DRUG DEVELOPMENT RESEARCH 72 (2011)

DDR

Research Overview

In Silico Structure-Activity-Relationship (SAR) Models From Machine Learning: A Review

Xia Ning* **and George Karypis**

Department of Computer Science and Engineering, University of Minnesota, Twin Cities, Minnesota

Strategy, Management and Health Policy				
Enabling Technology, Genomics, Proteomics	Preclinical Research	Preclinical Development Toxicology, Formulation Drug Delivery, Pharmacokinetics	Clinical Development Phases I-III Regulatory, Quality, Manufacturing	Postmarketing Phase IV

ABSTRACT In this article, we review the recent development for in silico Structure-Activity-Relationship (SAR) models using machine-learning techniques. The review focuses on the following topics: machine-learning algorithms for computational SAR models, single-target-oriented SAR methodologies, Chemogenomics, and future trends. We try to provide the state-of-the-art SAR methods as well as the most up-to-date advancement, in order for the researchers to have a general overview at this area. Drug Dev Res, 2010. © 2010 Wiley-Liss, Inc.

Key words: structure-activity-relationship (SAR); machine learning; chemogenomics

INTRODUCTION

Small organic molecules, by binding to different proteins, can be used to modulate (inhibit/activate) protein function for therapeutic purposes and to elucidate the molecular mechanisms underlying biological processes. This chemical genetics approach [Tolliday et al., 2006; Kawasumi and Nghiem, 2007] of perturbing living biological systems has been gaining momentum because it provides key advantages over the approaches based on molecular genetics. However, experimental high-throughput screening (HTS) techniques [Jona and Snyder, 2003; Bulseco and Wolf, 2003; Inglese et al., 2006] for identifying compounds that bind selectively and with high affinity to the various protein targets are hindered by a number of problems associated with hit identification and limited chemical diversity [Oprea and Gottfries, 2001; Lipinski and Hopkins, 2004; Dobson, 2004]. Therefore, in silico methods that computationally study the relationship between compound structures and their properties against protein targets have distinguished themselves by their efficiency and accuracy, and have quickly become popular options for initial hit compound identification. Such methods are formalized as in silico Structure-Activity-Relationship (SAR) modeling.

The pioneering work of Hansch et al. [1962, 1963], which demonstrated that the biological activity of a chemical compound can be mathematically expressed as a function of its physicochemical properties, led to the development of in silico quantitative methods for modeling SARs. Since that work, many different approaches have been developed for building SAR models [Bravi et al., 2000; Agrafiotis et al., 2007]. These in silico models have become an essential and effective tool for computationally predicting the biological activity of a compound against a certain protein target from its molecular structures. They play a critical role in drug and chemical probe discovery by informing the initial screens, design, and optimization of compounds with the desired biological properties in an efficient fashion.

*Correspondence to: Xia Ning, Department of Computer Science and Engineering, University of Minnesota, Twin Cities, MN 55455. E-mail: xning@cs.umn.edu

Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/ddr.20410

In this article, we discuss the SAR modeling problem formulation, challenges, and approaches. We give a short review on most recent advances of in silico SAR modeling from two aspects. The first is single-target oriented schemes, e.g., the conventional SAR methods. We discuss machine learning and data mining techniques that are particularly developed and applied for single-target-based SAR model learning. The second aspect is domain (i.e., biological and chemical) knowledge incorporation, which falls within the category of chemogenomics and is able to produce more sophisticated and powerful SAR models by utilizing binding information from protein families. In the end, we give a brief discussion on the methods that may go beyond current state-of-the-art methodologies.

CHARACTERISTICS AND CHALLENGES FOR IN SILICO SAR MODELS

The typical setup for in silico SAR model learning is as follows. Given a set of compounds from experimental results (e.g., bioactivity assays), which show a certain level of binding affinity against a protein target under consideration, a computational SAR model is built from such compounds so that the model learns/captures the structural properties of the compounds that are causally related to their bioactivity. This in silico SAR model is then applied to predict the bioactivity of unseen compounds. The above setting of in silico SAR model learning falls into the standard setting of supervised learning. Supervised learning is a machine-learning technique, which learns knowledge from fully labeled/annotated data [Kotsiantis, 2007]. Therefore, in principle existing supervised learning algorithms could be applied directly on chemical data. However, in silico SAR modeling problem has its special characteristics and challenges such that extra care needs to be taken in order to better tackle the problem with supervised learning. Such features include the following aspects.

Compound Representation

Compound representation is a popular topic in chemical information learning. In order to computationally learn knowledge from compounds or proteins, the first step is to represent such instances in a meaningful format that encodes compound properties so that computational algorithms can manipulate them and learn knowledge from them. Many in silico SAR methods represent compounds using various descriptors, as they represent a convenient and computationally efficient way to capture key characteristics of the compounds' structures. These descriptors include physicochemical property descriptors [Bravi et al., 2000; Bajorath, 2002], topological descriptors derived from the molecular graph of a compound [Daylight; MDL; Deshpande et al., 2005; Wale et al., 2007; Rogers et al., 2005], and 2D and 3D pharmacophore descriptors that capture interactions important to protein-ligand binding [Sheridan et al., 1996; Davies, 1996; Stiefl et al., 2006]. Among them, hashed 2D descriptors corresponding to subgraphs of various sizes and types (e.g., paths, trees, rings) are the most common descriptors and include the fingerprints from Daylight Inc. and Chemaxon Inc. and the extended connectivity fingerprints [Hert et al., 2004]. A good review on compound representation is Wale et al. [2010]. For many compound representations, they may suffer from very high feature dimensionality and thus the curse of dimensionality. Due to this, dimension selection and reduction methods can be applied explicitly or implicitly.

