

The goal of this research is to develop and evaluate a new framework for the virtual discovery of new pharmaceuticals. The basic idea is to utilize large existing pharmaceutical databases as input for a new type of structure/activity correlation methodology. In order to accomplish this, large sets of new and traditional descriptors are being generated using machine intelligence tools. These descriptors are then used to create improved Quantitative Structure-Activity Relationship (QSAR) models that characterize and predict important biological responses. Once the descriptors have been generated and predictive models have been built, thousands of new potential molecules, chemically similar to those of the benchmark data set, are scanned from large databases and evaluated for their chemical properties based on the predictive model. The aim is to target a few novel molecules with potentially attractive pharmaceutical properties and drug-like ADME physical characteristics that can then be screened by more traditional means. Computationally intelligent data mining techniques are vital for extracting the information necessary to select these novel molecules from tens of thousands of possible candidates. This research is leading to the development of novel machine learning paradigms such as semi-supervised learning with capacity control. These algorithms are used to generate superior QSAR models by using molecular data with both known (labeled) and unknown (unlabeled) biological responses. Transduction methods allow unlabeled data to constrain and refine the resulting predictive property models. This project involves the development of an infrastructure of computationally intelligent computer codes that allow for the virtual design of novel pharmaceuticals and the improvement of existing pharmaceuticals. The proposed methodology is applicable to most pharmaceuticals for which a database of responses is available. The ultimate pay-off of this methodology is expected to lead to methods that support the rapid invention of new drugs in response to new society-threatening diseases where a very fast response is warranted.

In our proposal, we identified the following subtasks that will help bring about our ultimate objective of a system for discovering and screening novel pharmaceuticals with desirable properties.

- (1) An enhanced TAE methodology for the generation of spatially-resolved electron density-based molecular descriptors.
- (2) Construction of TAE molecular training and validation datasets
- (3) Development of a learning-based software system to predict molecular bio-responses.
- (4) Development of novel semi-supervised learning models utilizing both labeled and unlabeled data.
- (5) Development, validation, and interpretation of molecular bio-response models.
- (6) Dissemination to promote cross-disciplinary research.
- (7) The development of studio-mode educational laboratory modules and undergraduate/graduate research experience that will be an integral part of Rensselaer's innovative multidisciplinary Information Technology degree program.

Substantial progress has been made on all of these tasks:

### **(1)-(2) Enhancements to TAE Methodology and Dataset Creation**

Since the task (1) of improvement of the TAE methodology and (2) dataset creation are intimately linked we described activities in these areas together

During the first year of the project, we have obtained and examined five new datasets and developed three new types of electron density-derived descriptors – EDDFA, GAU and

WCD. Two of the datasets were derived from published materials, and the other three were obtained through sources in the pharmaceutical industry. Electron Density-Derived Field Analysis (EDDDFA) descriptors were originally described in a Ph.D. thesis produced in the Breneman group (Christopher Whitehead, currently at Parke-Davis Pharmaceuticals). The use of EDDFA descriptors augments the CoMFA 3-D QSAR technique by providing more information about the electronic environment surrounding each molecule in a dataset. EDDFA descriptor generation originally required HF/SCF wavefunctions information, but a new procedure has been developed that allows the very rapid TAE/RECON program to produce these descriptors at a pace that will allow virtual screening of large databases to be practical.

Gaussian-smoothed surface property (GAU) descriptors were developed as adaptations of a previous class of Transferable Atom Equivalent (TAE) molecular surface property descriptors described by us [Brene97]. The original TAE descriptors were created from ten or twenty-bin fixed-range histograms describing the distributions of ten different electronic properties on electron density-derived Van der Waals molecular surfaces. The properties could be assembled using either our TAE/RECON program, or a post-SCF procedure that utilized Gaussian94/98 or Jaguar wavefunctions. The new GAU descriptors use adaptive scaling of properties prior to molecular surface histogram generation, and a smoothing algorithm designed to reduce sampling noise.

A new class of Wavelet Coefficient Descriptors (WCDs) has been developed that encodes molecular surface property information into a small set of wavelet coefficients. Initial experimentation has involved the use of ab initio electronic wavefunctions to provide the properties to be encoded. This procedure has proven to be the most effective way of capturing this information without having to increase the dimensionality of the modeling problem.

We are currently modifying the RECON/TAE procedure to allow very rapid reconstruction of molecular surface property distributions using wavelets.

