Issues in Computational Modeling of Molecular Structure-Property Relationships in Real-World Settings

Rahul Singh
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332
rsingh@ece.gatech.edu

ABSTRACT

The problem of modeling structure-property relationships is a fundamental one in contemporary biology and drug discovery. An accurate model can not only be used to predict the behavior of a molecule and understand how structural variations may influence molecular property, but also to identify regions of molecular space that hold promise in context of a specific investigation. However, a variety of factors contribute to the difficulty of constructing robust structure activity models in real-world problems. These include conceptual issues related to how well the true biological problem is accounted for in the computational solution, algorithmic issues associated with determining the proper molecular descriptors, access to small quantities of data (possibly on tens of molecules only) due to the high cost and complexity of the experimental process, and the complex nature of bio-chemical phenomena underlying the data. This paper attempts to address this problem from the rudiments. We first identify and discuss the salient computational issues that span (and complicate) structure-property modeling formulations. We then consider a specific problem: that of modeling intestinal drug absorption, where many of the aforementioned factors play a role. In addressing them, our solution uses a novel characterization of molecular space based on the notion of surface-based molecular similarity. This is followed by identifying a statistically relevant set of molecular descriptors, which along with an appropriate machine learning technique is used to build the structure-property model. To ensure robustness, we propose simultaneous use of both ratio and ordinal error-measures for model construction and validation. The applicability of the approach is demonstrated through the predictive capability of the model in real world situations.

Keywords: Structure-Property Modeling, Molecular Similarity, Machine Learning, Data Mining, ADME, Drug Discovery

1. INTRODUCTION

The recent past in human history has been witness to several significant events in the evolution of our understanding of biology and medicine. These include among others, the first modern-day commercial drug Aspirin, the elucidation of the structure of the DNA, understanding the cell-cycle, development of rational drug design (especially against well identified targets like ACE-inhibitors), cloning of proteins, and most recently, the sequencing of the human genome and mapping of the genomic DNA [12]. Considering the fact that all known commercial drugs today, interact with no more than 500 distinct targets [17], Genomics promises to provide a proliferation of targets that may not only lead to newer or improved therapeutics, but also open exciting avenues like individualized medicine. Somewhat simultaneously, recent developments in industrial robotics, combinatorial chemistry, and high-throughput screening have significantly increased the number of lead compounds synthesized in pharmaceutical drug-discovery settings [10] [17]. Taken together, these factors may be assumed to point to both advancements in eradication and treatment of diseases as well as a significant reduction in the time to market (currently approximately 14 years on average per drug) and cost (currently 100-897 million dollars per drug [7]) of drug discovery. Unfortunately, the trends from pharmaceutical science and industry differ considerably. A recent detailed study involving the pharmaceutical sector [17] accessed the impact of genomics on biopharmaceutical drug development. Broadly speaking, [17] found that the cost and number of failures in drug discovery can be expected *to increase* in the immediate future. This startling result can be explained due to two factors. First, once a target is identified, it needs to be validated to establish its role in a disease. Moreover, its interactions with other genes/targets have to be identified as well, for example, by elucidating the pathways it is involved in. However, validation is a complex, non-standardized process and the advancements in Genomics have primarily increased our capabilities in identifying new targets, not in validating them. This has typically resulted in many insufficiently validated targets being considered for drug discovery. Second, newer targets often require that newer classes of molecules be designed to interact with them. However, owing to their structural novelty little historical data is available on the pharmacokinetics (influence of the human biological system on the drug molecule), pharmacodynamics (influence of the drug molecule on the human body), or toxicity of such compounds. These properties are essential for a successful drug but are typically tested for, in the later stages of drug discovery due to the associated time and cost. This can leads to late stage attrition when the pharmacology of the molecules is found to be unsuitable.

