# Molecular Data Mining Tool: Advances In HIV Research

The ability to predict biological activity based on molecular structure is leading researchers to breakthroughs in the most complex challenges of medicine. Using a combination of artificial intelligence tools, Dr. Wayne Danter of Critical Outcome Technologies (London, Ontario, Canada) has developed a method to predict whether specific molecular structures are effective against a disease. Currently under study is the HIV1 virus.

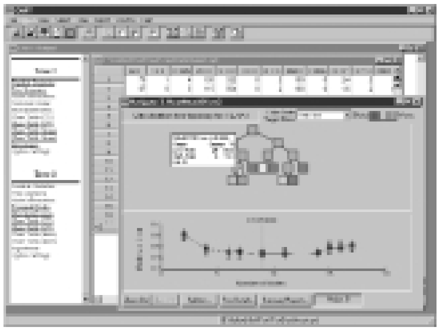

**Figure 1: The optimal CART tree. Red nodes contain greatest concentration of the "High Risk" group and blue nodes concentrate the "Low Risk Group." Hovering the mouse over a node displays its contents.**

Pharmaceutical companies may have as many as a million molecules in their databases. Modeling each molecule and predicting its effectiveness using standard statistical methods is virtually impossible because of the enormous number of variables. Dr. Danter uses CART® (Classification and Regression Trees), a software package from Salford Systems to help build models that isolate the most important variables. Working with public domain, molecular HIV data, Danter trains CART and complementary systems to predict if a given molecular structure is biologically active against a disease. Says Dr. Danter, "Once we have such a model, we can screen almost any molecule with a molecular weight up to 1700 daltons (an atomic mass unit). It's an area called molecular mining. We've developed it as a generic tool, so that if there is a specific target biological activity, we can screen for it."

To build a model, CART generates a binary decision tree based on yes/no answers. It generates nodes until it has created the largest tree that fits the data. This ensures that the node-generating process is not halted too soon and important structures are not overlooked.

## Pruning Decision Trees

Upon creating the structure, the system prunes back the tree and uses a self-test procedure to ensure that the model is not over-fitting — that is, finding patterns that apply only to training data. This produces a smaller, optimal-sized tree. The tree's terminal nodes become the model used for the remainder of the research process.

A list of important variables is automatically produced and is used to develop the model, ranked by importance. This is crucial because many of the variables turn out to be relatively unimportant. "You may have a couple of hundred input variables, but a subgroup of those variables are the most important ones and the only ones we really need to use," says Dr. Danter. Using all the variables throughout the analysis would make the process needlessly cumbersome — possibly skewing the results.

To satisfy Dr. Danter's specialized modeling needs in his HIV research, he inputs the results into another Salford Systems product, MARS® (Multivariate Adaptive Regression Splines), then into a neural network program from Ward Systems Group, NeuroShell® Classifier. MARS is a non-parametric regression procedure that extends Dr. Danter's work by improving the accuracy of predictions. NeuroShell® Classifier then categorizes a molecule's activity based on patterns derived from CART and MARS.

## Honing The Data

The results are honed to specific research needs using a proprietary algorithm Dr. Danter developed called CHEMSAS™. This process decomposes complex molecular structures into key elements, teaching artificial neural networks (ANN) to relate structure to function. The ability to analyze molecular structure and predict effectiveness helps Dr. Danter look for existing drugs to battle diseases like HIV, as well as to develop potential new medications. Analyzing chemical structures, CHEMSAS™ utilizes hybrid ANN systems to predict the *in vitro* response of HIV1 to potential anti-viral drugs.

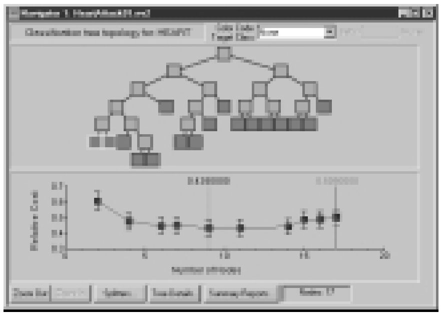

**Figure 2: The overtrained maximal tree has a relative error rate of .505 (red line); the optimal tree relative error is .435 (green line). The highlighted nodes on the left of tree contribute least to performance and will be the first to be pruned away.**

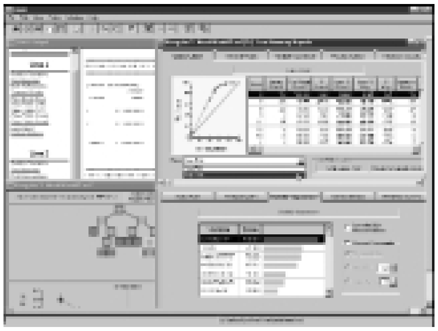

**Figure 3: Summary reports include a variable importance ranking, gains and lift charts and tables, misclassification reports, and an overall summary of all trees grown in a session.**

The results to date are impressive. In a recent study conducted by Dr. Danter, he analyzed 311 drugs with known *in vitro* activity against the HIV1 virus. The system correctly classified more than 96% of the molecules.

One of the great strengths of a data-mining tool like CART is its ability to pick out the significant variables – even when they are hidden among hundreds or thousands of irrelevant variables. It also clearly identifies complex interactions among study variables, and permits Dr. Danter to obtain more accurate results in minutes – rather than days.

## Mining In Other Areas

During the past several months, Dr. Danter has also used CART in developing models to study central nervous system receptors, anti-arthritic medications, and antibiotics, among others. As an artificial intelligence tool, CART's role in predicting specific biological activity continues to be vital to his research at Critical Outcome, Inc. To view detailed study results and modeling procedures, review their research at *www.critical outcome.com*.

**Richard Burnham can be reached at (651) 773-0619 or at *published@att.net***

**Dr. Wayne Danter, MD, FRCPC is an Associate Professor of Medicine and Director, LRI Neural Computing Lab at the University of Western Ontario London Ontario, Canada and can be reached at (519) 851-0035 or *wdanter@criticaloutcome.com***

**Salford Systems (*www.salford-systems.com*) can be reached at (619) 543-8880 or *info@salford-systems.com***