Applications of Statistical Geometry to the Functional Analysis of Protein Mutants

Majid Masso
Ph.D. Dissertation Defense
Four-Body Statistical Potential

- Protein structures are represented as discrete sets of points in 3D, each corresponding to an amino acid (aa).
- Delaunay tessellation of a protein structure yields an aggregate of space-filling, non-overlapping, irregular tetrahedra (simplices) that each define a quadruplet of nearest-neighbor aa’s.
- A four-body statistical potential function is derived via tessellation of a training set of structures, assigning a log-likelihood score to all possible quadruplets of aa’s.
Computational Mutagenesis Methodology

- **Total Potential** or **Topological Score** of a protein structure, a global measure of sequence-structure compatibility, is obtained by summing the scores of all the simplices in the tessellation.

- **Individual Residue Potential** or **Residue Environment Score** of each aa in a protein structure is obtained by locally summing the scores of only the simplices that use the aa’s point representation as a vertex; the scores of all the aa’s form a *Potential Profile* vector.

- Assumption: minor structural differences and similar tessellations between each mutant and the wild-type (wt) protein.

- Approach: the total potential and potential profile of every mutant can be derived from the tessellation of the wt structure.
Computational Mutagenesis Methodology

Based on the methodology, each mutant is characterized by a scalar *Residual Score* and a vector *Residual Profile*:

- **Residual Score** – difference between mutant and wt total potentials
  - Measures the relative change in mutant sequence-structure compatibility from wt
- **Residual Profile** – difference between mutant and wt potential profiles
  - Quantifies environmental perturbations from wt at every aa position
  - Each component in the profile is referred to as an *environmental change (EC) score* for the corresponding aa position
Comprehensive Mutational Profile (CMP)

- At each residue position in a protein structure, a CMP score is obtained by calculating the mean of the 20 residual scores associated with all possible aa replacements (including the degenerate mutant obtained by substituting the wt aa with itself, with residual score 0)
- Mathematically,

\[
\text{CMP}_j = \frac{1}{20} \sum_{i=1}^{20} [(\text{mutant topological score})_{ij} - (\text{wt topological score})]
\]

\[
= \frac{1}{20} \sum_{i=1}^{20} (\text{mutant residual score})_{ij}
\]

\[
= \{\text{mean residual score}\}_{j}
\]

where index \(i\) refers to the 20 aa’s, and index \(j\) refers to the position in the 1° sequence of the protein
CMP Example: HIV-1 Protease

- PDB ID: 3phv (monomer, 99 aa’s)
- Functional as a homodimer
  - Interface: P1-T4 and C95-F99
  - Catalytic triad: D25-T26-G27
  - Flap region: M46-V56
CMP Example: HIV-1 Protease

3phv Comprehensive Mutational Profile vs. Potential Profile

Residue Types
- Hydrophobic
- Charged
- Polar

R^2 = 0.8961
HIV-1 Protease Experimental Data

• Synthesis and analysis of 536 single site missense mutants
  ➢ 200 mutants provided by R. Swanstrom (UNC)

• Each mutant placed in one of 3 phenotypic categories, positive, negative, or intermediate, based on activity (ability to process the Pol polyprotein)

• Residual scores of the mutants can be used to elucidate the structure-function relationship in HIV-1 protease
How significant are the differences in class-pair means?
Pos-Neg: $p = 1.65 \times 10^{-11}$; Int-Neg: $p = 9.90 \times 10^{-6}$; and Pos-Int: $p = 0.086$. 
Structure-Function Correlations Based on Residual Scores: Bacteriophage T4 Lysozyme

- Experimental data: 2015 single site mutants generated by introducing the same 13 aa replacements at 163/164 positions - all but M1 (PDB ID: 3lzm)
- Four mutant activity classes: high, medium, low, negative
- Investigators recommend data analysis using only two classes (active = high + med, inactive = low + neg): $p = 0.0003$

![Graphical representation of mutant activity and structure-function correlation](image-url)
Structure-Function Correlations Based on Residual Scores: *E. coli Lac* Repressor

- Experimental data: 4041 single site mutants generated by introducing the same 13 aa replacements at positions 2-329 (PDB ID: 1efaB)
- Four mutant activity classes based on degree of repression of β-galactosidase: fully active (greater than 200-fold), moderate (20 to 200-fold), low (4 to 20-fold), inactive (less than 4-fold)
- Investigators suggest combining moderate + low = intermediate
- Recent computational studies using this data set define two classes: unaffected (fully active) and affected (all other classes combined)
- All 328 *lac* repressor residue positions were annotated and clustered into 15 groups based on their structural locations, functional roles, and level of tolerance to mutations
How significant are the differences in class-pair means?
full-inter: \( p = 4.64 \times 10^{-7} \); full-inactive: \( p = 1.95 \times 10^{-36} \); and inter-inactive: \( p = 6.57 \times 10^{-10} \).
Lac Repressor: CMP vs. Potential Profile