It is increasingly being recognized that computational approaches can play a significant role in biology and drug discovery, not only at the level of data management, sequence/structure comparison and analysis, but also in modeling behavior of molecules and other bio-chemical systems *in-silico* [20]. This could, for example be, characterization of the relationship between the structure of a molecule and its properties like Binding, Localization, or Expression. The above illustration typifies an important research direction of *in-silico* biology called structure-property modeling (also known as structure-activity modeling). A structure-property model captures the relationship between the bio-chemical properties of a molecule and its physicochemical description [4], [13] by essentially envisaging the biochemical property Φ of a molecule M_i as the function of its "chemical constitution" [4, 18]:

$$\Phi = f(M_i) \qquad (1)$$

The basic elements needed for the development of a structure-property model are:

- A set of parameters describing the molecular structure [9].
- Verified assay (experimental) results describing the bio-chemical property of interest [9].
- The learning formulation. This can be (a) *Classification*, involving estimation of class decision boundaries, (b) *Regression*, requiring estimation of an unknown continuous function from noisy samples, or (c) *Probability density estimation.*
- A statistical or machine learning technique (e.g. multivariate regression, discriminant analysis, neural networks, or support-vector machine), to establish a relationship between the chemical description and bio-chemical property measurements, with respect to the aforementioned forms of the learning formulation.

The concept of structure-property modeling significantly predates the idea of *in-silico* biology [4, 14]. However, it has become an integral part of this paradigm today, especially for predictive modeling. Typically, the properties which are modeled are those that are either too expensive or too time consuming to determine experimentally for large sets of molecules. In the context of increased late stage failures in drug discovery, one possibility is to build structure-property models that can be used to predict critical late stage behavior in terms of the pharmacokinetics and/or pharmacodynamics of a drug molecule. Such models can then be used to identify early on in the drug discovery process, the set of molecules that will potentially minimize the probability of late stage failures due to poor pharmacological attributes.

The issues and approaches considered in this paper are based on our experience in real world pharmaceutical settings. We identify the salient challenges that are encountered in developing complex structure-property models and present our approach in addressing them. For specificity, we focus in this paper on the problem of modeling human intestinal absorption of small (drug) molecules. However, the general principles described here are equally applicable both for large molecules and other structure-property relationships. We start this paper in Section 2, by providing a brief overview of the drug-discovery process and cover the basic biology behind the specific structure-property model we will consider. A brief review of some recent efforts in the computer science community towards solving such problems is also presented. We follow this by enumerating the primary challenges encountered while developing structure-property models in real world situations in Section 3. An approach, investigated by us, that holds promise in context of the issues identified by us in Section 3, is covered in Section 4. The experimental results from our work, based on modeling human intestinal absorption using real-life data is presented in Section 5. Finally, the conclusions from this research are presented in Section 6.

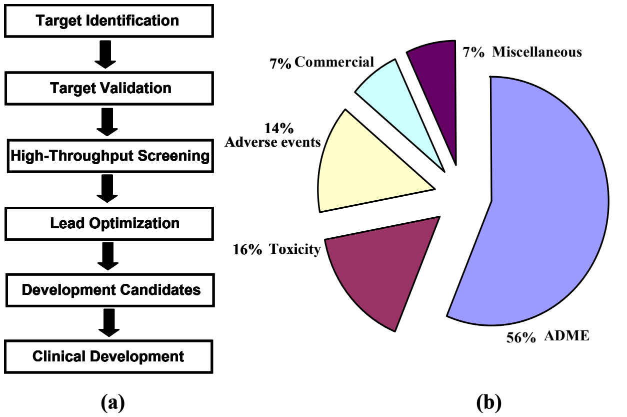

Figure 1: (a) Stages in a typical drug discovery process. (b) Primary causes for late-stage failures

2. BACKGROUND

The process of drug discovery is a complex and multi-stage one, with the goal of discovering therapeutically useful products called drugs. Generally, but not exclusively (antibody-based therapeutics is one exception), a drug is a small molecule that interacts with a target, typically a protein or an enzyme, in a therapeutically beneficial manner. The primary stages of a standard drug discovery process are shown in Figure 1 (a). This process begins by determining potentially interesting targets. The identification of a target for drug discovery is done by screening gene databases using homology (sequence similarity) to known targets, co-location (which organ/tissue the gene is expressed in), or other criteria like similar expression profiles that can indicate a biological connection between the gene and targeted symptoms or phenotypes. Once identified, the target is validated using molecular biology techniques like gene-knockouts and/or computational approaches like expression analysis along with pathway elucidation. A validated target is then screened against a large number of small molecules (potentially in millions) to determine which of them (called *hits*) interact with the target. The hits are further analyzed and optimized in terms of properties like binding potency, pharmacokinetics (PK), pharmacodynamics (PD), and efficacy, to come up with the molecules that are most suited to undergo clinical trials. In typical settings, the number of *hits* against a target represents a two to three order of magnitude reduction in the number of molecules being considered at the start. During the lead optimization stage, the initial focus is on maximizing potency by minimizing the concentration at which the drug-target interaction occurs. This further reduces the candidate list of molecules to (typically) a few hundred. Of these, after testing for PK/PD, only tens of molecules or fewer may remain to be considered for clinical trials.

