF) code using identity as property. The RDF code is usually calculated with a higher resolution than the 3D autocorrelation vectors and therefore better suited for the representation of the shape of a compound.

The selected descriptors were again combined and evaluated for their suitability to discriminate highly active (class 2), medium active (class 1), and inactive (class 0) compounds. The best descriptor combination comprised component 13 (corresponding to the distance interval from xxx to yyy A) of the radial distribution function for the property identity (RDF_ident13), the molecular dipole moment (DIPOLE_M), and the 3D autocorrelation vector for π-charges (AC3D_qpi). The RDF code using identity as property (RDF_ident) reflects the distribution of the interatomic distance distribution within the considered molecule. RDF vector components for distances greater than the molecule’s diameter are always zero. The distribution itself reflects the overall shape. The molecular dipole moment (DIPOLE_M) indicates whether strong or weak electrostatic interactions occur between the ligand and the protein. 3D autocorrelation vectors considering π-charges

(AC3D_qpi) carry information whether molecules contain hydrophobic substructures such as a phenyl ring. The properties of the descriptors are summarized in Table 2. In this work, the descriptors for the training set were always scaled to zero mean and unit variance, while the descriptors for the test set were scaled using the scaling parameters of the training set. This insured that all descriptors have comparable ranges.

Activity Classification Using SOMs—In order to determine the suitability of the descriptors to group the three activity classes correctly, the 80 compounds of the training set were projected through a SOM. This experiment led to a distinguished clustering of compounds according to their AChE inhibition values as shown in Fig. (3).

Out of 40 neurons forming the torus, 36 were occupied. Overall, only five neurons were occupied by mixed activity classes.

The clustering quality of the resulting SOM was assessed by a k-nearest neighbor (*knn)*-type analysis [28]. In this experiment, each compound is projected through the trained SOM into its winning neuron. The occupancy of the neighborhood of the winning neuron is analyzed by evaluating the class distribution of its four orthogonal neurons. If the majority of compounds belong to the same class as the considered compound, the classification is assumed to be correct. If the majority belongs to different classes, the classification is presumably wrong. In all other cases the classification is drawn, thus providing no activity prediction. For the SOM depicted in Fig. (3) the *knn*-analysis results are provided in Table 2.

As can be seen from Table 2, 63 of the 80 compounds (78.7%) were predicted correctly. Only 14 compounds (17.5%) were predicted incorrectly. For three compounds (2.4%) the decision is drawn. The classification accuracy of the individual classes varies between 73.9% for the highly actives and 80.9% for the medium actives. The *knn*-analysis confirms the good clustering quality of the SOM and corresponds to the visual assessment of the map.

Visualization of Activity Clusters Using SPINNE—The data structure of the 80 training set compounds and the selected descriptors was further analyzed with the program SPINNE [29]. This software can be used for visualizing multi-dimensional data on a 2D plane using nonlinear mapping (NLM) [30]. One potential problem of such visualization is that the dimension reduction induces inaccuracies: two points that are displayed close to one another on the 2D plot might not necessarily be neighbors in the original high-dimensional descriptor space. To reduce misinterpretations of the 2D projection, we superimposed a minimal spanning tree (MST) [31] on the map. This MST was calculated in the high-dimensional space (all 47 descriptors) using Euclidean distance metric and displayed using the 2D coordinates found by NLM. The MST connects points that are nearest neighbors in the high-dimensional space. The plots generated by SPINNE display colored connecting lines using a continuous color scale that encodes the Euclidean distance in the high-dimensional space: from red for large distances to violet for the shortest ones.

The SPINNE projection shows a separation of the training data into several regions that are related to the activity classes as shown in Fig. (4). One compound placed at the bottom left of the plot appears as an outlier. It is connected to its nearest neighbor by a red line, which means the largest Euclidean distance (see Supplementary Material Fig. (S-1)). This compound is the only nitro derivative of the training set. The presence of high partial charges on the nitro group shifts the AC3D part of its descriptor away from the other compounds’ descriptor.

Evaluation of Predictability of Descriptors

SOM—We evaluated the suitability of our selected descriptors to accurately predict highly active AChE inhibitors from our test set (233 compounds). A SOM was trained using the auto-scaled descriptors of our training set (80 compounds) employing the software SONNIA. The resulting clustering is given in Fig. (5A and 5B).

All neurons were investigated for their content. The occupancy was analyzed using a confusion matrix (Table 3).

From the confusion matrix it is obvious that it is not trivial to separate the three activity classes from each other. Especially the activity thresholds defining our activity classes may have a major impact on the resulting classification. It is therefore not surprising that only one third of all class 1 inhibitors were predicted correctly, the rest was nearly equally distributed among class 0 and 2. The prediction of class 0 compounds yielded better results. First, most compounds predicted as class 0 (51.7%) were true inactive molecules. The majority of inactive compounds was matched to class 0 neurons (or empty neurons for which no activity prediction was performed). As we aimed to identify new AChE inhibitors using the descriptors validated in this experiment, the most interesting question for our study was which compounds would be assigned to class 2 neurons. Although the overall prediction of class 2 compounds was poor (most highly actives were placed in class 0, class 1, or empty neurons), the occupancy of class 2 neurons showed a clear accumulation of class 2 compounds (72.2%). We therefore concluded that the parameters used were suitable for detecting highly active ligands from a compound set and correctly assign them to class 2 neurons. The limit of this method is that not all class 2 compounds would be assigned to the correct neurons. Accordingly, we would miss actives from the dataset. As we aim at seeking a fast approach to find new lead structures, the outcome of the SOM prediction showed promising classification of highly actives. Leave-one out cross-validation of the model using a counter propagation neural network with a binary representation of the class membership in three output layers yielded 88.6% correct predictions.

