Structural Analysis of Proteome Vulnerability to Oxidative Damage in *Deinococcus radiodurans*

Matthew Robinson[1,2], Roger Chang[1], Joel Sher[1], Devon Stork[1], Pamela Silver[1]

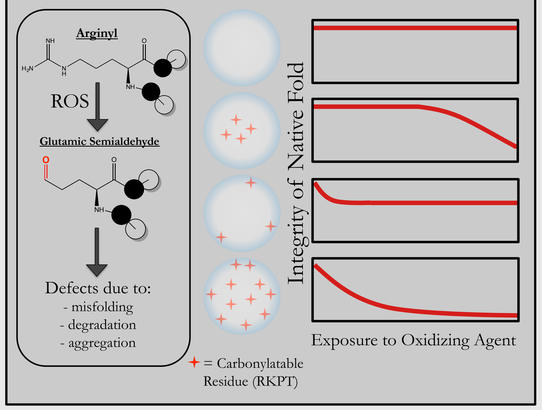
[1]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA.
[2]Department of Chemistry, Yale University, New Haven, CT 06520-8170, USA.

FAS Center for Systems Biology

Contact: Matt Robinson
matthew.robinson@yale.edu

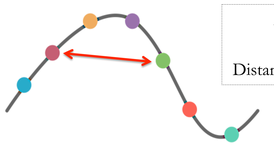
Introduction

Protein damage by reactive oxygen species (ROS) affects molecular function and cellular phenotype. The most critical protein damage by ROS is irreparable carbonylation of arginine, lysine, proline, and threonine side chains (RKPT), leading to misfolding, aggregation, and proteolysis. Empirical data suggests intrinsic vulnerability to ROS varies with tertiary structure, but previous studies have been only in sequence space. Here we investigate how distribution of RKPT in 3D structures determines vulnerability. A preliminary carbonylation site (CS) prediction model was derived by training a support vector machine (SVM) on a training set of proteins with measured CS. With this model, we can study the structural proteome of the radioresistant *D. radiodurans* in comparison to the non-resistant *E. coli*. We hypothesize that *D. radiodurans* robustness to ROS exposure may be due in part to evolved intrinsic protection of key proteins.



The Approach

The Need for a Structure-Based Approach:



Actual Distance
≠
Distance in Sequence Space

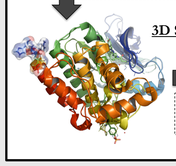
3D Protein Models give:
- better metrics of distance
- improved information about contacts
- ability to asses depth and tertiary structure

The Overall Method

1. Training set from previous studies
2. *D. radiodurans* proteome
3. *E. coli* proteome

PDB Structures
Homology Modeling
***Ab-initio* Folding**

3D Structure



Machine Learning using an SVM

The Model
A Structure-Based Predictor of Carbonylation Sites and Protein Oxidation

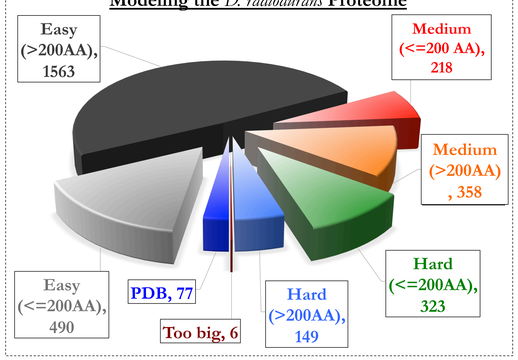
Application of Model to Proteomes

Insight into the radioresistance of *D. radiodurans* and the attendant susceptibility of *E. coli*

Obtaining Structures

Very few PDB structures for *D. radiodurans*
⇒ Need for high–throughput modeling

Modeling the *D. radiodurans* Proteome



For Easy and Medium:
- Homology modeling ideal
- Methods used:
  - I-TASSER
  - ProtMod

For Hard/Large Proteins:
- *Ab-Initio* folding useful
- Methods used:
  - QUARK
  - EVfold

Ultimately, every single prediction method was run on every single protein in the *D. radiodurans* proteome, *E. coli* proteome, and training set. Subsequently, all models (including PDB structures) are tested by multiple evaluation methods such as TM-score. The structure with the best overall score is then chosen for use in the machine learning model.

Training the Model

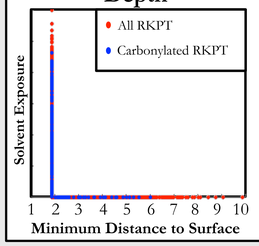
The Training Set

Organism	Unique Proteins	Carbonylation Sites R	K	P	T
Cow	1	4	7	10	5
E. Coli	18	8	6	37	24
Human	84	35	88	31	28
Mouse	2	2	1	0	0
P. angustum	55	8	15	33	38
Rabbit	1	2	2	1	0
Rat	8	1	6	1	1
Yeast	101	42	51	33	23
Total	**270**	**102**	**176**	**146**	**119**

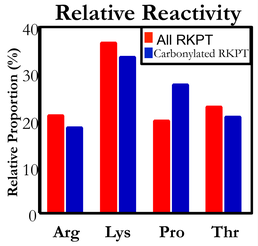
The training set data was compiled from 13 previously published protein mass-spectrometry studies.