The various stages of drug discovery are essentially filters to weed out unsuitable molecules, with the low-throughput or more expensive experimental stages occurring later in the pipeline. Such a staging minimizes average cost and time. However, failures when they do happen in the later stages become exceedingly critical, both in monetary terms and even more importantly, in terms of time. In Figure 1 (b), we present some of the primary causes that lead to late-stage failures. As can be noted, a significant majority of these is due to poor ADMET

(Absorption, Distribution, Metabolism, Excretion, and Toxicity) characteristics of the molecules. The first of these properties, *Absorption*, refers to the intestinal permeability of a (potential) drug molecule, and is particularly important for oral-drug entry into the body. Intestinal permeability is influenced by both passive factors like diffusion due to concentration gradient and active factors like monocarboxylic acid carrier (for transporting salicylic acid) and efflux systems like P-glycoprotein. Other factors that can also play a role include electrostatic interactions between the drug and the lipid surface and partitioning into and across the lipid phase. While a detailed analysis of the biology behind intestinal absorption is beyond the scope of this paper (the interested reader is referred to [23]), it can be inferred from recent research in biology that several transport mechanisms related to intestinal absorption have been identified. However, their relative influence on in-vivo drug absorption is yet to be determined. From the computational perspective, this implies that there is as of now, insufficient knowledge to develop analytical (also known as "hard" [25] or "first-principle" [5]) models for intestinal absorption. The alternative is to develop "soft" models for such properties by using formal or informal fitting techniques to match a mathematical model's behavior to that of the observed (experimental) data.

A review of recent literature shows that different research efforts in the computer science community have started to explore this problem. Based on frequent subgraph discovery, an algorithm is proposed in [6] to identify all topological and geometric sub-structures that are present in a dataset and distinguish compounds that constitute hits. Another approach based on inductive logic programming, is used in [22] for classifying chemical compounds. A number of researchers have employed string-based molecular representation to discover frequently occurring substrings (corresponding to sub-structures) [21]. In [27] the idea of graph-based substructure characterization is extended to closed frequent graphs: A graph ***G*** is defined to be a closed graph, if it does not consist of any subgraph ***g*** having the same support as that of ***G*** itself. In this approach, the problem associated with having an exponential number of frequent subgraphs in a graph is avoided. In [28] a support vector machine is used to determine important features that define drug activity. Other recent techniques [2, 8] have considered graph-based structure characterization of molecules. It may be noted that graph theoretic approaches towards characterizing molecular properties have a long history of research in computational chemistry. A review of these results can be obtained from [11] and references therein. The primary inadequacy of such approaches lies in that they can not properly represent the 3D nature of a molecule and associated molecular properties. For a review of various molecular descriptors and their representational capabilities, we refer the reader to [21].

3. CHALLENGES IN STRUCTURE-PROPERTY MODELING

A variety of issues typically encountered during real-world drug discovery, complicate the task of building accurate structure-property models. The primary challenges that a modeling process needs to consider are:

- *Broad applicability:* The modeling framework should be applicable to various classes of molecules. It may be noted, that this criterion does not necessarily imply that a specific model has to predict biological activity across different structural classes. Rather, the criterion seeks to promote techniques that are general in nature.

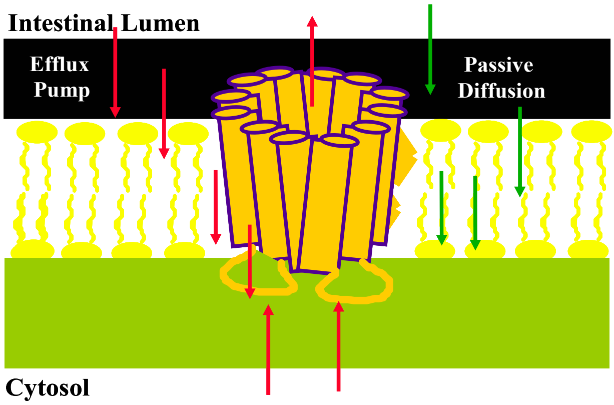

Figure 2: Schematic representation of factors (see text) involved in intestinal membrane permeability.

