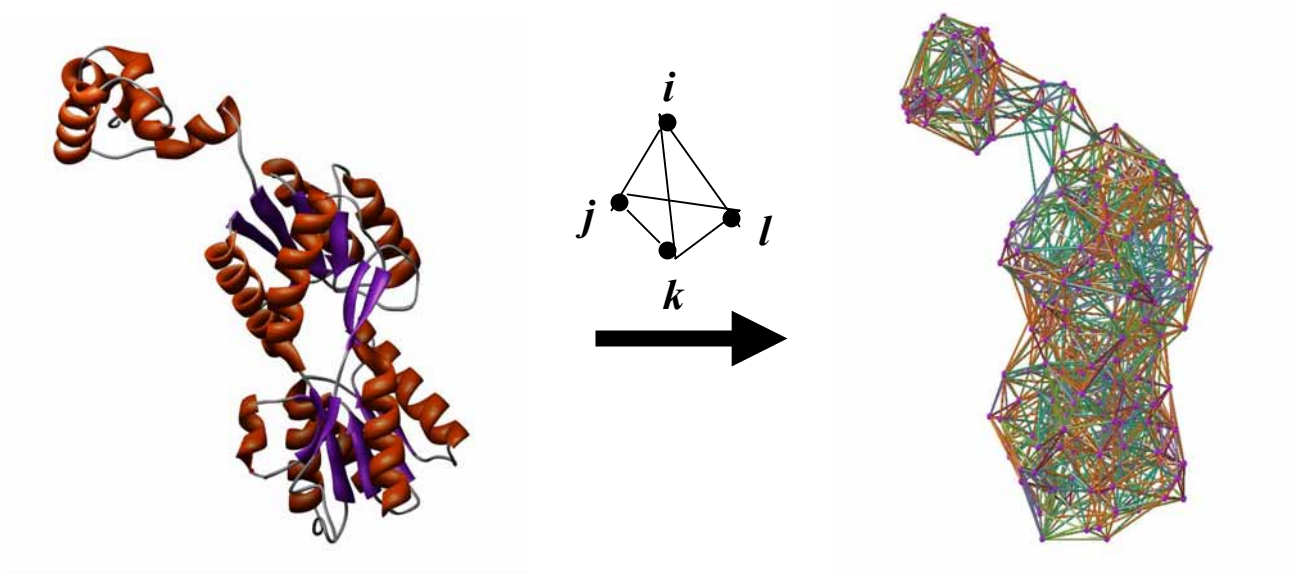
Functional Analysis of the *Escherichia coli Lac* Repressor: A Computational Mutagenesis Approach



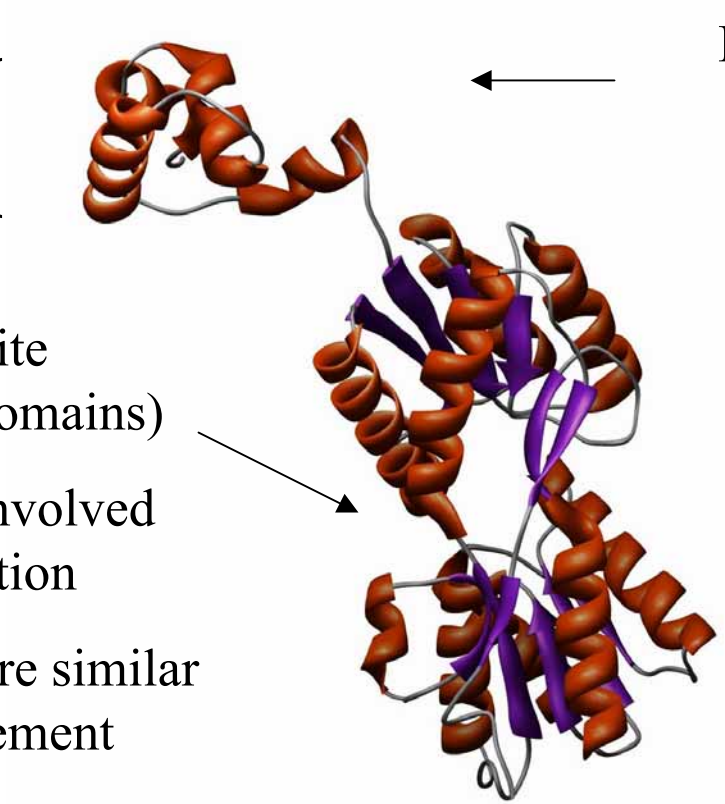
Majid Masso, Ph.D.

Laboratory for Structural Bioinformatics

George Mason University

http://binf.gmu.edu/mmasso mmasso@gmu.edu

Lac Repressor: Structure and Function



Headpiece
(residues 1-59)

DNA binding domain
HTH motif

N-terminal
core domain
(residues 61-160
and 293-320)

Sugar binding site
(cleft between core domains)

Both core domains involved
in homodimerization

Both core domains share similar
2° structure arrangement

C-terminal
core domain
(residues 165-290)

Not shown: Leucine zipper tetramerization domain
(residues 330-360) beyond core domain

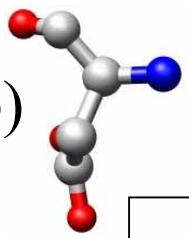
Lac Repressor Experimental Mutagenesis Data

- UCLA researchers (Jeffery H. Miller lab) introduced the same 13 amino acid substitutions at positions 2-329
- 4041 non-degenerate single point mutants

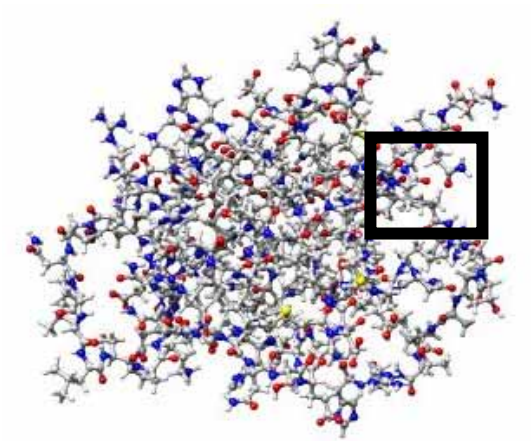
  223 self-substitutions (control)
- Full activity (> 200-fold repression of β-galactosidase); moderate (20 to 200-fold); low (4 to 20-fold); and inactive (less than 4-fold)
- 2267 full activity mutants; 253 moderate; 355 low; 1166 inactive
- Researchers suggest combining moderate and low

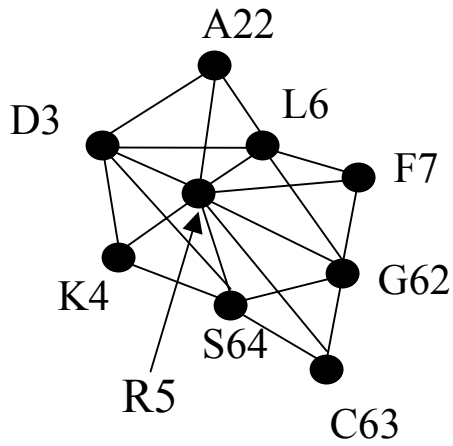
  (i.e., 608 intermediate)

Delaunay Tessellation of Protein Structure

D (Asp)  → ● C_α coordinates in 3D

Abstract each amino acid to a point
Atomic coordinates – Protein Data Bank (PDB)





Delaunay tessellation: 3D "tiling" of space into non-overlapping, irregular tetrahedral simplices. Each simplex objectively defines a quadruplet of nearest-neighbor amino acids at its vertices.