### **(3) Development of Learning Based Software**

The QSAR learning task is typically very difficult because we must produce a model based on a relatively few number of molecules with known bio-activities using a very large number of descriptors or features. So we are specifically designing a learning-based software system for problems of these type know as Strip Miner.

- **STRIPMINER**

StripMiner is a shell for machine learning that manages and integrates the execution of several hundreds or thousands of neural networks, genetic algorithm-driven supervised regression clustering models, or support vector machine models. We currently have developed in-house codes for neural networks (MetaNeural), GA clustering, and Support Vector Machines that are compatible (or will be compatible) with the StripMiner shell. All the software was developed in standard C to facilitate the use of most computer platforms and operating systems.

StripMiner can run in several modes (currently only modes 1 and 2 are implemented):

- 1) Bootstrap mode
- 2) Sensitivity Analysis mode
- 3) Ensemble Method mode

Bootstrapping: selects different subsets for training and validation to improve statistical confidence.

Sensitivity Analysis: determines the most or the least important descriptors by tweaking the inputs of a machine learning model. In the current mode of operation, we typically drop the five least sensitive descriptors and repeat the building of a predictive model for the remaining descriptors in an iterated fashion.

Ensemble Methods such as boosting and bagging are a way to combine several learning models to improve the performance by a weighted voting scheme. In boosting learning emphasis is on patterns that are difficult to learn or conflicting with the model.

Currently a prototype of StripMiner was successfully implemented and tested for bootstrapping with neural networks and sensitivity analysis. We are investigating ensemble methods using column generation (described below) but they are not integrated into StripMiner at this time.

Since we do not a priori know the best learning methodology to be used with QSAR, we are considering many types of methodologies for inclusion in StripMiner. Here we will describe the supervised and unsupervised methods being used and leave the semi-supervised methods for the next section.

- **SUPERVISED SCALED REGRESSION CLUSTERING WITH GAs**

SSRCGA is a multi-dimensional regression analysis tool that can be used as an alternative to Artificial Neural Networks. Supervised Scaled Regression Clustering with Genetic Algorithms (SSRCGA) is a hybrid model based on a GA-based semi-supervised clustering algorithm augmented with local learning. The local learning in this method is supervised in the sense that the prediction quality is incorporated as a penalty term added to the fitness function of the Genetic Algorithm (GA). SSRCGA offers certain advantages related to robustness, generalization performance, feature selection, explanative behavior, and the additional flexibility of defining the fitness function with or without regularization constraints. The SSRCGA methodology has the following features: (i) local learning, (ii) minimizing cluster dispersion with GAs, (iii) clustering with GAs for a variable number of clusters and (iv) supervised regression clustering, (v) scaled supervised regression clustering and feature selection.

In addition to estimating the cluster centers SSRCGA also determines scaling factors for each dimension. The scaling factors act as multipliers for each descriptor (dimension) and are normalized in the phenotype. The scaling factors are used as feature selectors. SSRCGA is generally applied multiple times and from the original large descriptor space a few descriptors (the ones with the smaller scaling factors) drop out from the model. In order to boost validation confidence the SSRCGA model is applied in a bootstrapping mode.

- **GENERAL PURPOSE LIBRARY FOR GENETIC ALGORITHMS**

A floating-point general GA library was developed and can be used in a similar fashion to MIT's GALib. It is a stand-alone library, which is transparent to the general user. The reason for the in-house development of such a library is to facilitate the code integration and overcome some of the user's limitations of GALib.

- **FEATURE SELECTION**

A key component of StripMiner is Feature Selection. An appropriate feature selection method is essential for the success of DDASSL. The datasets being considered typically have between 500 and 1000 molecular property descriptors. Several novel feature selection methods have been implemented or are currently being developed. These methods are summarized below:

1. Sensitivity analysis: the inputs for a machine learning model are tweaked one at a time within their allowable range (while holding the other inputs at their mean value). The most sensitive descriptor inputs are those for which the output from the model shows the largest variation. This process was successfully implemented, tested and applied with neural networks. The best results were obtained by eliminating a few of the least sensitive descriptors at a time and iterating the procedure.
2. GA-based feature scaling: a GA scales the descriptors and tries to drive them to zero after scaling. The sum of the scaling factors is normalized and the most relevant descriptors after scaling are retained. This method was successfully implemented for GA-based regression clustering.
3. Direct feature selection with GAs (in development): a specified number of the most relevant features is selected by supplying the performance from a machine learning model as cost function to the GA.
4. Feature selection via the correlation matrix (in development): Features are selected by specifying a given number of features and defining an appropriate objective function that needs to be maximized. A GA could be applied here and the cost function consists of a main term (sum of the correlation coefficients for selected features) and a penalty term (penalty related to how inter-correlated selected features are).
5. Column generation approaches based on Support Vector Machines (in development). Using column generation methods from mathematical programming, we can practically optimize support vector machines based on large dictionaries of kernel functions. The capacity control inherent in the Support Vector Machine can then be used to select subsets of the features.
6. The visualization tools of the Viscovery software for generating Self-Organizing Feature Maps (SOMs) were successfully applied for feature selection.

