Artificial Intelligence Approaches for Rational Drug Design and Discovery

Włodzisław Duch[1,2,*], Karthikeyan Swaminathan[3] and Jarosław Meller[1,3]

[1]*Department of Informatics, Nicolaus Copernicus University, Grudziądzka 5, 87100 Toruń, Poland;* [2]*School of Computer Engineering, Nanyang Technological University, Singapore and* [3]*Division of Biomedical Informatics, Children's Hospital Research Foundation, 3333 Burnet Avenue, Cincinnati, OH 45242, USA*

Abstract: Pattern recognition, machine learning and artificial intelligence approaches play an increasingly important role in rational drug design, screening and identification of candidate molecules and studies on quantitative structure-activity relationships (QSAR). In this review, we present an overview of basic concepts and methodology in the fields of machine learning and artificial intelligence (AI). An emphasis is put on methods that enable an intuitive interpretation of the results and facilitate gaining an insight into the structure of the problem at hand. We also discuss representative applications of AI methods to docking, screening and QSAR studies. The growing trend to integrate computational and experimental efforts in that regard and some future developments are discussed. In addition, we comment on a broader role of machine learning and artificial intelligence approaches in biomedical research.

Key Words: QSAR, rational drug design, docking, artificial intelligence, machine learning, pattern recognition, neural networks, support vector regression.

INTRODUCTION

Artificial Intelligence (AI) is defined here in very broad terms as a field that deals with the design and application of algorithms for analysis of, learning from and interpreting data. Thus, broadly defined AI encompasses many branches of statistical and machine learning, pattern recognition, clustering, similarity-based methods, logics and probability theory, as well as biologically motivated approaches, such as neural networks, evolutionary computing or fuzzy modeling, collectively described as "computational intelligence" [1-5]. Typical applications of AI methods involve selection of relevant information, data modeling, classification and regression, optimization and prediction. In this review, we focus on those aspects of AI methodology that are relevant for drug design and discovery.

Many AI problems involve capturing complex relations between relevant descriptors (attributes used to represent objects of interest or similarities between these objects), and observed outcomes (e.g. biological activity of a chemical compound [6]). Multiple examples of such relationships are often available as a result of experimental studies, e.g., on complex molecular systems and processes. The advantage of AI approaches is that they can be applied to learn from examples and develop predictive models even when our understanding of the underlying molecular processes is limited, or when computational simulations based on fundamental physical models are too expensive to carry out (for general overview of machine learning in science see, e.g. [7]). Canonical examples of successful applications of statistical and machine learning methods include gene prediction from the primary DNA sequence [8-10], and prediction of secondary structures from the amino acid sequence [11].

In the realm of drug design, AI techniques are being used to classify candidate compounds in terms of their activity and other properties. For example, a predictive model capable of approximating the strength of binding for candidate molecules may be developed based on experimentally measured binding affinities for a range of substrates. Such predictors can subsequently be used for fast *in silico* screening and identification of potential drugs with desired properties [12]. Moreover, AI methods are also being used to identify informative biomarkers that correlate with tested outcomes. Food and Drug Administration (FDA) allows for monitoring the effectiveness of drugs using such biomarkers. For example, changes in cholesterol levels may be used to measure the effectiveness of some drugs because low cholesterol level correlates well with a healthy cardiovascular system. On the other hand, traditional drug development process takes many years and is very costly, relying on clinical evaluation of efficacy and safety of new drugs, their influence on clinical symptoms and mortality rates.

The structure of the review is as follows. In the next section, general concepts pertaining to machine learning and more broadly AI are introduced, with focus on learning from examples with known outcomes (supervised learning). Several AI approaches that are widely used in the field of drug design, including regression and classification problems and techniques to solve them, are briefly revisited. In the third section, the issues of finding appropriate representation of the problem, model complexity as well as understanding and interpretation of the results of data analysis are discussed. In the next part, an overview of applications of AI methods to QSAR, docking and other problems in drug

*Address correspondence to this author at the Department of Informatics, Nicolaus Copernicus University, Grudziądzka 5, 87100 Toruń, Poland; E-mail: wduch@is.umk.pl

1381-6128/07 $50.00+.00

design and discover is given, followed by discussion of future directions and challenges in the field.

OVERVIEW OF AI APPROACHES

There are many alternative methods and formulations for learning a predictor from data and for other relevant applications of AI methods considered here. Moreover, there is a variety of existing implementations and software packages that can be applied, e.g., when solving specific data mining and analysis problems in the context of drug design and discovery. Consequently, it is often difficult (especially for a non-specialist) to assess the usefulness and limitations of a particular method for the problem at hand. One of the goals of this review is to provide the reader with a conceptual and practical framework to better navigate this field. We start by informally introducing some central concepts, including supervised and unsupervised learning, classification and regression, as well as feature selection and aggregation. Next, a brief non-technical introduction to selected AI methods is provided, with emphasis on underlying ideas, advantages and limitations of different approaches.

In order to introduce some basic concepts, let us assume that our goal is to predict various characteristics of candidate compounds, such as their toxicity or affinity for binding to their targets. These characteristics will be represented by "target variables", with values indicating the type (class) or continuous attributes of candidate compounds. Specifically, if the target variable to be predicted has a few symbolic or numerical values the problem is of the *classification* type, and if the target value is continuous or has many numerical values the problem is of the *regression* type. If each data sample is given a label or has an associated target value, then *supervised learning* techniques for classification or regression can be used to develop a predictor. If no such information is available for known data samples, then *unsupervised learning* techniques are used to discover interesting structures in data, e.g., clusters or patterns. The essence of supervised learning approach, which is the focus of this review, is to learn from known examples in order to subsequently make predictions for new instances of data. In the case of classification problems, the training examples are assigned class labels (e.g. active vs. inactive for a chemical compound considered as a potential drug) and the task is to train a system that can be used to classify new data points.

Another crucial consideration is the choice of an appropriate *representation* of the problem at hand. The most common representation is based on a vector **X** of numerical and/or symbolic values X_i of a set of attributes (or features) that describe objects (such as molecules or amino acid residues) or states to be classified. If large amount of information is provided, then relevant features should be first identified (*feature selection*) or extracted (*feature aggregation* or transformation). For example, an amino acid residue in a protein may be represented in terms of its physico-chemical properties or one can take into account the evolutionary context of this residue, e.g., by considering patterns of substitutions at that position in homologous proteins [11].

Most pattern recognition methods of classification and regression are formulated assuming a common vector space representation of all input attributes. It is important to realize, however, that many problems in drug design cannot be naturally represented using a common vector space. In particular, the molecules considered may be quite diverse, implying vectors of different size. Therefore, alternative approaches have been proposed that rely on similarities between classified molecules or their structural descriptor, e.g., defined in terms of trees or graphs to capture structural relations between chemical groups or individual atoms. Although methods based on structural representations are not yet common, in recent years several new approaches with numerous potential applications have been formulated [13-15].

There are many learning algorithms that use different underlying formulations and models to construct a specific discriminatory function and the resulting decision boundaries to classify the data. In general, there are two groups of classification methods: those that partition the "feature space" using hyperplanes (i.e. a plane in high-dimensional spaces) or non-linear surfaces into regions containing data from a single class, and those that use prototypes and similarity evaluations to define such regions (frequently localized around the prototypes). For example, a simple linear "classifier" may be obtained in the "feature space" by finding a hyperplane that separates two classes of training vectors, or, alternatively, the most similar labeled prototypes may be used to classify a new data point. In the following, we present examples of both approaches.