![Graph showing the relationship between CMP scores and residue environment scores. The graph is divided into four quadrants: Quadrant 1, Quadrant 2, Quadrant 3, and Quadrant 4. The graph includes data points for hydrophobic, charged, and polar residue types. The $R^2$ value is 0.81.]
Distribution of \textit{Lac} Repressor Residue Positions

Apply chi-square test with 18 df: $\chi^2 = 51.11$, so reject null hypothesis that no association exists between structural/functional groups and quadrant locations, with $p < 0.0001$
Characterizing Structural or Functional Roles of *Lac* Repressor Residues Based on Residual Scores and Residue Environment Scores

![Graph showing mean scores for different categories and conditions.](image-url)
Mutant Residual Profiles: Motivation

- Residual profile vectors encode much more sequence and structure information about the mutants than residual scores; hence, they may prove to be more useful for classification and inference for mutants belonging to different activity classes.

- Nonzero components (EC scores) of a mutant residual profile identify the mutated position(s) as well as all of their topological nearest-neighbors based on tessellation (i.e., all positions that participate in simplices with the mutated positions).

- For any single site mutant, the EC score at the residual profile component corresponding to the mutated position is precisely the residual score of the mutant.

- A consequence of the above is that all 19 single site mutants at a particular position have residual profiles with identical arrangements of zero and nonzero components (only the EC scores at any given nonzero component differ among the 19 residual profiles).
HIV-1 Protease Dataset: Residual Profiles of the Experimental Mutants

In each of the 536 rows, the initial three components identify the mutant. This is followed by the 99-dimensional residual profile. The final component is the mutant activity class.
Supervised Classification

• Algorithms: Neural Network (NN), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF)
• Implementations available with the Weka suite of machine learning tools: http://www.cs.waikato.ac.nz/ml/weka/
• Training set: Residual profile vectors for the mutants of a protein that have been studied experimentally, along with the activity class of each mutant (i.e., supervised)
• Each mutant (represented as a residual profile + activity class) is referred to as an instance; each component of the residual profiles is referred to as an attribute
Model Performance: HIV-1 Protease Mutants

Support Vector Machines (SVM)  Decision Trees (DT)  Neural Networks (NN)

AUC = 0.7810  Standard Error = 0.0195
Control AUC = 0.5104

AUC = 0.8324  Standard Error = 0.0172
Control AUC = 0.5059

AUC = 0.7764  Standard Error = 0.0197
Control AUC = 0.4925

Neg AUC = 0.7810
Int AUC = 0.6731
Pos AUC = 0.7389
Overall AUC = 0.7631

Neg AUC = 0.8324
Int AUC = 0.6632
Pos AUC = 0.7732
Overall AUC = 0.7904

Neg AUC = 0.7764
Int AUC = 0.6814
Pos AUC = 0.7282
Overall AUC = 0.7489

Int-Neg AUC = 0.7511
Int-Neg AUC = 0.7726
Pos-Neg AUC = 0.7412
Pos-Int AUC = 0.5889
Overall AUC = 0.7336

Int-Neg AUC = 0.7655
Int-Neg AUC = 0.7558
Pos-Neg AUC = 0.8192
Pos-Int AUC = 0.6522
Overall AUC = 0.7421

False Positive Rate (1 – Specificity)
# AUC Summary for HIV-1 Protease ROC Curves

<table>
<thead>
<tr>
<th></th>
<th>Pos (1-against-1)</th>
<th>Int (1-against-1)</th>
<th>Neg (1-against-1)</th>
<th>Others Combined (1-against-all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos</td>
<td>---</td>
<td>0.6522 (SVM)</td>
<td>0.8182 (SVM)</td>
<td>0.7389 (SVM)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5869 (DT)</td>
<td>0.8414 (DT)</td>
<td>0.7732 (DT)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.6225 (NN)</td>
<td>0.7877 (NN)</td>
<td>0.7282 (NN)</td>
</tr>
<tr>
<td>Int</td>
<td></td>
<td>---</td>
<td>0.7558 (SVM)</td>
<td>0.6731 (SVM)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.7726 (DT)</td>
<td>0.6632 (DT)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.7511 (NN)</td>
<td>0.6814 (NN)</td>
</tr>
<tr>
<td>Neg</td>
<td></td>
<td></td>
<td>---</td>
<td>0.7810 (SVM)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8324 (DT)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.7764 (NN)</td>
</tr>
</tbody>
</table>