Sparse Active Compound Set

Sparse active compound set means that compared to the entire chemical space, labeled compounds (i.e., active and inactive compounds), with respect to the target under consideration, only occupy an extreme small subspace. This leads to two outcomes. The first is that many times we may not have a rich set of training information to build a very representative model from. For example, some confirmatory assays only verify less than 20 compounds, which may not allow a learning algorithm to learn sufficient knowledge from. SAR models learned from such small compound sets may suffer from low quality, and therefore model quality improvement becomes a challenge in in silico SAR model learning. The second outcome is that there exists a large set of unlabeled data (i.e., compounds that are not known to be either active or inactive due to limited experimental facilities and/or restricted screening libraries), which, once well explored, potentially give additional useful information for better model learning. This leaves a huge space for semi-supervised learning (i.e., a machine learning technique that learns knowledge from both fully labeled/annotated data and not labeled/annotated data [Zhu, 2005]) methodology to be applied in this domain.

Co-Existence of Biological Space

Co-existence of biological space with chemical space can be another source of information for SAR models, since SAR models learn the relationship between chemicals and a certain biological. Biological space is well structured because proteins can be evolutionarily related such that they form different protein families [Kunin et al., 2003]. Given this, it is

expected that information from biologicals can serve as auxiliary knowledge.

MACHINE LEARNING ALGORITHMS FOR CONVENTIONAL IN SILICO MODELS

Over the years, many machine-learning algorithms have been developed and applied to build conventional single-target-oriented in silico SAR models. Such algorithms contribute as a significant component for SAR model learning.

Support Vector Machines (SVMs) [Vapnik, 1998] are widely applied learning algorithms to do classification and regression (SVR) [Smola and Olkopf, 1998] for SAR modeling. The main idea of SVM is to construct a maximum-margin hyperplane that linearly separates the two classes of training instances in a certain (high or infinite dimensional) space. The hyperplane is then used to classify new instances by looking at which side of the hyperplane the new instances fall within in the space. The instances are mapped to this space through a kernel, which is positive semidefinite and intuitively can be considered as a similarity measure. The advantage of such kernel methods is that they directly give the instance similarity within that mapped space without asking the users to explicitly design such space or understanding the structure of the space. SVMs achieve the stated-of-the-art performance among current classification methods in many application domains, and for SAR model learning, they are also the most popular and successful options [Darnag et al., 2010; Byvatov et al., 2003; Deshpande et al., 2005; Wale et al., 2007].

Partial Least-Squares (PLS) regression [Rosipal and Krämer, 2006a] is an early but still popular technique for SAR model learning, originally developed for chemometrics [Otto, 2007]. The main idea of PLS is to model the response relationship between compound structures and compound bioactivities by projecting them into a latent space, where a subset of compound structures (principal components) best describe compound bioactivities in a linear manner [Hasegawa and Funatsu, 2000; Rosipal and Krämer, 2006b; Roy and Roy, 2008; Zhou et al., 2007]. PLS has proven to be useful in situations where the number of compound structural features is much greater than the number of bioactivity values and high multicollinearity among the variables exists, both of which are common for SAR. Recently there are modified PLS methods for SAR modeling [Bennett et al., 2003; Rosipal and Krämer, 2006b], in which kernel partial least squares (kernel PLS) is a kernel method for PLS that introduces on-linear mapping through kernels [Lapinsh et al., 2005; Deng et al., 2004]. Many other learning algorithms are applcable for SAR modeling and include Neural Networks (NN) [Tetko et al., 2001; Livingstone and Manallack, 2003; Guha and Jurs, 2005], Decision Tree [Sussman et al., 2003], Recursive Partitioning [Chen et al., 1998; Rusinko et al., 1999; An and Wang, 2001], Linear Discriminant Analysis (LDA) [Otto, 1999], Bayesian models [Xia et al., 2004; Mccallum and Nigam, 1998], and Random Forest [Zhang and Aires-de Sousa, 2007; Breiman, 2001]. Hughes-Oliver et al. [2008] implement and compare many popular learning methods for SAR modeling.

In recent years, a new class of kernel-based techniques has been developed that builds SAR models by operating directly on the molecular graphs [Raymond and Willett, 2002; Kashima et al., 2003; Le et al., 2003, 2004; Ralaivola et al., 2005; Menchetti et al., 2005]. These kernels are computed by using powers of adjacency matrices [Kashima et al., 2003; Ralaivola et al., 2005], Markov random walks on the underlying graphs [Kashima et al., 2003; Ralaivola et al., 2005], optimal assignment between atoms and bonds of two graphs [Froehlich et al., 2005; Kozaz et al., 2007], maximum common subgraph [Raymond and Willett, 2002; Le et al., 2003, 2004], and weighted substructure matching [Menchetti et al., 2005]. The advantage of these techniques is that they determine the similarity between compound pairs by directly analyzing their molecular graphs and eliminate the step of descriptor generation. However, such methods are inherently less descriptive as it is hard to identify the different features of the molecular graphs that might be important for activity.