Linear Discriminant Analysis and Support Vector Machines

The name "Linear Discriminant Analysis" (LDA) refers to the one of the oldest groups of classification methods. These methods find a hyperplane in the vector space **X** that separates vectors of one group (e.g., toxic substances) from another group (non-toxic substances). Two-class problems are sufficient in most cases because if more than two classes are defined a single class may be separated from the remaining ones. A hyperplane **W** for N features X_i may be found using many algorithms [3-5], providing combination of feature values:

$$Y = \mathbf{W} \cdot \mathbf{X} + W_0 = \sum_{i=1}^{N} W_i X_i + W_0 > 0$$

for vectors **X** from the first group, and $Y<0$ for vectors of the second group. Results on a new data that have not been used for determining coefficients **W**, will usually be better if the hyperplane decision border **W** is placed as far from the data as possible, increasing the margin between the decision boundary and data points. This idea is explored explicitly in linear support vector machines (SVMs), which may be regarded as generalizations of LDA classifiers and are based on a learning algorithm that selects all vectors close to the decision boundary to "support" the orientation of the hyperplane **W**. When the information contained in features is not

sufficient to achieve linear separability, a tradeoff between the margin of separation and the misclassification error needs to be specified.

An alternative approach is to add more features, either by measuring or calculating new properties, or combining existing features (e.g. taking products or ratios). Extended features spaces may lead to some data becoming linearly separable. A simple trick to avoid explicit consideration of new features is to define a "kernel function" $K(\mathbf{W},\mathbf{X})$ that computes the combination (scalar product) $\mathbf{W}\cdot\mathbf{X}$ in some high-dimensional space. Various classifiers known as "kernel machines", including non-linear versions of SVMs, use this trick to achieve separability in the extended space. The separating hyperplane may subsequently be mapped back into the original feature space in order to obtain a non-linear separating hypersurface (see Fig. **1**). Popular choices for kernel functions are low-order polynomial functions $K(\mathbf{W},\mathbf{X})=(\mathbf{W}\cdot\mathbf{X})^n$ or Gaussian functions $K(\mathbf{W},\mathbf{X})=\exp(-|\mathbf{W}-\mathbf{X}|^2)$. More sophisticated kernels, reflecting some prior knowledge about the specific problem analyzed, may be designed [16]. Due to their efficiency and overall excellent performance, such classifiers have achieved great popularity in recent years [17].

Neural Networks

Another solution that goes beyond LDA is provided by neural networks (NNs). NNs can generate arbitrary non-linear decision boundaries by addition of many simple functions. This is typically achieved by a multi-stage transformation, which may be represented graphically as a network (directed graph) of interconnected layers of "computing" nodes that integrate input signals from previous layers. In particular, the input features (attributes), X_i , are represented by individual nodes in the input layer and are subsequently transformed into a new set of features using several hyperplanes, $\mathbf{W}_k$, corresponding to the hidden layer nodes (here for simplicity we assume that only one hidden layer is used). In other words, the inputs for the hidden layer nodes are linear combinations of the original N inputs X_i, with the coefficients of the linear combination, W_{ik}, associated with connections between the input node i and hidden layer node k. The hidden layer nodes transform these signals further, using some functions $h_k(\mathbf{X})=\sigma(\mathbf{W}_k\cdot\mathbf{X}+W_{k0})$, where the scalar functions $\sigma(x)$ are usually chosen to be logistic functions i.e. they have a sigmoidal shape with output bounded by maximum and minimum values. As a result, the outputs are in general non-linear functions of inputs (see Fig. **2**).

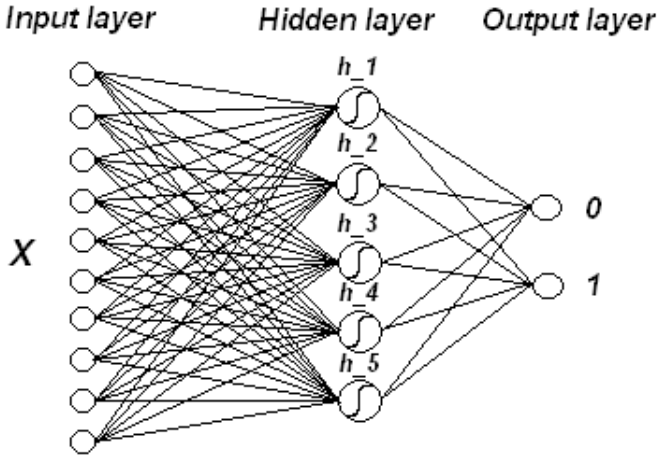

Fig. (2). An example of a multilayer perceptron with one hidden layer and a binary output layer for two-class classification problems.

There is a distant analogy between NNs and the activity of biological neurons that sum input signals weighted by the strength of synaptic connections and send output signals that are bounded by some maximum values. For that reason NNs' nodes are called artificial neurons or perceptrons. A number of these nodes connected to the same input form a layer that transforms the input vector **X** to the vector of hidden layer activities **H**. NNs with sigmoidal-shape activation functions are called "multilayered perceptrons" (MLP). In general, neural networks are basically function-mapping systems for classification and regression that can learn how to associate numerical inputs with arbitrary outputs, changing their internal parameters. Many training algorithms have been devised to find parameters **W** that fit inputs **X** to the desired outputs [3-5].

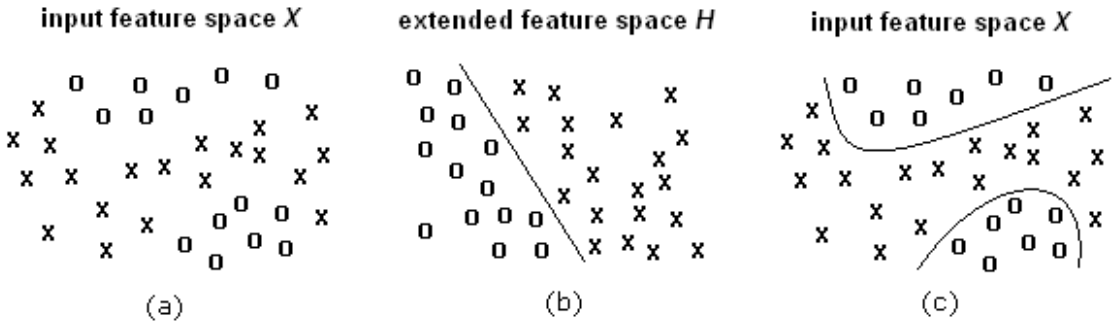

Fig. (1). Kernel-based approaches for classification: a) data distribution in the feature space X where no separating hyperplane exists; b) data distribution in the extended feature space H with one of the possible separating hyperplanes shown; c) non-linear decision boundary in the original feature space.

Self-Organizing Maps

Another type of neural network has been inspired by the self-organizing developmental processes that lead to the topographical organization of auditory, visual and somatosensory cortex [1-2]. The Self-Organized Mapping (SOM, also called the Kohonen network) is a grid with nodes that adjust their parameters to the incoming data. After the presentation of a vector **X** to node *i*, with parameters $\mathbf{W}_i$, the vector that is most similar to **X** is selected as a "winner" representing node *i*. At the same time, parameters of grid nodes that are close to the winning node *i* are adjusted to make them more similar to $\mathbf{W}_i$. As a result of many presentations of input data, parameters of a whole group of nodes become clustered around the peaks of the data density in the feature space. Therefore, SOM can be regarded as a clustering method and may also be used to visualize the structure of highly dimensional data, as reflected in activations of the nodes in the grid.

Similarity Based Approaches

Kernel-based approaches use implicitly the notion of similarity between objects, e.g., in order to find the decision boundary in the feature space. There is a whole group of methods that use similarity (or dissimilarity, i.e., distance) to known samples (prototypes) directly. For example, the *k*-nearest neighbors (*k*-NN) method [4] compares new cases to all known reference cases (prototypes) and assigns each point to the class represented by the majority of its *k* nearest neighbors, providing flexible decision boundaries. Variants of this method use different similarity functions, e.g., weighing contributions of the reference vectors, selecting and optimizing reference vectors, adjusting cost functions for different types of errors, and other parameters and procedures [18]. Similarity-based models may be used for classification, regression, clustering and to address the problem of missing feature values. They do not require an explicit numerical representation; just the distance matrix $D(\mathbf{X},\mathbf{Y})$ between objects **X**, **Y**. Therefore, they may be used in conjunction with any method for evaluation of similarity in chemical applications [19].

