CRITICAL OUTCOME TECHNOLOGIES INC.

Critical Outcome Technologies Inc.

Combining CART and Other Machine Learning Technologies: Evolving New Strategies for Drug Discovery and Pre-Clinical Drug Development

WR Danter MD, FRCPC,

President and CSO

Powered By

SALFORD SYSTEMS

Proud Partner

DALTON Pharma Services

Accelerated Drug Discovery And Development

CRITICAL OUTCOME TECHNOLOGIES INC.

OVERVIEW

- CART and TreeNET → our experience
  - Which is the “best” choice?
- *In silico* Drug Discovery:
  - HIV Integrase inhibitors
  - Why New Drugs fail
- A Glimpse of the Future of Drug Discovery
  - Improving the odds of success
- The Support Vector Like Machine (SVLM)
  - Combining the best of the best

CRITICAL OUTCOME
TECHNOLOGIES INC

Basic Assumptions

- Molecular Structure --> Biological Activity
- Relationship is **Specific, Complex** and **Highly Non Linear**
- Specific Relationships can be accurately represented by an **optimal set of mathematical models**
- Entry point for all potential drugs remains ***in vitro*** screening
- Computer Hardware and **Software** exists that permit high dimensional/highly non-linear mathematical modeling in a optimal development cycle
- With sufficient and representative data ***in silico*** models can accurately **simulate** ***in vitro*** and ***in vivo*** methods

CRITICAL OUTCOME TECHNOLOGIES INC.

Computational Power

- 5 Dedicated Dual Xeon processor work stations – isolated/LAN
  - 3.43 gigahertz dual processor speed
  - 4 gigabytes of RAM/machine
  - High end server connection
- 4 additional research machines
- New system development -> few days
- New system deployment -> few minutes
- Custom Applications -> few weeks

CRITICAL OUTCOME TECHNOLOGIES INC.

The Modeling Options

For Advanced Molecular Structure-Activity Modeling:

CART or MART/TreeNET ->combined with

Artificial Neural Networks: PNN or GRNN or

Multivariate Adaptive Regression Splines or

Genetic Algorithm or

Support Vector Machine or

Generalized Additive-Support Vector Like Machine or

Combining the above into “Super Ensembles”

CART

- Widely used and validated Data Mining Tool in business beginning in 1980s
- More recent application to Life Sciences and Biotechnology
- Applicable to Classification and Regression problems with sufficiently large training set : Tj(x) → F*(x) as N → infinity!
- Some disadvantages but overall a robust and valuable Tool when used appropriately
- CART has been our Data Mining tool of choice since 1998

CART: Advantages

- Fast, recursive binary splitting, backward pruning algorithm
- Handles continuous, binary and categorical variables
- Handles missing values easily via surrogates
- Training Scales up well → n*N*log N
- Testing very fast → scales as Log J (# of regions)
- Always produces model summary
- Invariant to monotonic predictor variable transformations → gives the same tree model
- High degree of resistance to outliers and irrelevant predictors
- Relative variable importance provided
- Ease of use → just a few tunable parameters

CART: Limitations

- Tree Models are **Regionally Constant:**
  - Predictions constant within a given region R_j
  - Predictions can be sharply discontinuous across regions
  - Leads to potential Instability where small changes in predictors lead to dramatic changes in predictions
  - Instability likely worse for larger trees as well

- Trees **Fragment** the available data:
  - At each split daughter nodes contain fewer observations
  - Terminal nodes may contain <10 observations
  - Tree accuracy will suffer if target function requires substantial number of observations in all regions
  - Small datasets at a real disadvantage

CART: Bootstrap Re-sampling

- Most limitations addressed by Bagging many trees (hundreds)
- Lose some of the advantages of Trees but often dramatic increases in accuracy
- Generally avoid ARCing
- We only use Ensemble models based on Bagging:
  - Regression → mean of all predictions
  - Classification → simple plurality voting system

CRITICAL OUTCOME TECHNOLOGIES INC

Boosted Trees: TreeNET

- Prediction depends on linear combination of **many** small trees
  - $F_m(x) = \Sigma F_{m-1}(x) + (v * a_m) * T_m(x)$
- Constant Tree size (m = 4-10 Regions)
- Regularization →
  - "Early Stopping"/Test set
  - "Shrinkage" → v ~0.05-0.1 (Learning Rate)
  - Re-sampling