Another class of methods builds SAR models by operating directly on the structure of the chemical compounds and automatically identifying a small number of chemical substructures that relate to their biological activity using approaches based on inductive logic programming [King et al., 1992; Muggleton and De Raedt, 1994; King et al., 1996] (e.g., Golem [Muggleton and Feng, 1992] and WARMR [King et al., 2001; Dehaspe et al., 1998]). The high computational requirement of these approaches led to the development of various heuristic methods [Klopman, 1998; Gonzalez et al., 2001; Matsuda et al., 2002; Nicolaou et al., 2002] that either restrict the search space or the type of rules being discovered. However, the savings in computational time of such restrictions come at the expense of failing to identify the best rules in the cases in which the critical substructures are not just linear chains [Deshpande et al., 2005; Wale et al., 2007].

CHEMOGENOMICS FOR IN SILICO SAR MODELS

A different aspect of current SAR modeling is what information can be utilized for an SAR model

construct a protein target and how it can be used. Conventional (single-target-oriented) SAR modeling usually uses information from the target itself (i.e., its sequence, structures, etc.) and information from its own experimentally determined compounds (i.e., compound weight, compound structure, etc.).

A new perspective, chemogenomics [Frye, 1999; Caron et al., 2001; Klabunde, 2007; Harris and Stevens, 2006; Guba, 2006; Rognan, 2007; Gaither, 2007] takes advantage of the concept of protein family, and utilizes information from proteins that belong to a family as additional knowledge so as to build better SAR models. In this way, the lack of known ligands for a given target can be compensated by the availability of known ligands for other targets. Chemogenomics has been becoming a prominent methodology for SAR learning [Klabunde, 2007; Rognan, 2007; Strömbergsson and Kleywegt, 2009].

Chemogenomics

Chemogenomics-based approaches leverage SAR information from proteins within a same family as a new target. These proteins share key characteristics with the new target and their SAR information is used in order to filter libraries for focused screening to aid in the identification of active compounds for the new target [Caron et al., 2001; Bredel and Jacoby, 2004; Harris and Stevens, 2006; Guba, 2006; Rognan, 2007; Gaither, 2007]. Chemogenomics methods usually organize the members of a family into groups such that within each group ligands have similar binding patterns and key characteristics of the ligand-binding sites are conserved (e.g., similar amino acid composition, physicochemical properties, or structures). For a new protein, the most relevant group is usually identified by comparing the ligand-binding part of the sequence or structure to that of the proteins in each group.

One of the early efforts in this area has been the work of Frye [1999] who proposed the SARAH framework for organizing proteins in each family based on their SAR similarity to create SAR homologous clusters. Frye's work advocated that such clusters will help in establishing correlations between sequence conservation and SAR homology, thus, making it possible to predict the cluster membership of a new protein based on its sequence. Since then, there have been a number of Chemogenomics efforts that have primarily focused on kinases [Vieth et al., 2004; Hu et al., 2005; Birault et al., 2006; Kellenberger et al., 2006; Hoppe et al., 2006], and GPCRs [Jacoby et al., 1999; Jacoby, 2001; Frimurer et al., 2005; Surgand et al., 2006]. Some of these approaches identify the right subset of family members using similarity search, either with respect to sequence [Frimurer et al., 2005; Surgand et al., 2006] or structure [Hu et al., 2005; Kellenberger et al., 2006; Hoppe et al., 2006], whereas other approaches employ machine-learning techniques to estimate and analyze the ligand-target affinity within each family [Bock and Gough, 2002, 2005; Vieth et al., 2004; Jacob and Vert, 2008]. Even though chemogenomics-based approaches have been successfully used to identify lead compounds [Nguyen et al., 2003; Eguchi et al., 2003; Klabunde and Jger, 2006; Martin et al., 2007], the methods that were developed are to a large extent specific to kinases and GPCRs and have a significant manual component. Moreover, the quality of the focused libraries that they create is a function of the diversity in the original library.

Machine-Learning Methods for Chemogenomics-Based SAR Models

Over the past few years, various chemogenomics-based in silico approaches have been developed that differ on how they formulate the chemogenomics framework so as to fit the available information into the framework and learn knowledge from there systematically. Basically, there are two components that these methods need to deal with. The first one is data representation, which includes representing information involved in both biological space and chemical space, and their known relationship. The second is the learning algorithms to explore/learn biological-chemical relationship for SAR. Of course, these two aspects are closely correlated since data representation has to fit and better serve learning schemes, but we discuss them independently for the sake of simplicity.

Data Representation for Chemogenomics-Based SAR

Data representation for Chemogenomics-based SAR can be different from single-target-oriented methods, if protein properties are involved in learning. In this case, the Chemogenomics-based data representation requires three ingredients of efforts. The first is the features for the targets (e.g., protein structures [Lindström et al., 2006], amino acid sequence [Jacob and Vert, 2008], binding site descriptors [Strömbergsson et al., 2008; Deng et al., 2004], etc.). This part is new from conventional single-target oriented SAR methods, but has already been well studied in structural biology independently, and thus in principle a good feature representation from their studies can be applied directly for chemogenomics. The second is ligand representation that is in single-target SAR learning. A novel yet critical requirement is on explicit target-ligand complex/relationship representation (i.e., representing the binding event/relationship between biologicals and chemicals). In chemogenomics-based

methods, the complexes are represented/modeled by protein-ligand fingerprints [Weill and Rognan, 2009], protein and ligand descriptor concatenation [Bock and Gough, 2005], protein space and ligand space tensor product and kernel fusion [Jacob and Vert, 2008], among others.