Logical Rules

Similarity based methods may also be used to elucidate the structure of data. In particular, the nearest neighbor approach may be represented in terms of decision rules, e.g., if $D(\mathbf{X},\mathbf{R}_1) < D(\mathbf{X},\mathbf{R}_2)$ then **X** has the same class as the prototype $\mathbf{R}_1$ (and the class of the prototype $\mathbf{R}_2$ otherwise). Alternatively, a threshold rule may be introduced: if the distance $D(\mathbf{X},\mathbf{R}_1) < \rho_1$ then **X** has the same class as $\mathbf{R}_1$. In the latter case, the decision boundaries are localized around the prototype $\mathbf{R}_1$. This type of rules may have strong exploratory power, provided that the number of prototypes is limited and the distance function is rather simple. Thus, finding informative rules may require selecting prototypes and features in order to simplify similarity measures [20]. Decision trees try to achieve the same result splitting the data using subsets or intervals of single feature values, and thus partitioning the feature space into hyperboxes. Each of these boxes has a class label, providing classical prepositional logic rules. Many other methods to extract logical rules from data are reviewed in [21-24]. If the problem has an inherent logical structure they frequently provide the best and the most intuitive results. Furthermore, symbolic information may be handled in a natural way in this framework by using probabilistic data-dependent similarity measures and based on them logical rules. However, mixing features of different types makes the interpretation difficult in some applications.

Graphical Probabilistic Models

Many classification and regression models that transform features characterizing objects to class labels or numerical values ignore the fact that local interdependencies between some features may exist. For example, some features may be first used in order to generate intermediate descriptors, with the goal of improving the final classification. A family of probabilistic graphical models may be used to capture such dependencies in an explicit way, directly reflecting the input data structures [25-26]. In particular, Bayesian networks allow one to incorporate conditional probabilities, with network connections representing variables that determine the probability distribution of a variable associated with a node. Hidden Markov Models (HMMs) provide one example of a successful realization of probabilistic graphical models, with applications in sequence modeling and other areas of bioinformatics [27-28] It should be noted, however, that (as with NNs) the choice of an appropriate topology of the graph and the optimization of parameters (e.g. transition and emission probabilities) may be computationally quite demanding for these models.

Inductive Logic Programming

Structured data may be described in terms of a set of relationships that can be expressed using logical formulas. Inductive Logic Programming (ILP) is a subfield of AI techniques that use the background knowledge to derive logical rules from positive and negative facts stored typically in relational databases. The language of predicate logic allows ILP approach to incorporate complex concepts, e.g., representing all atoms, bonds, and other properties of a molecule. ILP has a great potential, especially for symbolic data. In addition, ILP has interesting connections with graphical models and stochastic grammars and may be combined with statistical approaches [29-30]. Unfortunately, even though ILP can inductively learn (in principle) any function, the space of all logical theories is extremely large, causing problems with the efficiency of learning.

Evolutionary Computing

Many drug design and discovery approaches involve solving global optimization problems. For example, sampling conformational space in order to find the optimal docking structure for a protein substrate involves finding a global maximum of a scoring function. Several biologically-inspired computational intelligence techniques are used to solve such optimization problems. For example, widely used evolutionary methods are based on the concept of "evolving"

improved (according to a fitness or scoring function) solutions from less accurate ones. Thus, a population of "structures" (or, at a deeper level, population of "genes" that control programs creating "structures") is modified mimicking "mutation" and "recombination" processes observed in biological systems. Evolutionary operators that represent these processes may frequently be tailored to a specific problem at hand. A fitness function is defined to evaluate each solution and to select a pool of the fittest (best scoring) solutions, which is subsequently expanded by producing an "offspring" generation [1-2, 31].

Because of their intuitive appeal and ability to solve hard optimization problems, evolutionary computing techniques have become popular among pharmaceutical companies. For example, Axys Pharmaceuticals, Daylight Chemical Information Systems, Nanodesign and Tripos offer products and services that are based on evolutionary algorithms. They are used to fit ligands into protein binding sites, design useful ligands, explore molecular conformations and assess similarity indices for optimal matching. Evolutionary Molecular Design (EMD) developed by Nanodesign, identifies the desired activity of a receptor and searches for the appropriate ligand structures using genetic algorithms.

Support Vector Regression

Regression problems occur very frequently in the field of drug design. In particular, many applications of quantitative structure-activity relationship (QSAR) approaches involve solving (multiple) regression problems. Let us assume that observed binding affinities (or other numerical measures of activity), Y_k^{obs}, are given for a number of ligands represented as vectors, X_k, in some feature space. The problem is to find a mapping Y that approximates the observed affinities, such that the differences between observed, Y_k^{obs}, and predicted values, $Y(X_k)$, are minimized. If the mapping takes the form of a linear combination of input features, $Y(X_k) = \sum W_i X_{ki} + W_0$, and the sum of squared errors between the predicted and observed values is minimized, then one obtains a classical least square (LS) regression problem [5] that may be solved using many standard packages.

Support Vector Regression (SVR) approach offers another solution to the regression problem. SVR is closely related to the SVMs for classification, offering similar advantages. In particular, SVR problems may be solved using mathematical programming techniques which guarantee to find optimal solutions in polynomial time. Thus, SVR approach is numerically very efficient and can be applied to large-scale problems. In addition, SVR offers flexibility of the model, especially in conjunction with kernel approaches. In particular, the so-called ε-insensitive SVR model assumes that the error measure M(r), where $r = \left| Y_k^{obs} - Y(X_k) \right|$, is set to zero if $r < \varepsilon$ and increase linearly otherwise (see Fig. **3**). This allows one to define the error function in a flexible way, reflecting the expected level of errors by varying error bars, ε, for different types of training examples that may differ in their characteristics [32].

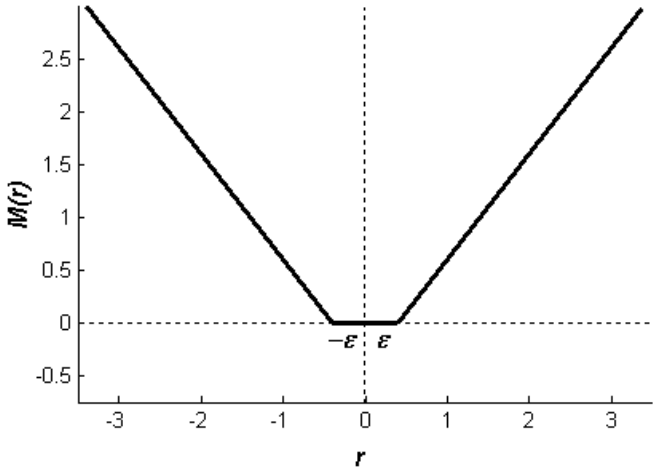

Fig. (3). Support Vector Regression model with ε-insensitive definition of the error measure M(r).

REPRESENTATIONS AND AI MODELS FOR DRUG DESIGN

Among issues to be considered when applying AI techniques are: the *complexity of the model*, as roughly defined by the type of its discriminatory function (e.g., linear vs. non-linear) and the number of free parameters to be optimized; design of appropriate (representative and non-redundant) *training and control sets*; and careful *validation* of the results. Examples of parameters to be optimized are the weights of connections between the nodes ("neurons") in NNs, or the coefficients defining a separating hyperplane in case of SVM. Typically, free parameters are optimized with the goal of minimizing the misclassification error in the training. Other criteria may involve estimates of generalization capabilities in order to avoid over-fitting [3-5]. The importance of the choice of an appropriate model and representation for the problem at hand are illustrated here using several well established problems and examples of applications of AI methods in structural bioinformatics. Structural bioinformatics deals primarily with protein and other macromolecular structure and involves, for example, protein structure and function prediction. Therefore, prediction methods developed in structural bioinformatics are relevant for drug design and discovery and are being integrated with modeling and docking techniques.