TreeNET Advantages

- Maintains all of the advantages of single CART tree except interpretability
- Scales well to large problems → n*N*LogN
- Often dramatic increase in accuracy
- Model Predictions still regionally constant but many more/smaller regions→ smaller discontinuities → improved accuracy
- Much more stable → averaging large number of trees
- Degree of data fragmentation reduced by using small trees

CRITICAL OUTCOME TECHNOLOGIES INC.

CART or TreeNET?

A Few Heuristics: for Drug Discovery

- CART + Ensemble Bootstrap Re-sampling likely better:
  - Some missing values → surrogates
  - Mixture of strong > weak predictors
- MART/TreeNET likely better:
  - If data is quite limited → 10 fold cross-validation
  - Combining many weak predictors → common
- Both are strong data mining tools – choice often dependent on the data set in question
- Actual differences tend to be small and NS

CRITICAL OUTCOME TECHNOLOGIES INC.

HIV Integrase Inhibitors

The Search for Orally Effective, Non-Toxic Integrase Inhibitors for the Treatment and Possible Prevention of HIV Infection Using *In Silico* Methods

CRITICAL OUTCOME TECHNOLOGIES INC.

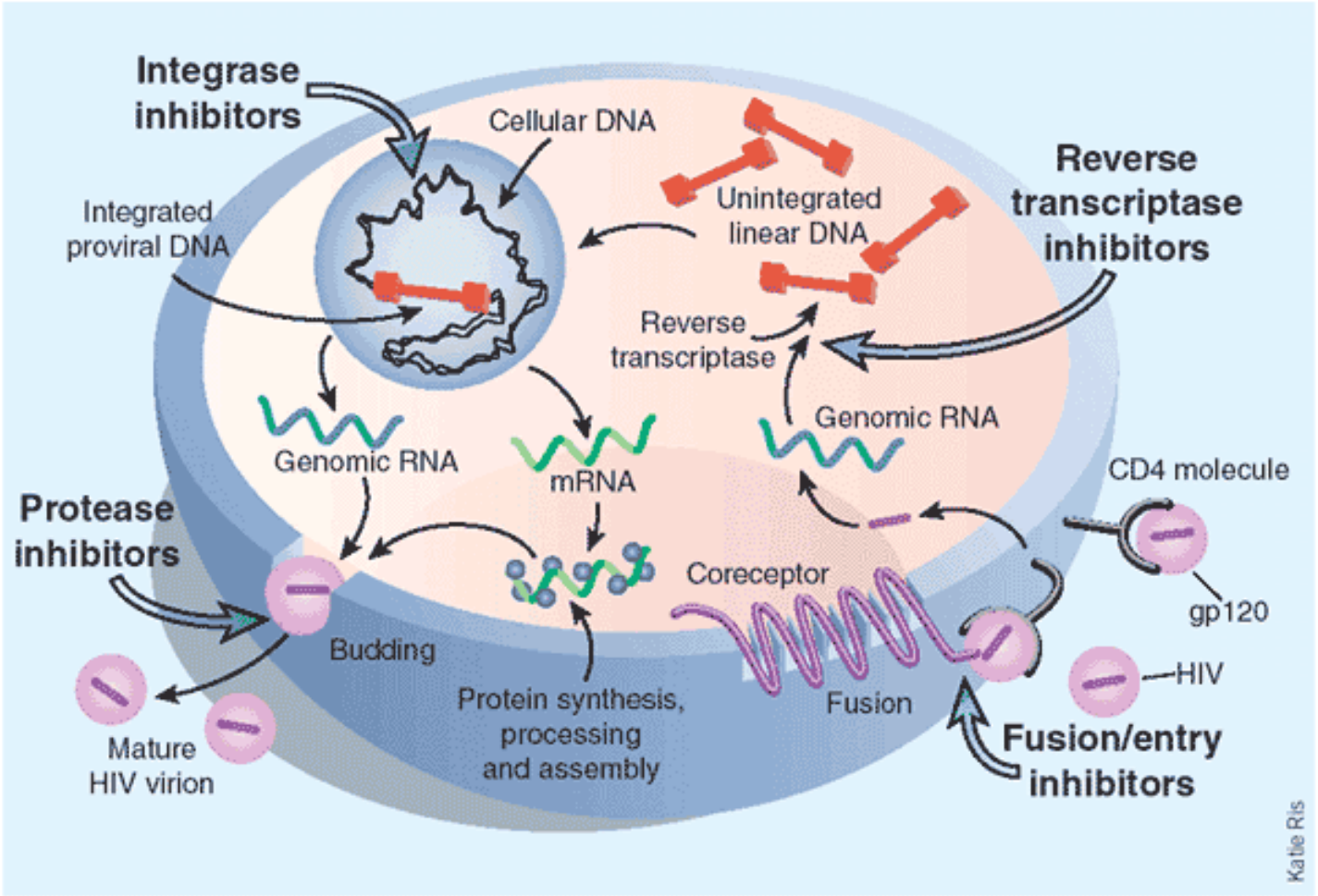
Why HIV Integrase As a Target?