Activity Prediction of the External Dataset by Virtual Screening

Based on the results shown in Table 3, we concluded that the descriptors selected in previous steps were applicable for selecting putative AChE inhibitors from an external dataset. Two methods were used for virtual screening: novelty detection and counterpropagation (CPG) neural network. Both methods are described in more detail in the experimental section.

Novelty Detection—The 107 highly active compounds belonging to class 2 in our training and test set were used to develop a SOM novelty detector and to screen the full DIOS dataset (see Datasets). Acidic compounds were filtered from the dataset. The 10 top-ranked DIOS compounds are compiled in Table 4. They show structural scaffolds which are characterized by high chemical diversity; accordingly, the predicted compounds show a wide distribution as secondary metabolites in botanical systematic. They cover, e.g., a neolignan (**5**) known from different Pinaceae species as well as from Sibirian ginseng, a secoiridoid glucoside (**6**; isolated from *Fraxinus ornus*, Oleaceae), a minor furanocoumarine (**7**) from *Ficus carica* leaves (Moraceae), some diterpenoids (**8**, **9**, **10**, **14**), a caffeic acid glycoside (**11**) from Forsythia species (Oleaceae), and abundant flavanoidglycosides (**12**, **13**).

Counterpropagation (CPG) Neural Network—The training set and the large DIOS database were combined and projected with a counterpropagation neural network [32]. Compounds from the DIOS database colocalized with highly active compounds from the training set are expected to show also AChE inhibiting activity. A network with 60 × 40 neurons (2400 neurons) was trained for 100 epochs using the descriptors identified before. The class membership was specified in a single output layer (0: inactive, 1: intermediate, 2: highly active, 3: unknown). A CPG with 2206 (91.9%) occupied neurons resulted from the training as indicated in the Supplementary Material – Fig. (S-2). 139 compounds from the DIOS database have a predicted value of the class membership between 1.5 and 2.5 and are considered to include highly active AChE inhibitors.

In order to facilitate the selection of compounds for *in vitro* testing, the 139 hits were ranked according to their score which is calculated from the weight in their winning neuron in the output layer by equation 1:

$$\text{score}=1-\left[abs\,(2-\text{weight})\cdot 2\right]\quad (1.5<\text{weight}<2.5) \qquad \text{Eq.(1)}$$

The score ranges from 0 (weight = 1.5 or 2.5) to 1 (weight is 2) and was used to rank the compounds. A score of almost 0 does not mean that the compound is not predicted to be an AChE inhibitor; only the level of confidence is lower. The top-ranked structures predicted by counterpropagation network show less structural diversity than the top-ranked compounds obtained by novelty detection (Table 4). None of the top-ranked compounds contained a negatively ionizable substructure. The scaffolds among the top 10 in the hit list (Table 5) fall into two main types, namely phenolic compounds and terpenoids. Among the first group are chalcones, benzophenones, and a xanthon derivative. Within the terpenoid group are monoterpenes, like pinol or trans-sabinol, and diterpenes. The top listed natural compound even contains both typical structural features, i.e. a rare diterpene oxophenol isolated from *Tetraclinis articulata* Mast. (Cupressaceae) [33].

Evaluation of VS Hits

Novelty detection ranked all compounds from the external dataset DIOS according to their predicted inhibitory activity of AChE. From the CPG-based virtual screening experiment, 139 hits were returned from DIOS. From each of the VS approaches, the top-ranked compounds were submitted to docking experiments. The analysis of availability and accessibility information as well as the docking score led to the selection of three products for biological testing: forsythoside A (**11**), (+)-sesamolin (**25**), and propolis (**18**, **20**, **26**, and **27**) (Fig. 6).

Forsythoside A (**11**) is a caffeic acid glycoside which is found mainly in *Forsythia suspensa* (Oleaceae; weeping forsythia or weeping golden bell). The plant was reported to be used for treating bacterial infections and infections of the upper respiratory tract. Forsythia fruits are broadly used in traditional Chinese medicine to treat fever. Additionally, it is described to have diuretic properties and serves as cardiovascular tonic [34]. Compound **11** (forsythoside A) itself interferes with the arachidonic acid metabolism which confirms its use against inflammation-related diseases [35].

The lignan derivative (+)-sesamolin (**25**) is one of the main components found in sesame seed oil derived from *Sesamum inidicum* seed. This compound shows antioxidative, anti-inflammatory, and neuroprotective effects. Also a significant inhibition of nitrogen monoxide (NO) production *via* the inducible nitric oxide synthase mRNA and protein expression was detected [36]. Interestingly, NO released by activated microglia cells in the central nervous system are suggested to contribute to neurodegeneration (among other parameters such as cytokines and reactive oxygen species) [37].

Four components of the final hitlists originate from propolis. Propolis is a resinous substance that bees collect from tree buds (e.g., poplar, birch, oak, fir, or spruce) or other botanical sources. It is used as a sealant for unwanted open spaces in the hive. Its main components include organic acids, flavonoids, cumarins, aromatic aldehydes, and vitamins. In our hitlist, we detected the propolis ingredients pinostrobin chalcone (**18**), pinocembrin chalcone (**20**), and two ester derivatives (**26**) and (**27**). As the purified compounds were not commercially available, a propolis dichloromethane extract was used for testing. Table 6 shows the calculated properties of all selected test compounds.

In Vitro Inhibition of AChE

Forsythoside A (**11**) and (+)-sesamolin (**25**) were tested for their AChE-inhibiting activity by using a spectrophotometric enzyme assay with Ellman's reagent [38]. Compounds **11** and **25** showed dose-dependent and long-lasting inhibitory effects on AChE with IC_{50} of 39.5 ± 8.5 μg/mL (63.2 ± 13.6 μM) and 44.1 ± 11.7 μg/mL (119.0 ± 31.6 μM), respectively. For comparison, the clinically used AChE inhibitor galanthamine showed an IC50 of 3.2 ± 1.0 μM. The data are presented in Fig. (7).