Counting Amino Acid Quadruplets

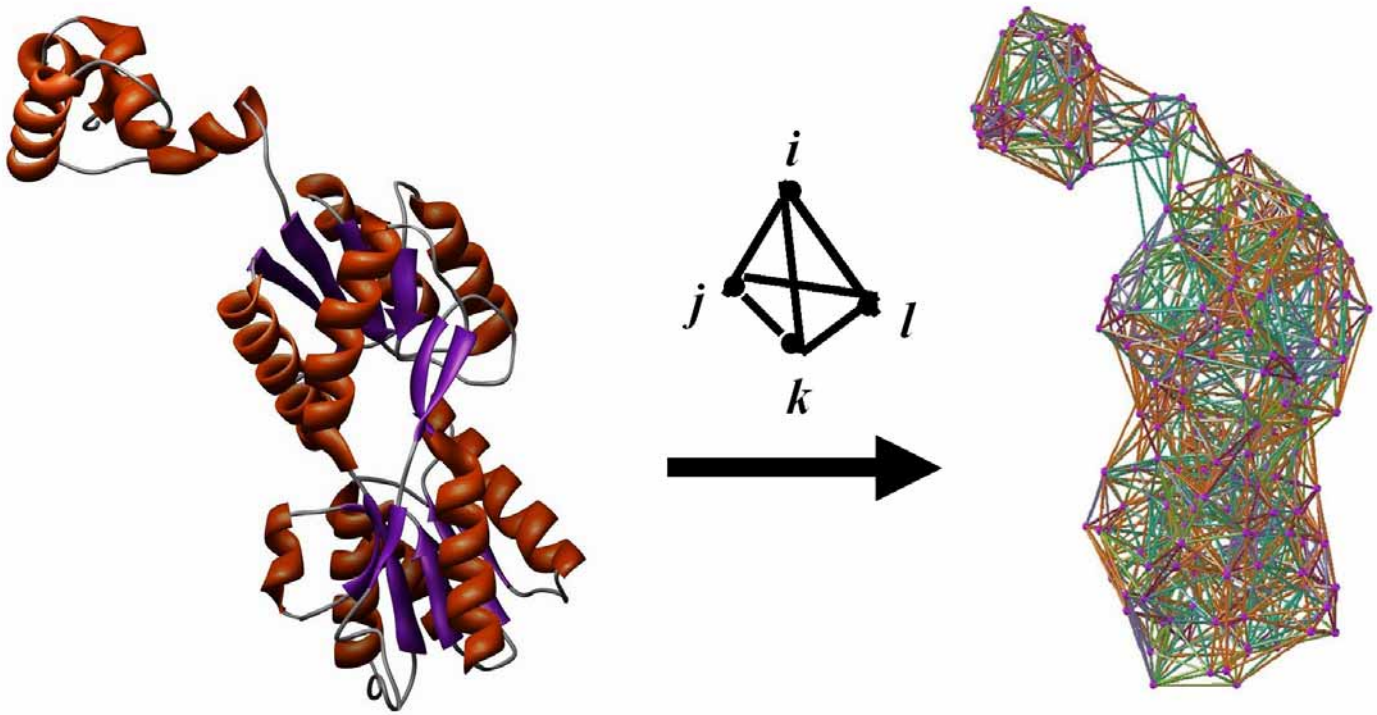
Ordered quadruplets: $20^4 = 160{,}000$ (too many)

Order-independent quadruplets (our approach):

C D E F	$\binom{20}{4}$
C C D E	$20 \cdot \binom{19}{2}$
C C D D	$\binom{20}{2}$
C C C D	$20 \cdot 19$
C C C C	20

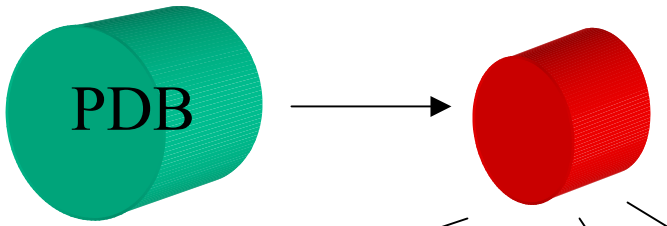
Total: 8,855 distinct unordered quadruplets

Delaunay Tessellation: *E. coli Lac* Repressor



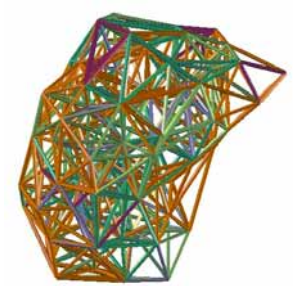
- Ribbon diagram (left) is based on structural coordinates located in the PDB accession file 1efa, chain B (residue positions 2 – 331)
- Each of the 330 amino acid residues is represented as a point in 3D, using the C_α coordinates
- Tessellation (right) is performed by using a 12Å edge-length cutoff on the allowed simplices (“true” quadruplet interactions)

Four-Body Statistical Potential

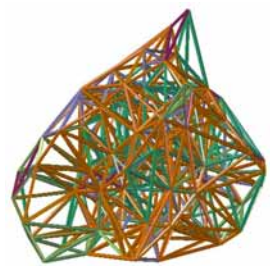


Training set: over 1,400 diverse high-resolution x-ray structures

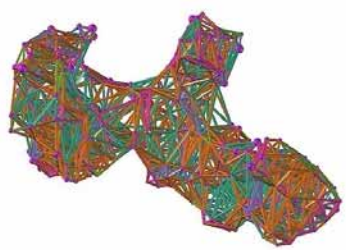
Tessellate



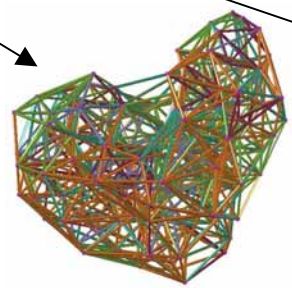
1bniA
barnase



1jli
IL-3



1rtjA
HIV-1 RT



3lzm
t4 lysozyme

Pool together all simplices from the tessellations, and compute observed frequencies of simplicial quadruplets