Machine-Learning Algorithms for Chemogenomics-Based SAR

Machine-learning algorithms for chemogenomics-based SAR learn models from target-ligand relationship, not from only targets or ligands in isolation. Since there are two parties involved in chemogenomics-based methods (i.e., proteins and compounds), their relationship is handled in different styles. The first processes the two parties step by step. Klabunde and Jager [2006] first identify the family (e.g., GPCR) or subfamily information (e.g., purinergic GPCR), and then pool together the ligands for all family proteins and learn a family-level SAR model from the ligands. SAR models learned from such a fashion are never specific to a certain single protein target, but actually for the entire family. Such methodology is superior to other chemogenomics-based methods (discussed later) mostly in efficiency, but the model learned may become too general and noisy, if not problematic, when the diversity of ligands targeting at different proteins within a family reaches above a certain threshold. In effect, ligand diversity is highly desired for drug discovery. Another concern would be that ligands for an entire family may suffer from low selectivity. An alternative on such a problem was implemented by Frimurer et al. [2005], where protein targets were clustered based on ligand binding site similarity, and then cluster-level SAR models were learnt.

A different approach to utilize protein family and ligand information is to directly couple them into protein-ligand pairs, and models are learned from such pairs. Bock and Gough [2005] proposed a support vector regression method to predict compound activity against orphan GPCRs using target-ligand complexes, which are represented by concatenating target representation and compound representation. Targets were represented by their physicochemical properties of their primary structures (i.e., surface tension, isoelectric point, accessible surface area, etc.) and compounds were described using a 2-D molecular connectivity matrix supplemented by chemical properties, and the matrix factorized using singular value decomposition (SVD) with singular values are used to represent the compounds. Erhan et al. [2006] showed how the same concept can be cast in the framework of Neural Networks (NNs) and Support Vector Machines (SVMs). In particular, they show that a given set of receptor descriptors can be combined with a given set of ligand descriptors in a computationally efficient framework, offering in principle a large flexibility in the choice of the receptor and ligand descriptors. Similarly, in Strömbergsson et al. [2008], 3D structures of proteins around binding sites were used to describe proteins and then used to support vector regression to train a model and predict enzyme-ligand interactions. Jacob and Vert [2008] proposed a set of kernels on protein-ligand complexes and then used SVMs for chemogenomics-based SAR learning demonstrating that such methods introduce significant improvement over conventional methods.

Multi-Task Learning (MTL)

Multi-Task Learning (MTL) [Caruana, 1993; Evgeniou et al., 2005; Bonilla et al., 2007] is a machine-learning methodology that can learn multiple related tasks simultaneously within a single model by implicitly transferring knowledge across different tasks to boost general performance. A key requirement for the applicability of multi-task learning is that the tasks under consideration are related and share a common representation. There has been theoretical proof showing that MTL improves performance and is of interest to the chemoinformatics community as it naturally conforms to the structures of chemogenomics-based methodology. For the latter, multiple proteins from a same family, with respect to the target under consideration, and their ligands are used simultaneously for a single SAR model learning. This can be considered as a multi-task learning system, in which each protein and its own ligands correspond to a single learning task. Since the proteins are related (they are from a same family and share common characteristics), chemogenomics-based methods learn all such tasks in parallel and collaboratively family-related information in terms of ligand binding properties can be transferred and shared across multiple proteins. Kernel methods in Jacob and Vert [2008] are essentially examples of multi-task learning. Erhan et al. [2006] predicted target-ligand interactions inside a family of targets by employing Collaborative Filtering (CF) [Goldberg et al., 1992] technology, a specific form of multi-task learning, by applying Neural Networks (NN) and a kernel-based ordinal regression method named JRank in which a set of new kernels on targets and compounds (i.e., identity kernel, Gaussian kernel, correlation kernel, and quadratic kernel) are also developed. Geppert et al. [2009] designed 6 methods to predict compound activity, including linearly combining multiple linear SVM models and multi-task SVM models. Nigsch et al. [2008] applied Laplacian-modified Naïve

Bayesian classifier and a perceptron-like learning algorithm called Winnow [Nigsch and Mitchell, 2008] to perform multi-class classification to predict ligand-target interactions. They augmented the features of compounds by creating an additional so-called Orthogonal Sparse Bigrams (OSBs) from original features. Weill and Rognan [2009] proposed a novel protein-ligand fingerprint and then applied Random Forest (RF) and Naïve Bayesian (NB) to perform interactions between GPCRs and their ligands, an effect that multi-tasks learning in a chemogenomic framework with the fingerprint being generated by concatenating the GPCR cavity descriptor and GPCR ligand descriptor.