Additionally, a dichloromethane crude extract of propolis was prepared and analyzed by LC-DAD-MS to check whether the virtually predicted metabolites (**18**, **20**, **26**, and **27**) could be identified as constituents of the extract under investigation. Molecular masses and UV-spectra corresponding to those of chalcons **18** and **20** were assigned to the main constituents of the investigated propolis extract [39]; however, the molecular masses of **26** and **27** could not be identified doubtlessly in the extract. In the enzyme test, this propolis extract showed a significant AChE-inhibiting effect with an IC_{50} of 152.2 ± 48.7 μg/mL. Thus, we assume compounds **18** and **20** being the secondary metabolites responsible for the measured effect. However, more detailed phytochemical efforts, including isolation and structure elucidation, are necessary to verify this assumption.

CONCLUSION

In this study, we successfully applied ligand- and structure-based virtual screening methods to identify novel acetylcholinesterase (AChE) inhibitors from natural sources. The complementary use of different approaches enabled us to prioritize hits for biological testing. Additionally, the molecular descriptors used for model building and virtual screening reflect well the binding peculiarities from the AChE binding pocket. Intriguingly, the virtual screening methods (novelty detection and counterpropagation neural networks) returned no common top-ranked compounds although they were trained with the same active compounds and the same set of descriptors. However, both approaches led to the identification of biologically active compounds/mixtures (propolis). This fact points towards the need to employ different, complementary virtual screening techniques to fully exploit a compound database for active hits.

EXPERIMENTAL SECTION

Structure Representation and Descriptor Calculation

Descriptors were calculated based on the 3D structure of compounds computed by CORINA [22]. ADRIANA.Code 2.0 [23] was used to represent the molecules in this dataset at different levels of sophistication. The simplest descriptors used within this hierarchical structure representation scheme are global molecular properties such as the molecular dipole moment μ. Topological autocorrelation vectors reflect the constitution of molecules. Descriptors considering interatomic distances rely on the 3D structure of molecules.

CORINA [22] is included in ADRIANA.Code 2.0 and generates low-energy 3D conformations. Conformational flexibility was not considered within this study.

Topological autocorrelation vectors [40] were calculated according to eq. 2 considering a set of different atomic properties (identity, σ-, π-, total-charge, σ-, π-, lone-pair electronegativity, effective atom polarizability).

$$a(d) = \frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{N} p_j p_i \delta\left(d_{t,ij} - d\right) \quad \delta = \begin{cases} 1 \forall d_{t,ij} & = d \\ 0 \forall d_{t,ij} & \neq d \end{cases} \quad \text{Eq.(2)}$$

Here, $d_{t,ij}$ is the topological distance between atoms i and j (i.e. the number of bonds for the shortest path in the structure diagram), N is the number of atoms in the molecule, and p_i and p_j are properties of atoms i and j, respectively.

Spatial autocorrelation vectors were computed for the same set of properties mentioned before according to eq. 3.

$$A(d_l, d_u) = \sum_{i=1}^{N}\sum_{j=1}^{N} \delta\left(d_{ij}, d_l, d_u\right) p_j p_i$$

with

$$\delta\left(d_{ij}, d_l, d_u\right) = \begin{cases} 1 & \forall d_l < d_{ij} \leq d_u \\ 0 & \forall d_{ij} \leq dl \vee d_{ij} > d_u \end{cases} \quad \text{Eq.(3)}$$

where p_i and p_j are the atomic properties of atoms i and j possessing an Euclidian distance d within the boundaries d_l (lower) and d_u (upper), and N is the number of atoms within the molecule. Surface autocorrelation vectors can be derived using eq. 3. In this case p_i and p_j are the value of the potential of points i and j lying on the surface and d_{ij} is their distance.

For an ensemble of atoms, the RDF-code can be interpreted as the probability distribution of the individual interatomic distances r regarding the respective atomic properties:

$$g(r) = \sum_{i=1}^{N-1}\sum_{j>i}^{N} p_i p_j e^{-B\left(r - r_{ij}\right)^2} \quad \text{Eq.(4)}$$

where p_i and p_j are the atomic properties of atoms i and j, N is the number of atoms within the molecule, r_{ij} the distance between atoms i and j, and B is a smoothing factor [41, 42].

Self-Organizing Neural Networks and Counterpropagation Neural Networks

Self-organizing neural networks are well suited to project high-dimensional objects to a two-dimensional plane. The projection preserves the topology of the high-dimensional space. Therefore, this method can be used for similarity perception. Self-organizing maps were introduced by Kohonen [24] and their application in drug design and in industrial pharmaceutical research was reviewed in several publications [9,43,44]. SONNIA [25] version 4.2 was used within this study to train the SOM and counter propagation neural networks [32,45-47].

Novelty Detection

Novelty detection refers to a class of machine learning techniques which attempt to identify patterns that do not belong to the space covered by a given dataset [48, 49]. An implementation of novelty detection based on SOMs has been recently successfully applied to ligand-based virtual screening [50, 51]. These methods have led to the development of the software LISA (LIgand-based virtual Screening Application) for virtual screening [52]. In this work, the virtual screening protocol described in reference [50] was used and two experiments were performed.

SPINNE Plot

The program SPINNE [29] can be used for the visualization of multi-dimensional data on a 2-dimensional plane using NLM [30]. Generally, this dimension reduction induces inaccuracies: Two points that are displayed close to one another on the 2D plot might not necessarily be neighbors in the original high-dimensional descriptor space. In order to reduce misinterpretations of the 2D projection, a MST [31] is superimposed onto the map. This MST was calculated in the high dimensional space (all 47 descriptors) using Euclidean distance metric. It was displayed using the 2D coordinates identified by NLM. The MST correctly connects points that are nearest neighbors in the high-dimensional space. In order to reflect the Euclidean distance in the high-dimensional space, the plots generated by SPINNE display colored connecting lines using a continuous color scale. This scale ranges from red for large distances to violet for the smallest ones.

