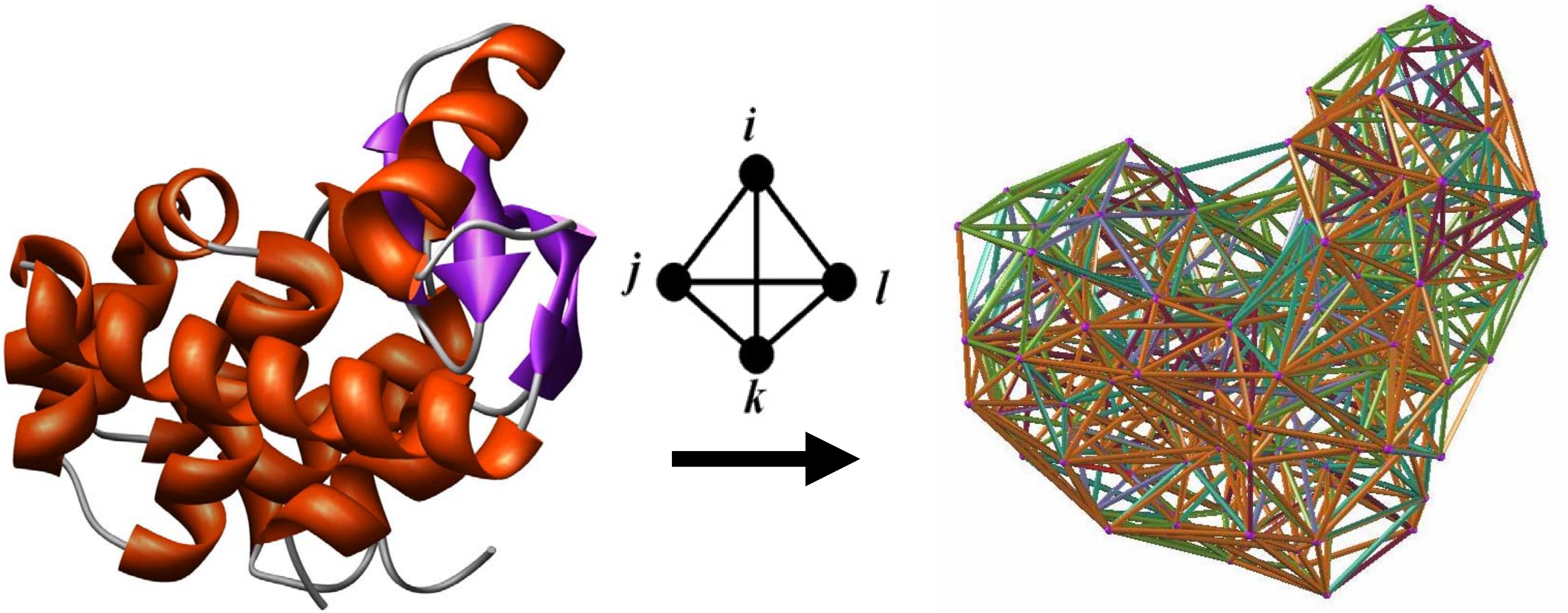
Structure-Based Prediction of Protein Activity Changes: Assessing the Impact of Single Residue Replacements



Majid Masso (mmasso@gmu.edu)

Laboratory for Structural Bioinformatics, School of Systems Biology

George Mason University, Manassas, Virginia

Slides available for download at http://binf.gmu.edu/mmasso

What Constitutes a Consequence of Single Residue Replacements in Proteins?

- Relative activity change; relative stability change; relative change in inhibitor binding energy to a target protein; neutrality versus disease association of protein mutations; etc. …

- No universally applicable formulas for inferring one mutant property based on knowledge of any other property
  - Example: We previously developed models for predicting stability change upon mutation, but these cannot be used to infer activity change