- *Interpretability:* The results of a classification or prediction must be interpretable in structural terms so that either new molecules can be synthesized or structural aspects introduced/consolidated in existing ones to obtain the desired bio-chemical behavior.
- *Performance:* The calculation of descriptors must not be a rate-limiting step. This is critical because structure-property models can be used during initial stages of drug discovery, where millions of molecules are typically involved. Also, the models can be used on virtual libraries that consist of molecules that have been designed, but not synthesized. The size of such virtual libraries can easily exceed those of the real ones by one or two orders of magnitude.
- *Training set size:* Many biological properties like intestinal absorption are expensive to determine experimentally. Therefore, the model construction (learning) stage may not have access to large amounts of training data, especially if the property being modeled is typically investigated in the later stages of a drug discovery process.
- *Influence of experimental conditions:* Determination of many biological properties depends on experimental protocols involving among others, experimental conditions and choice of reagents. For example, *Verapamil* a P-glycoprotein substrate permeates the human jejunum at higher concentrations [23]. The permeation rate, in this case, can not be fully explained through molecular structure or derived properties alone. A structure-property model that does not (either explicitly or implicitly) account for such factors is difficult to validate using commonly available experimental data and is of limited use.
- *Descriptor selection:* A molecule can be represented using various types of descriptors. For example, it may be represented using simple physical-chemical properties like molecular weight, number of atoms, or its octanol-water partition coefficient. Alternatively, it could be represented by graph-based descriptors that seek to characterize molecular substructures as subgraphs. Molecules may also be represented by complex surface-based descriptors that can represent surface properties and intra-molecular interactions

like superposition-effects. It is important to note that different descriptors have different representational capabilities [21]. Furthermore, there is no current scientific agreement on a general set of descriptors. Therefore, it is imperative to select descriptors in a manner that takes into account the fundamental nature of the bio-chemical phenomenon being modeled.

- *Dimensionality of the descriptor set:* Typically molecular descriptors tend to be complex and high-dimensional. For example, the surface based descriptor described in [15] consists of 265 elements, corresponding to the points around a molecule where specific surface properties are measured. Similarly, the sub-structure based descriptor used for representing molecules in the ISIS system [19] has close to a thousand relevant features. In our context, the high-dimensionality of molecular descriptors leads to two repercussions: First, given the typically small number of molecules for which data is available for model building, issues related to model overfitting can become significant. Second, recent results [1, 3] suggest that under reasonable assumptions, for a wide variety of data distributions, the ratio of the distances of the nearest neighbor to the farthest neighbor is almost constant in high-dimensional spaces. This implies that the concept of nearest-neighbor may be ill-posed in high dimensional spaces since the distinction between distances to different data point does not exist. This can cause fundamental problems in learning formulations and algorithms.
- *Management of modeling errors*: Errors in modeling which are manifested during operation/prediction are either numeric (typical in a regression formulation) or misclassifications (false positives or false negatives). Generally, it can not be assumed that such errors influence the overall process uniformly. For example, the effect of misclassifying a promising molecule (a false negative), leading to its rejection is typically considered worse than having a "reasonable" number of false positives. Also, numeric errors that incorrectly change the rank-order of a molecule are undesirable even if their magnitudes are small, since this can change the prioritization of the molecules. Therefore, proper error modeling and handling needs to be an inseparable part of structure-property modeling.

4. MODELING STRUCTURE-PROPERTY RELATIONS

The proposed modeling approach consists of the following stages: (i) descriptor design, (ii) determination of the characteristic molecules, (iii) descriptor generation, (iv) feature selection, (v) model construction, and (vi) model application/prediction. In the following we describe each one of these stages.