3 Major HIV Enzymes/Targets identified:

- Reverse Transcriptase -> RTIs
- HIV Protease -> HIV PIs
- Current Therapy (HAART) -> combines RTI + PI
- HIV Integrase: Incorporates Viral DNA into host DNA
  - No agents currently on the market
  - Few candidates in clinical trials
  - Future optimal therapy likely to be → RT Inhibitor + Protease Inhibitor + Integrase Inhibitor
  - Potential as post exposure prophylaxis therapy
- Main problem -> identifying orally available non-toxic molecules that are BOTH potent inhibitors of HIV Integrase and effective anti-retrovirals
- Huge potential unmet economic and medical need -> viral eradication

CRITICAL OUTCOME
TECHNOLOGIES INC.

The Replication Cycle of HIV



CRITICAL OUTCOME TECHNOLOGIES INC.

Creating *In Silico* Assays

First Need to Create 3 New Assays or *In Silico* Activity Filters:

1. **HIV Integrase:** **711** compounds found with *in vitro* **IC50** data for 3'Processing and Strand Transfer or unspecified "Integrase Inhibition" activity (umol/l)
2. NCI/DTP/NIAID: **2246** compounds identified with *in vitro* **anti HIV EC50** (umol/l) and
3. NCI/DTP: **1785** compounds with *in vitro* activity -> **Confirmed Active (CA) or Confirmed Inactive (CI)** -> cytoprotection assay (NOTE: CMA molecules excluded!)

CRITICAL OUTCOME TECHNOLOGIES INC.

Formulating the Problem

Problems Essentially Reduced to:

2 Regression/Function Approx. Problems:

HIV Integrase Inhibition (IC50) - umol/l

Anti-retroviral Activity (EC50) - umol/l

and

1 Classification Problem:

Confirmed Active vs Confirmed Inactive

CRITICAL OUTCOME TECHNOLOGIES INC.

HIV Integrase Problem (IC50)

- 711 molecular structures decomposed into 216 (largely proprietary) predictors
- Data set randomly divided into:
  - Training Set → 561 (~79%)
  - Test Set → 150 (~21%)
- CART(Bagged) → Hybrid model
  - CART output added to Training Set and Hybrid model developed as usual.
- All subsequent hybrid models developed on the Training Set and evaluated on the Test Set of 150 molecular structures in a 2 stage evaluation

CRITICAL OUTCOME TECHNOLOGIES INC.

HIV Integrase (IC50) Results

Base Learner → CART/Bagging/300 Trees

Method	Test N	R^2	95% CI	p value	SD	95%CI	p value
CART	150	0.937	±0.039		0.310	±0.608	
CART+MARS	150	0.942	±0.037	NS	0.298	±0.584	NS
CART+TreeNET	150	0.939	±0.038	NS	0.310	±0.608	NS
CART+MLP	150	0.933	±0.040	NS	0.320	±0.628	NS
CART+GRNN+GA	150	0.938	±0.039	NS	0.310	±0.608	NS
CART+GA	150	0.937	±0.039	NS	0.310	±0.608	NS
CART+SVLM	150	0.937	±0.039	NS	0.314	±0.615	NS

Combining Hybrids into Super Ensemble provided no improvement over the CART+MARS model

CRITICAL OUTCOME TECHNOLOGIES INC.