Docking

Complementary to the ligand-based virtual screening methods, docking was used to aid in the selection of biological testing material. GOLD 3.1 [53] was employed to dock the hits from the combined hitlist into the AChE active site gorge using the PDB entry 1w6r. The ligands and the protein were prepared according to reference [54]. Visual inspection and investigation of the docked binding orientations were performed by importing the docked compounds into LigandScout 2.0 [27] and automatically determining all suggested protein-ligand interactions.

Selection of Virtual Hits for Biological Testing

A combined hitlist was built from the hitlists from the two different screening approaches. For all top-ranked hits from these predictions, a literature search was performed collecting information on herbal origin, already known activity on acetylcholinesterase, pharmacological activity, and phytochemical isolation/commercial availability. General guidelines for selecting virtual hits from natural sources for biological evaluation have been described [2, 3].

AChE Enzyme Assay

The AChE inhibitory activity was determined using a spectrophotometric method with Ellman’s reagent in a 96-well microplate assay as previously described [38, 55]. AChE derived from electric eel (test concentration 0.022 U/mL each), acetylthiocholine iodide, and 5,5′-dithiobis-(2-nitrobenzoic acid) were purchased from Sigma-Aldrich Chemie GmbH; positive control: galanthamine.HBr (Tocris Cookson Ltd.).

Origin of Natural Materials

Forsythoside A was purchased from Advanced Technology and Industrial Co., Ltd, Tai Kok Tsui, Kln., Hong Kong. (+)-Sesamolin was derived from Fine Chemicals, Industrial Research Ltd., Gracefield, New Zealand. Propolis was purchased from a local distributor. A voucher specimen is deposited in the Herbarium of the Institute of Pharmacy/ Pharmacognosy, University of Innsbruck, Austria.

Statistical Analysis

The percentage of the enzyme inhibition was calculated by comparing the rates for the sample to the blank (containing 1% DMSO; $n = 4$) and analyzed with the Student's t-test. The IC_{50} values were determined with Probit analysis. For statistical processing, the SPSS 11.5 program package was used.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was granted by a young talents grant of the Leopold-Franzens-University Innsbruck (J.M.R.) and by the Austrian Science Fund (FWF, project No. P18379). We also thank Philipp Schuster for assistance with preparing the manuscript.

REFERENCES

[1]. Newman DJ, Cragg GM. Natural products as sources of new drugs over the last 25 years. J. Nat. Prod. 2007; 70:461–477. [PubMed: 17309302]

[2]. Rollinger JM, Langer T, Stuppner H. Strategies for efficient lead structure discovery from natural products. Curr. Med. Chem. 2006; 13:1491–1507. [PubMed: 16787200]

[3]. Rollinger JM, Langer T, Stuppner H. Integrated in silico tools for exploiting the natural products' bioactivity. Planta Med. 2006; 72:671–678. [PubMed: 16783689]

[4]. Rollinger, JM.; Stuppner, H.; Langer, T. Natural Compounds as Drugs. Frank, P.; Amstutz, R., editors. Vol. 1. Birkhäuser Verlag; Basel: 2008. p. 213-249.

[5]. Ekins S, Mestres J, Testa B. *In silico* pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br. J. Pharmacol. 2007; 152:9–20. [PubMed: 17549047]

[6]. Ekins S, Mestres J, Testa B. *In silico* pharmacology for drug discovery: applications to targets and beyond. Br. J. Pharmacol. 2007; 152:21–37. [PubMed: 17549046]

[7]. Kirchmair J, Distinto S, Schuster D, Spitzer G, Langer T, Wolber G. Enhancing drug discovery through *in silico* screening: strategies to increase true positives retrieval rates. Curr. Med. Chem. 2008; 15:2040–2053. [PubMed: 18691055]

[8]. Langer, T.; Hoffmann, RD. Pharmacophores and Pharmacophore Searches. Wiley-VCH; Weinheim: 2006.

[9]. Gasteiger J, Teckentrup A, Terfloth L, Spycher S. Neural networks as data mining tools in drug design. J. Phys. Org. Chem. 2003; 16:232–245.

[10]. Perry EK, Tomlinson BE, Blessed G, Bergmann K, Gibson PH, Perry RH. Correlation of cholinergic abnormalities with senile plaques and mental test scores in senile dementia. Br. Med. J. 1978; 2:1457–1459. [PubMed: 719462]

[11]. Silman I, Sussman JL. Acetylcholinesterase: "classical" and "non-classical" functions and pharmacology. Curr. Opin. Pharmacol. 2005; 5:293–302. [PubMed: 15907917]

[12]. Zhang X. Cholinergic activity and amyloid precursor protein processing in aging and Alzheimer's disease. Curr. Drug Targets CNS Neurol. Disord. 2004; 3:137–152. [PubMed: 15078189]

[13]. Schumacher M, Camp S, Maulet Y, Newton M, MacPhee-Quigley K, Taylor SS, Friedmann T, Taylor P. Primary structure of torpedo californica acetylcholinesterase deduced from its cDNA sequence. Nature. 1986; 319:407–409. [PubMed: 3753747]

[14]. Axelsen PH, Harel M, Silman I, Sussman JL. Structure and dynamics of the active site gorge of acetylcholinesterase: synergistic use of molecular dynamics simulation and X-ray crystallography. Protein Sci. 1994; 3:188–197. [PubMed: 8003956]

[15]. Sussman JL, Harel M, Frolow F, Oefner C, Goldman A, Toker L, Silman I. Atomic structure of acetylcholinesterase from Torpedo californica: a prototypic acetylcholine-binding protein. Science. 1991; 253:872–879. [PubMed: 1678899]

[16]. Bourne Y, Taylor P, Marchot P. Acetylcholinesterase inhibition by fasciculin: crystal structure of the complex. Cell. 1995; 83:503–512. [PubMed: 8521480]

