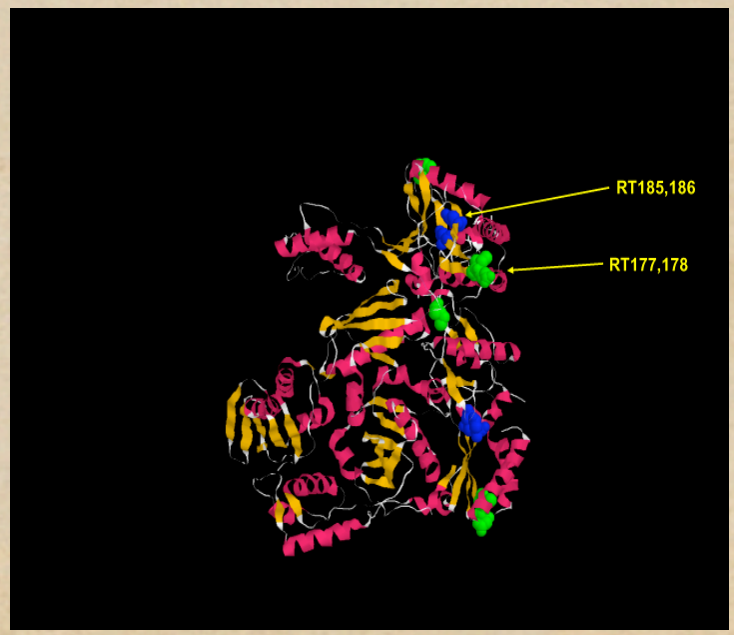
Sequence Based Prediction of HIV-1 Replication Capacity

Mark Segal
UCSF



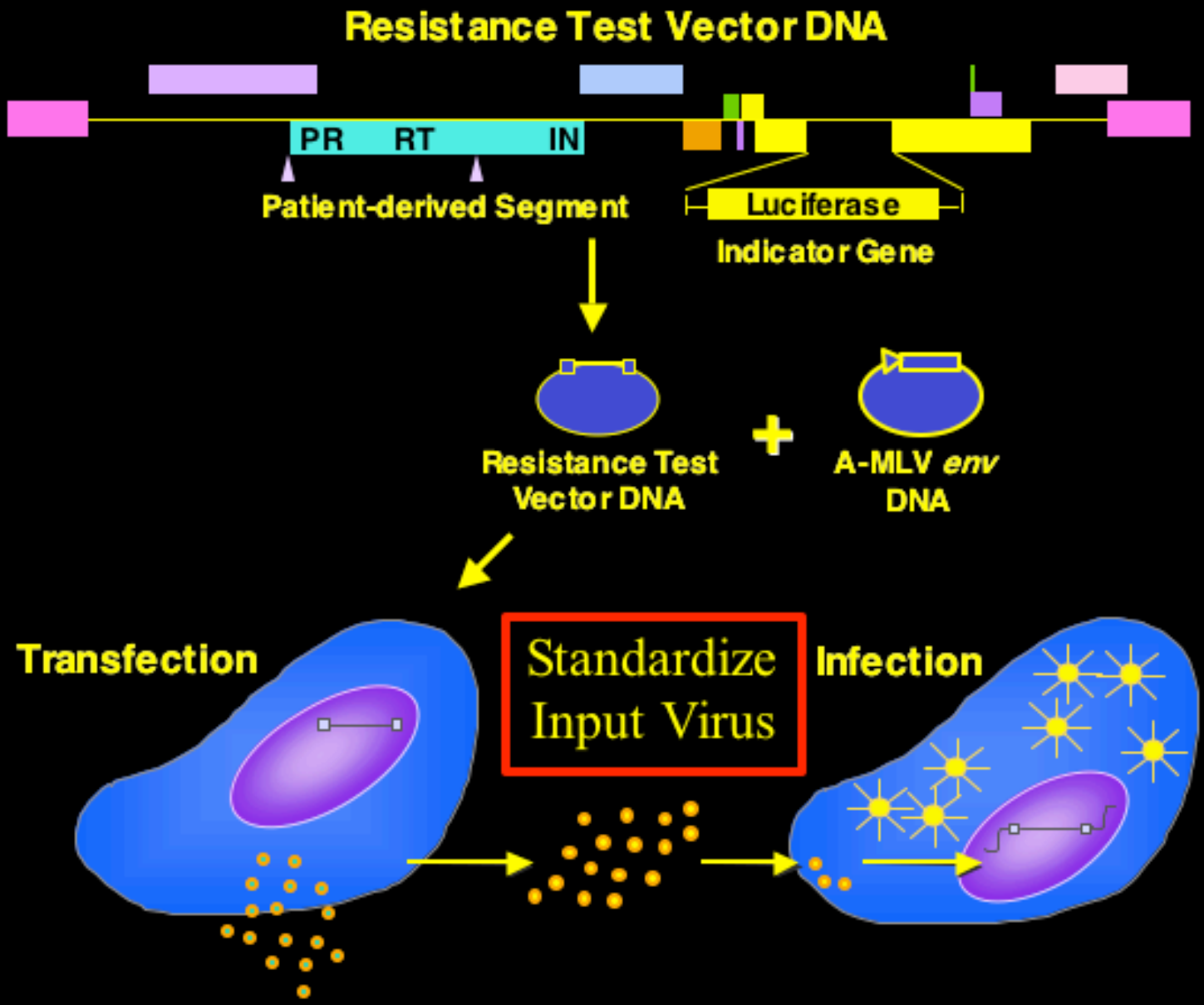
Outline

- HIV-1 Replication Capacity
  - Geno - Pheno Approaches & Issues
  - Tree-Structured Methods & Results
- Random Forests
- Logic Regression
- Conclusions