Four-Body Statistical Potential

- Knowledge-based, modeled after inverse Boltzmann law:

  p_i = Frequency (feature i) α e $^{-\text{ Energy (feature } i) / KT}$, i.e., $E_i \alpha -KT \ln p_i$; and Potential (feature i) = $E_i - E_{ref} = \Delta E_i = -KT \ln(p_i / p_{ref})$

- For amino acid quadruplet (i,j,k,l), a log-likelihood score (interaction "pseudo-energy") is given by $s(i,j,k,l) = \log(f_{ijkl} / p_{ijkl})$

- f_{ijkl} = observed proportion of training set simplices whose four vertex residues are i,j,k,l

- p_{ijkl} = rate expected by chance (multinomial distribution, based on training set proportions of residues i,j,k,l)

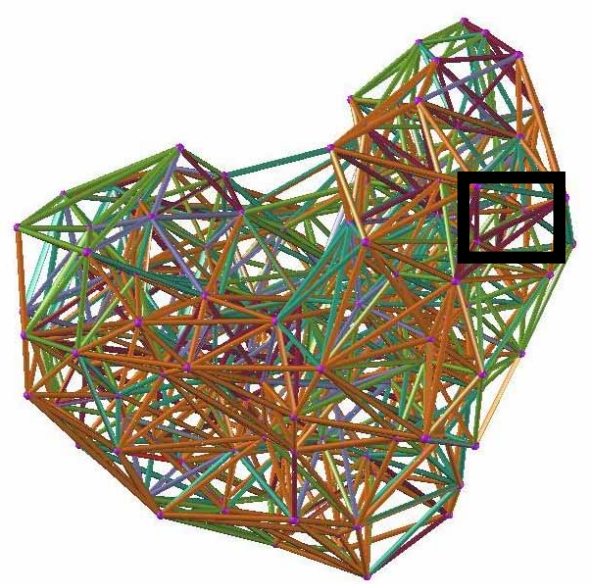
- Four-body statistical potential: the collection of 8855 quadruplet (or simplex) types and their respective log-likelihood scores

Four-Body Statistical Potential

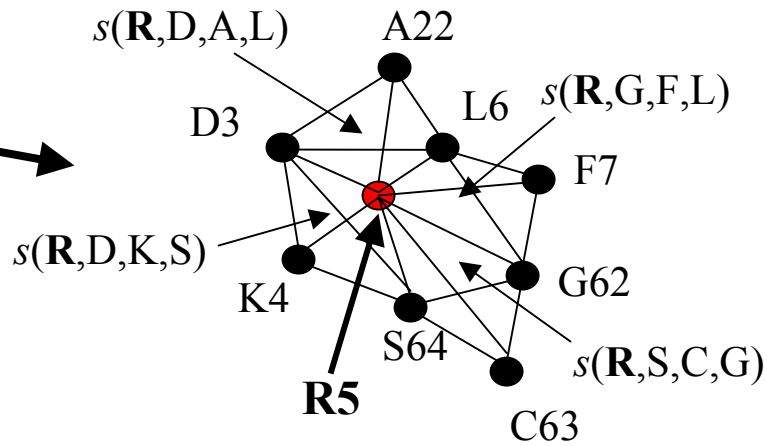
Amino Acid Quadruplet	"Pseudo-Energy" Log-likelihood s(i,j,k,l)
CCCC	3.29042538
CCCH	2.09542785
CCCS	1.96177162
CCCG	1.84022021
CCCI	1.79961166
CCCF	1.77139046
CCCT	1.76378293
CCCP	1.74840641
ACCC	1.74777711
CCCW	1.74711265
CCHH	1.70747111
CCCN	1.69741431
HHHH	1.61473339
.	.
.	.
.	.
HMNP	0.000221495
DGGY	0.000178988
DRSV	9.45855E-05
EHHV	4.979E-06
LRYY	-6.29797E-05
DGKP	-9.73563E-05
NPSS	-0.000100914
IPRW	-0.000136526
MMRT	-0.000168007
GLLP	-0.000294376
EKNT	-0.000312593
EKQR	-0.000343148
.	.
.	.
.	.
HKKW	-0.66398714
KKKP	-0.66875323
CDEQ	-0.67215257
CKKW	-0.75315166
CDDM	-0.76390474
HHKK	-0.85974
CKKR	-0.88002907
CIKR	-0.90372634
CHKW	-0.94458122
CEEE	-1.02439761
HKKM	-1.14234339

Application 1: Protein Topological Score (TS)

- Obtained by summing the log-likelihood scores of **all** simplicial quadruplets defined by the protein tessellation
- Global measure of protein sequence-structure compatibility
- Total (empirical or statistical) potential of the protein



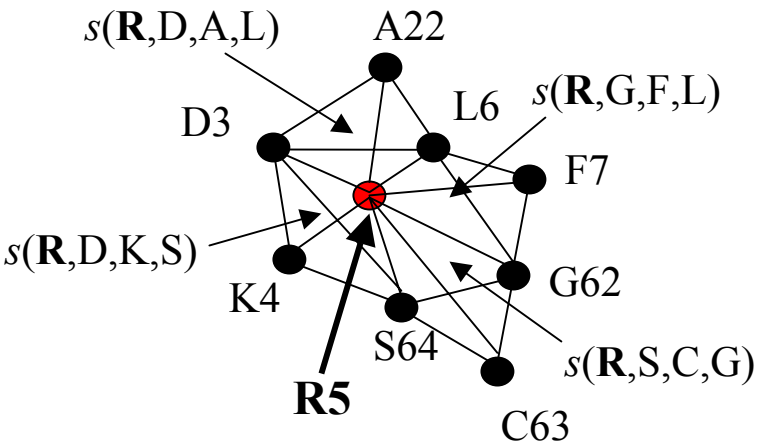
TS = $\sum_{\hat{\mathbf{i}}} s(\hat{\mathbf{i}})$, sum taken over **all** simplex quadruplets $\hat{\mathbf{i}}$ in the entire tessellation.



Close-up view of **only** the four simplices that use **R** at position **5** as a vertex