- Most disparate signals stem from residual profiles of the positive and negative mutants, followed closely by the intermediate and negative mutants
  - consistent with biological notion that fully active and inactive mutants display the greatest differences in structural and functional properties, while partially active and inactive mutants display significant, albeit less dramatic differences
- Residual profiles of mutants in the positive and intermediate classes display the least divergent signals
  - reflects the fact that both classes contain mutants that are more or less functionally active and display at most minimal structural changes from wt
Reliability/Reproducibility of Model Predictions

- 60/40 split test option => model learned with 60% of the mutants is used to predict the activity classes of the remaining 40%; 60 runs => expect approx. 24 predictions/mutant
- Apply two-class decision tree learning (default costs)
- For each mutant, \( n_c (n_i) = \) total # of correct (incorrect) predictions
- Mutant reliability metric: \( m = (n_c - n_i) / (n_c + n_i) \)
- \( m = 0 \) => equal # of correct and incorrect predictions; \( m = 1 \) => all predictions correct; \( m = -1 \) => all predictions incorrect

<table>
<thead>
<tr>
<th>residue</th>
<th>wt</th>
<th>sub.</th>
<th>sum</th>
<th>proportion</th>
<th>wt</th>
<th>sub.</th>
<th>sum</th>
<th>proportion</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>TYR</td>
<td>2</td>
<td>15</td>
<td>17</td>
<td>0.0159</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0125</td>
<td>0.00</td>
</tr>
<tr>
<td>LYS</td>
<td>20</td>
<td>27</td>
<td>47</td>
<td>0.0438</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.0125</td>
<td>0.29</td>
</tr>
<tr>
<td>PHE</td>
<td>20</td>
<td>21</td>
<td>41</td>
<td>0.0382</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.0125</td>
<td>0.33</td>
</tr>
<tr>
<td>HIS</td>
<td>6</td>
<td>29</td>
<td>35</td>
<td>0.0243</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.0125</td>
<td>0.52</td>
</tr>
<tr>
<td>ASP</td>
<td>18</td>
<td>22</td>
<td>40</td>
<td>0.0373</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0.025</td>
<td>0.67</td>
</tr>
<tr>
<td>ILE</td>
<td>92</td>
<td>28</td>
<td>120</td>
<td>0.1119</td>
<td>8</td>
<td>0</td>
<td>6</td>
<td>0.075</td>
<td>0.67</td>
</tr>
<tr>
<td>VAL</td>
<td>35</td>
<td>40</td>
<td>75</td>
<td>0.0790</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>0.05</td>
<td>0.71</td>
</tr>
<tr>
<td>ARG</td>
<td>27</td>
<td>45</td>
<td>72</td>
<td>0.0672</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>0.05</td>
<td>0.74</td>
</tr>
<tr>
<td>CYS</td>
<td>21</td>
<td>11</td>
<td>32</td>
<td>0.0299</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.025</td>
<td>0.84</td>
</tr>
<tr>
<td>MET</td>
<td>14</td>
<td>18</td>
<td>32</td>
<td>0.0299</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0.025</td>
<td>0.84</td>
</tr>
<tr>
<td>ASN</td>
<td>24</td>
<td>22</td>
<td>46</td>
<td>0.0429</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0.0375</td>
<td>0.87</td>
</tr>
<tr>
<td>LEU</td>
<td>52</td>
<td>37</td>
<td>89</td>
<td>0.0830</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>0.075</td>
<td>0.90</td>
</tr>
<tr>
<td>GLU</td>
<td>12</td>
<td>29</td>
<td>41</td>
<td>0.0382</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0.0375</td>
<td>0.88</td>
</tr>
<tr>
<td>GLY</td>
<td>98</td>
<td>29</td>
<td>127</td>
<td>0.1185</td>
<td>7</td>
<td>4</td>
<td>11</td>
<td>0.375</td>
<td>1.16</td>
</tr>
<tr>
<td>SER</td>
<td>2</td>
<td>40</td>
<td>42</td>
<td>0.0392</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>0.05</td>
<td>1.28</td>
</tr>
<tr>
<td>THR</td>
<td>32</td>
<td>35</td>
<td>67</td>
<td>0.0625</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>0.0875</td>
<td>1.40</td>
</tr>
<tr>
<td>GLN</td>
<td>17</td>
<td>25</td>
<td>42</td>
<td>0.0392</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>0.0625</td>
<td>1.60</td>
</tr>
<tr>
<td>PRO</td>
<td>28</td>
<td>23</td>
<td>51</td>
<td>0.0475</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>0.0875</td>
<td>1.84</td>
</tr>
<tr>
<td>ALA</td>
<td>10</td>
<td>33</td>
<td>43</td>
<td>0.0491</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>0.075</td>
<td>1.87</td>
</tr>
<tr>
<td>TRP</td>
<td>6</td>
<td>16</td>
<td>22</td>
<td>0.0205</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>0.0625</td>
<td>3.05</td>
</tr>
<tr>
<td>SUM</td>
<td>536</td>
<td>536</td>
<td>1072</td>
<td>1.00</td>
<td>40</td>
<td>40</td>
<td>80</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