PERSPECTIVES ON THE FUTURE

Given the maturity of various data mining and machine-learning techniques on large-scale real-life problems over the years, as well as the advanced knowledge and experimental testing/validation on system biology and chemical medicine, the question now becomes what will be the next step for current in silico SAR model learning strategies? Over the years, a number of methods have been developed for improving accuracy of SAR models. One major effort is to utilize possible additional information from what has been learned beyond the known ligands of the targets under consideration. The inspiration is that it is very typical in domains of biology and chemistry that unlabeled data (e.g., compounds whose druggablity is unknown, proteins whose binding site structures are uncertain, etc.) are always overwhelmingly dominating, whereas labeled information is always extremely sparse, leaving considerable space for computational techniques to take the major rule on information learning and utilization. One early method adopted approaches based on active learning and iteratively expanded the set of training compounds used for learning SAR models [Warmuth et al., 2003]. In this approach, experimentally determined ligands for the target were used to build an initial SVM-based SAR model. Unlabeled compounds close to the decision boundary of the SVM model were then selected and treated as additional positive training examples for learning a new SVM model. This process was repeated multiple times until the performance of the learned model cannot be improved further. A critical challenge of active learning is that such strategies suffer from the incestuous training bias problem (i.e., when the newly added data are actually not precisely predicted, this introduces significant noise so as to make the predictions intractable).

Until now, single-target-oriented SAR model learning and chemogenomics-based methods have been dominating and considered as standard in silico SAR methodology. A very natural attempt is to go beyond the concept of "protein family" and consider the entire biological space. The main idea is that if proteins even from different families somehow share some similarities in terms of their binding site structures and ligand similarities, and so on, then they can be used to learn each other's SAR models [Ning et al., 2009]. These authors first identify a set of related proteins from entire biological space with respect to the protein under consideration, and utilize such proteins and their known ligands as an additional complementary source of information to build a SAR model for the target protein. The key difference of this methodology from conventional chemogenomics is that it is not restricted by protein families and thus explores a larger space so as to increase the possibility of finding the most informative and related protein-ligand interactions. From this perspective, chemogenomics can be considered as a special case of Ning et al.'s [2009] methodology. Ning et al. [2009] developed different measures for related protein selection based on protein sequence similarity and binding ligand similarity, which was then applied to multi-task, semi-supervised learning from label propagation and classifier ensembles [Swanson and Tsai, 2003; Shen and Chou, 2006] so as to utilize related proteins and their ligands for better SAR model learning. These methods achieve better results than chemogenomics schemes. There is insufficient literature discussing methods that go beyond protein families for SAR modeling, Nonetheless, we believe that intuitively there would be subtle signals, not well identified or understood, that are shared across different proteins, given that protein-ligand binding interaction is very local to the protein-binding sites, and the underlying principles of natural evolution should illuminate studies as to how to collaboratively explore biological and chemical space. With the further development of protein structure crystallography and chemical screening methods, we may expect to gain better understanding from both biological and chemical space so as to improve connectivity.

REFERENCES

Agrafiotis D, Bandyopadhyay D, Wegner J, vanVlijmen H. 2007. Recent advances in chemoinformatics. J Chem Info Model 47: 1279–1293.

An A, Wang Y. 2001. Comparisons of classification methods for screening potential compounds. San Jose, California: Proceedings of the IEEE International Conference on Data Mining; November 2001. p 11–18.

Bajorath J. 2002. Integration of virtual and high-throughput screening. Nat Rev Drug Discov 1:882–894.

Bennett KP, Bennett KP, Embrechts MJ, Embrechts MJ. 2003. An optimization perspective on kernel partial least squares regression. In: Advances in learning theory: methods, models and applications. Belgium: IOS press. p 227–250.

Birault V, Harris CJ, Le J, Lipkin M, Nerella R, Stevens A. 2006. Bringing kinases into focus: efficient drug design through the use of Chemogenomic toolkits. Curr Med Chem 13:1735–1748.

Bock JR, Gough DA. 2002. A new method to estimate ligand-receptor energetics. Mol Cell Proteom 1:904–910.

Bock J, Gough D. 2005. Virtual screen for ligands of orphan G protein-coupled receptors. J Chem Info Model 45:1402–1414.

Bonilla E, Agakov F, Williams C. 2007. Kernel multi-task learning using task-specific features. In: Proceedings of the 11th International Conference on Artificial Intelligence and Statistics. San Juan, Puerto Rico; March 2007. p 21–24.

Bravi G, Green EGD, Hann V, Mike M. 2000. Modeling structure-activity relationship. In: Bohm H, Schneider G, editors. Virtual screening for bioactive molecules, vol. 10. Weinheim: Wiley-VCH. p 81–116.

Bredel M, Jacoby E. 2004. Chemogenomics: an emerging strategy for rapid target and drug discovery. Nat Rev Genet 5:262–275.

Breiman L. 2001. Random forests. In: Machine learning. Netherlands: Kluwer Academic Publishers 45:5–32.

Bulseco DA, Wolf DE. 2003. Fluorescence correlation spectroscopy: molecular complexing in solution and in living cells. Methods Cell Biol 72:465–498.

Byvatov E, Fechner U, Sadowski J, Schneider G. 2003. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. J Chem Info Comput Sci 43: 1882–1889.

Caron PR, Mullican MD, Mashal RD, Wilson KP, Su MS, Murcko, MA. 2001. Chemogenomic approaches to drug discovery. Curr Opin Chem Biol 5:464–470.

Caruana RA. 1993. Multitask learning: a knowledge-based source of inductive bias. In: Proceedings of the Tenth International Conference on Machine Learning. University of Massachusetts, Amherst: Morgan Kauffmann. p 41–48.

Chen X, Rusinko A, Young SS. 1998. Recursive partitioning analysis of a large structure-activity data set using three-imensional descriptors. J Chem Info Comput Sci 38:1054–1062.