The Initial Features:

Depth



Relative Reactivity



Expanding the Feature Set

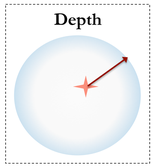
Features for machine learning were extracted from 3D protein structures using **UCSF Chimera**, a powerful program for the analysis of molecular structures. By writing a **Python script** to interface with Chimera, we were able to the obtain the following information:

Sequence-based Features

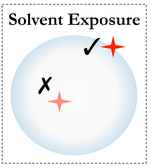
- Relative reactivity of the R/K/P/T residue
- Reaction rate of the residue with hydroxyl radicals
- Hydrophobicity of the residue using the Kyte-Doolittle scale
- Net charge of the residue
- Disorder score of the reside as predicted by DISOPRED
- Probability that residue is part of a binding site, also predicted by DISOPRED
- 5 summary attributes per residue derived from the AAindex by Atchley et al. that roughly correspond to:
  - Polarity
  - Secondary structure
  - Molecular volume
  - Codon Diversity
  - Electrostatic Charge

Structure-Based Features

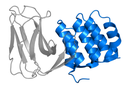
Depth



Solvent Exposure



Is residue in alpha helix or beta sheet



Metal Binding Sites as predicted by FINDSITE-metal prediction software

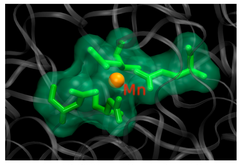
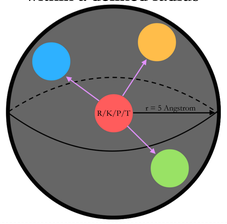
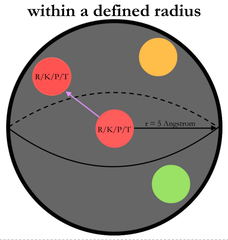


Figure from FINDSITE-metal Manual -- Brylinski M and Skolnick J (2010).

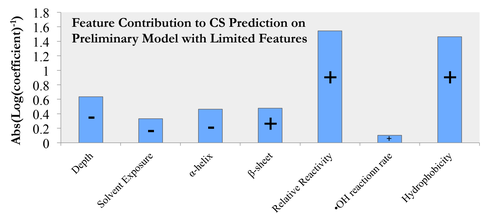
Number of amino acid contacts within a defined radius



Number of RKPT contacts within a defined radius



Using the last two features, we can thus define a local **spherical region** with further features including:
- Average hydrophobicity in region
- Net charge in region
- Average polarity in region
- Number of local metal binding sites
- Average molecular volume in region
- Number of local surface methionines/cysteines in region (perhaps important because surface methionines/cysteines can protect proteins from oxidation)



Machine Learning Performance

Performance of a Preliminary Model

	This Study	CSPD	CarSPred
True Positive CP	248	88	245
False Negative CP	0	160	3
True Positive CS	140	20	100
False Negative CS	371	491	411
True Negative CS	16792	20630	19221
False Positive CS	4297	459	1868
Specificity	99%	100%	99%
Sensitivity	27%	4%	20%
Accuracy	78%	96%	89%
Precision	3%	4%	5%
G mean	0.52	0.20	0.44
F1 Score	0.57	0.40	0.81

Due to **experimental constraints**, true negatives and false positives are not good measures of performance. Thus, the goal should not be perfect accuracy, but rather a good mix of sensitivity and precision.

Sensitivity:
= TP/(TP + FN)

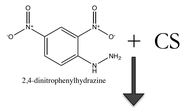
Precision:
= TP/(TP + FP)

F1 Score:
= Harmonic mean of Sensitivity and Precision

Future Directions

Validation using Redox Proteomics
We plan to use tandem mass spectrometry on *D. radiodurans* following irradiation in order to identify specific carbonylation sites and oxidized proteins. Detection is achieved through derivatization of carbonylated sites.

DNPH labeling

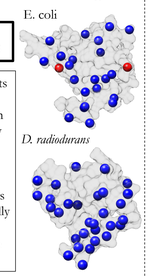


Derivative compound with mass shift detectable in MS

Pairwise Comparison of *D. radiodurans* and *E. coli* orthologs

RKPT
Predicted CS

Preliminary results suggest that 50S ribosomal protein L18 is specifically protected in *D. radiodurans*. Thus suggesting that some key proteins may be intrinsically protected from oxidative damage by ROS.



Literature References

- Lv H, Han J, Liu J, Zheng J, Liu R, Zhong D (2014) CarSPred: A Computational Tool for Predicting Carbonylation Sites of Human Proteins. PLoS ONE 9(10): e111478. doi:10.1371/journal.pone.0111478
- Maisonneuve E, Ducret A, Khoueiry P, Lignon S, Longhi S, Talla E, et al. (2009) Rules Governing Selective Protein Carbonylation. PLoS ONE 4(10): e7269. doi:10.1371/journal.pone.0007269
- Brylinski, M. and Skolnick, J. (2011), FINDSITE-metal: Integrating evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level. Proteins, 79: 735–751. doi:10.1002/prot.22913
- W. R. Atchley, J. Zhao, A. D. Fernandes and T. Druke, Proc. Natl. Acad. Sci. U. S. A., 2005, 102, 6395–6400.

Acknowledgments

This work was supported by the Gwill York and Paul Maeder Research Award for Systems Biology and the FAS Center for Systems Biology. Additional funding is provided by the Gordon and Betty Moore Foundation GBMF 2550.04 Life Sciences Research Foundation postdoctoral fellowship.