

Automated Design and Discovery of Novel Pharmaceutical using Semi-Supervised Learning in Large Molecular Databases

Mark J. Embrechts (embrem@rpi.edu)
Department of Decision Sciences and Engineering System

Curt Breneman (brenec@rpi.edu)
Department of Chemistry

Kristin P. Bennett (bennek@rpi.edu)
Department of Mathematics

[Rensselaer Polytechnic Institute](#), Troy NY 12180

Contact Information

[Mark J. Embrechts](#)
DSES, CII5217
Rensselaer Polytechnic Institute
Troy, NY 12180
Phone: (518) 276-4009 Fax : (518) 276-8227
Email: (embrem@rpi.edu) [URL](#)

WWW PAGE

<http://www.drugmining.com>

List of Supported Students and Staff

Graduate Students - Mathew Sundling, Muhsin Ozdemir, Fabio Arcienegas, Robert Kewley, Jr., Neil Eklund, Ayhan Demiriz, Dirk De Vogelaere, Bo Jiang, Quiong Luo, Mighu Song, Wei Deng, Larry Lockwood, Dechuan Zhuang, Jinbo Bi, Michinari Momma, Abigail Michels, and Robert Bress.

Undergraduate Students - Pieter De Temmerman, David Goldstein, Andres de la Guardia, Bill Katt

Project Award Information

Virtual design of pharmaceuticals using data mining. This is a three-year NSF funded project from 09/01/1999 - 08/31/2002. This is the second year of the project. The award number is IIS-9979860.

Keywords

Drug Discovery, Virtual Design, Novel Pharmaceuticals, Molecular Databases, Support Vector Machines, Machine Learning, QSAR, QSPR, Descriptor Generation, Wavelet Descriptors,

Shape-Specific Property Descriptors, Descriptor selection, Transferable Atom Equivalent (TAE) Descriptors, RECON, Neural Networks, GAPLS, Sensitivity Analysis, Partial Least Squares, Evolutionary Computing, Data Mining

Project Summary

This research results in 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. A large set of new and traditional descriptors is used to create improved Quantitative Structure-Activity Relationship (QSAR) models that characterize and predict important biological responses. Once the descriptors have been determined and a predictive model has been built, thousands of new potential molecules, chemically similar to those of the benchmark data set, are scanned from large databases and are evaluated for their chemical properties based on the predictive model. The aim is to target a few novel molecules with potentially attractive pharmaceutical properties that can then be tested further in the traditional way in the laboratory. Computationally intelligent data mining techniques are vital to extract the information necessary to select these novel molecules. This research develops and applies novel machine learning paradigms for solving inference problems in high dimensions with few data points. These algorithms predict desired biological responses and generate QSAR models using both known (labeled) and unknown (unlabeled) biological responses. This project involves the development of an infrastructure of computationally intelligent computer codes that allow for the virtual design of novel pharmaceuticals or the improvement of existing pharmaceuticals. The proposed methodology is applicable to most pharmaceuticals for which a database of bioactivities is available. The ultimate pay-off of this methodology is the rapid invention of new drugs for new or known society threatening diseases where a very fast response is warranted.

Publications

- 1.K. P. Bennett and C. Campbell, "Support Vector Machines: Hype or Hallelujah?" *SIGKDD Explorations*, 3:1, 2001.
- 2.A. Demiriz, K. P. Bennett and J. Shawe-Taylor, "Linear Programming Boosting via Column Generation", *Machine Learning*, 2001, to appear.
- 3.Ayhan Demiriz, Kristin P. Bennett, and Mark J. Embrechts, "Semi-Supervised Clustering with Genetic Algorithms," Accepted for publication by the International Journal of Smart Engineering System Design, 2001, to appear.
- 4.G. Raetsch, A. Demiriz, K.P. Bennett, "Sparse Regression Ensembles in Infinite and Finite Hypothesis Spaces", *Machine Learning*, 2001, to appear.
- 5.Tugcu, N., Mazza, C.B., Moore, J.A., Breneman, C.M., Sanghvi, Y.S., Cramer, C.M., "High Throughput Screening and Quantitative Structure Efficacy Relationship Models for Designing Displacers for Antisense Oligonucleotide Purification in Anion Exchange Systems, 2000, PNAS, under review.
- 6.Mazza, C.B., Sukumar, N., Breneman, C.M. and Cramer, S.M., "Prediction of Protein Retention in Ion-Exchange Systems Using Molecular Descriptors Obtained from Crystal Structure", 2000, Nature Biotechnology, under review.

7. Breneman, C. M., Embrechts, M., Lockwood, L., Sundling, M., Sukumar, N. "Wavelet Coefficient Descriptors in QSPR, QSAR and ADME", 2000, Journal of Molecular Graphics and Modelling (Symposium Proceedings Issue), submitted.

8. K.P. Bennett, A. Demiriz and J. Shawe-Taylor, "A Column Generation Algorithm for Boosting", Proceedings of the Seventeenth International Conference on Machine Learning, P. Langley Editor, pp 65-72, 2000.