AI Techniques for Structural Bioinformatics

Finding representations capable of capturing the underlying principles and correlations is critical for the success of applications of any AI technique. In order to illustrate the above point, let us revisit some classical problems in structural bioinformatics, such as the prediction of secondary structure of an amino acid residue in a protein. In fact, secondary structure prediction was one of the first successful applications of machine learning techniques in the field of

protein structure prediction. In their pioneering work, Rost and Sander [11] demonstrated the importance of multiple alignment representation and used a neural network to train a successful classifier capable of assigning each residue to one of the three classes (helix, beta strand or coil) with over 70% classification accuracy.

Subsequent developments pushed the accuracy of secondary structure prediction to about 80%, which proved to be sufficient for many important applications, e.g., to protein folding simulations and prediction of protein 3D structure [33]. At the same time, however, these studies showed clearly that the multiple alignment representation is far more important than the type of a classifier used. In particular, differences in accuracy between top performing methods, based on NNs, SVMs or HMMs, are not statistically significant (see, e.g. [33]). Since then many different pattern recognition and machine learning techniques have been devised to improve protein structure and function prediction. For example, AI-based protein annotation protocols are being used to provide proteome-wide prediction and annotation of membrane domains, solvent accessibility, protein-protein interactions, post-translational modifications and other attributes that can be used to facilitate drug design and discovery (see, e.g., [34-40]).

As a further illustration of some of the issues arising when applying machine learning and AI techniques, let us consider the problem of predicting which amino acid residues can undergo phosphorylation due to the enzymatic activity of protein kinases. In fact, kinases are targeted by many rational drug design efforts since phosphorylation plays an important role in cancer and other diseases by modulating structure and function of specific proteins, e.g., by affecting their interactions with co-factors. The computational prediction of phosphorylation and other post-translational modification sites, both from the structure and the primary amino acid sequence is an active field of research [41-46]. Examples of methods for phosphorylation site prediction are NetPhos [44], a NN-based predictor, Scansite [46], a sequence-motif based predictor, and DISPHOS [42], which uses indicators of intrinsic disorder, in addition to sequence information. It should be also noted that some "negative" sites might include those which have not (yet) been reported as phosphorylated, suggesting the importance of the so-called "one-class" machine learning protocols (using positive examples only) for prediction of phosphorylation [17]. Next, we specifically address some of the above considerations in the context of the drug design field.

Representation and Model Selection for Drug Design

Learning from examples may be simple if relations between attributes and outcomes are almost linear. However, in drug design the number of parameters that may have influence on biological activity is typically high and relations may be strongly non-linear. Moreover, with limited number of experimental data, the predictive ability of statistical learning systems may suffer from the "curse of dimensionality" [3-5]: if m points are sufficient to obtain a reasonable approximation in each dimension, then m^N points are required in N dimensions. For example, if m=10 points are sufficient to approximate the relation for each parameter, then for just N=20 parameters the number of examples that are required is equal to 10^{20}. With a small number of examples a very large number of N-parameter functions may perfectly fit the data. This problem is addressed in computational learning theory, where methods of selection of appropriate data models are provided. The choice of appropriate model is also related to the learning algorithm used and the representation of data. For example, model selection is very important for QSAR approaches that rely on solving explicitly the underlying (multiple) regression [6].

As mentioned before, information about chemical compounds and other structured objects may be better represented in the form of labeled graphs, rather than vectors. One common approach to represent molecular structures is to use "fingerprints", i.e. long bit-strings (100-1000) encoding yes/no answers about the presence or absence of various features, including substructures within the molecular structure of a chemical compound [20]. For each atom of the chemical compound, a depth-first search for substructures may be used. Each bit set to one in the string represents the presence of particular substructure in the search tree, or several bits are assigned to substructures using hashing techniques.

"Molecular holograms" is a variant of fingerprinting techniques that uses integers to denote the number of fragments of a particular type. Some chemical databases include parameters to generate useful fingerprint strings. The Tanimoto distance between pairs of bit strings generated by molecular fingerprinting is widely used to measure similarity in database searches. However, recent calculations on five bio-logical activity classes showed strong influence of the compound class-specific effects on the results [47]. Although molecular fingerprinting captures some structural similarities, based on it representations of chemical structures lack the information about the geometry and global properties of the molecule. On the other hand, coupling the universal encoding with classification tasks should increase the efficacy in this context. For a review of other approaches to measure chemical similarity the reader is referred to [19]. Chemical information may also be encoded into kernel functions $K(\mathbf{R},\mathbf{S})$ in SVMs [48]. These kernels effectively measure similarity between structures **R** and **S**. Several such kernels have been proposed recently, based on adjacency matrices, sequences of labels in subgraphs, the number of shared walks in subgraphs, and may be combined with molecular fingerprinting. Many variants of such kernels have been constructed and proved to be quite useful in screening for drugs that inhibit cancer cell growth [49].

Trees and directed acyclic graphs may also be used to represent chemical entities, often in conjunction with sequential processing, fragment after fragment, by neural nodes with recurrent connections [13-16, 48-52]. Recurrent connections are needed to preserve information about graph fragments that have already been processed. Each node s of the graph has some features **X** associated with it, and the function implemented by a neuron depends on $\mathbf{W}\cdot\mathbf{X}+h(\mathbf{X})$, where the extra term represents the sum of weighted outputs

from nodes that are connected to *s*. Such generalized recurrent neurons are used to transform input graphs **G** into vectors **X**, and conventional NNs are used in this space for the classification. To use this approach for drug design all molecular structures must be decomposed, and become fragments of directed acyclic graphs (DAGs). The number of recurrent connections is equal to the maximum number of bonds for a given atom represented by the node. However, the decomposition process is not unique and depends on the starting point and orientation (for covalent bonds) chosen initially.

Inductive Logic Programming (ILP) is another approach that may be used to address some of the issues in model selection, especially in the context of symbolic attributes of molecular systems. ILP uses an expressive language and enables construction of new, interesting features [29]. However, due to computational costs and other difficulties in learning, ILP-based models did not show clear advantage over QSAR in applications to mutagenicity, cancerogenicity, toxicology and biodegradability. Summary and bibliography of ILP applications to 3D protein structure, discovery of neuropeptides, chemical compound structure elucidation, diagnosis, mutagenesis, pharmacophore discovery for ACE inhibition, carcinogenicity, toxicology and other areas may be found at http://www-ai.ijs.si/~ilpnet2/apps/.

AI AND QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS

QSAR Approach to Drug Design

In the past, drug identification was done largely through random experimentation and it was clearly not very effective [54]. Furthermore, the mechanism of action of a successful drug would typically remain obscured. It was in the 1960s that alternative approaches to drug design were beginning to be explored, with focus on Quantitative Structure-Activity Relationships (QSAR). The idea behind QSAR approaches is to use the known responses (activities) of simple compounds (structures) to predict the responses of complex compounds, made from different combinations of basic modules. Only compounds predicted to have desired properties would then be tested.

QSAR typically relies on electronic, hydrophobic and steric attributes of a molecule, as well as structural, quantum mechanical and other descriptors. A classical example is the Hammett's study of the effect that different substituents have on the ionization of the carboxylic group in benzoic acid [55]. The effect on ionization was measured in terms of equilibrium (dissociation) constant for the reaction. Hammett discovered that equilibrium constants were higher for more electron-withdrawing substituents. This led to the formulation of linear free energy relationships (LFERs) that motivated subsequent development of many similar approaches for quantitatively describing structure-activity relationships[56].

Hydrophobicity and lipophilicity play a vital role in many biological processes, especially in receptor-ligand interactions. Therefore, QSAR methods often employ hydrophobicity-related descriptors [57]. Another important set of QSAR descriptors are those related to steric effects, such as the molar refraction (MR) index, various parameters accounting for the shape of a compound [57] and descriptors indicating the presence or absence of certain structural features. Another QSAR approach that has gained a lot of momentum is the use of quantum-chemical descriptors, which theoretically can account (at least in principle) for other properties, both electronic and geometric [58]. The increase in the number of parameters required the use of AI approaches to obtain correlations between the molecular and other attributes and observed activities. A typical QSAR study, would involve Hammett's constants, partition coefficients, molar refractivity and many other descriptors. Statistical and machine learning techniques, such as multiple linear regression (MLR), principal component analysis (PCA) or partial least squares (PLS) [59] would then be used to solve the problem. It should be mentioned that MLR is still one of the most widely used AI techniques in QSAR studies [57-59].