[17]. Choudhary MI, Nawaz SA, Zaheer-ul-Haq, Azim MK, Ghayur MN, Lodhi MA, Jalil S, Khalid A, Ahmed A, Rode BM, Atta-ur-Rahman, Gilani A.-u.-H. Ahmad VU. Juliflorine: a potent natural peripheral anionic-site-binding inhibitor of acetylcholinesterase with calcium-channel blocking potential, a leading candidate for Alzheimer's disease therapy. Biochem. Biophys. Res. Commun. 2005; 332:1171–1179. [PubMed: 16021692]

[18]. Wong DM, Greenblatt HM, Dvir H, Carlier PR, Han Y-F, Pang Y-P, Silman I, Sussman JL. Acetylcholinesterase complexed with bivalent ligands related to huperzine A: Experimental evidence for species-dependent protein-ligand complementarity. J. Am. Chem. Soc. 2003; 125:363–373. [PubMed: 12517147]

[19]. Rydberg EH, Brumshtein B, Greenblatt HM, Wong DM, Shaya D, Williams LD, Carlier PR, Pang Y-P, Silman I, Sussman JL. Complexes of alkylene-linked tacrine dimers with torpedo californica acetylcholinesterase: Binding of bis(5)-tacrine produces a dramatic rearrangement in the active-site gorge. J. Med. Chem. 2006; 49:5491–5500. [PubMed: 16942022]

[20]. Rollinger JM, Steindl TM, Schuster D, Kirchmair J, Anrain K, Ellmerer EP, Langer T, Stuppner H, Wutzler P, Schmidtke M. Structure-based virtual screening for the discovery of natural inhibitors for human rhinovirus coat protein. J. Med. Chem. 2008; 51:842–851. [PubMed: 18247552]

[21]. Gasteiger J. Of molecules and humans. J. Med. Chem. 2006; 49:6429–6434. [PubMed: 17064061]

[22]. Corina. Version 3.0. Molecular Networks GmbH; Erlangen, Germany: www.mol-net.com

[23]. ADRIANA.Code. Molecular Networks GmbH; Erlangen, Germany: www.mol-net.com

[24]. Kohonen T. Self-organized formation of topologically correct feature maps. Biol. Cybern. 1982; 43:59–69.

[25]. SONNIA. Molecular Networks GmbH; Erlangen, Germany: www.mol-net.com

[26]. Greenblatt HM, Guillou C, Guénard D, Argaman A, Botti S, Badet B, Thal C, Silman I, Sussman JL. The complex of a bivalent derivative of galanthamine with Torpedo acetylcholinesterase displays drastic deformation of the active-site gorge: implications for structure-based drug design. J. Am. Chem. Soc. 2004; 126:15405–15411. [PubMed: 15563167]

[27]. Wolber G, Langer T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J. Chem. Inf. Model. 2005; 45:160–169. [PubMed: 15667141]

[28]. Fix E, Hodges JLJ. Discriminatory analysis - nonparametric discrimination: consistency properties. Int. Stat. Rev. 1989; 57:238–247.

[29]. Bienfait B, Gasteiger J. Checking the projection display of multivariate data with colored graphs. J. Mol. Graph. Mod. 1997; 15:203–215.

[30]. Sammon JWJ. A nonlinear mapping for data structure analysis. IEEE Trans. Comput. 1969; C-18:401–409.

[31]. Gower JC, Ross JS. Minimum spanning trees and single linkage cluster analysis. Appl. Stat. 1969; 18:54–64.

[32]. Zupan, J.; Gasteiger, J. Neural Networks in Chemistry and Drug Design. 2nd ed. Wiley-VCH; Weinheim: 1999.

[33]. Chow Y-L, Erdtman H. Totarolone, a new diterpene oxophenol from Tetraclinis articulata. Acta Chem. Scand. 1960; 14:1852–1853.

[34]. Forsythia. www.drugs.com/npp/forsythia.html

[35]. Ozaki Y, Rui J, Tang YT. Antiinflammatory effect of forsythia suspensa vahl and its active principle. Biol. Pharm. Bull. 2000; 23:265–367. [PubMed: 10706400]

[36]. Hou RC-W, Chen H-L, Tzen JTC, Jeng K-CG. Effect of sesame antioxidants on LPS-induced NO production by BV2 microglial cells. Neuroreport. 2003; 14:1815–1819. [PubMed: 14534426]

[37]. Jeng KCG, Hou RC-W. Sesamin and sesamolin: nature's therapeutic lignans. Curr. Enzyme Inhibit. 2005; 1:11–20.

[38]. Ellman GL, Courtney KD, Andres VJ, Featherstone RM. A new and rapid colorimetric determination of acetylcholinesterase activity. Biochem. Pharmacol. 1961; 7:88–95. [PubMed: 13726518]

[39]. Bankova V, Popov S, Bocari G, Haxhialushi E. Phenolics in Albanian poplar buds and their relationship to propolis. Fitoterapia. 1994; 65:326–330.

[40]. Terfloth L, Bienfait B, Gasteiger J. Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. J. Chem. Inf. Model. 2007; 47:1688–1701. [PubMed: 17608404]

[41]. Hemmer MC, Steinhauer V, Gasteiger J. Deriving the 3D structure of organic molecules from their infrared spectra. Vib. Spectrosc. 1999; 19:151–164.

[42]. Hemmer MC, Gasteiger J. Prediction of three-dimensional molecular structures using information from infrared spectra. Anal. Chim. Acta. 2000; 420:145–154.

[43]. Selzer P, Ertl P. Applications of self-organizing neural networks in virtual screening and diversity selection. J. Chem. Inf. Model. 2006; 46:2319–2323. [PubMed: 17125175]

[44]. Terfloth L, Gasteiger J. Neural networks and genetic algorithms in drug design. Drug Discov. Today. 2001; 6:102–108. [PubMed: 11166258]

[45]. Hecht-Nielsen R. Counterpropagation networks. Appl. Opt. 1987; 26:4979–4984. [PubMed: 20523476]

[46]. Hecht-Nielsen R. Applications of counterpropagation networks. Neural Netw. 1988; 1:131–139.