Application 2: Residue Environment Scores

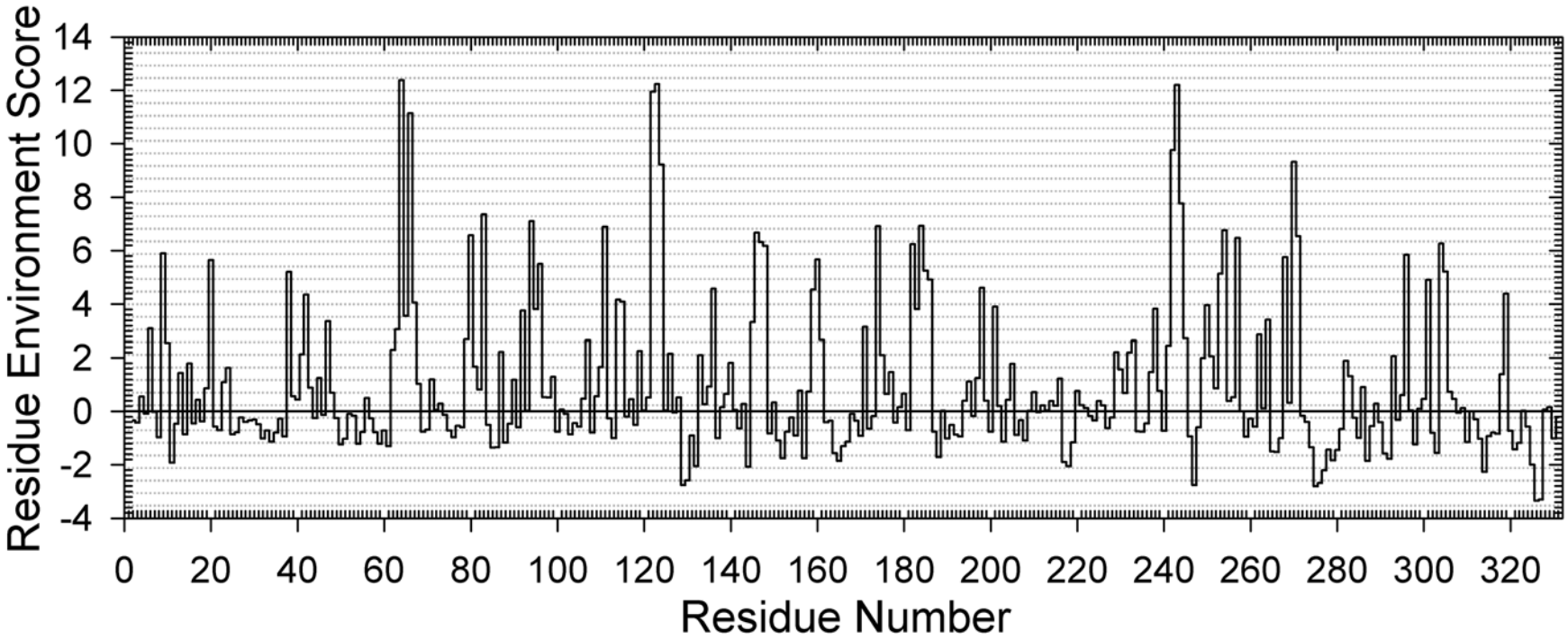
- For each amino acid position, locally sum log-likelihood scores $s(i,j,k,l)$ of only simplices that use the position as a vertex



Example: $q_5 = q(\mathrm{R5}) = \sum_{(i,j,k,l)} s(i,j,k,l)$, sum is taken **only** over all simplex quadruplets (i,j,k,l) that use R5

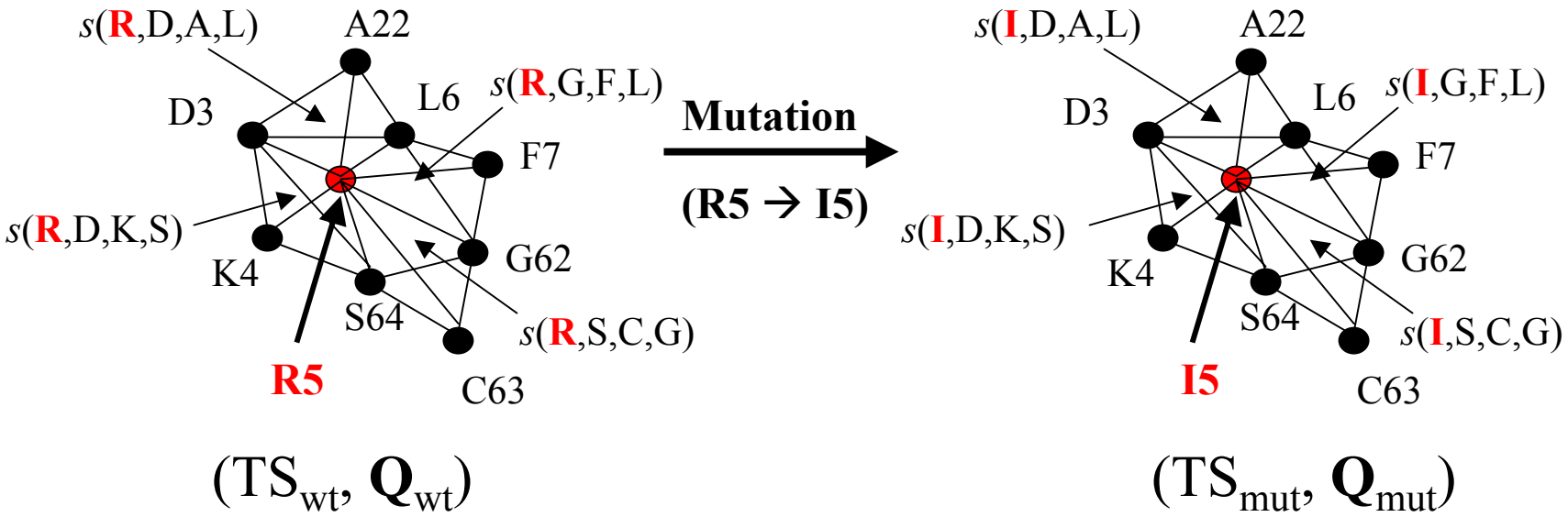
- The scores of all the amino acid positions in the protein structure form a **Potential Profile** vector $\mathbf{Q} = < q_1, q_2, q_3, \ldots, q_N >$
  (N = length of primary sequence in the solved structure)

Potential Profile: *E. coli Lac* Repressor



Computational Mutagenesis Methodology

- Observations:
  - Few solved mutant structures to compare with solved wild type (wt) structure
  - Mutant and wt protein structure tessellations are very similar or identical
- Approach:
  - Obtain topological score (TS_{mut}) and potential profile vector ($\mathbf{Q}_{mut}$) for any mutant protein by using the wt structure tessellation as a template
  - Simply change the residue label at a given point and re-compute



(TS_{wt}, $\mathbf{Q}_{wt}$)

(TS_{mut}, $\mathbf{Q}_{mut}$)

Computational Mutagenesis Methodology

- **Scalar "Residual Score"** of a mutant:

  (mutant – wt) topological score difference = $TS_{mut} - TS_{wt}$
  (empirical measure of relative structural change due to mutation)

- **Vector "Residual Profile"** of a mutant:

  $\mathbf{R} = \mathbf{Q}_{mut} - \mathbf{Q}_{wt}$ = (mutant – wt) potential profile vector difference
  (environmental perturbation score for every position in structure)