Oftentimes, the structure of the target protein is not known. In such cases, the potential drugs may be analyzed using experimental techniques and common structural features called pharmacophores identified [61]. Such pharmacophore (or ligand)-based methods include the 3D-QSAR techniques, for instance [62]. Examples of popular 3D-QSAR methods are the comparative molecular field analysis (CoMFA) [63], the comparative molecular similarity indices analysis (CoMSIA) [64] and GRID [65]. For a more comprehensive list of representative methods and programs within this category, the reader is referred to [66]. The basic idea behind CoMFA is that the biological activity of molecules is related to its electrostatic and steric interactions. The molecules (ligands) that are being studied are aligned structurally on a 3D grid. Using a probe atom, electrostatic and steric fields are determined at every point in the grid. CoMSIA, on the other hand, also takes into account hydrophobic parameters [66]. Such obtained descriptors are then analyzed using statistical methods, such as partial least squares (PLS), to obtain correlations between activity and the fields leading to a 3D-QSAR model of the ligand [58][66-67]. GRID is similar to CoMFA and may also be used to determine the interaction energies between the probe and the ligand. In addition, GRID can also be used to calculate hydrogen bonding energies [66].

3D-QSAR methods have been employed to design anti-HIV-1 drugs [68], matrix metalloproteinase inhibitors [69], therapeutic agents for Alzheimer's and Parkinson's diseases [70] and anti-tuberculosis agents [71], to name a few. Furthermore, 3D-QSAR has been applied along with molecular modeling and molecular dynamics in the design of pteridine-derived therapeutic agents [72], indolomorphinan derivatives [73] and in vaccinology studies [74]. A detailed review on the applications of CoMFA and CoMSIA has been presented by Bordas *et al.* [67]. In 4D-QSAR, the fourth dimension represents an ensemble of conformations, orientations, or protonation states for each molecule [75]. This reduces the bias that may come from the ligand alignment, but requires identification of the most likely bioactive conformation and orientation (or protonation state), frequently obtained using

evolutionary algorithms [31]. The 5-D QSAR carries this one step further, allowing for changes in the receptor binding pocket and ligand topology [76]. Adding solvation effects leads to 6D-QSAR, which allowed, in combination with flexible docking, for relatively accurate identification of the endocrine-disrupting potential associated with a drug candidate [77].

AI Methods in QSAR

Ligands may be represented by multiple structural and other descriptors. Thus, selection of key descriptors is an important step in any QSAR study. Another important step is the identification of patterns (predictive fingerprints or combinations of features) that correlate with activity. Furthermore, compounds exhibiting promising properties may be compared with other candidates in order to identify other potential drugs that share critical features. Therefore, it is evident that the AI approaches to feature-selection, pattern-recognition, classification and clustering can be applied to the problems posed above [78-80].

In fact, various clustering methods (e.g., hierarchical divisive clustering, hierarchical agglomerative clustering, non-hierarchical clustering and multi-domain clustering) have been applied to such problems [78]. For example, clustering receptor proteins, based on their structural similarity, has been shown to improve docking studies and drug design [80]. Applications of clustering techniques and genetic algorithms towards predicting molecular interactions have been reviewed in [62]. The role of feature selection in QSAR has also been reviewed recently [79]. Furthermore, another method called consensus *k*-nearest neighbor (kNN) QSAR has been developed towards predicting estrogenic activity [81]. The concept behind this approach is that the activity can be estimated by averaging activities over *k*-nearest neighbors. Multiple models, each making use of different sets of descriptors, are then used to make the consensus prediction [81].

Many other studies have made use of machine learning techniques to address similar problems. In particular, NNs have been widely used to solve many problems in drug design. A comprehensive review of the applications of NNs in variety of QSAR problems, has been presented by Winkler [82]. The review discusses how NNs can be applied to the prediction of physicochemical, toxicological and pharmacokinetic parameters. In another study, the NN methods were compared with statistical approaches [83]. Self-organized maps (SOM) have been used for studying molecular diversity and employed in drug design [84]. In particular, the SOM-based method of comparative molecular surface analysis (CoMSA) has been presented in detail [84].

In recent years, SVMs have become relatively widely used. For example, Zhao *et al.* made use of SVMs for predicting toxicity and found that this method yielded improved performance compared to multiple linear regression and radial basis function NNs [85]. A new method called least squares support vector machine (LSSVM) was employed to screen calcium channel antagonists in a QSAR study [86]. SVMs were also used (providing accuracies competitive with that of other QSAR approaches) to predict oral absorption in humans involving molecular structure descriptors [87] and to calculate the activity of certain enzyme inhibitors [88], as well as many other investigations of similar type.

Furthermore, evolutionary QSAR techniques that employ genetic algorithms are being developed for docking and related studies. One example is the Multi-objective genetic QSAR (MoQSAR) that has been used to study neuronal nicotinic acetylcholine ion channel receptors (nAChRs) [70]. Other examples involve the use of genetic algorithms for prediction of binding affinities of receptor-ligands [89] and in classification-based SAR [91]. Another technique similar to evolutionary methods and inspired by biological phenomena is Particle Swarm Optimization (PSO) [92]. This technique has been employed in many QSAR studies [93-95] and for biomarker selection [96]. Bayesian networks have also been used to solve different problems in the context of drug design, for examples see [97] and [98].

AI in Predictive Toxicology

Identification of potential toxic effects of candidate drugs using bioassays is a costly and time consuming procedure that often requires animal testing. Attrition rates due to the drug toxicity have already reached over 20%, and are quickly rising [99]. The problem of estimating the toxicity, mutagenicity and carcinogenicity of potential and existing drugs has been approached from three main perspectives: physical simulations using molecular modeling techniques, expert systems capable of reasoning about the domain, and data mining systems based on AI techniques. Reviews of these approaches are presented in several recent books [100-102].

AI methods learn from data, and the quality of results is determined by availability of databases for training. The Distributed Structure-Searchable Toxicity (DSSTox) Database Network created by the U.S. Environmental Protection Agency's Computational Toxicology Program (http://www.epa.gov/nheerl/dsstox/) created a public data foundation for predictive toxicology research. Another database initiative, Vitic toxicity database, supported by a number of pharmaceutical and chemical companies, has been initiated by the Health and Environmental Sciences Institute (HESI), as part of the International Life Sciences Institute (ILSI), and is being managed now by the Lhasa Limited (http://www.lhasalimited.org). These databases store various in-house toxicology data that may be re-analyzed using different techniques.

The availability of such databases and well annotated, large data sets makes it possible to develop and evaluate novel approaches. In particular, a number of challenges in data analysis have been proposed. Conclusions from predictive toxicology challenges [100] and recent work in this area [103-104] are very encouraging. Accuracy of computational approaches in some cases is in the range of 80-95%, comparable to *in vivo* assessments.

Docking and AI Methods

The goal of docking methods is to determine the mode and strength of binding between a ligand and a receptor molecule (typically protein). Traditional docking studies that attempt to determine the binding between a few potential ligands and receptors have been extended in recent years to high-throughput docking (HTD), in which large-scale *in silico* screening of potential drugs for known receptors is employed [105] A variety of methods have been used to solve docking problems, involving improvements in terms of both: search algorithms and scoring functions.

The search for the optimal binding conformation in docking methods leads, in general, to a global optimization problem. Many optimization algorithms are being used to find good solutions to this hard problem, including gradient-based minimization, Molecular Dynamics protocols, Monte Carlo approaches, genetic algorithms and evolutionary programming techniques, fragment based methods, point complementary methods, tabu and systematic searches [105]. Scoring functions are equally important part of the docking protocol and can range from simple force fields used in MD to specifically optimized potentials [105]. For an excellent review of docking methods and programs the reader is referred to Taylor *et al*. [105]. There are many programs for docking, as listed in [106], for instance. Examples of applications of docking protocols to drug design in diabetes and cancer and also in QSAR have been discussed in [106]. Another recent review discusses in detail latest improvements in docking methods [107].