[47]. Zupan J, Novic M, Gasteiger J. Neural networks with counter-propagation learning strategy used for modelling. Chemom. Intell. Lab. Syst. 1995; 27:175–187.

[48]. Markou M, Singh S. Novelty detection: a review - part 1: statistical approaches. Signal Process. 2003; 83:2481–2497.

[49]. Markou M, Singh S. Novelty detection: a review - part 2: neural network based approaches. Signal Process. 2003; 83:2499–2521.

[50]. Hristozov D, Oprea TI, Gasteiger J. Ligand-Based Virtual Screening by Novelty Detection with Self-Organizing Maps. J. Chem. Inf. Model. 2007; 47:2044–2062. [PubMed: 17854167]

[51]. Hristozov D, Oprea T, Gasteiger J. Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. J. Comput. Aided Mol. Des. 2007; 21:617–640. [PubMed: 18008169]

[52]. LISA. Molecular Networks GmbH; Erlangen, Germany: Development prototype for novelty detection and similarity searches

[53]. GOLD 3.1. The Cambridge Crystallographic Data Centre - CCDC; Cambridge:

[54]. Schuster D, Maurer E, Laggner C, Nashev LG, Wilckens T, Langer T, Odermatt A. The discovery of new 11b-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening. J. Med. Chem. 2006; 49:3454–3466. [PubMed: 16759088]

[55]. Rollinger JM, Hornick A, Langer T, Stuppner H, Prast H. Acetylcholinesterase inhibitory activity of scopolin and scopoletin discovered by virtual screening of natural products. J. Med. Chem. 2004; 47:6248–6254. [PubMed: 15566295]

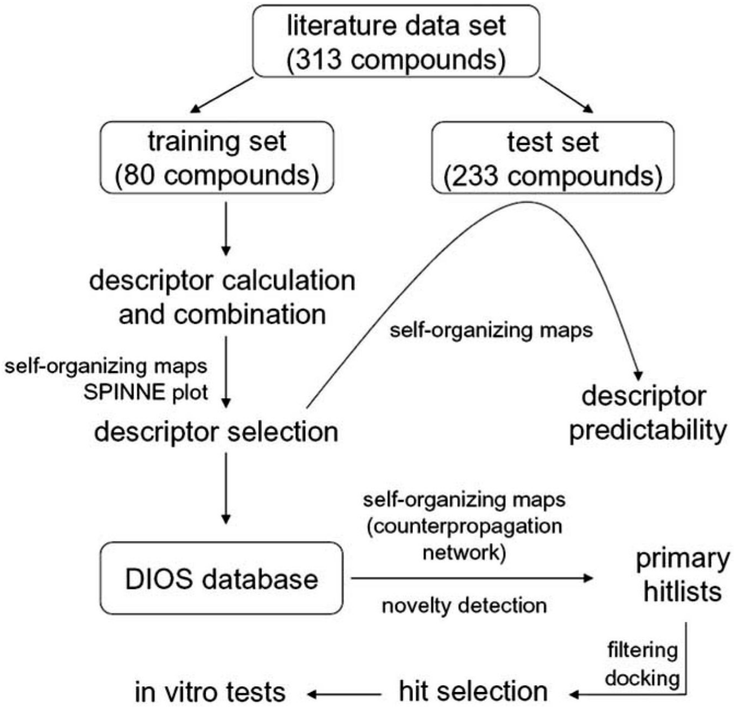


Fig. (1).

Workflow followed for the identification of novel AChE inhibitors.

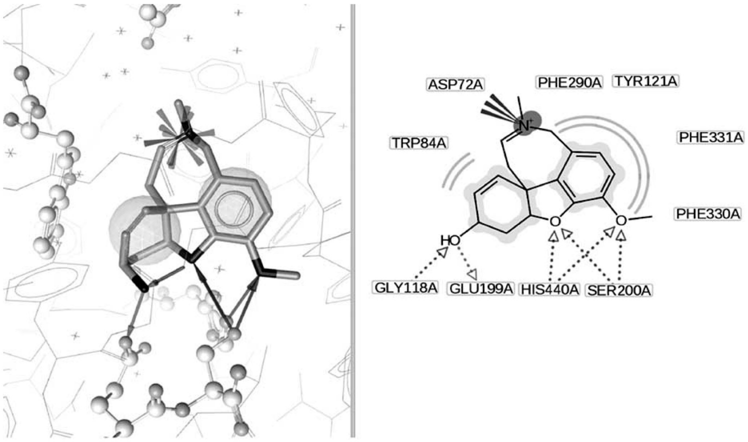


Fig. (2).
Chemical interactions of a highly active galanthamine derivative in *Tc* AChE as determined by LigandScout. Left: 3D Visualization of protein-ligand interactions. Right: 2D visualized interactions of the ligand to the surrounding amino acids; chemical features: positively ionized – star; hydrophobic – spheres; hydrogen bonds - arrows.

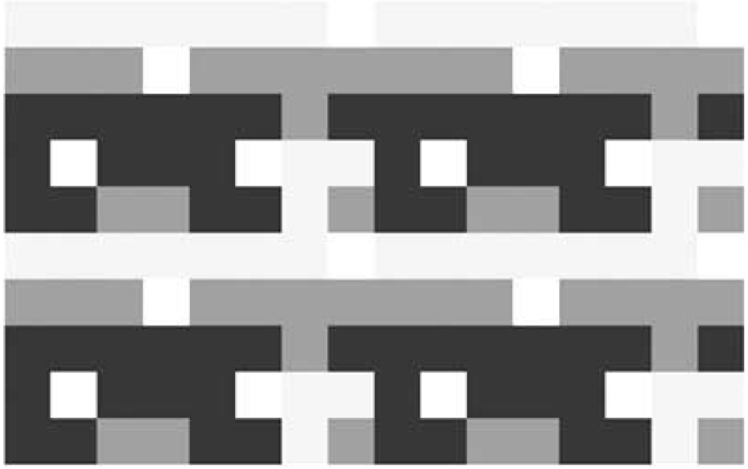


Fig. (3).