Darnag R, Mazouz EM, Schmitzer A, Villemin D, Jarid A, Cherqaoui D. 2010. Support vector machines: Development of qsar models for predicting anti-hiv-1 activity of tibo derivatives. Eur J Med Chem 45:1590–1597.

Davies EK. 1996. Molecular diversity and combinatorial chemistry: Libraries and drug discovery. Am Chem Soc 118:309–316.

Dehaspe L, Toivonen H, King RD. 1998. Finding frequent substructures in chemical compounds. In: Agrawal R, Stolorz P, Piatetsky-Shapiro G, editors, Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. New York, NY: AAAI Press. p 30–36.

Deng Z, Chuaqui C, Singh J. 2004. Structural interaction fingerprint (sift): a novel method for analyzing three-dimensional protein-ligand binding interactions. J Med Chem 47:337–344.

Deshpande M, Kuramochi M, Wale N, Karypis G. 2005. Frequent substructure-based approaches for classifying chemical compounds. IEEE Transact Knowledge Data Eng 17:1036–1050.

Dobson CM. 2004. Chemical space and biology. Nature 432: 824–828.

Eguchi M, McMillan M, Nguyen C, Teo J-L, Chi EY, Henderson WR, Kahn M. 2003. Chemogenomics with peptide secondary structure mimetics. Comb Chem High Throughput Screen 6:611–621.

Erhan D, L'Heureux P-J, Yue SY, Bengio Y. 2006. Collaborative filtering on a family of biological targets. J Chem Inform Model 46:626–635.

Evgeniou T, Micchelli CA, Pontil M. 2005. Learning multiple tasks with kernel methods. J Mach Learn Res 6:615–637.

Frimurer TM, Ulven T, Elling CE, Gerlach L-O, Kostenis E, Hgberg T. 2005. A physicogenetic method to assign ligand-binding relationships between 7tm receptors. Bioorg Med Chem Lett 15:3707–3712.

Froehlich H, Wegner JK, Sieker F, Zell A. 2005. Optimal assignment kernels for attributed molecular graphs. In: Proceedings of the 22nd International Conference in Machine Learning. Bonn, Germany: ACM Press. p 225–232.

Frye S. 1999. Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. Chem Biol 6:R3–R7.

Gaither LA. 2007. Chemogenomics approaches to novel target discovery. Expert Rev Proteom 4:411–419.

Gärtner T. 2002. Exponential and geometric kernels for graphs. In NIPS*02 workshop on unreal data. Principles of modeling nonvectorial data, 2002.

Geppert H, Humrich J, Stumpfe D, Gartner T, Bajorath J. 2009. Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors. J Chem Info Model 49:767–779.

Goldberg D, Nichols D, Oki BM, Terry D. 1992. Using collaborative filtering to weave an information tapestry. Commun ACM 35: 61–70.

Gonzalez J, Holder L, Cook D. 2001. Application of graph based concept learning to the predictive toxicology domain. In: PTC, Workshop at the 5th PKDD. Freiburg, Germany, September 2001; 3–5.

Guba W. 2006. Chemogenomics strategies for g-protein coupled receptor hit finding. Ernst Schering Res Found Workshop 58: 21–29.

Guha R, Jurs PC. 2005. Interpreting computational neural network QSAR models: a measure of descriptor importance. J Chem Info Model 45:800–806.

Hansch C, Maolney PP, Fujita T, Muir RM. 1962. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. Nature 194: 178–180.

Hansch C, Muir RM, Fujita T, Maloney PP, Geiger F, Streich M. 1963. The correlation of biological activity of plant growth-regulators and hloromycetin derivatives with hammett constants and partition coefficients. J Am Chem Soc 85:1824–2817.

Harris CJ, Stevens AP. 2006. Chemogenomics: structuring the drug discovery process to gene families. Drug Discov Today 11: 880–888.

Hasegawa K, Funatsu K. 2000. Partial least squares modeling and genetic algorithm optimization in quantitative structure activity relationships. SAR QSAR Environ Res 11:189–209.

Hert J, Willet P, Wilton, D, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A. 2004. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. Organ Biomol Chem 2:3256–3266.

Hoppe C, Steinbeck C, Wohlfahrt G. 2006. Classification and comparison of ligand-binding sites derived from grid-mapped knowledge-based potentials. J Mol Graph Model 24:328–340.

Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. 2005. Binding moad (mother of all databases). Proteins 60:333–340.

Hughes-Oliver JM, Brooks AD, Welch WJ, Khaledi MG, Hawkins D, Young SS, Patil K, Howell GW, Ng RT, Chu MT. 2008. Chemmodlab: a web-based Cheminformatics modeling laboratory. Cheminformatics 2:1–18.

Inglese J, Auld DS, Jadhav A, Johnson RL, Simeonov A, Yasgar A, Zheng W, Austin CP. 2006. Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. Proc Natl Acad Sci USA 103: 11473–11478.

Jacob L, Vert J-P. 2008. Protein-ligand interaction prediction: an improved Chemogenomics approach. Bioinformatics 24: 2149–2156.

Jacoby E. 2001. A novel Chemogenomics knowledge-based ligand design strategy: application to g protein-coupled receptors. Quant Struct-Activ Relat 20:115–123.