- Denote $\mathbf{R} = < EC_1, EC_2, EC_3, \ldots, EC_N >$

  $EC_i = q_{i,mut} - q_{i,wt}$ = relative environmental change at position i

- Geometric property: mutation at position i => EC_i = residual score

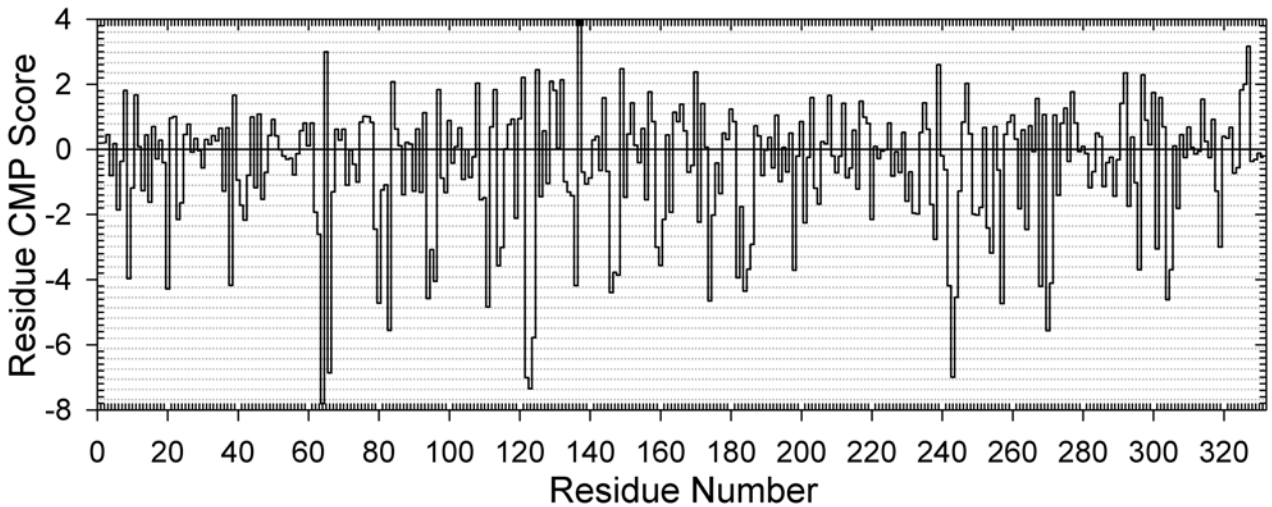
Comprehensive Mutational Profile (CMP)

- At each position, the **CMP score** is the mean of the residual scores associated with all possible amino acid substitutions

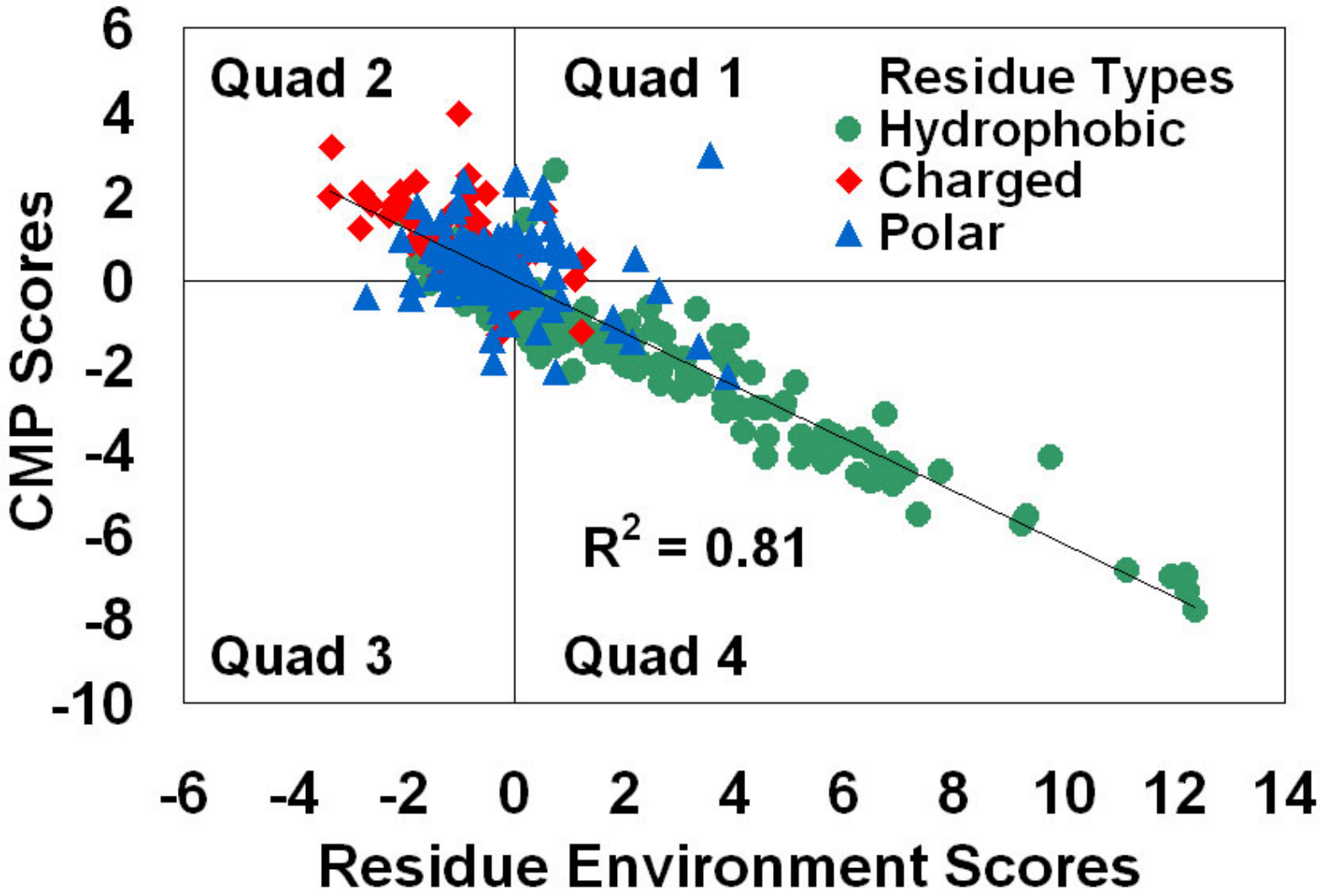
- Computationally,

$$\begin{aligned} \mathrm{CMP}_j &= \tfrac{1}{20}\sum_{i=1}^{20} \left[(\text{mutant topological score})_{ij} - (\text{wt topological score})\right] \\ &= \tfrac{1}{20}\sum_{i=1}^{20} (\text{mutant residual score})_{ij} \\ &= \{\text{mean residual score}\}_j \end{aligned}$$

where index *i* refers to the 20 amino acids, and index *j* refers to the primary sequence position