SOM of the training set (80 compounds). A toroidal network was chosen, i.e. the plane of projection consisted of the surface of a torus. In order to reflect that this surface has no beginning and no end, four identical maps were put together like tiles. Activity classes are color-coded: highly actives – light grey; medium actives – grey; inactives – dark grey. Highly active compounds are clustered into a hook-shaped area, which is surrounded by an area of medium activity dropping into a pool of inactives.

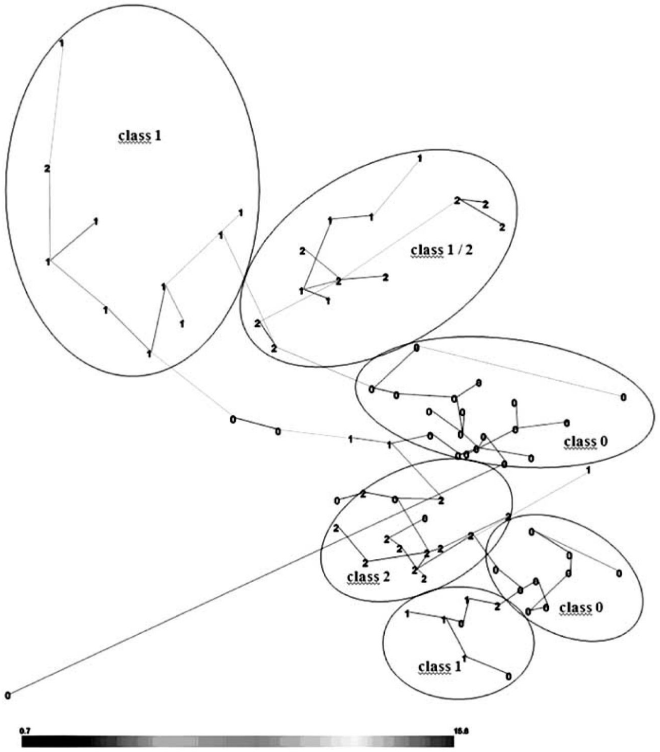


Fig. (4).

SPINNE 2D representation of the training set compounds using scaled descriptors described in Table 1. The lines connect nearest neighbors based on the similarity of the compounds in the original descriptor space. Each point is represented by its activity class: class 0 – inactive; class 1 – medium active; class 2 – highly active. Activity clusters are highlighted. A colored version is available as Supplementary Material.

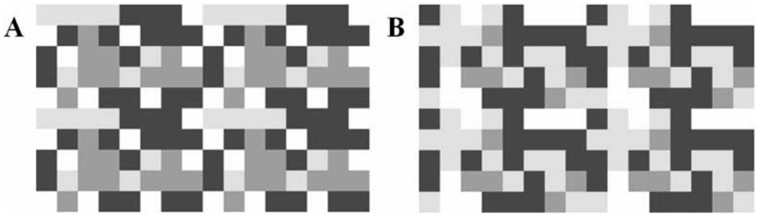


Fig. (5).

A: Tiled SOM of the training set (scaled descriptors) (80 compounds). (For explanation of tiling see Fig. 3) Activity classes are color-coded: highly actives – light grey; medium actives – grey; inactives – dark grey. **B**: Tiled SOM of the test set (233 compounds). In both SOMs, the clustering of compounds from the same activity classes is clearly visible.

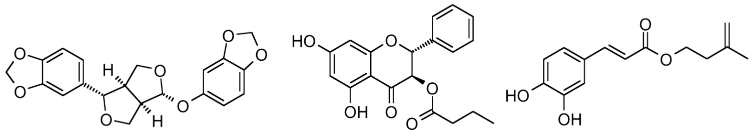


Fig. (6).

Virtual hits selected for biological testing (compounds **14** and **16** are shown in Table **5**).

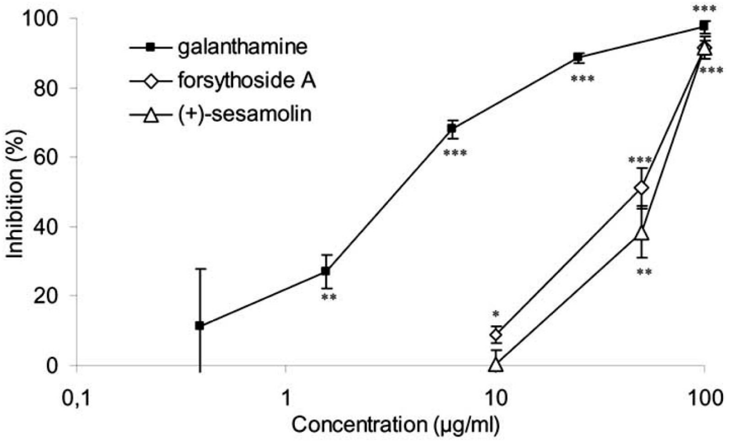


Fig. (7).

AChE enzyme assay: inhibitory effects of the reference compound galanthamine, forsythoside A (**7**), and (+)-sesamolin (**21**) on AChE [a].

[a] Statistical analysis: *** p<0.001, ** p<0.01, * p<0.05 Student's test of absorption data after 30 min in comparison with medium control, n = 4.