In analogy to pharmocophore methods, there are many steps in which AI methods come into play in the context of docking, including feature selection and extraction, classification and regression for the design of scoring functions and the identification of putative binding sites [107]. Several recent examples include applications of probabilistic Naïve Bayes methods to improve scoring functions for docking [108], applications of NNs to virtual screening in combination with docking methods [109], and combination of kNN and docking methods to achieve improved results in QSAR [110]. In general, one may observe a growing tendency to combine different techniques and use consensus-based methods, such as the lateral validation [111]. In addition, AI methods are being used to address the problems of absorption, distribution, metabolism elimination and toxicology (ADMET) in pharmacokinetics [112]. Many drugs have failed to reach the market because of their ADMET properties. Therefore, prediction methods are being used in order to identify these properties early in the drug design pipeline/ process [113].

DISCUSSION

Artificial Intelligence (AI) is broadly defined here as a field that deals with the design and application of algorithms for analysis of, learning from and interpretation of data. AI integrates many branches of statistical and machine learning, pattern recognition, logics and probability theory as well as biologically motivated approaches, such as neural networks, evolutionary computing or fuzzy modeling, collectively described as "computational intelligence" [1,2]. In the last decade, the barriers between these fields started to soften, with algorithms that use inspiration from many sources being applied to various problems, including drug design and discovery which is the focus of this review.

Drug design poses many challenging problems in terms of selection of relevant information, data modeling, classification, prediction, and optimization [3-5] that stimulate the development and applications of tailored AI approaches. In fact, many AI methods reviewed here have only been formulated in recent years. Various challenges, such as the predictive toxicology challenge and the feature selection challenge, show on difficult, real life problems advantages of new approaches over the established statistical and pattern recognition methods [114,115]. Some of these new approaches have been summarized here.

The overall impact of computational methods on drug design, testing and discovery will certainly grow even further in the future. Already now many results show that computational methods are indispensable in drug design and pre-clinical evaluations. Efforts to combine predictive, data-driven techniques with molecular modeling and simulations are likely to bring further progress in this field. Use of ontologies and analysis of symbolic, as well as textual data to build complex models of biological organisms, is another growing trend. For example, the EcoCyc model (http:// ecocyc.org/) of the *Escherichia coli* bacterium includes the entire genome, transcriptional regulation, transporters, and metabolic pathways. Other organisms are being annotated in a similar, integrated manner (see, e.g., http://biocyc.org/), with the potential to result in new approaches and enhanced tools for drug design.

Among challenges in this field, one should consider the danger of hampering both basic and applied research by the growing tendency to patent specific findings, algorithmic solutions and even general ideas pertaining to drug discovery and design. For example, Axys Pharmaceuticals has a patent on a NN-based approach to designing new compounds and modeling their activity. Furthermore, Health Discovery Corporation owns over 80 patents for application of SVMs (which is also a patented technique [116]) and other AI algorithms to biomedical problems, such as the discovery of biomarkers [117]. Even techniques such as the principal component analysis, known for about 100 years, have been patented in applications to text analysis [118]. These unfortunate attempts to patent widely used and developed mostly in academic settings methods, as well as their obvious applications, may slow down the rate of scientific discovery in the field of drug design.

On the other hand, recent surge of interest in life sciences accelerates the development of new approaches. AI and related methods play already an essential role in mining and analysis of vast amounts of data being generated as a result of recent advances in genomics. Examples of new experimental techniques and applications generating massive amounts of data for analysis include: high throughput sequenc-

ing and genotyping for large-scale studies of genomic variations, and genome-wide studies of gene expression profiles with microarrays. Such studies hold the promise of elucidating further the role of genetic variation in human disease, identifying novel drug targets and enabling personalized interventions with specifically optimized drugs and treatments. In this context, AI methods are being used to find correlations between patterns of genetic variations and expression profiles with clinical and other phenotypes and to identify predictive fingerprints of disease states, progression and results of therapeutic interventions. Such applications pose both challenges and opportunities for AI methods, stimulating their further development.

REFERENCES

[1] Engelbrecht AP. Computational intelligence: An introduction. New York: J. Wiley; 2003.

[2] Konar A. Computational intelligence; principles, techniques and applications. Berlin: Springer 2005.

[3] Duda RO, Hart PE, Stork DG. Pattern classification, New York: J. Wiley, 2nd ed 2001.

[4] Webb A. Statistical pattern recognition. New York: J. Wiley 2002.

[5] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer 2001.

[6] Devillers J. Eds. Neural Networks in QSAR and Drug Design an essential reference source for those on the frontiers of this field. Academic Press 1996.

[7] Mjolsness E, DeCoste D. Machine learning for science: State of the art and future prospects. Science 2001; 293: 2051-5

[8] Burge C, Karlin S. Finding the genes in genomic DNA. Curr Opin in Struct Biol 1998; 8: 346-54.

[9] Krogh A, Mian S, Haussler D. A hidden Markov model that finds genes in *E. coli* DNA. Nucl Acid Res 1994; 22(22): 4768-78.

[10] Solovyev VV, Salamov AA, Lawrence CB. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucl Acid Res 1994; 22: 5156-63.

[11] Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993; 232: 584-99.

[12] Kitchen DB, Stahura FL, Bajorath J. Computational techniques for diversity analysis and compound classification. Mini Rev Med Chem 2004; 4(10): 1029-39.

[13] Frasconi P, Gori M, Sperduti, A. A general framework for adaptive processing of data structures. IEEE Trans Neural Netw 1998; 9(5): 768-85.

[14] Sperduti A, Starita A. Supervised neural networks for classification of structures. IEEE Trans Neural Netw 1997; 8(3): 714-35.

[15] Sperduti A. In: Zurada J, Cloete I Eds, A tutorial on neurocomputing of structures. Knowledge-based neurocomputing. MIT Press 2000; 117-54.

[16] Gärtner T, Lloyd JW, Flach PA. Kernels and distances for structured data. Mach Learn 2004; 57(3): 205-32

[17] Schölkopf B, Smola AJ. Learning with kernels. MIT Press 2002.

[18] Duch W. Similarity based methods: a general framework for classification, approximation and association. Control Cybern 2000; 29(4): 937-68.

[19] Nikolova N, Jaworska J. Approaches to measure chemical similarity - A review. QSAR Comb Sci 2003; 22: 1006-26.

[20] Flower DR. On the properties of bit string-based measures of chemical similarity. J Chem Inf Comput Sci 1998; 38: 378-86.

[21] Duch W, Blachnik M, Fuzzy rule-based systems derived from similarity to prototypes. Lect Notes Comput Sci 2004; 3316: 912-7.

[22] Duch W, Grąbczewski K, Heterogeneous adaptive systems. IEEE World Congress on Computational Intelligence, Honolulu, May 2002; 524-9.

[23] Duch W, Adamczak R, Grąbczewski K. A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. IEEE Trans Neural Netw 2001; 12: 277-306.

[24] Duch W, Setiono R, Zurada JM. Computational intelligence methods for understanding of data. Proc IEEE 2004; 92(5): 771-805.

[25] Jordan M, Sejnowski TJ Eds. Graphical models. Foundations of neural computation. MIT Press, 2001.

[26] Jensen FV, Jensen FB. Bayesian networks and decision graphs. Springer Verlag, 2001.

[27] Baldi P, Brunak S. Bioinformatics: The machine learning approach (Adaptive computation and machine learning series), MIT Press, 2nd ed, 2001.

[28] Yanover C, Weiss Y. Approximate inference and protein-folding. Adv NIPS 2002; 15: 1457-64.

[29] Srinivasan A, King RD. Feature construction with inductive logic programming: A study of quantitative predictions of biological activity aided by structural attributes. Data Min Knowl Discov 1999; 3(1): 37-57.

[30] Page D, Srinivasan A. ILP: A short look back and a longer look forward. J Mach Learn Res 2003; 1: 1-16.