HIV-1 Replication Capacity

- Outcome: measure of viral fitness (cts)
- Predictors: amino acid sequence from
  - protease (codons 4 - 99) and
  - reverse transcriptase (38 - 223)
- 336 linked RC : PRO/RT records

Replication Capacity Assay



Petropoulos CJ et al AAC

Problem Features / Methods Used

- Distinguished from standard regression problems by the nature of amino acid sequence data:
  - high dimensional (here 282 positions)
  - unordered categorical covariates (amino acids)
  - between-site dependence
  - interactions anticipated
- Various techniques that have been applied:
  - Artificial Neural Networks (Milik et al., 1998; Resch et al., 2001)
  - Prediction Based Classification (Foulkes, DeGruttola, 2002).
  - Tree-Structured Methods (Segal et al., 2001; Beerenwinkel et al., 2002).

Critique: LMs and ANNs

- Linear Models: Difficulties in interpretting linear combinations of unordered categorical covariates.

  Requires computing, examining, grouping indicator coefficients.

  These proliferate when interactions required ⇒ fitting prohibitive.

- ANNs: Effective when high signal-to-noise ratio and prediction, not interpretation, is the goal.

- Plots of connection weights are used to identify important sites

  ⇒ profound identifiability concerns.

- Devices for avoiding indicator encoding of amino acids:

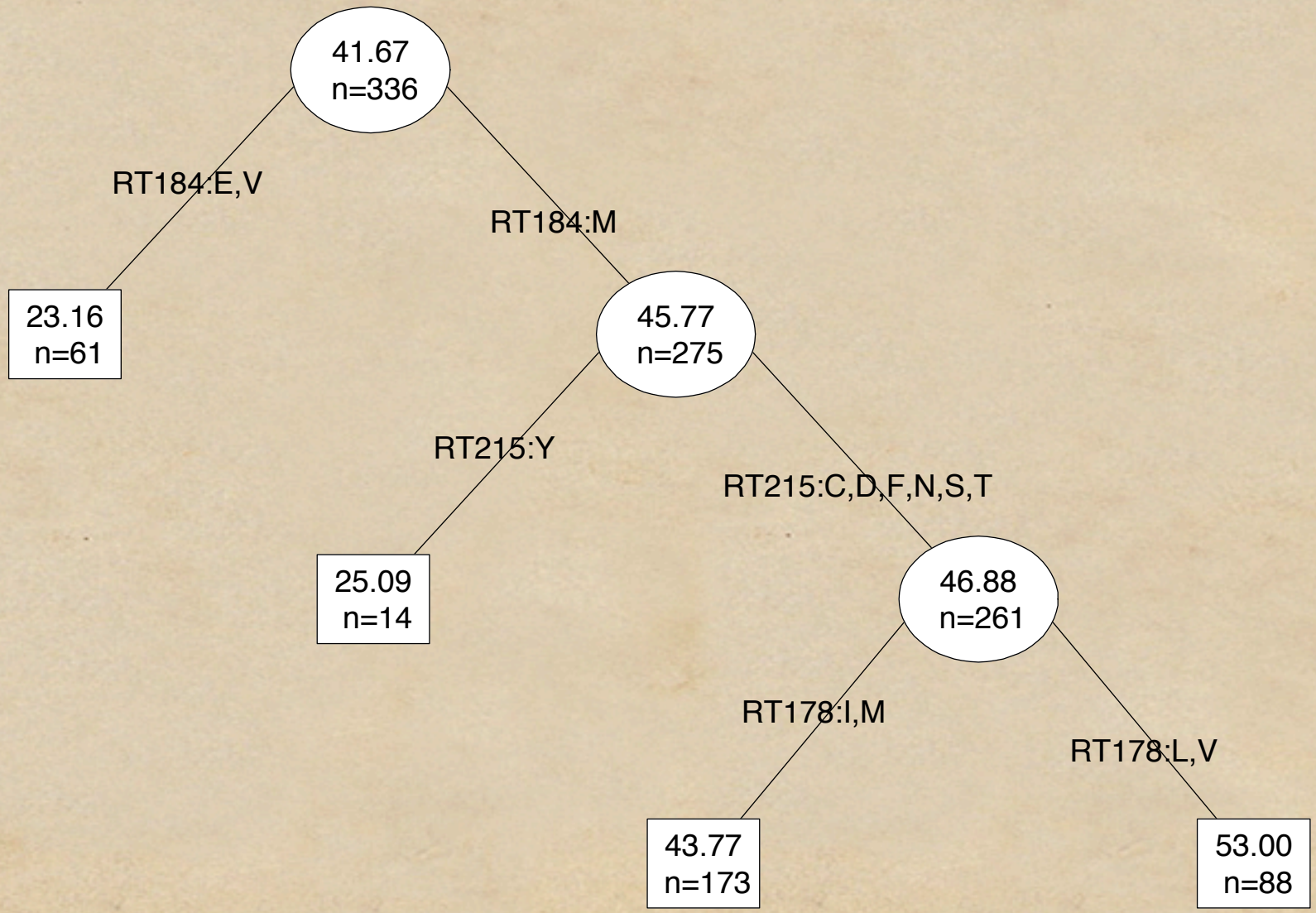
  use of biophysical properties (Milik et al., 1998) or