Anti-Retroviral Problem (EC50)

- 2246 molecular structures decomposed into 216 (largely proprietary) predictors
- Data set randomly divided into:
  - Training Set → 1796 (~80%)
  - Test Set → 450 (~20%)
- CART(Bagged) → Hybrid model
  - CART output added to Training Set and Hybrid model developed as usual.
- All subsequent hybrid models developed on the Training Set and evaluated on the Test Set of 450 molecular structures in a 2 stage evaluation

CRITICAL OUTCOME TECHNOLOGIES INC.

Anti-Retroviral (EC50) Results

Base Learner → CART/Bagging/300 Trees

Model	Test N	Test R^2	Improve	Upper	Lower	pvalue	Test SD	Test 95%CI
CART	450	0.838	0.000	0.872	0.804		0.442	0.866
CART+MARS	450	0.880	0.042	0.910	0.850	NS	0.397	0.778
CART+TreeNET	450	0.898	0.060	0.926	0.870	0.048	0.352	0.689
CART+MLP	450	0.879	0.041	0.909	0.849	NS	0.381	0.747
CART+GRNN+GA	450	0.876	0.038	0.906	0.846	NS	0.387	0.759
CART+GA	450	0.885	0.047	0.914	0.856	NS	0.372	0.730
CART+SVLM	450	0.874	0.036	0.905	0.843	NS	0.293	0.574

Combining Hybrids into Super Ensemble provided no improvement over the CART+TreeNET model

CART+SVLM may be best model overall??

CRITICAL OUTCOME TECHNOLOGIES INC.

Anti-Retroviral Activity

- Classification Problem: Active vs Inactive *in vitro*
- 1785 molecular structures decomposed into 216 (largely proprietary) predictors
- Active → 605 (33.9%); Inactive → 1180 (66.1%)
- Data set randomly divided into:
  - Training Set → 1435 (~80%)
  - Test Set → 350 (~20%)
- CART(Bagged) → Hybrid model
  - CART output added to Training Set and Hybrid model developed as usual.
- All subsequent hybrid models developed on the Training Set and evaluated on the Test Set of 350 molecular structures in a 2 stage evaluation

CRITICAL OUTCOME TECHNOLOGIES INC.

Classification Test Results

Model	Test N	Accuracy	Sensitivity	Specificity	PPV	NPV
CART	350	88.57%	73.55%	96.51%	91.75%	87.35%
CART+TreeNet	350	87.14%	85.19%	88.02%	76.03%	93.01%
CART+ANN(PNN)+GA	350	89.14%	78.50%	94.76%	88.79%	89.30%
CART+SVM	350	88.57%	73.55%	96.51%	91.75%	87.35%
CART+SVLM	350	88.57%	89.96%	85.95%	81.89%	92.38%
TreeNET	350	89.14%	78.51%	94.76%	88.79%	89.30%
TreeNET+CART	350	86.29%	77.86%	91.32%	84.30%	87.34%
TreeNET+ANN(PNN)+GA	350	86.00%	85.95%	86.03%	76.47%	92.06%
TreeNET+SVM	350	90.57%	81.82%	95.20%	90.00%	90.83%
TreeNET+SVLM	350	88.86%	82.65%	92.14%	84.75%	90.95%

Combining Hybrids into Super Ensembles provided no improvement over the TreeNET+SVM model

Base Learner → CART/Bagging/300 Trees

CRITICAL OUTCOME
TECHNOLOGIES INC.

Summary of Our Experience

- Both CART and TreeNET are powerful data mining tools for Drug Discovery
- Whether CART or TreeNET is “better” depends on data set to be modeled
- CART or TreeNET hybrid models produce variable and min/modest improvements
- Hybrid SVM models may be optimal for difficult classification problems
- Super Ensembles produce no additional improvement over single best hybrid

CRITICAL OUTCOME
TECHNOLOGIES INC.