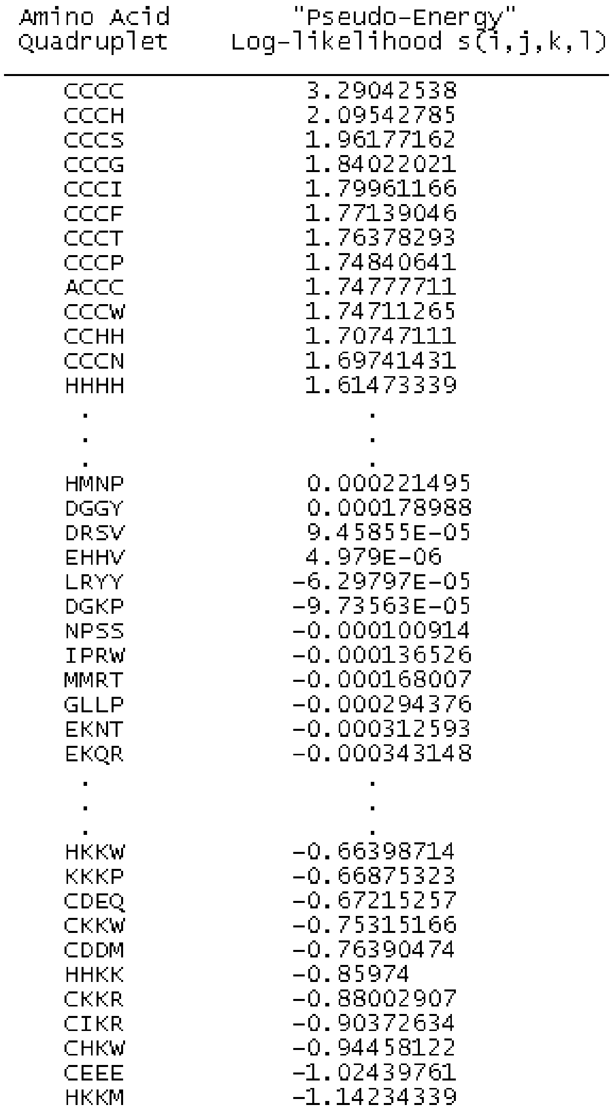
- Here we report on the development of a model for predicting activity change upon mutation

Mutant Dataset for Model Training

- 8561 single residue replacements in 7 diverse proteins
- Mutant activity experimentally determined and reported qualitatively: 5251 unaffected (U) and 3310 affected (A)

Protein	Source	Function	Mutant Data	PDB Code	SCOP Class
PR	HIV-1	proteinase	U: 218 A: 294	3phvA	all β
RT	HIV-1	transferase	U: 170 A: 196	1rtjA	α / β
lys	phage T4	hydrolase	U: 1364 A: 638	3lzmA	α + β
GVP	phage f1	DNA binding (replication)	U: 130 A: 221	1gvpA	all β
barn	*E. coli*	RNase	U: 643 A: 34	1bniA	α + β
lac	*E. coli*	DNA binding (regulation)	U: 2256 A: 1773	1efaB	all α
IL-3	human	signaling (growth factor)	U: 395 A: 229	1jliA	all α