[31] Devillers J Ed. Genetic algorithms in molecular modeling. Academic Press: New York, 1996.

[32] Wagner M, Adamczak R, Porollo A, Meller J. Linear regression models for solvent accessibility prediction in proteins. J Comput Biol 2005; 12(3): 355-69.

[33] Przybylski D, Rost B. Alignments grow, secondary structure prediction improves. Proteins 2002; 47: 197-205.

[34] Krogh A, Larsson B, von Heijne G, Sonnhammer E. Predicting transmembrane protein topology with a hidden markov model: Applications to complete genomes. J Mol Biol 2001; 305(3): 567-80.

[35] Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein-protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002; 269: 1356-61.

[36] Blom N, Sicheritz-Ponten T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics 2004; 4: 1633-49.

[37] Bigelow HR, Petrey DS, Liu J, Przybylski, D, Rost, B. Predicting transmembrane beta-barrels in proteomes. Nucleic Acids Res 2004; 32: 2566-77.

[38] Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks based regression. Proteins 2004; 56: 753-67.

[39] Cao B, Porollo A, Adamczak R, Jarrell M, Meller J. Enhanced recognition of protein transmembrane domains with prediction-based structural profiles. Bioinformatics 2006; 22 (3): 303-9.

[40] Brunak S, Hansen JE, Lund O, Tolstrup N, Gooley AA, Williams KL. NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. Glycoconj J 1998; 15: 115-30.

[41] Berry EA, Dalby AR, Yang ZR. Reduced bio basis function neural network for identification of protein phosphorylation sites: comparison with pattern recognition algorithms. Comput Biol Chem 2004; 28: 75-85.

[42] Dunker AK, Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, *et al*. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 2004; 32: 1037-49.

[43] Yao X, Zhou FF, Xue Y, Chen GL. GPS: a novel group-based phosphorylation predicting and scoring method. Biochem Biophys Res Commun 2004; 325: 1443-8.

[44] Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 1999; 294: 1351-62.

[45] Yaffe MB, Leparc GG, Lai J, Obata T, Volinia S, Cantley LC. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. Nat Biotechnol 2001; 19: 348-53.

[46] Yaffe MB, Obenauer JC, Cantley LC. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 2003; 31: 3635-41.

[47] Godden JW, Stahura FL, Bajorath J. Anatomy of fingerprint search calculations on structurally diverse sets of active compounds. J Chem Inform Model 2005; 45(6): 1812-9.

[48] Hammer B, Saunders C, Sperduti A Eds. Special issue on neural networks and kernel methods for structured domains. Neural Netw 2005; 18(8).

[49] Bianucci AM, Micheli A, Sperduti A, Starita A. In: Sztandera L, Cartwright H, Eds. A novel approach to QSPR/QSAR based on neural networks for structures, in soft computing approaches in chemistry. Studies in fuzziness and soft computing, Springer-Verlag 2003; 120: 265-96.

[50] Bianucci AM, Micheli A, Sperduti A, Starita A. Analysis of the internal representations developed by neural networks for structures applied to QSAR studies of benzodiazepines. J Chem Inform Comput Sci 2001; 41(1): 202-18.

[51] Ceroni A, Frasconi P, Pollastri G. Learning protein secondary structure from sequential and relational data. Neural Netw 2005; 18(8): 1029-39.

[52] Vullo, A. and Frasconi, P. Prediction of protein coarse contact maps. J Bioinform Comput Biol 2003; 1(2): 411-31.

[53] Ralaivolaa L, Swamidassa SJ, Saigoa H, Baldi P, Graph kernels for chemical informatics. Neural Netw 2005; 18 (8): 1093-110.

[54] Reddy MR, Parrill AL. Overview of rational drug design. rational drug design, ACS Symposium Series 719, American Chemical Society, Washington, DC 1999; 1-11.

[55] Hammett LP. Reaction rates and indicator acidities. Chem. Rev 1935; 17(1): 67-79.

[56] Hammett LP. Physical organic chemistry: Reaction rates, equilibria, and mechanisms. McGraw-Hill Book Co., New York, 2nd ed. 1970.

[57] Selassie CD. History of quantitative structure-activity relationships. Burger's medicinal chemistry and drug discovery, 6th ed 2003; 1: 1-48.

[58] Karelson M, Lobanov VS, Katritzky AR. Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev 1996; 96(3): 1027-44.

[59] Leonard JT, Roy K. QSAR by LFER model of HIV protease inhibitor mannitol derivatives using FA-MLR, PCRA, and PLS techniques. Bioorg Med Chem 2005; 14: 1039-46.

[60] Melagraki G, Afantitis A, Sarimveis H, Igglessi-Markopoulou O, Supuran CT. QSAR study on para-substituted aromatic sulfonamides as carbonic anhydrase II inhibitors using topological information indices. Bioorg Med Chem 2006; 14(4): 1108-14.

[61] Guner OF. History and evolution of the pharmacophore concept in computer-aided drug design. Curr Top Med Chem 2002; 2(12): 1321-32.

[62] Dror O, Shulman-Peleg A, Nussinov R, Wolfson HJ. Predicting molecular interactions in silico: I. A guide to pharmacophore identification and its applications to drug design. Curr Med Chem 2004; 11(1): 71-90.

[63] Cramer RD, DePriest SA, Patterson DE, Hecht P. In: Kubinyi H Ed, The developing practice of comparative molecular field analysis, in 3D QSAR in drug design: theory, methods and applications. ESCOM, Netherlands 1993; 443-85.

[64] Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 1994; 37: 4130-46.

[65] Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Am Chem Soc 1985; 28: 849-57.

[66] Akamatsu M. Curr state and perspectives of 3D-QSAR. Curr Top Med Chem 2002; 2(12): 1381-94.

[67] Bordas B, Komives T, Lopata A. Ligand-based computer-aided pesticide design. A review of applications of the CoMFA and CoMSIA methodologies. Pest Manag Sci 2003; 59(4): 393-400.

[68] Debnath AK. Application of 3D-QSAR techniques in anti-HIV-1 drug design – an overview. Curr Pharm Des 2005; 11(24): 3091-110.

[69] Kontogiorgis CA, Papaioannou P, Hadjipavlou-Litina DJ. Matrix metalloproteinase inhibitors: a review on pharmacophore mapping and (Q)SARs results. Curr Med Chem 2005; 12(3): 339-55.

[70] Nicolotti O, Altomare C, Pellegrini-Calace M, Carotti A. Neuronal nicotinic acetylcholine receptor agonists: pharmacophores, evolutionary QSAR and 3D-QSAR models. Curr Top Med Chem 2004; 4(3): 335-60.

[71] Nayyar A, Malde A, Jain R, Coutinho E. 3D-QSAR study of ring substituted quinoline class of anti-tuberculosis agents. Bioorg Med Chem 2006; 14(3): 847-56.

[72] Matter H, Kotsonis P. Biology and chemistry of the inhibition of nitric oxide synthases by pteridine-derivatives as therapeutic agents. Med Res Rev 2004; 24(5): 662-84.

[73] Li W, Tang Y, Zheng YL, Qiu ZB. Molecular modeling and 3D-QSAR studies of indolomorphinan derivatives as kappa opioid antagonists. Bioorg Med Chem 2006; 14(3): 601-10.

[74] Flower DR, McSparron H, Blythe MJ, Zygouri C, Taylor D, Guan P, *et al*. Computational vaccinology: quantitative approaches. Novartis Found Symp 2003; 254: 102-20; discussion 120-5, 216-22, 250-2.

[75] Vedani A, Briem H, Dobler M, Dollinger H, McMasters DR. Multiple conformation and protonation-state representation in 4D-QSAR: The neurokinin-1 receptor system. J Med Chem 2000; 43: 4416-27.

[76] Vedani A, Dobler M. 5D-QSAR: The key for simulating induced fit? J Med Chem 2002; 45: 2139-49.

[77] Vedani A, Dobler M, Lill MA. Combining protein modeling and 6D-QSAR - Simulating the binding of structurally diverse ligands to the estrogen receptor. J Med Chem 2005; 48: 3700-3.