***Descriptor design**:* The design of descriptors needs to take into account the fundamental bio-chemistry that we seek to model. This is one of the most crucial steps for ensuring that the computational solution targets the true biological problem and not a computational idealization of it. In our specific case, a recent study [26] on over 1100 drug candidates conducted at GlaxoSmithKline show that the most important molecular properties that influence oral bioavailability are: (i) molecular shape and its conformations (deformations) as defined by the number of rotatable bonds in a molecule and (ii) the polar-surface area of a molecule as defined by the number of H-bond donor and acceptor atoms. To account for these aspects, we define descriptors which are determined by placing a molecule inside a tessellated sphere at the surface of which properties like molecular geometry, and donor/acceptor fields can be measured. For example, the (donor or acceptor) field at point P_j due to an appropriate atom at position X_i having van der Walls radii r_i is defined as:

$$f(P_j, X_i) = \left(\frac{a^2}{2\pi r_i^2}\right)^{\frac{3}{2}} \exp\left(\frac{-a^2}{2r_i^2} \mid X_i - P_j \mid^2\right) \qquad (2)$$

Additionally, the molecular shape is estimated by computing the distance from the surface of the sphere to the *molecular surface* obtained by rolling a probe-atom over the molecule. For details on the descriptor design, we refer the reader to [21]. The computed octanol-water partition coefficient (clogP), which indicates the lipophilicity of a molecule, is used as an additional descriptor to geometry and donor/acceptor fields.

Determination of characteristic molecules**:* One of the primary challenges induced by the aforementioned descriptor is its high dimensionality, as the sphere encapsulating the molecule needs to be densely sampled to capture the relevant molecular properties. In our current research the sphere is tessellated at 2000 points. Other 3D molecular similarity techniques like [15] have used a lower tessellation frequency (265 points). However, this can lead to loss of resolution and still leads to unacceptable high-dimensionality of descriptors. Our strategy is based on a form of dimensionality reduction and uses a predefined set of molecules, called *characteristic molecules*. The idea behind selecting such a set is to tessellate and quantize the *d*-dimensional molecular descriptor space *D* into a finite subset *C* of the *d*-dimensional space. Formally, this process can be denoted as a mapping ***Q, which is defined as:

$$Q : D^d \to C, C = \{c_1, c_2, ..., c_m\} \wedge \forall j, c_j \in D^d \qquad (3)$$

In this approach, each characteristic molecule c_j is defined in the high-dimensional descriptor space. By selecting *m*, the number of characteristic molecules to be less than *d*, a dimensionality reduction can be achieved. In its essence, this idea is closely related to the concepts behind vector quantization [5].

***Descriptor generation**:* The descriptor generation process involves computing the similarity score of each input molecule with respect to the characteristic molecules. Computation of the similarity should take into account the 3D shape of the molecule, superposition-effects, and molecular conformations. To do this in a manner, that is not rate-limiting, we use the topologically-constrained histogram intersection technique proposed in [21]. This technique builds on the idea of histogram intersection [24], to compute the similarity of molecular property distributions. By utilizing constrains on how molecular properties are distributed on the surface of the sphere, [21] ameliorates the necessity of full pose optimization in 6-DOF. This leads to very rapid and accurate 3D surface-based similarity determination.

***Feature selection**:* In the feature selection step, the pair-wise correlation between all descriptors is used to determine the descriptor set that is least intra-correlated. This step removes redundant information in the descriptors and further reduces the descriptor dimensionality.

Model construction: In the model construction phase, a learning algorithm (backpropogation-based neural network with a single hidden layer), is used to build a structure-property model using the descriptors made available from the previous step for the training molecules. The training is stopped when the cross-validated error becomes lower than a predefined threshold. The evaluation of the model is based on using two measures. The first is a ratio-measure called cross-validated r^2 and shows how well the model predicts data not used during model construction. It is defined as:

$$r^2 = 1 - \frac{\sum_i (V_i - P_i)^2}{\sum_i (V_i - \bar{V})^2} \qquad (4)$$

Where V_i is the experimentally determined property of the molecule *i*, P_i is its predicted molecular property, and $\bar{V}$ is the mean experimental property value. The second measure is an ordinal measure called Kendall's τ, that shows how well the *ordering* of the data is preserved during prediction by the model. A perfect value of $\tau = 1$ is obtained when the predicted order coincides with the order as determined by actual experimental property values. This measure is computed, for *n* molecules as:

$$\tau = \frac{correct\_ordering - incorrect\_ordering}{n(n-1)/2} \qquad (5)$$

Using a combination of the above measures allows evaluation of a model both in terms of its numeric predictive accuracy, and in terms of how well it can maintain prioritization of molecules.

Model application/prediction: In the predictive setting, first the similarity scores for the input molecules is computed with respect to the characteristic molecule set. Additionally clogP (or other physico-chemical descriptors) are also calculated. The final feature set defining the input molecules is obtained by selecting the least correlated descriptors. This information is entered in the model to obtain the property predictions.