Why Do New Drugs Fail?

39% fail due to deficiencies in absorption, distribution, metabolism and elimination (ADME)

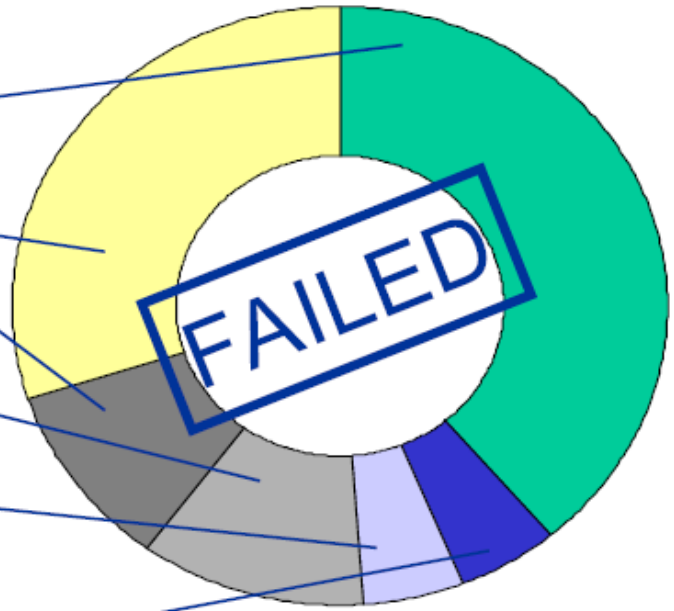
30% fail due to lack of efficacy

11% fail due to animal toxicity

10% fail due to adverse effects in man

5% fail due to commercial reasons

5% miscellaneous



Source: T.Kennedy, Drug Discovery Today, 2, 1997

CRITICAL OUTCOME TECHNOLOGIES INC.

Computerized Drug Discovery

- **Next Steps** >500,000 molecules searched, >400 unique compounds designed and profiled ->
- **Efficacy:** ALL 3 criteria must be met in order to proceed!
  - Integrase Inhibition IC50 -> <10 umol/l AND
  - Anti HIV EC50 -> <10 umol/l AND
  - Classification -> Confirmed Active AND
- **HIA% / BBB penetration** AND
- **Metabolic Stability** AND
- **CYP 450 Inhibition/Induction** - > NO AND
- **Pgp** -> Substrate -> NO AND
- **Potential Toxicity:**
  - Oral rat LD50 in mg/kg AND
  - IPR mouse LD50 in mg/kg AND
  - Human MRTDD mg/kg/day AND
  - Hepatotoxicity AND
  - hERG Inhibition assay AND etc.

CRITICAL OUTCOME TECHNOLOGIES INC.

Current Status – HIV Integrase Inhibitors

- *In silico* optimization completed →initial library of 13 compounds identified
- Traditional analytical chemistry completed → 170 new modified/optimized molecules
- New library + 13 original candidates → 183 underwent re-profiling/final optimization
- 12 compounds (re-optimized optimized) → Final library → awaiting synthesis
- *In vitro* HIV Integrase/anti-retroviral assays/acute toxicity → McGill University (Dr. Mark Wainberg) → Aug/Sept 2006

CRITICAL OUTCOME TECHNOLOGIES INC.

In Silico Drug Development Cycle

- **Based on our experience to date**

Stage of Development	Time Line
In Silico Drug Discovery	2-3 months
Drug Synthesis	2-3 months
Confirmatory Tests	2-3 months
PK/Tox Testing*	< 9* months
Phase 1 Human Trial	3-6 months
Our Development Cycle	< 24 months
Industry Cycle	~4-6 years

CRITICAL OUTCOME TECHNOLOGIES INC.

The Support Vector Like Machine

What do you get when you combine CART + the flexibility of Generalized Additive Models (GAM) + the best of SVMs + the evolutionary potential of a Genetic Algorithm?

The SVLM: a flexible tool for Classification and Regression Problems

CRITICAL OUTCOME TECHNOLOGIES INC.