m Metric Bins

HIV-1 Protease Mutants

536

40 Mutants Satisfying \( m < 0.5 \)
Assessment of the Statistical Significance for the Number of Correctly Classified Instances

- Random split: 436 HIV-1 protease mutants used as a training set for decision tree learning; remaining 100 mutants form a test set
- Training: 121 pos, 66 int, 249 neg; Testing: 19 pos, 18 int, 63 neg
- Result below based on two classes (similar method for 3 classes):

<table>
<thead>
<tr>
<th>Correctly Classified Instances</th>
<th>74</th>
<th>74%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrectly Classified Instances</td>
<td>26</td>
<td>26%</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

=== Detailed Accuracy By Class ===

<table>
<thead>
<tr>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.73</td>
<td>0.254</td>
<td>0.628</td>
<td>0.73</td>
<td>active</td>
</tr>
<tr>
<td>0.746</td>
<td>0.27</td>
<td>0.825</td>
<td>0.746</td>
<td>inactive</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>10</td>
<td>a = active</td>
</tr>
<tr>
<td>16</td>
<td>47</td>
<td>b = inactive</td>
</tr>
</tbody>
</table>
Assessment of the Statistical Significance for the Number of Correctly Classified Instances

- Training: 187 active, 249 inactive; Testing: 37 active, 63 inactive
- Let \( X = X_1 + X_2 + \ldots + X_{100} \), where each \( X_i \) is a Bernoulli random variable representing the outcome of a test set instance prediction.
- \( \mu = E(X) = 37 \cdot (187/436) + 63 \cdot (249/436) = 52 \)
- \( \sigma^2 = \text{Var}(X) = 100 \cdot (187/436) \cdot (249/436) = 24.5 \)
- So \( \sigma = 4.95 \), and \( p \)-value is

\[
P(X > 74; \mu = 52) = P \left( \frac{X - \mu}{\sigma} > \frac{74 - 52}{4.95} \right) = P(z > 4.44) \approx 1 - \Phi(4.44) = 4.42 \times 10^{-6}
\]

where \( \Phi \) is the cumulative dist. fn. for a standardized normal var.

Summary of Results Based on Two and Three Classes

<table>
<thead>
<tr>
<th>HIV-1 protease mutants (436 for training, 100 for testing)</th>
<th>2 classes</th>
<th>3 classes (1-against-all)</th>
<th>3 classes (1-against-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>expected number of correctly classified test mutants (std. dev.)</td>
<td>52 (4.95)</td>
<td>44 (4.64)</td>
<td>44 (4.64)</td>
</tr>
<tr>
<td>actual number of correctly classified test mutants</td>
<td>74</td>
<td>69</td>
<td>66</td>
</tr>
<tr>
<td>( p = 4.42 \times 10^{-6} )</td>
<td>( p = 3.57 \times 10^{-8} )</td>
<td>( p = 1.06 \times 10^{-6} )</td>
<td></td>
</tr>
</tbody>
</table>
Model Performance: T4 Lysozyme Mutants

Support Vector Machines (SVM)  
- AUC = 0.7107
- Standard Error = 0.0179
- Control AUC = 0.4987

Decision Trees (DT)  
- AUC = 0.7721
- Standard Error = 0.0168
- Control AUC = 0.4967

Neural Networks (NN)  
- AUC = 0.7471
- Standard Error = 0.0174
- Control AUC = 0.5236

False Positive Rate (1 – Specificity)

2 activity classes (active/inactive)

4 activity classes (1-against-all)

4 activity classes (1-against-1)
T4 Lysozyme Mutational Array

Training set mutants (n = 2015)  Predicted test set mutants (n = 1101)