- **SOM**

Self-organizing feature maps (SOMs) are neural network models based on unsupervised competitive learning. The Viscovery Software, a commercial SOM software package, was successful for QSAR applications for the prediction of bio-activities and feature selection. SOMs provide excellent visual insight to the domain expert but have the drawback that many parameter settings to tune to algorithm is heuristic.

- **BOOSTING VIA COLUMN GENERATION**

We proposed a linear program (LP) approaches to boosting and demonstrated their efficient solution using LPBoost, a column generation simplex method. We proved that minimizing the soft margin error function (equivalent to solving an LP) directly optimizes a generalization error bound. LPBoost can be used to solve any boosting LP by iteratively optimizing the dual classification costs in a restricted LP and dynamically generating weak learners to make new LP columns. Unlike gradient boosting algorithms, LPBoost converges finitely to a global solution using well-defined stopping criteria. Computationally, LPBoost finds very sparse solutions as good as or better than those found by ADABOOST using comparable computation. These same column generation techniques can be used to perform feature selection on more general learning problems. We are currently extending this methodology to the regression case.

#### **(4) Semi-Supervised Learning Methodologies**

We are examining approaches to semi-supervised learning that combine both supervised learning with labeled data and unsupervised learning with unlabeled data for use on high-dimensional problems. Once the TAE-EDDFA, GAU and WCD descriptors have been generated, the base problem can be viewed as a classification or regression problem using noisy high-dimensional data with few labeled and many unlabeled molecules. There are many other domains in which unlabeled data are abundant but labeled data are expensive to generate and therefore relatively scarce (e.g. medical diagnosis, web search, and database marketing). When the training data consist of relatively few labeled data points in a high-dimensional space, something must be done to prevent the classification or regression function from overfitting the training data. In this project, we propose using the information in the unlabeled data to prevent overfitting and guide model selection. We call this **semi-supervised learning** because it incorporates aspects of supervised and unsupervised learning. The following are the activities on semi-supervised learning to date. Note this work was jointly supported by NSF IRI-9702306

- **SEMI-SUPERVISED SUPPORT VECTOR MACHINES**

We examined mathematical models for semi-supervised support vector machines (S3VM). Given a training set of labeled data and a working set of unlabeled data, S3VM constructs a support vector machine using both the training and working sets. We use S3VM to solve the transductive inference problem posed by Vapnik. In transduction, the task is to estimate the value of a classification function at the given points in the working set. This contrasts with inductive inference, which estimates the classification function at all possible values. We propose a general SSS model that minimizes both the misclassification error and the function capacity based on all the available data. Depending on how poorly estimated unlabeled data are penalized, different mathematical models result. We examine several practical algorithms for solving these models. The first approach utilizes the S3VM model for 1-norm linear support vector machines converted to a mixed-integer program (MIP). A global solution of the MIP is found using a commercial integer programming solver. The second approach uses a non-convex quadratic program. Variations of block-coordinate-descent algorithms are used to find local solutions of this problem. Using this MIP within a local learning algorithm produced the best results. Our experimental study on these statistical learning methods

indicates that incorporating working data can improve generalization but we did not find significant improvements on the datasets tested.

- **SEMI-SUPERVISED CLUSTERING**

A semi-supervised clustering algorithm was proposed that combines the benefits of supervised and unsupervised learning methods. Data are segmented/clustered using an unsupervised learning technique that is biased toward producing segments or clusters as pure as possible in terms of class distribution. These clusters can then be used to predict the class of future points. For example in database marketing, the technique can be used to identify and characterize segments of the customer population likely to respond to a promotion. One benefit of the approach is that it allows unlabeled data with no known class to be used to improve classification accuracy. The objective function of an unsupervised technique, e.g. K-means clustering, is modified to minimize both the within cluster variance of the input attributes and a measure of cluster impurity based on the class labels. Minimizing the within cluster variance of the examples is a form of capacity control to prevent overfitting. For the output labels, impurity measures from decision tree algorithms such as the Gini index can be used. A genetic algorithm optimizes the objective function to produce clusters. Non-empty clusters are labeled with the majority class. Experimental results show that using class information improves the generalization ability compared to unsupervised methods based only on the input attributes. The results also indicate that the method performs very well even when few training examples are available. Training using information from unlabeled data can improve classification accuracy on that data as well.