Structure-Based Computational Mutagenesis

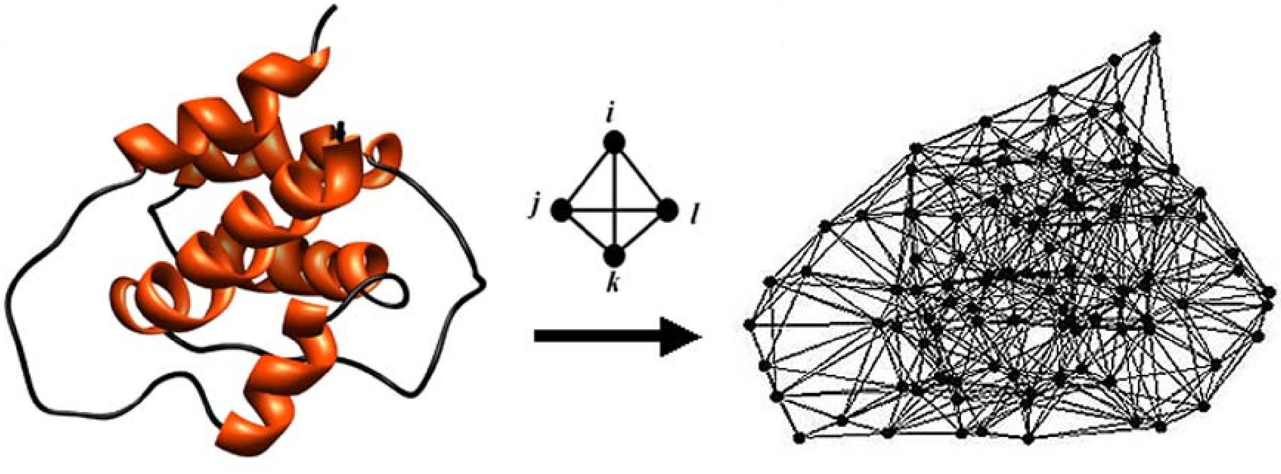


Amino Acid Quadruplet	"Pseudo-Energy" Log-likelihood s(i,j,k,l)
CCCC	3.29042538
CCCH	2.09542785
CCCS	1.96177162
CCCG	1.84022021
CCCI	1.79961166
CCCF	1.77139046
CCCT	1.76378293
CCCP	1.74840641
ACCC	1.74777711
CCCW	1.74711265
CCHH	1.70747111
CCCN	1.69741431
HHHH	1.61473339
.	.
.	.
.	.
HMNP	0.000221495
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

- Makes use of a four-body potential energy function that we previously developed
- Scores quantify the energy of interaction for every quadruplet of amino acid residues

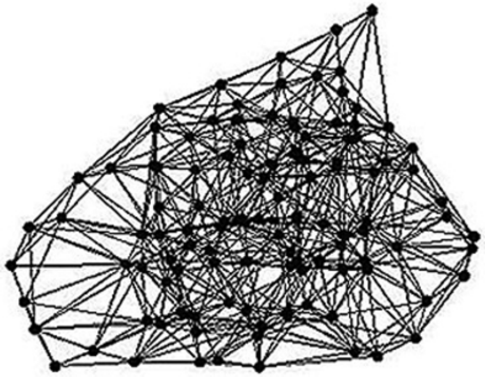
Structure-Based Computational Mutagenesis

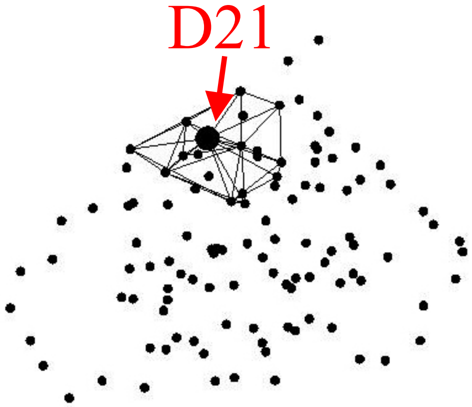
- The four-body potential and the computational mutagenesis technique utilize Delaunay tessellation of protein structure
- Creates a 3D tetrahedral tiling of the space occupied by a protein
- Each tetrahedron defines a residue quadruplet at the four vertices
- Tessellation also identifies a local structural neighborhood for every amino acid residue in a protein