GAMs: T.J. Hastie and R.J. Tibshirani,1990

- Linear Models have the form:
  - Y = a+Bx,
  - a is intercept, B is slope/coefficient
- Generalized additive models (GAMs) replace x_i with a function of x_i summed over all i →
  - **$Y = a+\Sigma B_i*f_i(x_i)$ for i = 1 to P**
- Requires iterative back fitting algorithm
- Can produce very flexible nonlinear models

CRITICAL OUTCOME TECHNOLOGIES INC.

Introducing:GA-SVLM

Combines features of GAMs and SVMs

1. **Function** → RBF → = $\exp(-(x_i-x')^2/(2*\sigma^2))$
   - x' → centroid → can use the mean of all x_i or search for an optimal value
   - σ^2 → width/scale → can use variance but very important to get this right
   - Epanechnikov and tri-cube kernels work too
   - If standardize data so mean = 0 and $\sigma = 1$

$$Y_j = \Sigma B_i * \exp(-(x_i-x_i')^2/(2*\sigma_i^2)) \text{ for } i = 1 \text{ to } P$$

   - **Problem**: must optimize B_i, x_i' and width (σ_i)

CRITICAL OUTCOME TECHNOLOGIES INC.

GA-SVLM

2. **SVM LOSS** → Vapniks ε-insensitive loss
   - $V_\varepsilon(d) = 0$ if $|d| < \varepsilon$ otherwise $= |d| - \varepsilon$
   - Ignores errors $< \varepsilon$
   - Resistant to outliers
   - Similar type of "SPARSE" solution relying only on a portion of examples
   - Promotes good generalization
   - ε is easily estimated from training data
     - $\varepsilon \sim 3 * \sigma_{noise} * sqrt(ln(N)/N)$

CRITICAL OUTCOME TECHNOLOGIES INC.

GA-SVLM

3. **SVM optimization:**
   - Minimize → $\frac{1}{2}*||B||^2+C*\Sigma V_\varepsilon(d_n)$
   - Regularization via minimizing $\frac{1}{2}*||B||^2$
   - Minimizes model complexity through smaller coefficients
   - Minimizes model error as $C*\Sigma V_\varepsilon(d_n)$ over all N examples
4. Dimension P (<=15) from CART/TreeNET data analysis – "variable importance"
5. Results in an optimized nonlinear equation (GA-SVLM) in P variables which can be examined and applied to new observations

Genetic Optimization

- Still have to determine B_i, x_i' and width (σ_i)
- Why not evolve the optimal solution using a Genetic Algorithm?
- Really just a function approximation, optimization problem with 3*P unknowns
- Think of unknowns as "chromosomes"
- Coefficients: $B_1, B_2, B_3, \ldots B_P$
- Centroids: $x'_1, x'_2, x'_3, \ldots x'_P$
- Width: $\sigma_1, \sigma_2, \sigma_3, \ldots \sigma_P$

CRITICAL OUTCOME
TECHNOLOGIES INC

Genetic Optimization

- Data → Training Set +
- Genetic Algorithm (GeneHunter, WSG)
  - Genetic Operations on chromosomes
  - Time/Generations +
- Fitness Function:
  - Minimize → $\frac{1}{2}*||B||^2+C*\Sigma V_\varepsilon(d_n)$
- = GA-SVLM solution
  - **C: $Y_j = \Sigma B_i * \exp(-(x_i-x_i')^2/(2*\sigma_i^2))>=0$ then 1 otherwise -1**
  - **R: $Y_j = \Sigma B_i * \exp(-(x_i-x_i')^2/(2*\sigma_i^2))$ = real value**

CRITICAL OUTCOME TECHNOLOGIES INC.

SVLM Status

- Still in development: Promising
  - Simple to construct and use
  - Performance always comparable to other methods
  - Often slightly better on difficult regression problems
- Development issues:
  - Automating estimates of SVM meta-parameters - epsilon and C
  - Optimizing choice of Kernel Function and Loss Function for data set

CRITICAL OUTCOME TECHNOLOGIES INC.

When All is Said and Done

"I found the secret to happiness, but the FDA won't let me release it."

CRITICAL OUTCOME TECHNOLOGIES INC.

The End..

Of Business as Usual

DALTON Pharma Services

Accelerated Drug Discovery And Development