- **GENERAL SEMI-SUPERVISED METHODOLOGIES**

We are supplying a test QSAR problem for a proposed workshop called “Unlabeled Data Supervised Learning Workshop” at the Neural Information Processing Systems Conference, Denver 2000. The Web site for the workshop is:

<http://q.cis.uoguelph.ca/~skremer/NIPS2000/>

The workshop is being organized by Stefan Kremer, Deborah Stacey, and Kristin Bennett. The goal of the workshop is to allow researchers to apply approaches to semi-supervised learning to a diverse set of benchmark problems with both labeled and unlabeled data. By including our dataset in this competition, the performance of a large set of semi-supervised learning methodologies will be benchmarked on the data constructed for this research. The competition will allow us to identify the most promising semi-supervised learning methodologies for drug design and promote inter-disciplinary research on this problem.

**(5) Development, validation, and interpretation of molecular bio-response models.**

Several benchmark datasets including the Merck CCK dataset, the NCI Developmental Therapeutics anti-cancer dataset, several HIV reverse-transcriptase inhibitor data sets and a tyrosine kinase dataset were analyzed.

**(6) Dissemination**

One of the primary goals of the project is to disseminate results, methodologies, and datasets in order to offer the highest probability of success of the project by both our team

and other researchers. The primary vehicles for dissemination to date are publications (described later), the DDASSL project web page (described later), the NIPS workshops (described above) and the following presentations:

Talks (past and scheduled) related to project with speaker:

1. "Soft Margin Boosting using Column Generation," West Coast Optimization Meeting organized by Terry Rockefeller, University of Washington, Seattle, September 1999. Kristin Bennett.
2. "Semi-Supervised Learning", AT&T Research Day, Florham Park, September 1999. Ayhan Demiriz.
3. "GA-based supervised regression clustering for the European EUFIT Data Mining Competition," September 2000. Dirk DeVogelaere (Catholic University of Leuven, Belgium), EUFIT 2000, Aachen, Germany.
4. "Support Vector Machines: Hype or Hallelujah?" Plenary Speech, ANNIE'99, Artificial Neural Networks in Engineering Conference, St. Louis, November 1999. Kristin Bennett.
5. "Semi-supervised Clustering using Genetic Algorithms," ANNIE'99 Artificial Neural Networks in Engineering Conference, St. Louis, November 1999. Kristin Bennett.
6. "Supervised Scaled Regression Clustering with Genetic Algorithms," ANNIE'99 Artificial Neural Networks in Engineering Conference, St. Louis, November 1999. Mark Embrechts.
7. "Semi-Supervised Clustering," Institute for Operations Research and the Management Science (INFORMS) Conference, Philadelphia, November 1999. Ayhan Demiriz.
8. "Drug-Design through Semi-Supervised Learning," Bioinformatics Workshop, Rensselaer Polytechnic Institute, November 1999. Curt Breneman.
9. "Soft Computing for the Virtual Design and Discovery of Pharmaceuticals", Invited Presentation at the Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute, November 1999. Mark Embrechts.
10. "Soft Margin Boosting using Column Generation," Support Vector Machine Workshop, Large Margin Classifiers Workshop, Neural Information Processing Systems Conference, Denver, CO, December 1999. Kristin Bennett.
11. "New Methods of Surface Descriptor Representation," Eastman Kodak, January 2000. Curt Breneman.
12. "Optimization of Molecular Properties using TAE Descriptors," GE Corporate R&D Center, Schenectady, NY, January 2000. Curt Breneman.
13. "Virtual Design of Pharmaceuticals with Semi-Supervised Learning," Invited Colloquium, Space and Naval Warfare Systems Center (SPAWAR), February 2000. Mark Embrechts.
14. "A Water Pollution Problem Solved: Comparison of GAdc with other Methods," Modelling, Identification and Control Conference - MIC 2000, Innsbruck, Austria, February 14 - 17, 2000. Dirk DeVogelaere\*/Mark Embrechts.