Jacoby E, Fauchere J-L, Raimbaud E, Ollivier S, Michel A, Spedding M. 1999. A three binding site hypothesis for the interaction of ligands with monoamine g protein-coupled receptors: Implications for combinatorial ligand design. Quant Struct-Activ Relat 18:561–572.

Jona G, Snyder M. 2003. Recent developments in analytical and functional protein microarrays. Curr Opin Mol Ther 5:271–277.

Kashima H, Tsuda K, Inokuchi A. 2003. Marginalized kernels between labeled graphs. In: Proceedings of the 20th International Conference in Machine Learning. Washington, DC. p 321–328.

Kawasumi M, Nghiem P. 2007. Chemical genetics: elucidating biological systems with small-molecule compounds. J Invest Dermatol 127:1577–1584.

Kellenberger E, Muller P, Schalon C, Bret G, Foata N, Rognan D. 2006. sc-pdb: an annotated database of druggable binding sites from the protein data bank. J Chem Inf Model 46:717–727.

King RD, Muggleton S, Lewis RA, Sternberg JE. 1992. Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. Proc Natl Acad Sci 89: 11322–11326.

King RD, Muggleton SH, Srinivasan A, Sternberg MJE. 1996. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenecity by inductive logic programming. Proc Natl Acad Sci 93: 438–442.

King, RD, Srinivasan A, Dehaspe L. 2001. Warmr: A data mining tool for chemical data. J Comput-Aided Mol Des 15:173–181.

Klabunde T. 2007. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. Br J Pharmacol 152:5–7.

Klabunde T, Jger R. 2006. Chemogenomics approaches to g-protein coupled receptor lead finding. Ernst Schering Res Found Workshop 58:31–46.

Klopman G. 1998. The multicase program ii. baseline activity identification algorithm (baia). J Chem Info Comput Sci 38: 78–81.

Kotsiantis SB. 2007. Supervised machine learning: a review of classification techniques. Informatica 31:249–268

Kozaz K, Kozak M, Stapor K. 2007. Kernels for chemical compounds in biological screening. Adaptive and Natural Computing Algorithms: Lecture Notes in Computer Science 4432:327–337.

Kunin V, Cases I, Enright A, de Lorenzo V, Ouzounis C. 2003. Myriads of protein families, and still counting. Genome Biol 4: 401.

Lapinsh M, Prusis P, Uhlen S, Wikberg JES. 2005. Improved approach for proteochemometrics modeling: application to organic compound–amine G protein-coupled receptor interactions. Bioinformatics 21:4289–4296.

Le SQ, Ho TB, Phan TTH. 2003. Heuristics for chemical compound matching. Genom Informat 14:144–153.

Le SQ, Ho T, Phan TTH. 2004. A novel graph-based similarity measure for 2d chemical structures. Genome Informat 15:82–91.

Lindström A, Pettersson F, Almqvist F, Berglund A, Kihlberg J, Linusson A. 2006. Hierarchical pls modeling for predicting the binding of a comprehensive set of structurally diverse protein-ligand complexes. J Chem Info Model 46:1154–1167.

Lipinski C, Hopkins A. 2004. Navigating chemical space for biology and medicine. Nature 432:855–861.

Livingstone D, Manallack D. 2003. Neural networks in 3d. QSAR. QSAR Comb Sci 22:510–518.

Martin RE, Green LG, Guba W, Kratochwil N, Christ A. 2007. Discovery of the first nonpeptidic, small-molecule, highly selective somatostatin receptor subtype 5 antagonists: a chemogenomics approach. J Med Chem 50:6291–6294.

Matsuda T, Motoda H, Yoshida T, Washio T. 2002. Mining patterns from structured data by beam-wise graph-based induction. In: Proceedings of the 5th International Conference on Discovery Science Discoveery (DS 2002), vol. 2534 of Lecture Notes in Computer Science. New York: Springer-Verlag. p 422–429.

Mccallum A, Nigam K. 1998. Employing EM and pool-based active learning for text classification. In: ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc. p 350–358.

Menchetti S, Costa F, Frasconi P. 2005. Weighted decomposition kernels. In: Proceedings of the 22nd International Conference in Machine Learning, vol. 119. Bonn, Germany. ACM International Conference Proceeding Series 119 ACM. p 585–592.

Muggleton S, De Raedt L. 1994. Inductive logic programming: Theory and methods. J Logic Program 19:629–679.

Muggleton SH, Feng C. 1992. Efficient induction of logic programs. In: Muggleton S, editor. Inductive logic programming, London: Academic Press. p 281–298.

Nguyen C, Teo J-L, Matsuda A, Eguchi M, Chi EY, Henderson WR, Kahn M. 2003. Chemogenomic identification of ref-1/ap-1 as a therapeutic target for asthma. Proc Natl Acad Sci USA 100: 1169–1173.

Nicolaou C, Tamura S, Kelley B, Bassett S, Nutt R. 2002. Analysis of large screening data sets via adaptively grown phylogenetic-like trees. J Chem Info Comput Sci 42:1069–1079.

Nigsch F, Mitchell JBO. 2008. How to winnow actives from inactives: Introducing molecular orthogonal sparse bigrams (mosbs) and multiclass winnow. J Chem Inform Model 48:306–318.

Nigsch F, Bender A, Jenkins JL, Mitchell JBO. 2008. Ligand-target prediction using winnow and naive bayesian algorithms and the implications of overall performance statistics. J Chem Inform Model 48:2313–2325.