  arbitrary numeric coding (Resch et al., 2001)

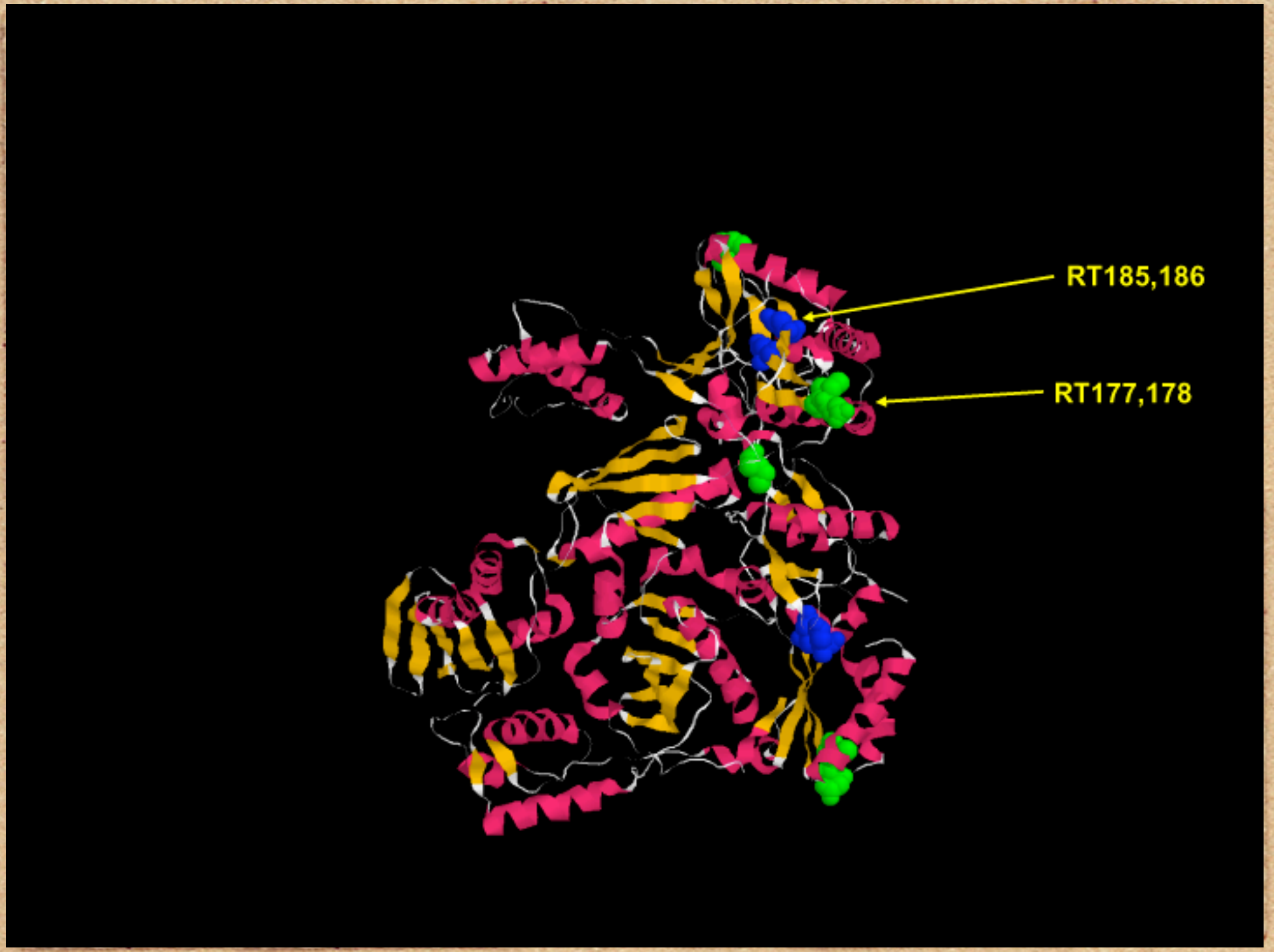
  ⇒ potential information loss, coding sensitive results.