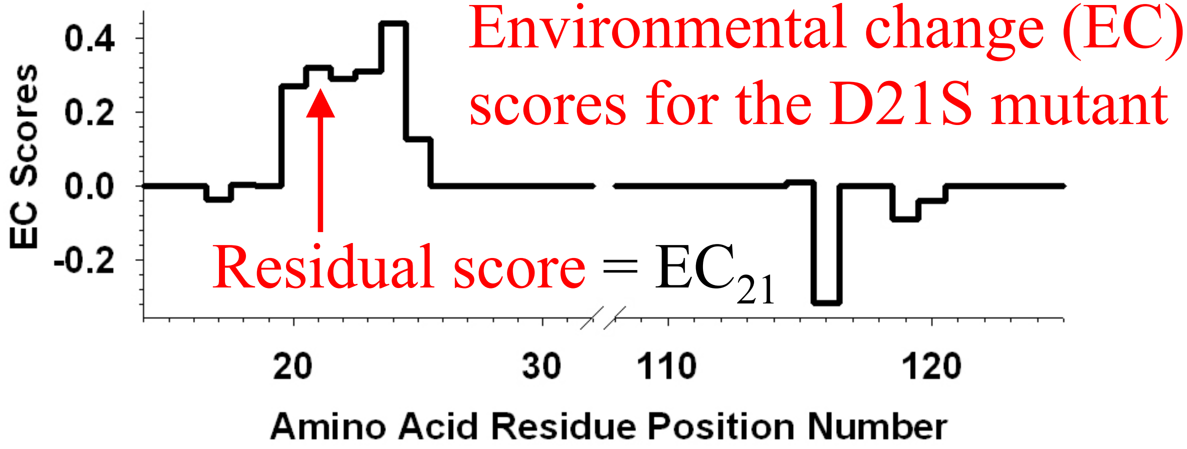
Computational Mutagenesis: IL-3 Example

IL-3 tessellation





D21 vertex is shared by 14 tetrahedra and has 11 neighbors



Environmental change (EC) scores for the D21S mutant

Residual score = EC_{21}

Representing Mutants via Common Attributes

- For a protein mutation at position *N*, nonzero EC scores occur only at *N* and its structural neighbors defined by tessellation

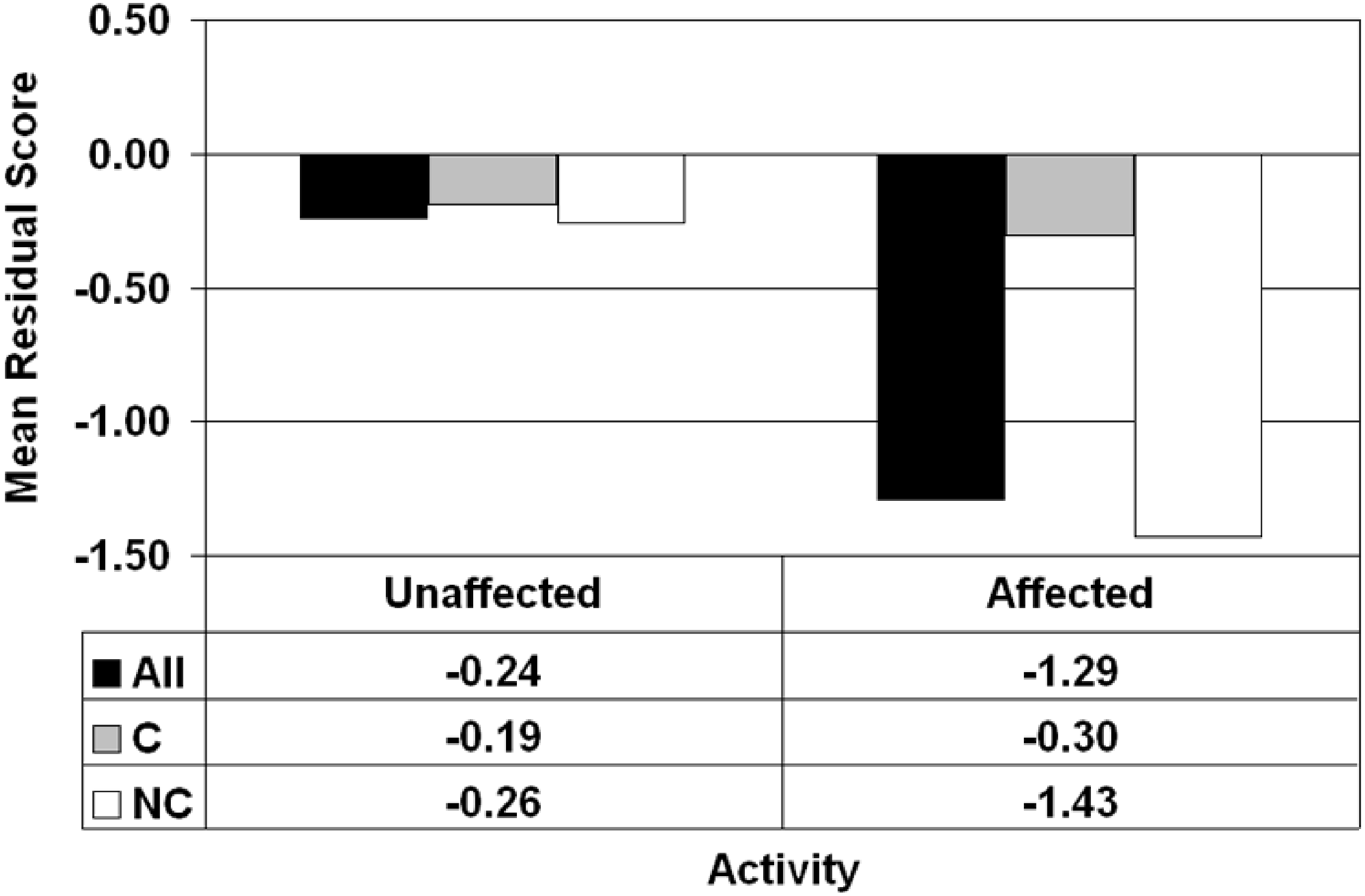
- Every position has at least 6 neighbors, can be ordered based on Euclidean distance from position *N* (tessellation edge-lengths)