Active  Inactive  Active  Inactive
T4 Lysozyme Prediction Results

- Predicted activities compared with exp. activity from 8 labs
- Exp. data obtained from ProTherm database
- Exp. activity \( \leq 5 \) inactive, and values \( > 5 \) active
- Result: 30/35 correct predictions, \(~86\%\)

<table>
<thead>
<tr>
<th>#</th>
<th>Mutant name</th>
<th>Predicted</th>
<th>Actual</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>E11M</td>
<td>inactive</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>E11N</td>
<td>inactive</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>D20N</td>
<td>inactive</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>D20T</td>
<td>inactive</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>S38D</td>
<td>active</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>N40D</td>
<td>active</td>
<td>124</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>A41D</td>
<td>active</td>
<td>105</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>A41V</td>
<td>active</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>I78M</td>
<td>active</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>L84M</td>
<td>active</td>
<td>104</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>P86D</td>
<td>active</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>P86I</td>
<td>active</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>P86T</td>
<td>active</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>L91M</td>
<td>active</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td>A93T</td>
<td>active</td>
<td>105</td>
<td></td>
</tr>
<tr>
<td>16.</td>
<td>A98V</td>
<td>inactive</td>
<td>80</td>
<td>+</td>
</tr>
<tr>
<td>17.</td>
<td>L99M</td>
<td>active</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>18.</td>
<td>L100M</td>
<td>active</td>
<td>105</td>
<td></td>
</tr>
<tr>
<td>19.</td>
<td>M102T</td>
<td>inactive</td>
<td>60</td>
<td>+</td>
</tr>
<tr>
<td>20.</td>
<td>N103M</td>
<td>active</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>21.</td>
<td>V111I</td>
<td>active</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>22.</td>
<td>N116D</td>
<td>active</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>23.</td>
<td>S171</td>
<td>inactive</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>24.</td>
<td>S17V</td>
<td>inactive</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>25.</td>
<td>L118M</td>
<td>active</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>26.</td>
<td>L121M</td>
<td>active</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>27.</td>
<td>N132I</td>
<td>active</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>28.</td>
<td>N132M</td>
<td>inactive</td>
<td>40</td>
<td>+</td>
</tr>
<tr>
<td>29.</td>
<td>L133M</td>
<td>active</td>
<td>106</td>
<td></td>
</tr>
<tr>
<td>30.</td>
<td>N144D</td>
<td>active</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>31.</td>
<td>A146T</td>
<td>active</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>32.</td>
<td>F153M</td>
<td>inactive</td>
<td>87</td>
<td>+</td>
</tr>
<tr>
<td>33.</td>
<td>G156D</td>
<td>active</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>34.</td>
<td>T157I</td>
<td>inactive</td>
<td>90</td>
<td>+</td>
</tr>
<tr>
<td>35.</td>
<td>N163D</td>
<td>active</td>
<td>193</td>
<td></td>
</tr>
</tbody>
</table>
**Lac Repressor Decision Tree Model Performance:**

Two Activity Classes (Unaffected/Affected)

- Accuracy: 78.67%
- AUC ± SE: 0.8023 ± 0.0068
- Control: activity labels randomly shuffled among the 4041 mutant residual profile vectors in the training set prior to applying decision tree learning
Learning Curve Example: Lac Repressor

- Stratified training sets randomly chosen with replacement in increments of 100 mutants
- At each training set size, mean 10 CV accuracy based on average of 10 runs using two-class decision tree supervised learning
- Error bars represent ±1 std. dev. from the mean
**Lac Repressor Mutational Array**

Training set mutants (n = 4041)  Predicted test set mutants (n = 2229)

- **Unaffected**
- **Affected**

---

**Legend:**
- Red: Unaffected
- Green: Affected
- Pink: Unaffected
- Light Green: Affected
Clinical Application: Prediction of Drug Resistance Protein Mutational Patterns

- Nearly 400 (single and multiple) mutants of HIV-1 protease, isolated and sequenced from over 4000 patients
- Monogram Biosciences PhenoSense assay:
  - High: 152 distinct mutational patterns assayed for NFV
  - Low: 84 patterns assayed for ATV
- Mutant fold change = IC$_{50}$(mutant) / IC$_{50}$(wt)
- Subscripts in table = no. of assayed mutants; fold change value in table = median value
- Individual fold changes all show small abs. dev. from median, reflecting assay consistency
- Clinical cutoffs (based on latest data, studies still underway):
  - 2 classes: Sensitive ≤ 10, Resistant > 10
  - 3 classes: S ≤ 2.5, 2.5 < I ≤ 10, R > 10
- Each of the 7 inhibitors uses a distinct training set; separate models are trained and their performance is evaluated for each drug
- For each inhibitor, the learned models are used to predict the susceptibility of the unassayed mutational patterns for the given drug
ROC Curves Based on Two-Class Training Sets
Factors Contributing to Classification Capability