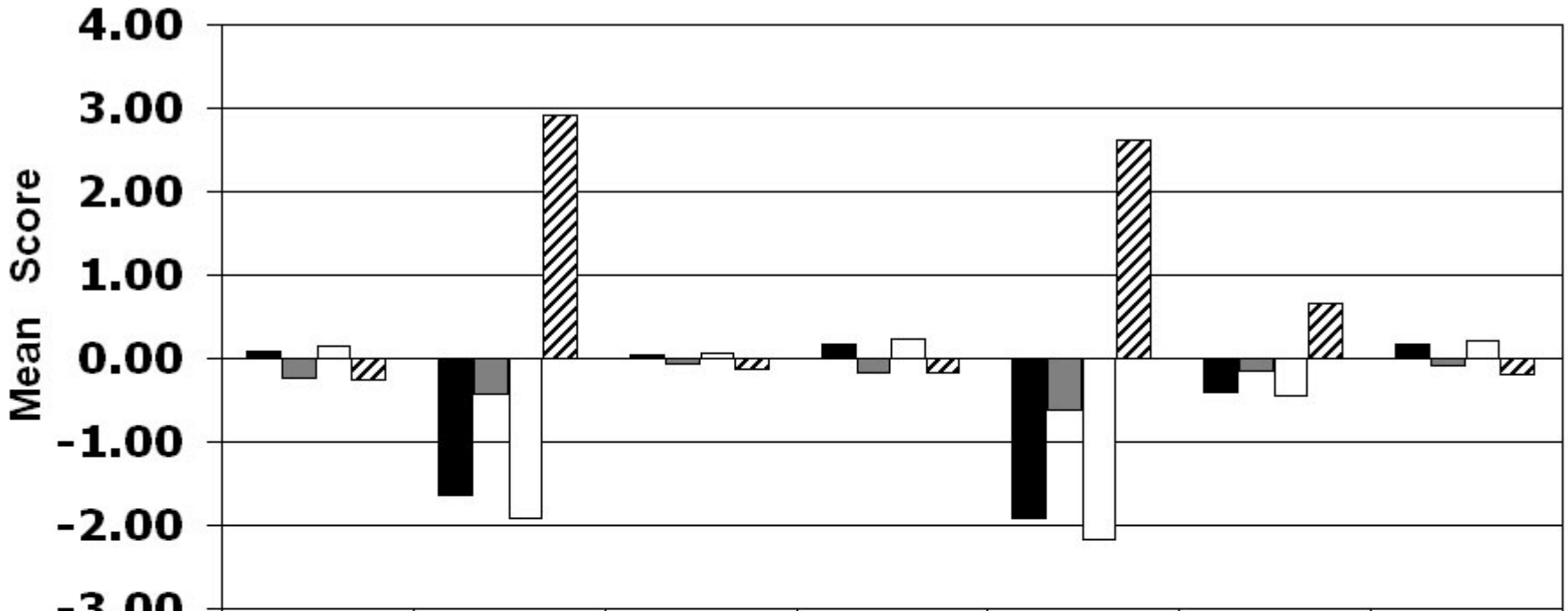


CMP – Potential Profile Correlation



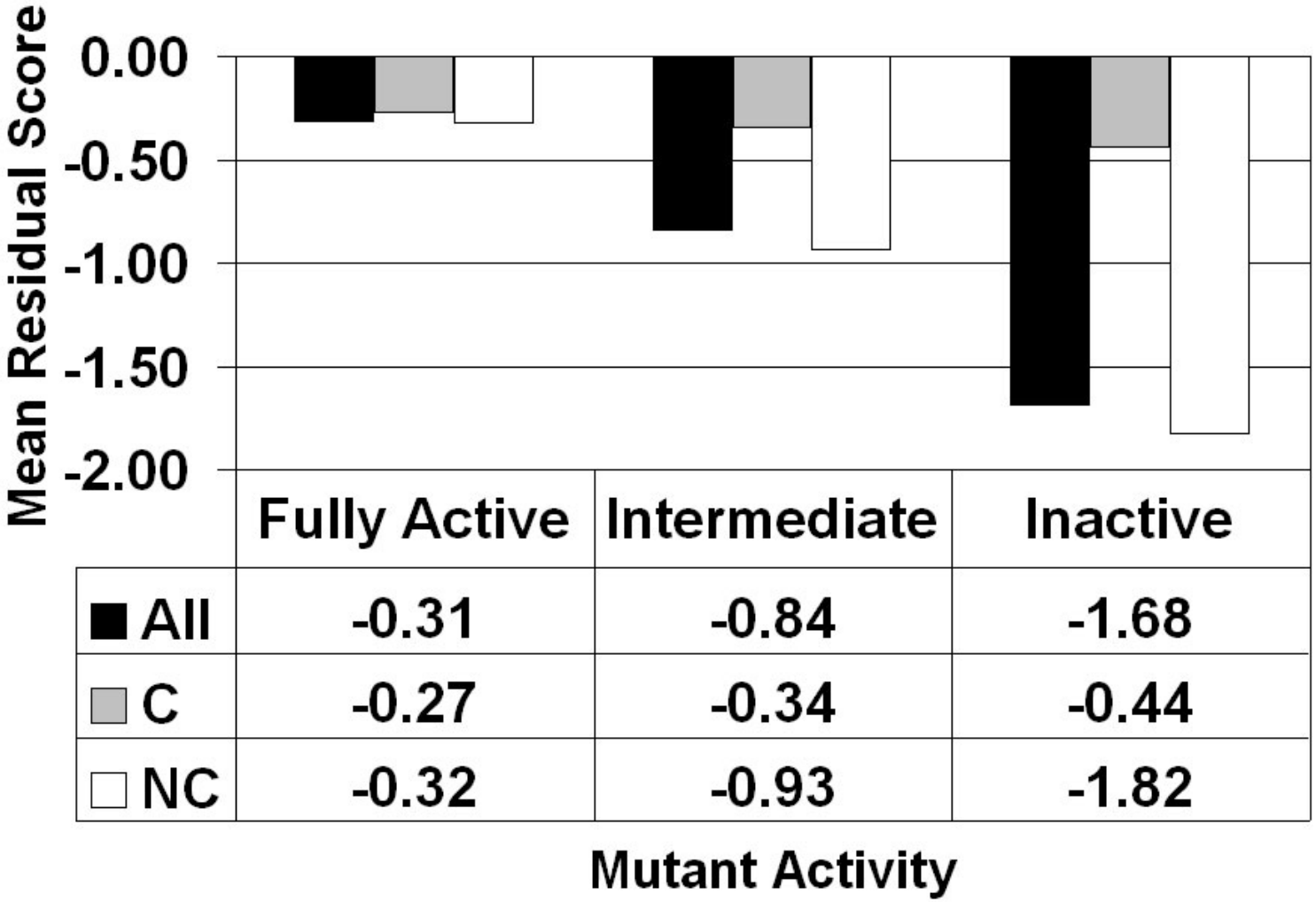
Distribution of *lac* repressor residue positions

Graph Quadrants	Residue Groups							
	Surface	Buried	DNA binding	IPTG binding	Stability	Interface	Spacers	Total
Q1	8	10	0	2	1	6	4	31
Q2	49	12	9	9	8	15	20	122
Q3	13	5	4	2	2	5	6	37
Q4	31	46	5	4	25	17	10	138
Total	101	73	18	17	36	43	40	328