Critique: TSMs

- Strengths of tree-structured methods:
  1. *exhaustively* handle *groups* of amino acids;
  2. can readily handle interactions,
  3. concerns re inadequacies of non-smooth (piecewise-constant) (*cf MARS*) response surface are *moot* with unordered categoric covariates,
  4. readily provide multiple solutions – important in view of strong between-position covariation *for reverse transcriptase, approximately 40% of all possible pairwise position correlations are* simultaneously *significant* ($p < 0.01$) *using the likelihood ratio / permutation testing approach of Bickel et al., (1996).*
- Primary deficiency of tree-structured methods: modest prediction performance compared with flexible methods (e.g., *ANNs, SVMs*).
- Solutions/refinements proposed: bagging, boosting, **Random Forests**.


41.67
n=336
RT184:E,V
RT184:M
23.16
n=61
45.77
n=275
RT215:Y
RT215:C,D,F,N,S,T
25.09
n=14
46.88
n=261
RT178:I,M
RT178:L,V
43.77
n=173
53.00
n=88

- RT184, RT215 primary drug resistance sites which are known to affect RC.
- Naturally occurring polymorphisms??
- RT178 sits on a loop that also holds two amino acids (D185/D186) critical for reverse transcriptase protein function: coordination of Mg^{2+} needed for binding the template.


RT185,186
RT177,178

- RT178 split primarily Isoleucine (I) versus Valine (V).
- While both are hydrophobic, there are volumetric and hydrogen bonding opportunity differences that may force a chain of structural changes along the loop containing RT185 and RT186.
- A similar effect has been described for the drug resistance substitution M184V.
- RT178 under HLA control.

On to Random Forests

Breiman

- Better the model fits, the more sound the inference
- Standard models tend to fit poorly
- Fit measured by prediction error (PE)
- Substantial gains in PE can be achieved by using ensembles of (simple) predictors

A random forest is a collection of tree predictors

$h(\mathbf{x};\boldsymbol{\theta}_k), \quad k = 1, \ldots, K; \quad \boldsymbol{\theta}_k \; iid$ random vectors.

For regression, the forest prediction is the

unweighted average over the collection: $\bar{h}(\mathbf{x})$.

As $k \to \infty$ the Law of Large Numbers ensures

$$E_{\mathbf{X},Y}(Y - \bar{h}(\mathbf{X}))^2 \to E_{\mathbf{X},Y}(Y - E_{\boldsymbol{\theta}} h(\mathbf{X};\boldsymbol{\theta}))^2 \equiv PE_f^*$$

the forest prediction error.

Convergence implies forests *don't* overfit.

Define average prediction error for a tree as

$$PE_t^* = E_{\boldsymbol{\theta}} E_{\mathbf{X},Y}(Y - h(\mathbf{X};\boldsymbol{\theta}))^2.$$

Assume $EY = E_{\mathbf{X}} h(\mathbf{x};\boldsymbol{\theta})\ \forall \boldsymbol{\theta}$. Then $PE_f^* \leq \bar{\rho} PE_t^*$ where $\bar{\rho}$ is weighted correlation between residuals for independent $\boldsymbol{\theta}', \boldsymbol{\theta}''$.