Factors

F1: values (magnitude and sign) of the non-zero components in each vector
F2: location of the non-zero components in each vector
F3: number of non-zero components in each vector

Controls

C1: multiply each non-zero vector component by a random number generated from the interval [-2, 2] (removes influence of F1, measures contributions of F2 and F3)
C2: randomly shuffle the components of each vector in C1 independently (removes influences of F1 and F2, measures contribution of F3)

Graphed ROC Example: RTV

- Apply Random Forest (RF) supervised classification
- Shuffled classes control: S, R class labels randomly shuffled among mutant vectors prior to RF learning

RF Results For All Inhibitors

<table>
<thead>
<tr>
<th>Inhibitor</th>
<th>Original Vectors</th>
<th>Control C1</th>
<th>Control C2</th>
<th>Shuffled Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFV</td>
<td>0.8371</td>
<td>0.7796</td>
<td>0.7404</td>
<td>0.4762</td>
</tr>
<tr>
<td>SQV</td>
<td>0.8740</td>
<td>0.7455</td>
<td>0.6629</td>
<td>0.4957</td>
</tr>
<tr>
<td>IDV</td>
<td>0.8220</td>
<td>0.7166</td>
<td>0.6552</td>
<td>0.3757</td>
</tr>
<tr>
<td>RTV</td>
<td>0.9154</td>
<td>0.8413</td>
<td>0.7278</td>
<td>0.5066</td>
</tr>
<tr>
<td>AFV</td>
<td>0.8156</td>
<td>0.6382</td>
<td>0.5349</td>
<td>0.4764</td>
</tr>
<tr>
<td>LPV</td>
<td>0.9214</td>
<td>0.8818</td>
<td>0.7163</td>
<td>0.4277</td>
</tr>
<tr>
<td>ATV</td>
<td>0.8129</td>
<td>0.7671</td>
<td>0.6826</td>
<td>0.5185</td>
</tr>
</tbody>
</table>
### RF AUCs Based on Three Susceptibility Classes

<table>
<thead>
<tr>
<th>Inhibitor</th>
<th>Ref. Class</th>
<th>No. of Mutants</th>
<th>Ref. Class AUC</th>
<th>Overall AUC (1-against-all)</th>
<th>Class Pairs</th>
<th>Class Pair AUC</th>
<th>Overall AUC (1-against-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFV</td>
<td>S</td>
<td>13</td>
<td>0.7939</td>
<td>0.8197</td>
<td>S-I</td>
<td>0.6923</td>
<td>0.7816</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>31</td>
<td>0.7457</td>
<td></td>
<td>I-R</td>
<td>0.8202</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>108</td>
<td>0.8441</td>
<td></td>
<td>R-S</td>
<td>0.8323</td>
<td></td>
</tr>
<tr>
<td>SQV</td>
<td>S</td>
<td>53</td>
<td>0.8892</td>
<td>0.8064</td>
<td>S-I</td>
<td>0.7142</td>
<td>0.7918</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>34</td>
<td>0.5938</td>
<td></td>
<td>I-R</td>
<td>0.6941</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>61</td>
<td>0.8530</td>
<td></td>
<td>R-S</td>
<td>0.9671</td>
<td></td>
</tr>
<tr>
<td>IDV</td>
<td>S</td>
<td>29</td>
<td>0.8864</td>
<td>0.7845</td>
<td>S-I</td>
<td>0.7791</td>
<td>0.8299</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>43</td>
<td>0.6729</td>
<td></td>
<td>I-R</td>
<td>0.7463</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>76</td>
<td>0.8088</td>
<td></td>
<td>R-S</td>
<td>0.9644</td>
<td></td>
</tr>
<tr>
<td>RTV</td>
<td>S</td>
<td>23</td>
<td>0.8909</td>
<td>0.8741</td>
<td>S-I</td>
<td>0.8312</td>
<td>0.8596</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>30</td>
<td>0.7549</td>
<td></td>
<td>I-R</td>
<td>0.8182</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>89</td>
<td>0.9100</td>
<td></td>
<td>R-S</td>
<td>0.9294</td>
<td></td>
</tr>
<tr>
<td>APV</td>
<td>S</td>
<td>54</td>
<td>0.7939</td>
<td>0.7562</td>
<td>S-I</td>
<td>0.7002</td>
<td>0.7959</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>56</td>
<td>0.6640</td>
<td></td>
<td>I-R</td>
<td>0.7773</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>35</td>
<td>0.8455</td>
<td></td>
<td>R-S</td>
<td>0.9101</td>
<td></td>
</tr>
<tr>
<td>LPV</td>
<td>S</td>
<td>22</td>
<td>0.8495</td>
<td>0.8883</td>
<td>S-I</td>
<td>0.7762</td>
<td>0.8711</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>26</td>
<td>0.7744</td>
<td></td>
<td>I-R</td>
<td>0.8684</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>64</td>
<td>0.9479</td>
<td></td>
<td>R-S</td>
<td>0.9688</td>
<td></td>
</tr>
<tr>
<td>ATV</td>
<td>S</td>
<td>21</td>
<td>0.7192</td>
<td>0.7306</td>
<td>S-I</td>
<td>0.5263</td>
<td>0.6918</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>29</td>
<td>0.6207</td>
<td></td>
<td>I-R</td>
<td>0.7277</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>34</td>
<td>0.8315</td>
<td></td>
<td>R-S</td>
<td>0.8214</td>
<td></td>
</tr>
</tbody>
</table>
Graphed RF ROC Example: RTV