15. "Geometry in Data Mining," Thompson Science Series, University of Puget Sound, Tacoma, WA, February, 2000. Kristin Bennett.
16. "Virtual Design of Pharmaceuticals with Semi-Supervised Learning," invited poster presentation at the INTEL Computing Continuum Conference March 2000, San Francisco, CA. Mark Embrechts.
17. "QSAR and Docking/Scoring Functions for the Analysis of Protein-DNA Interactions," New York State Health Labs, Wadsworth Lab, April 2000, Curt Breneman.
18. "Virtual Design of Pharmaceuticals with Semi-Supervised Learning," Poster presented at the Snowbird 2000 Learning Conference (by invitation), Snowbird, Utah, April 2000. Mark Embrechts.
19. "LP Boosting using Column Generation", Workshop on Selecting and Combining Models with Machine Learning Algorithms, Montreal, April 2000. Ayhan Demiriz.
20. "Virtual Design of Pharmaceuticals," Presentation for the DSES advisory board, Rensselaer Polytechnic Institute, May 2000. Mark Embrechts.
21. "Data Mining for the Virtual Design of Pharmaceuticals," Colloquium speaker for the Albany Chapter of the American Statistical Association (ASA), Albany, NY, April 2000. Mark Embrechts.
22. "Data Mining and Information Technology: Crossovers with Biotechnology," Washington Advisory Group, Blue Ribbon Panel on IT, May 2000. Curt Breneman.
23. "A Column Generation Approach to Boosting," International Conference on Machine Learning, June 2000. Kristin Bennett.
24. "Geometry, Duality, and Support Vector Machines," International Conference on Machine Learning, June 2000, Kristin Bennett.
25. "GA-Supervised Regression Clustering for the design of new pharmaceuticals," Invited Presentation at the "High-Throughput Technologies Conference," June 20, 2000, Wyndham Franklin Plaza Hotel, Philadelphia, PA. Mark Embrechts.
26. "Wavelet Coefficient Descriptors in Molecular Property Screening," Computational Chemistry Gordon Conference, July 1-8, 2000, Oxford, England. Curt Breneman.
27. "Column Generation via Boosting," International Symposium on Mathematical Programming, Atlanta, August 2000, Kristin Bennett.
28. "Predicting biomolecular recognition phenomena using the TAE/RECON method," The 220th National Meeting of the American Chemical Society, Washington, D.C., August 20-24, 2000. N. Sukumar\*/Curt Breneman.
29. "The use of 2D, 3D, TAE and wavelet coefficient descriptors (WCDs) for generating self-organizing Kohonen maps for QSAR, QSPR and ADME Analyses," ACS National Meeting, Washington, D.C., August 2000 Larry Lockwood\* / Curt Breneman.
30. "Predicting Biomolecular Recognition Phenomena using the TAE/RECON Method," ACS National Meeting, Washington, D.C., August 2000 Sukumar Nagamani\* / Curt Breneman.

31. "Wavelet Representations of Molecular Electronic Properties: Applications, ADME, QSPR and QSAR," August 2000 ACS National Meeting, Washington, D.C. Curt Breneman.
32. "Molecular Database Mining using Self-Organizing Maps for the Design of Novel Pharmaceuticals," ANNIE 2000 Artificial Neural Networks in Engineering Conference, St. Louis, November 2000. Mark Embrechts.
33. "Supervised Scaled Regression Clustering: an Alternative to Neural Networks," June 2000, IJCNN 2000, Como, Italy, Dirk DeVogelaere\*/Mark Embrechts
34. "The TAE/RECON Method in Large Database Mining, QSAR and ADME: A Progress Report," December 2000, Pacificchem, Honolulu, Hawaii. Sukumar Nagamani\* / Curt Breneman.
35. "GA Approaches for Successful QSAR Modelling," December 2000, Pacificchem, Honolulu, Hawaii. Mark Embrechts.

#### **(7) Educational Contributions**

Machine learning related course modules for drug design appropriate for undergraduate seniors were developed and posted on the course web page for the undergraduate course "Introduction to Computational Intelligence." Two modules were developed, one based on neural networks and a second module based on GA supervised regression clustering. The modules contain lecture notes introducing neural networks and genetic algorithms, a homework problem with a dataset. The course material, codes and user manual can be retrieved via the WWW. The Computational Intelligence course is project driven (teams of two students) and had 4 DDASSL-related course projects. Students prepared a project report and gave a presentation in class.

Curt Breneman is developing a new course in Computational Chemistry and Data Mining for the Spring semester of 2001.

Recon Tutorial – an on-line course of instructional materials concerning the use of the TAE/RECON electron density-based descriptor generator program package is now available via the WWW.

We established a weekly project seminar attended by all participants of the project. This is described below