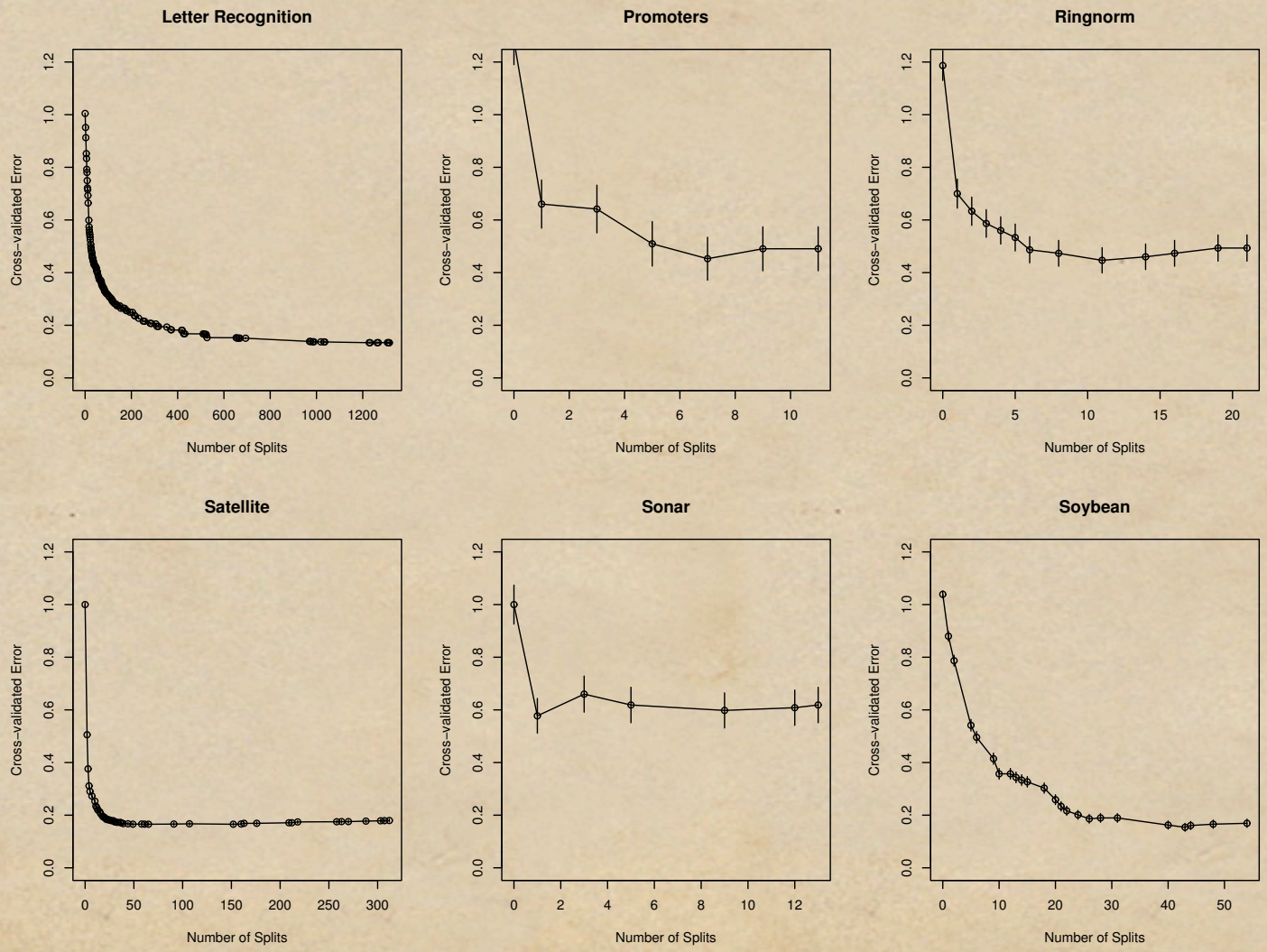
The inequality pinpoints requirements for accurate regression forests: low correlation between residuals and low error trees. Further, forests decrease PE_t^* by factor $\bar{\rho} \Rightarrow$ the randomization injected strives for low correlation.

- To keep error low, grow trees to maximum depth
  - *contols bias but not variance??*
  - *variance control by ensemble averaging*
- To keep correlation low randomize via
  1. Grow each tree on a bootstrap sample.
  2. Specify $m \ll p$ (number of covariates). At each node select m covariates and pick the best split based on these.

Bootstrapping allows for an internal (*oob*) test set estimate of PE_f^* to be carried along.

Empirically, RF proven to have very low PE^*_f.