1-against-all

- Sensitive AUC = 0.8909
- Intermediate AUC = 0.7549
- Resistant AUC = 0.9100

Overall AUC = 0.8741

1-against-1

- S-I AUC = 0.8312
- I-R AUC = 0.8182
- R-S AUC = 0.9294

Overall AUC = 0.8596
Publications


Selected Conference Presentations

Acknowledgements

Committee
Dr. Vaisman – Ph.D. Director
Dr. Grefenstette
Dr. Jamison
Dr. Royt

Structural Bioinformatics Group
Vadim Ravich
Ewy Mathe
Todd Taylor
Andrew Carr
Tariq Alsheddi
Greg Reck

Summer ’05 Interns
Dr. Saleet Jafri – Organizer
Kahkeshan Hijazi, Nida Parvez

Support Staff
Glenda Wilson, Chris Ryan,
Susan Beale

Software
Qhull – tessellation (Barber)
Glisten – tessellation visualization (Carr)
Chimera – ribbon diagrams (Ferrin)
Base Java programs – to generate raw
data based on tessellation (Lu)
Weka – machine learning (Witten, Frank)
Mutant Activity Class Distribution in $\mathbb{R}^{99}$

- Given $d(A,B) =$ mean Euclidean distance between all possible pairs of mutants (one from class A, and the other from class B), $d(\text{pos,pos}) < d(\text{pos,inter}) < d(\text{inter,inter}) < d(\text{pos,neg}) < d(\text{inter,neg}) < d(\text{neg,neg})$
- Order agrees with biological notions on impact of mutations
- Mutant pairs for which at least one of the mutants represents a NC substitution drive the order of the mean distances
Clustering Example: HIV-1 Protease

- Ron Shamir’s Expander software:
  http://www.cs.tau.ac.il/~rshamir/expander/expander.html
- Similar to k-means, but no a priori value of k needed; algorithm derives optimal number of clusters
- Leaves open the question of how well the residual profiles can be used to classify mutants with differing levels of activity
Test Options

• Use (partitioned) training set only—for assessing performance
  ➢ Tenfold cross-validation (10 CV): Stratified partitioning of the instances into ten equally sized subsets
    1. One subset is held out, while the other nine subsets (90% of the original instances) are combined to form a modified training set
    2. The supervised classification algorithm is used to learn a model with the modified training set; the learned model is used to predict the activity classes of the instances in the hold-out subset (the test set)
    3. The process is repeated ten times, whereby each subset serves once as a hold-out for prediction; hence, a single activity prediction is made for each instance
  ➢ Leave-one-out (or N CV, where N = size of full training set): Each subset consists of one instance; no stratification by definition; deterministic
  ➢ % split: Stratified partitioning of the instances into two (not necessarily equal) subsets; larger subset serves as a training set for model building, and smaller subset is a test set

• Use the full training set (for model building) and an independent test set (for example, to predict the activity classes of mutants that have not been studied experimentally, if performance as described above is acceptable)
Evaluation of Model Performance (Two Classes: P, N)

- Confusion matrix: tabulated number of test predictions (shown)
- Sensitivity = TP / (TP + FN), Specificity = TN / (TN + FP), and 1-Specificity = FP / (FP + TN)
- Sensitivity = True Positive Rate (TPR)
  1-Specificity = False Positive Rate (FPR)