Table 1

List of Descriptors Used for SOM: Global Molecular Properties, Atom Properties and Surfaces

Descriptor Abbreviation	Descriptor Name	Dimensionality
RDF_ident	radial distribution function using the property identity	34
DIPOLE_M	dipole moment	1
AC3D_qpi	spatial autocorrelation using the property π charge	12

Table 2

knn-Analysis of the SOM Shown in Fig. (3)

Class	Correct	Drawn	Wrong	Classification Accuracy [%]
0 – inactives	29	3	4	80.6
1 – medium actives	17	0	4	80.9
2 – highly actives	17	0	6	73.9
Total	**63**	**3**	**14**	**78.7**

Table 3

Confusion Matrix Comparing the Activities of the Test Set Compounds to their Prediction in the SOM Map

	Class 0 - Inactives	Class 1 – Medium Actives	Class 2 – Highly Actives	Sum	Correct Predictions
predicted as class 0	46	15	28	89	51.7 %
predicted as class 1	29	30	33	92	32.6 %
predicted as class 2	1	4	13	18	72.2 %
empty neurons	17	7	10	34	
∑	93	56	84	233	

Table 4

Top-Ranked Virtual Hits from the External Dataset (DIOS) Identified by Novelty Detection

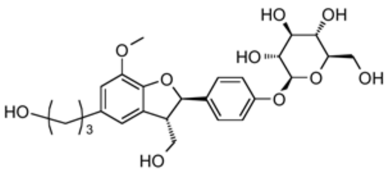
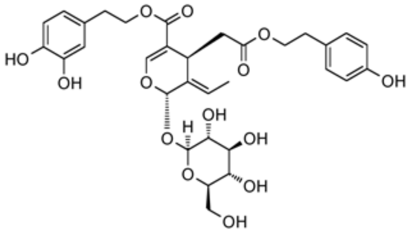
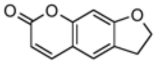
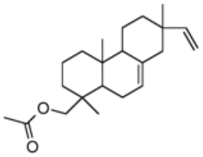
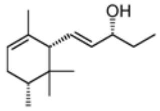
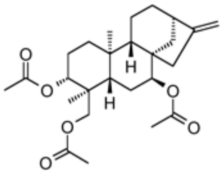
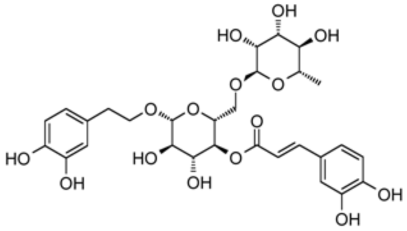
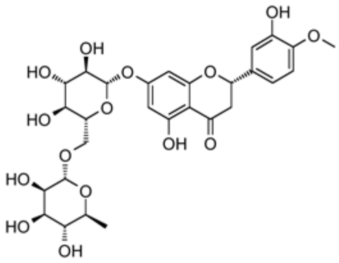
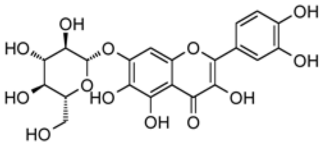
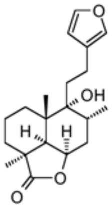
 rank 1: 4-O-β-glucopyranoside dihydrodehydrodiconiferyl alcohol (**1**)	 rank 2: hydroxyframoside B (**2**)
 rank 3: 4',5'-dihydropsoralen (**3**)	 rank 4: isopimarol acetate (**4**)
 rank 5: CAS124988-53-0 (**5**)	 rank 6: triacetylfolilol (**6**)
 rank 7: forsythoside A (**7**)	 rank 8: hesperidine (**8**)
 rank 9: quercetagitrin (**9**)	 rank 10: marrubin (**10**)

Table 5

Top-Ranked Virtual Hits from the External Dataset (DIOS) Identified by Counterpropagation Network

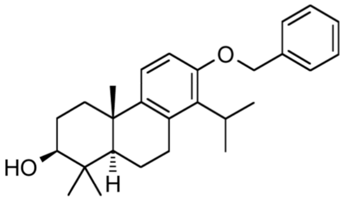
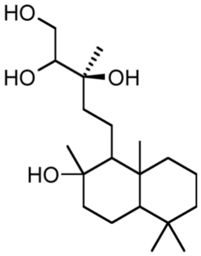
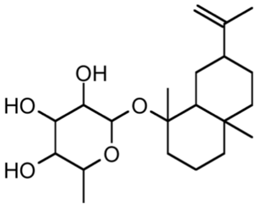
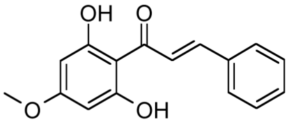
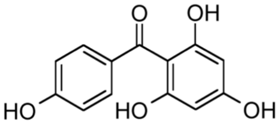
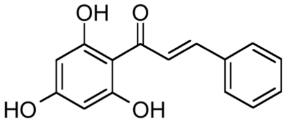
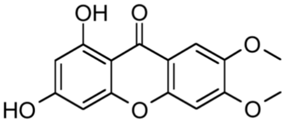
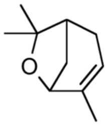
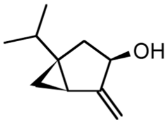
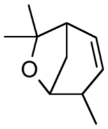
 rank 1: CAS107631-55-0 (**11**)	 rank 2: CAS58939-84-7 (**12**)
 rank 3: CAS132237-56-0 (**13**)	 rank 4: pinostrobin chalcone (**14**)
 rank 5: iriflophenone (**15**)	 rank 6: pinocembrin chalcone (**16**)
 rank 7: laxanthone I (**17**)	 rank 8: pinol (**18**)
 rank 9: trans-sabinol (**19**)	 rank 10: dehydro-1,8-cineole (**20**)

Table 6

Properties of Test Compounds

Compound	Nov.Det. Rank	CPG Rank	Docking Score (GoldScore)
7	9	not in top 100	36.63
14	not in top 100	4	52.98
16	not in top 100	6	53.26
21	14	not in top 100	60.71
22	not in top 100	45	70.08
23	33	not in top 100	51.26