Insensitive to only tuning parameter m. *BUT...*

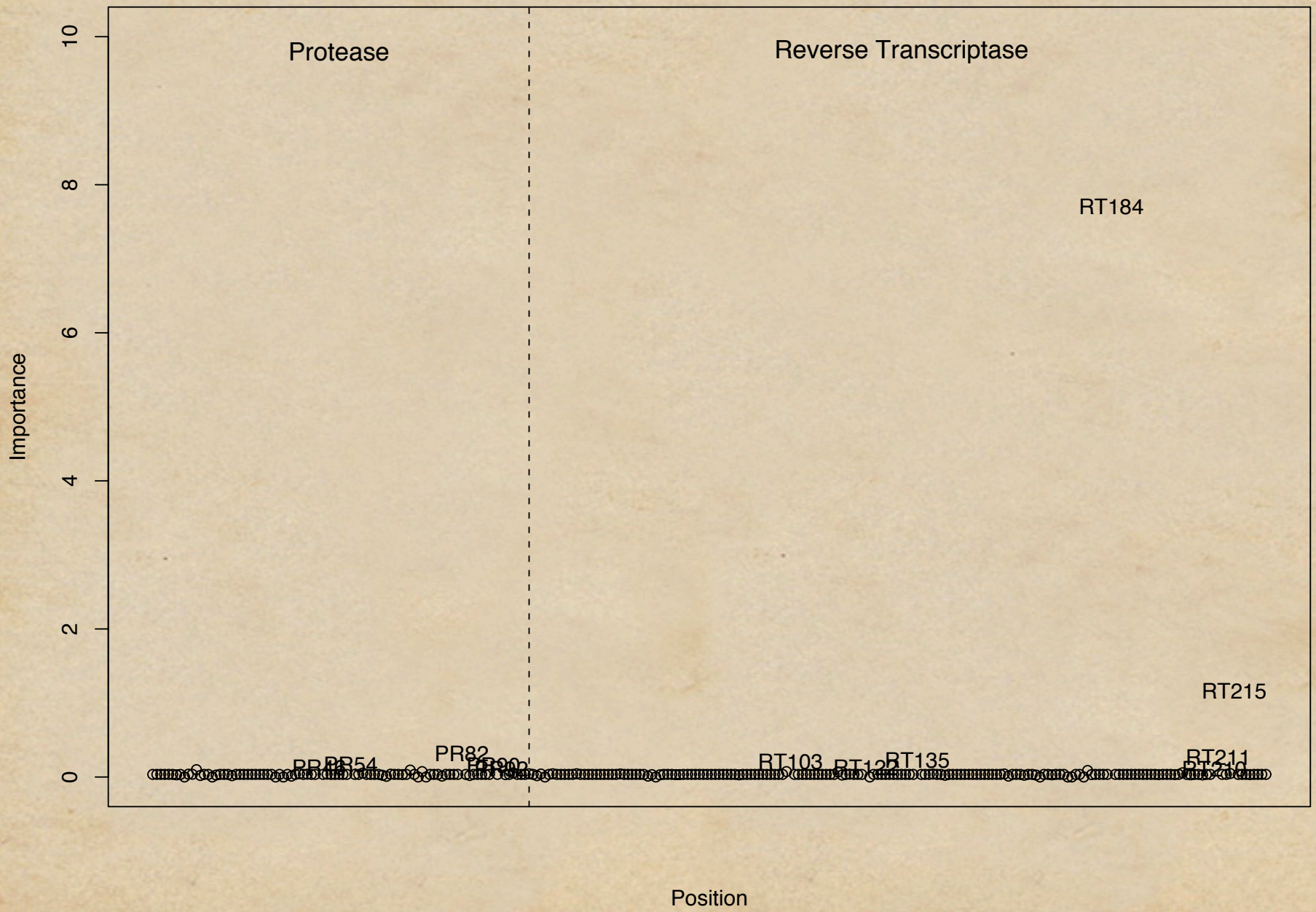


- Almost all UCI repository benchmark datasets exhibit this behaviour -- they are hard to overfit (using trees).
- For situations where overfitting arises, ease of RF exploration enhanced by addition of a tuning parameter governing (individual) tree depth.

Replication Capacity: Random Forest PE_f^*

# Splits per Tree	Minimun Node Size	# Covariates per Split (m) 10	20	100	282
Unlimited	5	589.7	590.4	**608.2**	602.9
	25	589.2	586.7	587.5	593.8
	50	594.0	583.7	582.1	584.2
5	5	602.9	592.9	575.6	578.6
	25	598.5	587.4	576.2	577.1
	50	592.4	588.4	581.2	581.6

Tree structured $PE_t^* = 575.5$


Protease
Reverse Transcriptase
RT184
RT215
PR82
RT103
RT135
RT211
Importance
Position
0
2
4
6
8
10

Logic Regression

- Ruczinski, Kooperberg, LeBlanc. *JCGS*, 2003.
- Intended for settings where most predictors are binary.
- Searches for Boolean combinations of predictors in the entire space of such combinations.
- Is completely embedded in a regression framework, with corresponding determination of model quality: *RSS*, log-likelihood, ...
- Distinguished by non-greedy search, generality.

Logic Regression Model Formulation

- $X_1, \ldots, X_k$ are 0/1 (False/True) predictors.
- Y is a response variable – here RC.
- Fit the model

$$g(E(Y)) = \beta_0 + \sum_{j=1}^{J} \beta_j \times L_j$$

where L_j is a Boolean combination of the covariates, e.g. $L_j = (X_1 \vee X_2) \wedge X_4^c$.

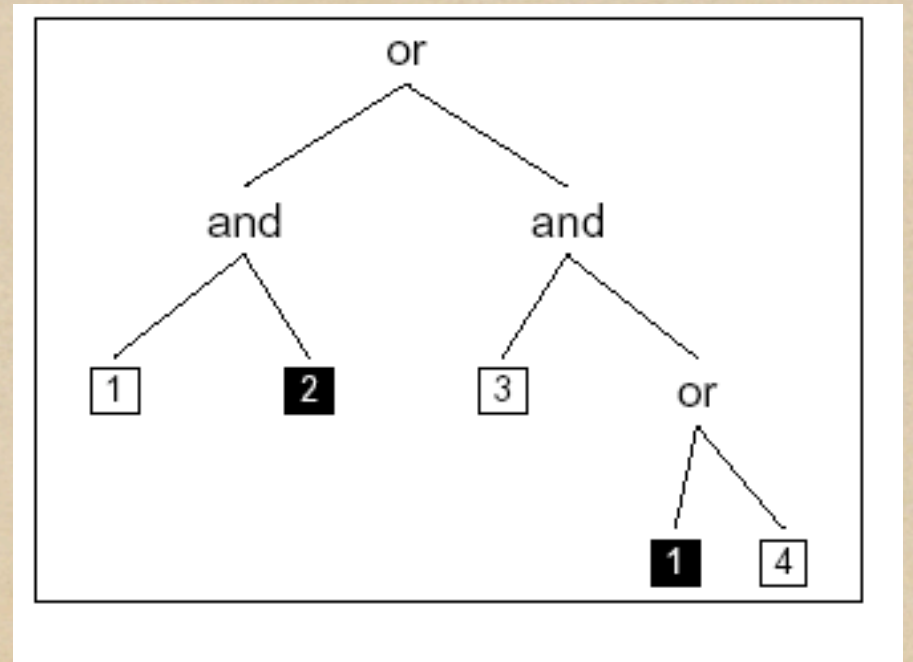
- Fix J and determine logic terms L_j and estimate β_j simultaneously.