- So, the 8561 mutants have 7 common EC values: residual score (EC score at *N*), and ordered EC scores of the 6 closest neighbors

- Calculate values for 20 additional sequence and structure features characterizing each mutant

- Result: each mutant represented as a 27-dimensional feature vector

Residual Scores Elucidate a Structure – Function Relationship



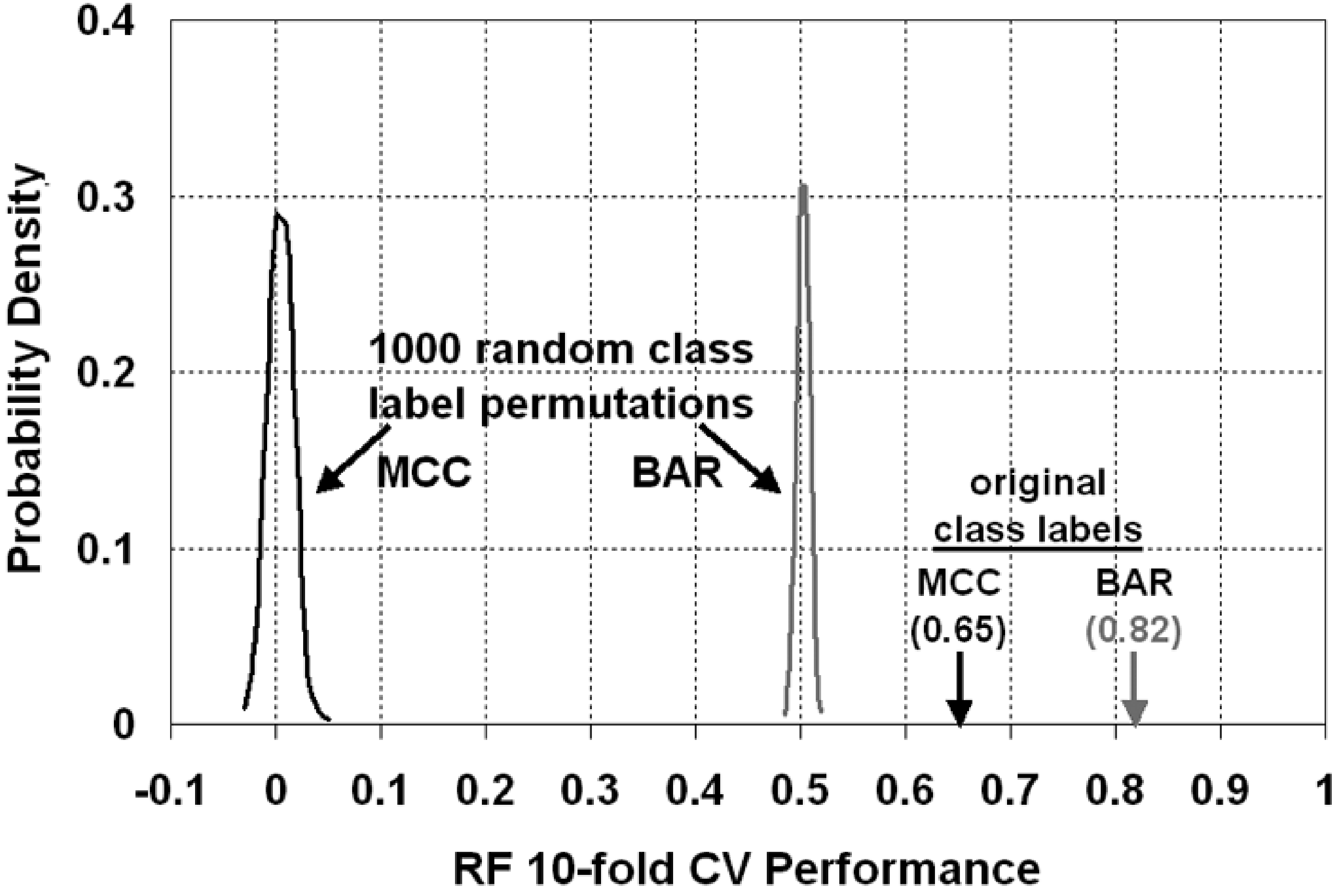
	Unaffected	Affected
■ All	-0.24	-1.29
▨ C	-0.19	-0.30
□ NC	-0.26	-1.43