9. K. P. Bennett and E. J. Bredensteiner, "Duality Geometry, and Support Vectors Machines". Proceedings of the Seventeenth International Conference on Machine Learning, P. Langley Editor, pp 57-64, 2000.

10. Fabio Arciniegas, Kristin Bennett, Curt Breneman, and Mark J. Embrechts, "Molecular Database Mining using Self-Organizing Maps for the Design of Novel Pharmaceuticals," in Intelligent Engineering Systems through Artificial Neural Networks: Smart Engineering System Design: Vol. 10, C. H. Dagli et al., Eds, ASME Press, pp. 477 – 482, 2000.

11. I.-N. Lee, S.-C. Liao, and M. Embrechts, "Data Mining Techniques Applied to Medical Information," Medical Information & The Internet in Medicine, Vol. 25, No 2., pp. 81 – 102, 2000.

12. Robert H. Kewley, Mark J. Embrechts and Curt Breneman, "Data Strip Mining for the Virtual Design of Pharmaceuticals with Neural Networks," IEEE Transactions on Neural Networks, Vol. 11, No 3, pp. 668 – 679, May 2000.

13. Yoshua Bengio, Joachim M. Buhman, Mark Embrechts, and Jacek M. Zurada, "Introduction to the Special Issue on Neural Networks for Data Mining and Knowledge Discovery," Guest Editorial, IEEE Transactions on Neural Networks, Vol. 11, No 3, pp. 545 – 549, May 2000.

14. M. J. Embrechts, D. Devogelaere, and M. Rijckaert, "Supervised Scaled Regression Clustering: An Alternative to Neural Networks," Proc. IEEE-INNS-ENNS International Conference (IJCNN 2000), Vol. 6, pp. 571-576, Como, Italy, 24 – 27 July, 2000.

Project Impact

The techniques developed for this project lead to new powerful data mining tools for the virtual design and discovery of pharmaceuticals. The use of machine intelligence in QSAR and molecular design will change the way new drugs are invented by minimizing the lengthy procedures for testing on humans and animals and allowing the near real-time virtual invention of drugs for society threatening diseases.

Goals, Objectives, and Targeted Activities

1. DATASETS -- Dataset selection and provision from both industrial and published sources on the basis of both intrinsic difficulty or lack previous success as well as biological and medical relevance. Development and analysis of large QSAR datasets with ~ 50,000-100,000 molecules.

2. DESCRIPTORS -- Development of rapidly calculable Wavelet Coefficient Descriptors (WCDs) that capture important features of molecular electron density distributions from either Transferable Atom Equivalent (TAE) reconstruction or from DFT or ab-initio wavefunctions.

Development of shape-specific electronic property descriptors. Promoting and supporting RECON code for TAE and wavelet descriptor generation.

3.DATAMINING TOOLS - Benchmarking, documenting and promoting StripMiner code for feature reduction with Genetic Algorithms and Sensitivity analysis. StripMiner incorporates bootstrapping and bagging for predictions based on support vector machines, neural networks, partial least squares and GA regression clustering for drug design . Creation of ensemble support vector machine approaches for drug design including automated feature and model selection.

Area Background

This project develops an infrastructure for rapid drug design using chemometric information from large molecular datasets. In a first phase, new descriptors were developed that are potentially related to biological activities. In a second phase, machine learning models were developed to predict these biological activities. In the third phase, the descriptors and modeling procedures will be validated and enhanced.

The project has two major components: development of molecular descriptors and creation of “strip mining” tools to predict bioresponses based on these descriptors. We have developed two new types of electron density-based descriptors as alternatives to traditional 2D and 3D property descriptors. The new descriptor types include a set of wavelet coefficient descriptors (WCDs), and a new type of shape-specific electronic property descriptors. The performance of the WCDs has been benchmarked against TAE descriptors and all other modern QSAR descriptors available in the open literature. The newer shape-specific descriptors are still under development, but will be the subject of an ACS meeting presentation in April. We have created a suite of inference and validation tools for bio-response prediction. Current StripMiner modules include neural network, GA-driven clustering, GA-clustering with Semi-supervised learning. A full cycle SVM QSAR methodology has been developed including feature selection, automated model selection, and robust ensemble predictions. Several benchmark datasets (CCK, NCI Developmental Therapeutics anti-cancer, HIV reverse-transcriptase inhibitor data sets and a tyrosine kinase dataset were analyzed in addition to the Lombardo ADME dataset. Standard formats for web dissemination of our datasets and results are being developed.

Area References

The DDASSL project activities and products are fully documented at www.drugmining.com.

Potential Related Projects

1) Development of screening and virtual library generation for rapid-responses to biological threats to humans, plants and animals. 2) Protein-stationary phase interaction modeling for bioseparations technology. 3) Displacement chromatography modeling and displacer design. 4) Molecular design techniques as applied to molecules of non-biological interest such as “Materials by Design” or specialization and optimization of industrial intermediates.