[78] Kitchen DB, Stahura FL, Bajorath J. Computational techniques for diversity analysis and compound classification. Mini Rev Med Chem 2004; 4(10): 1029-39.

[79] Walters WP, Goldman BB. Feature selection in quantitative structure-activity relationships. Curr Opin Drug Discov Devel 2005; 8(3): 329-33.

[80] Koch MA, Waldmann H. Protein structure similarity clustering and natural product structure as guiding principles in drug discovery. Drug Discov Today 2005; 10(7): 471-83.

[81] Asikainen AH, Ruuskanen J, Tuppurainen KA. Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds. SAR QSAR Environ Res 2004; 15(1): 19-32.

[82] Winkler DA. Neural networks as robust tools in drug lead discovery and development. Mol Biotechnol 2004; 27(2): 139-68.

[83] Douali L, Villemin D, Cherqaoui D. Comparative QSAR based on neural networks for the anti-HIV activity of HEPT derivatives. Curr Pharm Des 2003; 9(22): 1817-26.

[84] Polanski J, Gieleciak R. Comparative molecular surface analysis: a novel tool for drug design and molecular diversity studies. Mol Divers 2003; 7(1): 45-59.

[85] Zhao CY, Zhang HX, Zhang XY, Liu MC, Hu ZD, Fan BT. Application of support vector machine (SVM) for prediction toxic activity of different data sets. Toxicology 2006; 217(2-3): 105-19.

[86] Yao X, Liu H, Zhang R, Liu M, Hu Z, Panaye A, *et al*. QSAR and classification study of 1,4-dihydropyridine calcium channel antagonists based on least squares support vector machines. Mol Pharm 2005; 2(5): 348-56.

[87] Liu HX, Hu RJ, Zhang RS, Yao XJ, Liu MC, Hu ZD, *et al*. The prediction of human oral absorption for diffusion rate-limited drugs based on heuristic method and support vector machine. J Comput Aided Mol Des 2005; 19(1): 33-46.

[88] 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. J Chem Inf Comput Sci 2003; 43(6): 2048-56.

[89] Deng W, Breneman C, Embrechts MJ. Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods. J Chem Inf Comput Sci 2004; 44(2): 699-703.

[90] Sutherland JJ, O'Brien LA, Weaver DF. Spline-fitting with a genetic algorithm: a method for developing classification structure-activity relationships. J Chem Inf Comput Sci 2003; 43(6): 1906-15.

[91] Lu Q, Wu H, Yu R, Shen G. The lifetime of CFC substitutes studied by a network trained with chaotic mapping modified genetic algorithm and DFT calculations. SAR QSAR Environ Res 2004; 15(4): 279-92.

[92] Kennedy J, Eberhart RC. Particle swarm optimization. Proceedings of IEEE International Conference on Neural Networks, Piscataway, NJ 1995; 1942-8.

[93] Lin L, Lin WQ, Jiang JH, Shen GL, Yu RQ. QSAR analysis of substituted bis[(acridine-4-carboxamide)propyl]methylamines using optimized block-wise variable combination by particle swarm optimization for partial least squares modeling. Eur J Pharm Sci 2005; 25(2-3): 245-54.

[94] Shen Q, Jiang JH, Jiao CX, Huan SY, Shen GL, Yu RQ. Optimized partition of minimum spanning tree for piecewise modeling by particle swarm algorithm. QSAR studies of antagonism of angiotensin II antagonists. J Chem Inf Comput Sci 2004; 44(6): 2027-31.

[95] Shen Q, Jiang JH, Jiao CX, Lin WQ, Shen GL, Yu RQ. Hybridized particle swarm algorithm for adaptive structure training of multilayer feed-forward neural network: QSAR studies of bioactivity of organic compounds. J Comput Chem 2004; 25(14): 1726-35.

[96] Ressom HW, Varghese RS, Abdel-Hamid M, Eissa SA, Saha D, Goldman L, *et al*. Analysis of mass spectral serum profiles for biomarker selection. Bioinformatics 2005; 21(21): 4039-45.

[97] Caballero J, Fernandez M. Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks. J Mol Model 2006; 12(2): 168-81.

[98] Wang YH, Li Y, Yang SL, Yang L. An *in silico* approach for screening flavonoids as P-glycoprotein inhibitors based on a Bayesian-regularized neural network. J Comput Aided Mol Des 2005; 19(3): 137-47.

[99] Kola I, Landis J. Can pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov 2004; 3: 711-5.

[100] Helma C Ed. Predictive toxicology. Marcel Dekker, New York 2005.

[101] Cronin M, Livingstone D Eds. Predicting Chemical Toxicity and Fate. CRC Press, Boca Raton, Florida 2004.

[102] Benigni R Ed. Quantitative structure-activity relationship (QSAR) models of mutagens and carcinogens. CRC Press, Boca Raton, Florida 2003.

[103] Matthews E, Kruhlak N, Benz R, Contrera J. Assessment of the health effects of chemicals in humans: I. QSAR estimation of the maximum recommended therapeutic dose (MRTD) and no effect level (NOEL) of organic chemicals based on clinical trial data. Curr Drug Discov Technol 2004; 1: 61-76.

[104] Cheng A, Dixon S. *In silico* models for the prediction of dose-dependent human hepatotoxicity. J Comput Aided Mol Des 2004; 17: 811-23.

[105] Taylor RD, Jewsbury PJ, Essex JW. A review of protein-small molecule docking methods. J Comput Aided Mol Des 2002; 16(3): 151-66.

[106] Glen RC, Allen SC. Ligand-protein docking: cancer Res at the interface between biology and chemistry. Curr Med Chem 2003; 10(9): 763-7.

[107] Schneidman-Duhovny D, Nussinov R, Wolfson HJ. Predicting molecular interactions in silico: II. Protein-protein and protein-drug docking. Curr Med Chem 2004; 11(1): 91-107.

[108] Klon AE, Glick M, Thoma M, Acklin P, Davies JW. Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. J Med Chem 2004; 47(11): 2743-9.

[109] Sangma C, Chuakheaw D, Jongkon N, Saenbandit K, Nunrium P, Uthayopas P, *et al*. Virtual screening for anti-HIV-1 RT and anti-HIV-1 PR inhibitors from the Thai medicinal plants database: a combined docking with neural networks approach. Comb Chem High Throughput Screen 2005; 8(5): 417-29.

[110] Medina-Franco JL, Golbraikh A, Oloff S, Castillo R, Tropsha A. Quantitative structure-activity relationship analysis of pyridinone HIV-1 reverse transcriptase inhibitors using the k-nearest neighbor method and QSAR-based database mining. J Comput Aided Mol Des 2005; 19(4): 229-42.

[111] Doytchinova IA, Guan P, Flower DR. Quantitative structure-activity relationships and the prediction of MHC supermotifs. Methods 2004; 34(4): 444-53.

[112] Yamashita F, Hashida M. *In silico* approaches for predicting ADME properties of drugs. Drug Metab Pharmacokinet 2004; 19(5): 327-38.

[113] Davis AM, Riley RJ. Predictive ADMET studies, the challenges and the opportunities. Curr Opin Chem Biol 2004; 8(4): 378-86.

[114] Toivonen H, Srinivasan A, King RD, Kramer S, Helma C. Statistical evaluation of the predictive toxicology challenge 2000-2001, Bioinformatics 2003; 19(10): 1183-93.

[115] Guyon I, Gunn S, Nikravesh M, Zadeh L Eds, Feature extraction: foundations and applications. NIPS 2003 challenge on feature extraction, Springer Verlag 2005.

[116] Boser B, Guyon I, Vapnik V. Pattern recognition system using support vectors. US Patent 5,649,068, 1997.

[117] Health Discovery Corporation, http: //www.healthdiscoverycorp. com

[118] Deerwester; SC, Dumais ST, Furnas GW, Harshman RA, Landauer TK, et al. Computer information retrieval using latent semantic structure. U. S. Patent No. 4,839,853, 1989.