	Surface	Buried	DNA binding	IPTG binding	Stability	Interface	Spacers
All	0.08	-1.65	0.04	0.17	-1.91	-0.39	0.16
C	-0.23	-0.42	-0.07	-0.17	-0.62	-0.15	-0.08
NC	0.14	-1.91	0.07	0.24	-2.17	-0.44	0.21
M.R.E.S.	-0.26	2.91	-0.12	-0.17	2.61	0.65	-0.20

Experimental Mutants: Residual Scores Elucidate the Structure-Function Relationship



	Fully Active	Intermediate	Inactive
All	-0.31	-0.84	-1.68
C	-0.27	-0.34	-0.44
NC	-0.32	-0.93	-1.82

Mutant Residual Profiles as Feature Vectors for Decision Tree Classification and Prediction

- Training set: 4041 experimental mutants with known activity (fully active = "unaffected"; intermediate / inactive = "affected")
- Each feature vector includes three additional components: native residue, position number, and replacement residue
- Evaluating model performance: Tenfold cross-validation (10 CV), and random split (N% used for training, (100 – N)% are predicted)
- Performance measures:

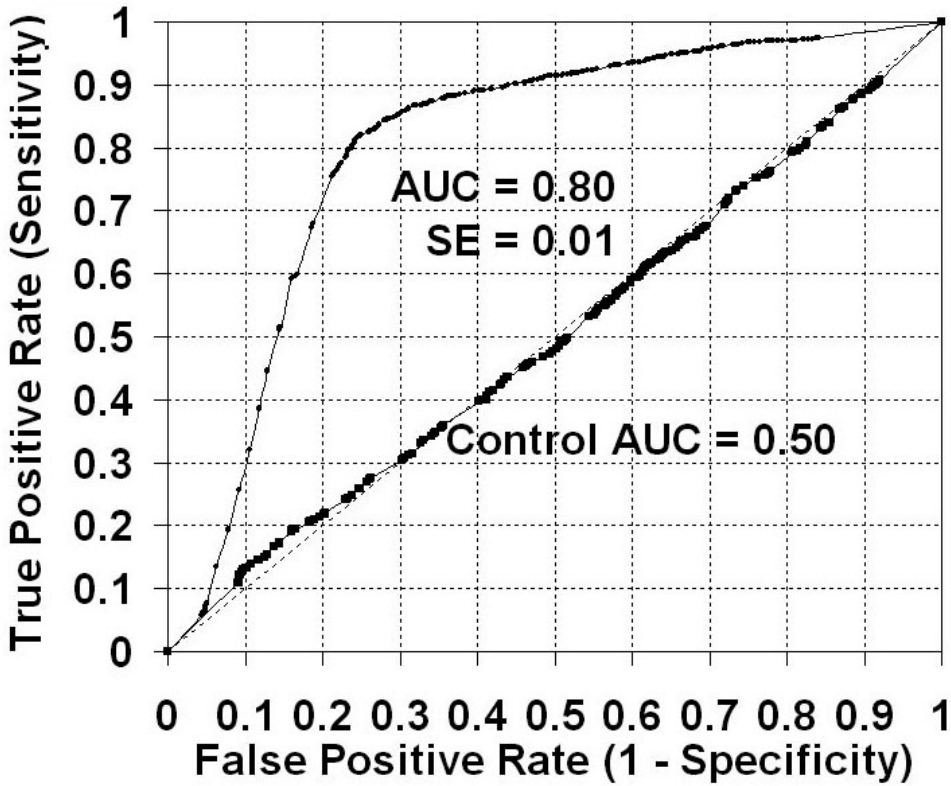
  $Q = (TP + TN) / (TP + FP + TN + FN)$

  $BER = 0.5 \times [FN / (FN + TP) + FP / (FP + TN)]$

  $MCC = (TP \times TN – FP \times FN) / [(TP+FN)(TP+FP)(TN+FN)(TN+FP)]^{½}$

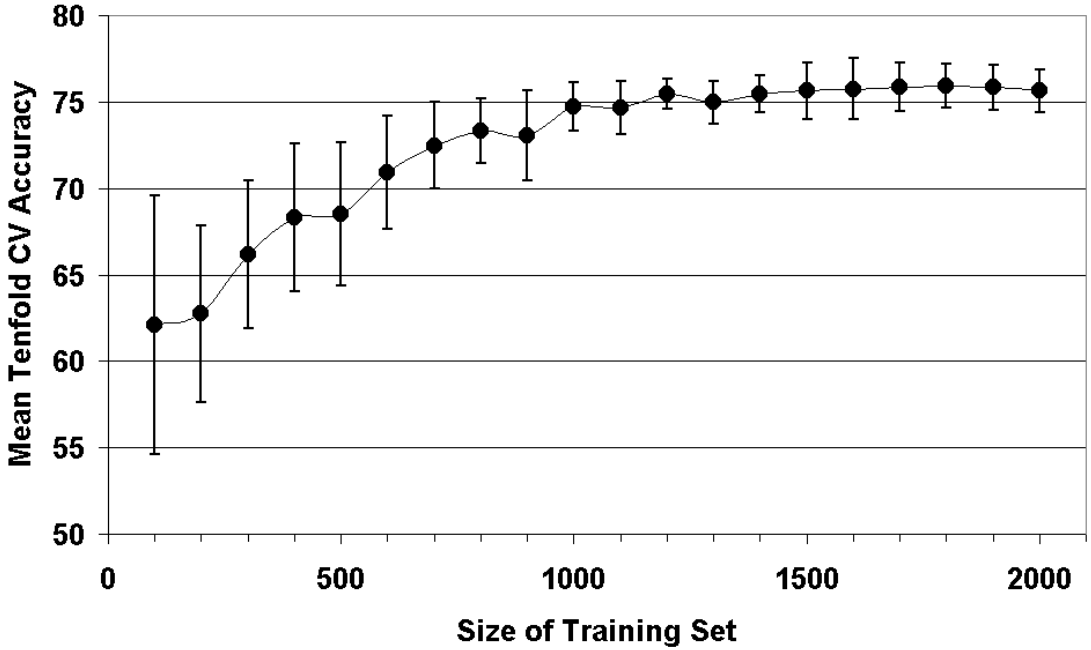
  AUC = Area under ROC (plot of *sensitivity* vs. 1 – *specificity*)