Logic Trees

- Boolean expressions can be represented as trees:

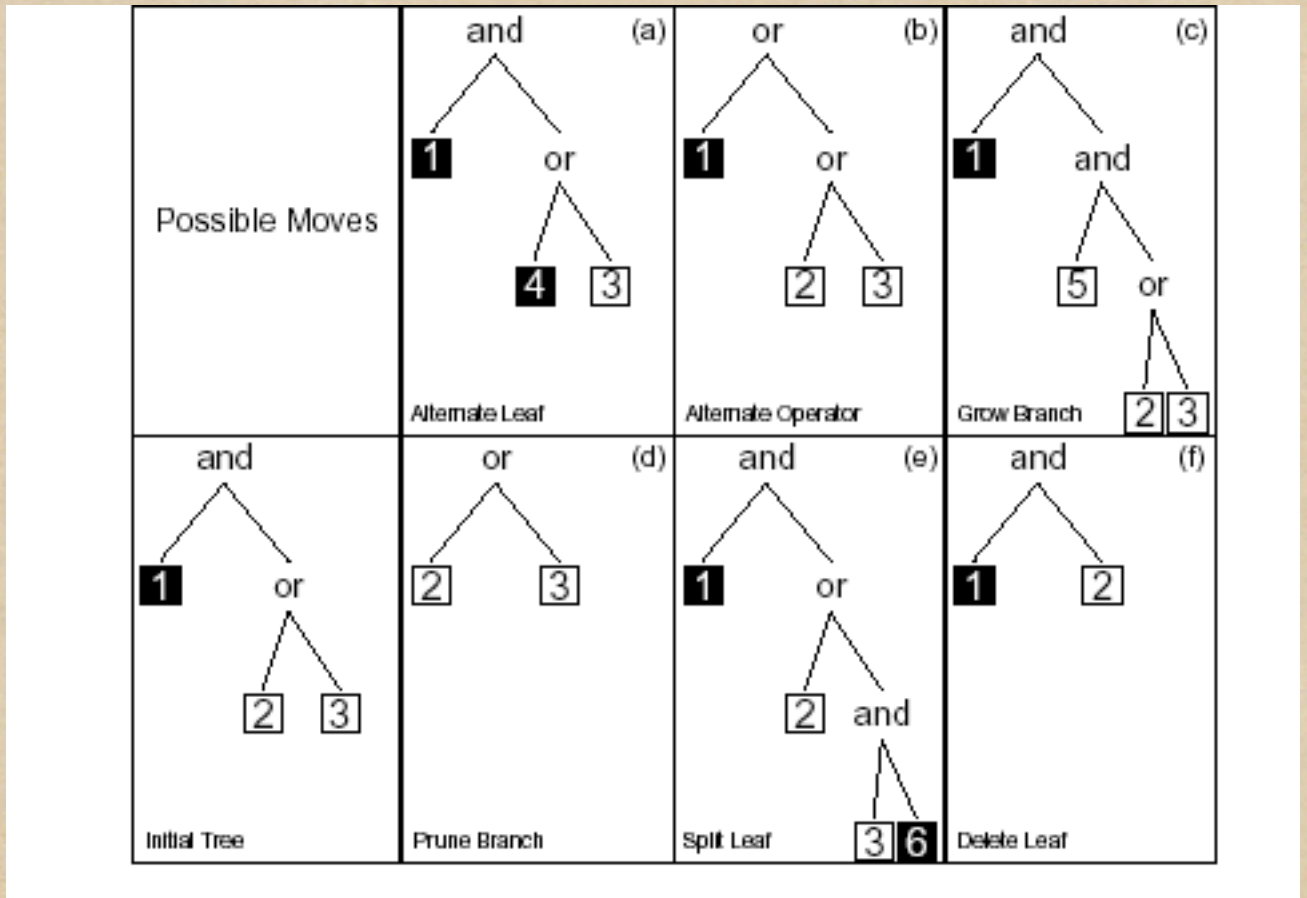
$$(X_1 \wedge X_2^c) \vee (X_3 \wedge (X_1^c \vee X_4))$$

corresponds to



Simulated Annealing:

The Move Set



Logic Regression Fitting

Select a scoring fn: RSS, log-likelihood,...

Pick the maximum number of Logic Trees. (J)

Pick the maximum number of leaves in a tree.