Ning X, Rangwala H, Karypis G. 2009. Multi-assay-based structure-activity relationship models: improving structure-activity relationship models by incorporating activity information from related targets. J Chem Inform Model 49:2444–2456.

Oprea TI, Gottfries J. 2001. Chemography: the art of navigating in chemical space. J Comb Chem 3:157–166.

Otto M. 1999. Chemometrics. Weinheim: Wiley-VCH.

Otto M. 2007. Chemometrics: statistics and computer application in analytical chemistry. Weinheim: Wiley-VCH Verlag GmbH.

Ralaivola L, Swamidassa SJ, Saigo H, Baldi P. 2005. Graph kernels for chemical informatics. Neural Networks 18:1093–1110.

Raymond JW, Willett P. 2002. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. J Comput-Aided Mol Des 16:521–533.

Rogers D, Brown R, Hahn M. 2005. Using extended-connectivity fingerprints with laplacian-modified bayesian analysis in high-throughput screening. J Biomol Screen 10:682–686.

Rognan D. 2007. Chemogenomic approaches to rational drug design. Br J Pharmacol 152:38–52.

Rosipal R, Krämer N. 2006a. Overview and recent advances in partial least squares. Subspace, Latent Structure and Feature Selection. Lecture Notes in Computer Science 3940:34–51.

Roy P, Roy K. 2008. On some aspects of variable selection for partial least squares regression models. QSAR Combin Sci 28:302–313.

Rusinko A, Farmen MW, Lambert CG, Brown PL, Young SS. 1999. Analysis of a large structure/biological activity data set using recursive partitioning. J Chem Inform Comput Sci 38:1054–1062.

Shen H-B, Chou K-C. 2006. Ensemble classifier for protein fold pattern recognition. Bioinformatics 22:1717–1722.

Sheridan RP, Miller MD, Underwood DJ, Kearsley SJ. 1996. Chemical similarity using geometric atom pair descriptors. J Chem Inform Comput Sci 36:128–136.

Smola AJ, Olkopf BS. 1998. A tutorial on support vector regression. Technical report. Statistics and Computing 14:199–222.

Stiefl N, Watson IA, Baumann K, Zaliani A. 2006. Erg: 2d pharmacophore descriptor for scaffold hopping. J Chem Info Model 46:208–220.

Strömbergsson H, Kleywegt G. 2009. A chemogenomics view on protein-ligand spaces. BMC Bioinform 10:S13.

Strömbergsson H, Daniluk P, Kryshtafovych A, Fidelis K, Wikberg JES, Kleywegt GJ, Hvidsten TR. 2008. Interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space. J Chem Inform Model 48: 2278–2288.

Surgand J-S, Rodrigo J, Kellenberger E, Rognan D. 2006. A chemogenomic analysis of the transmembrane binding cavity of human g-protein-coupled receptors. Proteins 62:509–538.

Sussman NB, Arena VC, Yu S, Mazumdar S, Thampatty BP. 2003. Decision tree SAR models for developmental toxicity based on an FDA/TERIS database. SAR and QSAR in Environmental Research 14:83–96.

Swanson R, Tsai J. 2003. Pretty good guessing: protein structure prediction at CASP5. J Bacteriol 185:3990–3993.

Tetko IV, Kovalishyn VV, Livingstone DJ. 2001. Volume learning algorithm artificial neural networks for 3D QSAR studies. J Med Chem 44:2411–2420.

Tolliday N, Clemons PA, Ferraiolo P, Koehler AN, Lewis TA, Li X, Schreiber SL, Gerhard DS, Eliasof S. 2006. Small molecules, big players: the national cancer institute's initiative for chemical genetics. Cancer Res 66:8935–8942.

Vapnik V. 1998. Statistical learning theory. New York: John Wiley.

Vieth M, Higgs RE, Robertson DH, Shapiro M, Gragg EA, Hemmerle H. 2004. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. Biochim Biophys Acta 1697:243–257.

Wale N, Ning X, Karypis G. 2010. Trends in chemical graph data mining, Managing and Mining Graph Data, Advances in Database Systems 40:581–606.

Wale N, Watson IA, Karypis G. 2007. Comparison of descriptor spaces for chemical compound retrieval and classification. Knowledge and Information Systems. Knowledge and Information System 14:347–375.

Warmuth MK, Liao J, Ratsch G, Mathieson M, Putta S, Lemmen C. 2003. Active learning with support vector machines in the drug discovery process. J Chem Inform Comput Sci 43:667–673.

Weill N, Rognan D. 2009. Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: Application to g protein-coupled receptors and their ligands. J Chem Info Model 49:1049–1062.

Xia X, Maliski EG, Gallant P, Rogers D. 2004. Classification of kinase inhibitors using a bayesian model. J Med Chem 47: 4463–4470.

Zhang Q-Y, Aires-de Sousa J. 2007. Random forest prediction of mutagenicity from empirical physicochemical descriptors. J Chem Inform Model 47:1–8.

Zhou Y-P, Cai C-B, Huan S, Jiang J-H, Wu H-L, Shen G-L, Yu R-Q. 2007. SAR study of angiotensin ii antagonists using robust boosting partial least squares regression. Anal Chim Acta 593:68–74.

Zhu X. 2005. Semi-supervised learning literature survey. Technical report, Computer Sciences, Madison: University of Wisconsin-Madison.