Compund ID	Predicted Permeability	Actual Permeability
000753	1.79	1.83
091217	0.09	0.20
322835	0.09	0.24
422025	2.83	3.01
489595	3.16	3.01
525792	3.15	3.06
531746	0.09	0.15
598738	0.24	0.26
696705	2.36	2.51
835218	0.08	0.11

Table 1: Predicted and measured permeability values for the molecules in the test set. All permeability values are in %flux units. The compound ID is a unique numeric identifier defined for each molecule. It has no relation to the molecular structure

5. EXPERIMENTAL RESULTS

In this section we present experimental evaluation of the proposed approach using a dataset of 30 compounds that were tested for human intestinal permeation using the Caco-2 assay. The assay protocol was designed to measure uni-directional flux and all compounds were analyzed at identical initial concentrations. The range of measured values was between 0.0% (no permeation) to 2.8% (maximum permeation) flux units. Model learning and testing were conducted in two steps. For learning, 20 of the 30 molecules were made available. The data for these molecules consisted of their chemical structures and their permeation as defined by the flux values. The remaining 10 molecules constituted the test data. It may be noted that the small number of molecules available during this experiment is a typical scenario that is encountered in many real-world structure-activity modeling setting.

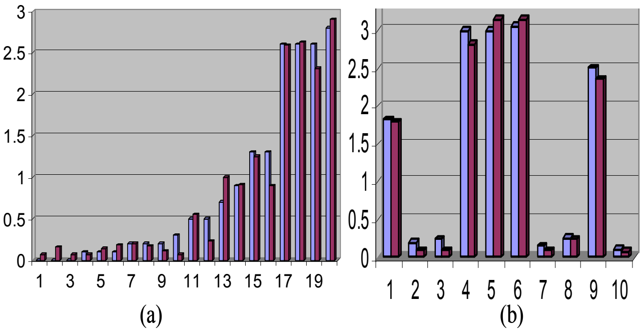

Figure 3: (a) Leave-one-out cross-validation results on the training set using the model. (b) Prediction performance on the test set. Predicted values shown as lighter shaded bars on the left. Actual (measured) values are darker bars on the right.

Model construction was done using the 20 training molecules. For each molecule in the training set, its corresponding structure was used to determine its surface-based similarity with respect to each of the 30 characteristic molecules. This information along with the computed octanol-water partition coefficient (clogP) constituted the (31 dimensional) descriptors. As part of the descriptor selection step, the complete cross-correlation matrix of the descriptors was computed and the top eight least correlated descriptors selected. A backpropogation network with a single hidden layer was used to learn the (empirical) mapping between the molecules as defined by the 8-dimensional feature vector and their permeability values. Learning was stopped when the cross-validated error became lower than a predefined threshold. Figure 3 (a) shows the performance of the learned model in a leave-one-out cross-validation setting for the training set. In this setting, one compound was randomly excluded from the training set and the remaining compounds used to learn a model that predicted the permeability for the excluded compound. For the model that was learnt, the cross-validated r^2 equaled 0.97 and the value for Kendall's τ was 0.65. We note, that our experience across experiments, suggests that values for Kendall's τ typically tend to be lower than those for cross-validated r^2, underlining thereby the necessity to use *both* these metrics in conjunction. For the test set, the descriptors were computed analogously to the training data. The raw prediction values for the test molecules, along with the experimentally determined values are presented in Table 1. In Figure 3(b) the same information is graphically presented. It may be noted, that in spite of the various computational challenges that accompanied the given problem, the proposed approach resulted in highly accurate structure-property models.

6. CONCLUSIONS

In this paper we considered the problem of building structure-property models for molecules having biological properties of interest. The paper presents various factor that cause complications for direct application of standard machine-learning or data mining algorithms to such problems. Based on these observations, we propose a general approach towards designing such models that ameliorates challenges associated with factors like lack of large quantities of data, high-dimensional descriptor spaces, and interpretability of resultant models. The efficacy of our approach is demonstrated by building structure-property models that demonstrate high predictability on real-world experimental data for human-intestinal absorption using Caco-2 cell lines.