Random Forest (RF) Model and Performance

- Evaluation: tenfold cross-validation (10-fold CV)
- ACC = accuracy (% correct); S/P = sensitivity/precision; BER = balanced error rate (BAR = 1 – BER); MCC = Matthew’s correlation coefficient; AUC = area under ROC
- ALL = combined dataset of all 8561 protein mutants

Data	ACC	S(U)	P(U)	S(A)	P(A)	BER	MCC	AUC
PR	0.83	0.74	0.83	0.89	0.82	0.18	0.64	0.89
RT	0.73	0.72	0.71	0.74	0.75	0.27	0.46	0.78
lys	0.82	0.88	0.87	0.71	0.73	0.21	0.59	0.89
GVP	0.74	0.72	0.62	0.75	0.82	0.27	0.45	0.78
barn	0.97	0.99	0.97	0.50	0.71	0.26	0.57	0.88
lac	0.84	0.86	0.85	0.81	0.82	0.16	0.67	0.92
IL-3	0.85	0.87	0.93	0.79	0.66	0.17	0.62	0.88
ALL	0.84	0.89	0.85	0.76	0.81	0.18	0.65	0.91

Statistical Significance of RF Model Trained Using the Combined Dataset



Protein-Specific Comparisons With Related Methods SIFT, MAPP, and Pmut

Protein / Method	ACC	S(U)	P(U)	S(A)	P(A)	BER	MCC
PR							
* AUTO-MUTE	0.83	0.74	0.83	0.89	0.82	0.18	0.64
SIFT	0.78	0.70	0.66	0.82	0.85	0.24	0.51
MAPP	0.76	0.62	0.89	0.92	0.68	0.23	0.55
Pmut	0.61	0.09	0.95	0.99	0.60	0.46	0.21
RT							
* AUTO-MUTE	0.73	0.72	0.71	0.74	0.75	0.27	0.46
MAPP	0.64	0.85	0.44	0.56	0.90	0.30	0.37
Pmut	0.56	0.05	0.90	0.99	0.55	0.48	0.15
lys							
* AUTO-MUTE	0.82	0.88	0.87	0.71	0.73	0.21	0.59
SIFT	0.63	0.59	0.82	0.72	0.45	0.35	0.29
MAPP	0.73	0.70	0.87	0.79	0.56	0.26	0.46
Pmut	0.52	0.42	0.77	0.74	0.37	0.42	0.15
lac							
* AUTO-MUTE	0.84	0.86	0.85	0.81	0.82	0.16	0.67
SIFT	0.68	0.78	0.70	0.57	0.66	0.33	0.35
MAPP	0.69	0.72	0.72	0.66	0.66	0.31	0.38
Pmut	0.61	0.77	0.66	0.36	0.49	0.44	0.14

***AUTO-MUTE is our method**

Performance of RF Model Trained Using the Combined Dataset on an Independent Test Set

- Obtained a diverse set of 248 single residue substitutions, each with known impact on activity, from Protein Mutant Database

- These mutations occur in 51 proteins not related to the 7 proteins used for model training

- The 51 proteins have 3D coordinate files in PDB – required in order for us to tessellate and generate 248 mutant feature vectors

- Comparison of RF model predictions with known impact on activity for 248 mutants: ACC = 0.84, MCC = 0.54, BER = 0.24

Conclusion

- Improved performance of our RF model due to:
  - training on a large and diverse dataset of mutants
  - use of structure-based attributes obtained from a computational mutagenesis technique relying on a four-body potential

- Public accessibility to RF model for making predictions: http://proteins.gmu.edu/automute/AUTO-MUTE_Activity.html (also accessible from my homepage: http://binf.gmu.edu/mmasso)

- Above activity prediction website provides access to all datasets, as well as mutant feature vectors, as downloadable text files