Initialize. $L_j = 0 \ \forall j$

Carry out the Simulated Annealing Algorithm:

-- **Propose a move.**

-- **Accept or reject the move, depending on scores and temperature.**

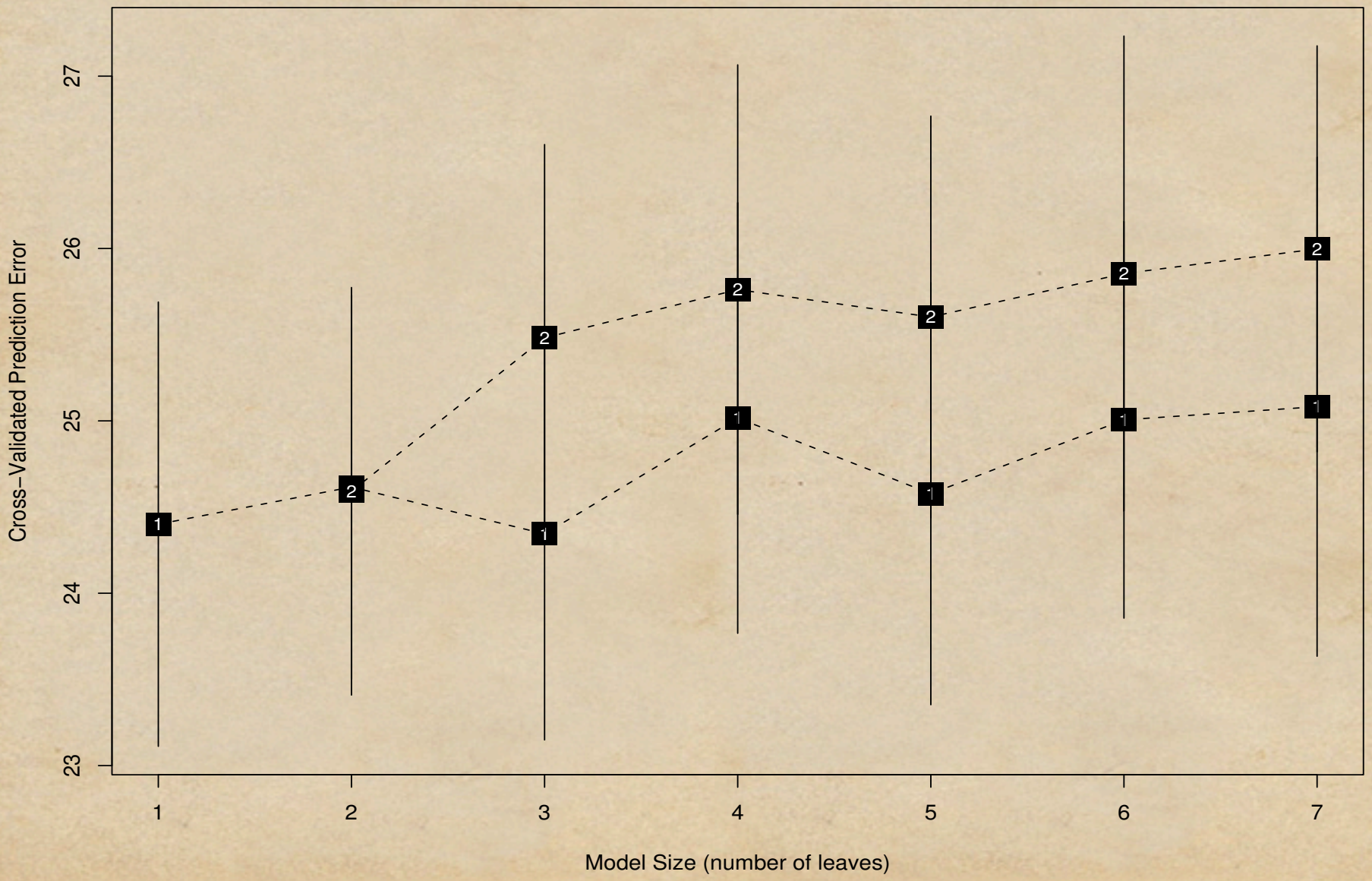
Model Selection & Size

- CV, randomization tests employed
- Requires measure of model size
- Presently taken as number of leaves
- Potentially problematic:
  - more complex models : fewer leaves
  - Boolean expressions non-unique

Adaptive DF

- Efron (86), Tibshirani and Knight (99), Ye (98), Efron et al (04)
- $\hat{\mu} = g(\mathbf{y}); \quad \mathrm{cov}(\mathbf{y}) = \sigma^2 \mathbf{I}$
- $df = \sum_{i=1}^{n} \mathrm{cov}(\hat{\mu}_i, y_i)/\sigma^2$

Logic Regression: RC



- Logic regression model with minimal cross-validation prediction error features one logic tree with three leaves.
- Variables used are RT184, RT215, RT178.
- Prediction error variance = 592.


Logic Regression ENPs for One Tree
Effective Number of Parameters
0
10
20
30
40
50
2
4
6
8
Number of Leaves

Conclusions

- TSM effective for evaluating genotype-phenotype association.
- RF may not realize prediction gains due to strong between site dependence.
- Adaptive degrees of freedom are a useful complement to logic regression.
- Structurally significant RT sites found.

Acknowledgements

- Jason Barbour
- Robert Grant
- Virologic Inc