Tenfold Cross-Validation Results



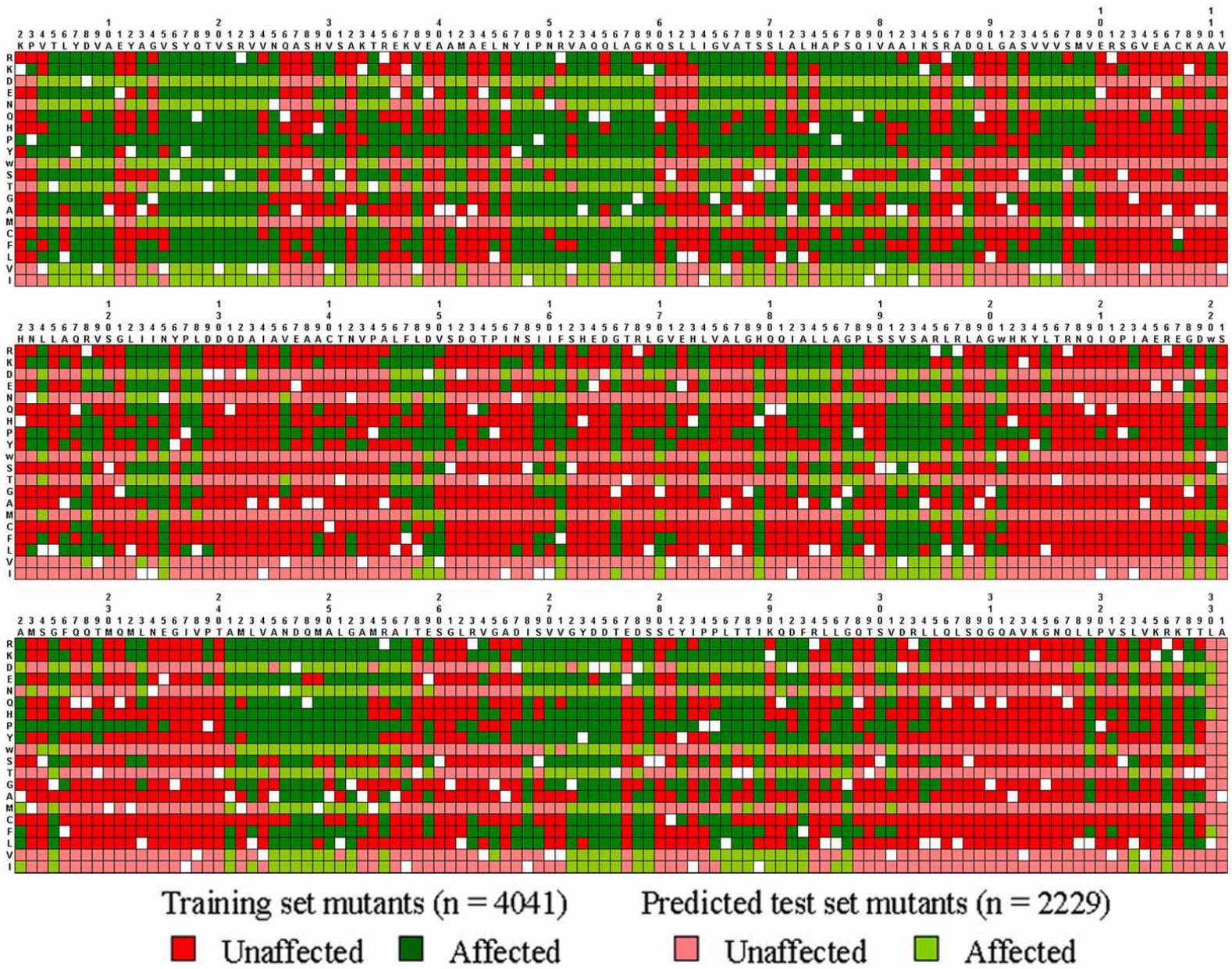
- 10 CV results: Q = 78.7%, BER = 0.22, MCC = 0.57, AUC = 0.80
- "Shuffled classes" random control results: Q = 51.1%, BER = 0.51, MCC = – 0.01, AUC = 0.50

Learning Curve



- Curve suggests that ~1200 mutant training set is optimal
- Hence, 30% of the 4041 mutants randomly selected for training
- Trained model used for predicting classes of remaining mutants
- Test set: 1316/1586 unaffected and 873/1243 affected correctly predicted, with Q = 77.4%, BER = 0.23, MCC = 0.54, AUC = 0.78

Lac Repressor Mutational Array



Conclusions and Future Directions

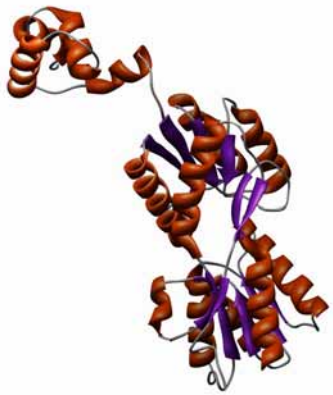
- Computational mutagenesis was developed through application of a four-body, knowledge-based, statistical contact potential
  - Residual scores of mutants with experimentally classified activity change elucidate the structure-function relationship
  - Mutant residual profiles serve as feature vectors for machine learning
- Future Aim: Develop a "universal" classification model to predict activity change of a residue replacement in any protein
  - Need common attribute set as feature vector components for all mutants
  - Instead of entire residual profile, use only EC scores at mutated position (i.e., residual score) as well as ordered EC scores at six nearest positions
  - Include additional information-rich common attributes
  - Already implemented for predicting stability change in mutants (see http://proteins.gmu.edu/automute)
  - Several candidate activity change mutant protein systems for training: *lac* repressor (4041), t4 lysozyme (2015), HIV-1 PR (536), IL-3 (629), …

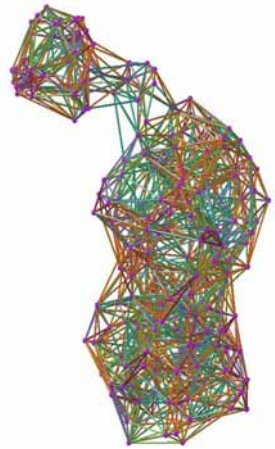
Acknowledgements and References

People

Iosif Vaisman (PI)

Kahkeshan Hijazi Nida Parvez





Software

AUTO-MUTE – server (Masso)
Qhull – tessellation (Barber)
Glisten – tessellation visualization (Carr)
Chimera – ribbon diagrams (Ferrin)
Ad hoc Java programs – potential (Taylor), residual profiles (Lu)
Weka – machine learning (Witten, Frank)

Publications

1. Masso M. and Vaisman I.I. Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach. *Biochem Biophys Res Comm* 2003; **305**: 322-326.
2. Masso M., Lu Z., and Vaisman I.I. Computational studies of protein structure-function correlations. *Proteins* 2006; **64**: 234-245.
3. Masso M. and Vaisman I.I. A novel sequence-structure approach for accurate prediction of resistance to HIV-1 protease inhibitors. *Proc IEEE BIBE* 2007; 952-958.
4. Masso M. and Vaisman I.I. Accurate prediction of enzyme mutant activity based on a multibody statistical potential. *Bioinformatics* 2007; **23**: 3155-3161.
5. Barenboim M., Masso M., Vaisman I.I., and Jamison D.C. Statistical geometry based prediction of non-synonymous SNP functional effects using random forest and neuro-fuzzy classifiers. *Proteins* 2008 (**in press**).