7. REFERENCES

1. C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space", ICDT, pp. 420-434, 2001
2. M. Berthold and C. Borgelt, "Mining Molecular Fragments: Finding Relevant Substructures of Molecules", Proc. ICDM, 2002
3. K. S. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, "When is Nearest Neighbor Meaningful?", ICDT, pp. 217-235, 1999
4. C. Brown et. al., Transactions, Royal Society of Edinburgh, Vol 25, pp. 151-203, 1868-1869
5. V. Cherkassky and F. Mulier, "Learning From Data", Wiley Inter-Science, 1998
6. M. Deshpande, M. Kuramochi, and G. Karypis, "Frequent Sub-Structure-Based Approaches for Classifying Chemical Compounds", ICDM 2003
7. "Big Trouble for Big Pharma", The Economist, December 4, 2003
8. I. Eidhammer, I Jonasses, and W. R. Taylor, "Structure Comparison and Structure Patterns", Journal of Computational Biology, Vol.7, 2000
9. Enslein K., "An Overview of Structure-Activity Relationships as an alternative to Testing in Animals for Carcinogenicity, Mutagenicity, Dermal and Eye Irritation, and Acute Oral Toxicity", Toxicology and Industrial Health, Vol. 4, No. 4, pp. 479-497, 1988
10. B. Flickinger, "Using Metabolism Data in Early Development", Drug Discovery and Development, September 2001
11. J. Galvez, J. V. Julian-Ortiz, R. Garcia-Domenech, "General Topological Patterns of Known Drugs", Journal of Molecular Graphics and Modeling, Vol. 20, pp. 84-94, 2001
12. Genome International Sequencing Consortium, "Initial Sequencing and Analysis of the Human Genome", Nature 409, 860-921, February 2001.
13. Grover M., Singh B., Bakshi M., and Singh S., "Quantitative Structure-Property Relationships in Pharmaceutical Research – Part 1", PSTT, Vol 3, No. 1, 2000
14. C. Hansch, P. P. Maloney, T. Fujita, R. M. Muir, "Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients", Nature, Vol. 194, pp. 178-180, 1962
15. A. N. Jain, K. Koile, and D. Chapman, "Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark", Journal of Medicinal Chemistry, Vol 37, pp. 2315-2327, 1994
16. King R. D., Muggleton S., Srinivasan A., Sternberg M., "Structure-Activity Relationships Derived by Machine Learning: The Use of Atoms and their Bond Connectivities to Predict Mutagenicity by Inductive Logic Programming", Proc. Nat. Acad. Sci., Vol 93, pp. 438-442, Jan 1996
17. Lehman Brothers and McKinsey&Company, "The Fruits of Genomics", January, 2001.
18. D. J. Livingstone, "The Characterization of Chemical Structures Using Molecular Properties. A Survey", Journal of Chemical Information and Computer Science, Vol 40, pp. 195-209, 2000
19. http://www.mdli.com
20. B. Palsson, "The Challenges of in silico Biology", Nature Biotechnology, Vol. 18, Nov. 2000
21. R. Singh, "A Computer Vision-Based Approach for Answering Molecular Similarity Queries", Proc. Asian Conference on Computer Vision, 2004
22. A. Srinivasan and R. King, "Feature Construction with Inductive Logic Programming: A Study of Quantitative Predictions of Biological Activity Aided by Structural Attributes", Knowledge Discovery and Data Mining Journal, Vol. 3, pp. 37-57, 1999
23. P. Stenberg, K. Luthman, and P. Artursson, "Virtual Screening of Intestinal Drug Permeability", Journal of Controlled Release, Vol. 65, pp. 231-243, 2000
24. M. Swain and D. Ballard, "Color Indexing", Int. J. of Comp. Vision, 7, 1, pp 11-32, 1991
25. J. H. Taylor, "Modeling & Simulation of Dynamic Systems – a Tutorial", Proc. Fourth Intr. Symp. On Math. Model. And Simul. in Agricul. And Bio-industries, June 2001
26. D. Veber, S. Johnson, H-Y Cheng, B. Smith, K. Ward, and K. Kopple, "Molecular Properties That Influence the Oral Bioavailability of Drug Candidates", J. Med. Chem., Vol 45, pp. 2615-2623, 2002
27. X. Yan and J. Han, "CloseGraph: Mining Closed Frequent Graph Patterns", Proc. ACM SIGKDD, 2003
28. H. Yu, J. Yang, W. Wang, and J. Han, "Discovering Compact and Highly Discriminative Features or Feature Combinations of Drug Activities Using Support Vector Machines", Proc. IEEE Computer Society Bioinformatics Conference, 2003