- Accuracy = (TP + TN) / (TP + FP + TN + FN); simple measure, but highly sensitive to class skew in test sets
- Default costs assigned prior to model building are 0 (TP, TN) and 1 (FP, FN); ↑ FP cost only → ↓ no. of FP’s → ↑ specificity; ↑ FN cost only → ↓ no. of FN’s → ↑ sensitivity
- ROC (Receiver-Operating Characteristic) Curve: Plot of TPR vs. FPR in unit square using 10 CV for a range of FN/FP cost ratios
- Area under ROC curve (AUC): performance measure that is insensitive to unequal class distributions in test sets
  - Perfect classifier: Piecewise linear ROC joining (0,0) to (0,1) and (0,1) to (1,1); AUC = 1.0
  - Random guessing model: Diagonal line ROC joining (0,0) to (1,1); AUC = 0.5
Application to Multiple (n > 2) Classes

• One-against-all approach (use all training set instances)
  1. Choose one class as a reference (class 1); combine all other classes together by re-labeling as non-reference (class 2)
  2. Apply ROC analysis to this two-class system
  3. Repeat n times so that each class serves as a reference once
  4. Overall AUC for the multi-class system is a weighted average of the two-class AUCs (each two-class AUC weight is the proportion of mutants belonging to the respective reference class in the training set); this method is sensitive to class skew in theory, but performs well in practice

• One-against-one approach (truncate the original training set)
  1. Choose one pair of classes; form a truncated training set consisting of only instances that belong to either of these two classes
  2. Apply ROC analysis to this two-class system
  3. Repeat n(n-1)/2 times, so that every pair of classes is considered
  4. Overall AUC for the multi-class system is a simple average of the two-class AUCs; this method remains insensitive to class skew in test sets
Factors Contributing to Classification Capability

Factors
F1: no. of non-zero components in each vector
F2: value (magnitude and sign) of the non-zero components in each vector
F3: location of the non-zero components in each vector
F4: no. of non-zero columns in each group of vectors (submatrix of the training set) representing all mutants generated by amino acid substitutions at the same position
F5: location of the non-zero columns in each group

Controls
C1: multiply each non-zero vector component by a different random no. generated from the interval [-2, 2] (removes influence of F2, measures contributions of F1 and F3)
C2: randomly shuffle the components of each vector independently (removes influence of F3, measures contributions of F1 and F2)
C3: composite of C1 followed by C2 (removes influences of F2 and F3, measures contribution of F1)
C4: randomly shuffle the columns within each group independently (removes influence of F5, measures contributions of F2 and F4)
C5: composite of C1 followed by C4 (removes influences of F2 and F5, measures contribution of F4)

Ten independent versions of each control training set were prepared, and two-class decision tree learning (default costs) was applied

<table>
<thead>
<tr>
<th>testing method</th>
<th>original vectors</th>
<th>C1 (mean)</th>
<th>C2 (mean)</th>
<th>C3 (mean)</th>
<th>C4 (mean)</th>
<th>C5 (mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 CV (10 runs)</td>
<td>76.8</td>
<td>72.6 (0.89)</td>
<td>54.6 (1.36)</td>
<td>53.5 (1.14)</td>
<td>73.0 (1.09)</td>
<td>65.8 (1.69)</td>
</tr>
<tr>
<td>60/40 split (100 runs)</td>
<td>75.3</td>
<td>71.0 (0.67)</td>
<td>54.8 (0.55)</td>
<td>53.4 (0.86)</td>
<td>70.3 (0.78)</td>
<td>63.6 (1.35)</td>
</tr>
<tr>
<td>536 CV</td>
<td>75.4</td>
<td>73.8 (3.24)</td>
<td>53.9 (3.74)</td>
<td>51.8 (3.86)</td>
<td>72.4 (2.12)</td>
<td>67.0 (2.32)</td>
</tr>
</tbody>
</table>
Decision Tree

• Default cost model learned from the training set of 536 experimental HIV-1 protease mutants (active/inactive)
Alternative Testing Approaches and Learning Curves

- Apply RF supervised learning to the 142 RTV mutants
- 100 stratified 66/34 random splits: accuracy = 83.2%, std. dev. = 4.7%
- 100 iterations of 10 CV: accuracy over 1000 folds = 84.3%, std. dev. = 9.5%
- Leave-one-out (142 CV): accuracy = 85.9%

- Learning Curves using the 142 RTV mutants and DT, SVM, and RF supervised learning
- Stratified training sets randomly chosen with replacement in increments of 20 mutants
- Mean 10 CV accuracy based on average of 10 runs
- Error bars = ±1 std. dev.