Sequence-Based Pattern Recognition and Structure/Function in Zinc Finger Proteins

Susan R. Atlas[1], Samuel H. Smith[2] and Laurie G. Hudson[3]

***Short Abstract* — We present a sequence-driven approach for predicting direct molecular targets of arsenic (As) in DNA repair pathways as a function of zinc finger interactions and structure. Unbiased string searches of the proteome are correlated with extensive structural bioinformatic data and literature annotations from online databases, followed by phylogenetic and machine learning analysis to further refine protein subfamilies and predict As binding activities. The methodology is generalizable to other protein families and pathways. Initial results are presented for the DNA repair pathway, a biologically-relevant focus of direct importance to arsenic epidemiology and cancer biology.**

***Keywords* — Structural motifs; machine learning; protein subfamilies; As-binding; DNA repair; epidemiology; cancer**

I. Introduction

The prediction of the protein structure and function based solely on amino acid sequence remains one of the great challenges of contemporary computational molecular biology. *Ab initio* approaches proceeding from the level of individual atoms represent the most fundamental starting point, but are not yet capable of predicting critical folds and secondary structure due to limitations in underlying force fields and simulation times. Coupled approaches, utilizing a combination of *de novo* prediction, machine learning techniques, and data mining, presently hold the greatest promise, particularly for specific biophysical applications.

Motivated by observations of synergy between low dose arsenite and ultraviolet radiation for DNA damage and inhibition of DNA repair targets [1], and the identification of specific cysteine-histidine (C3H1 and C4) zinc finger domains as direct targets for interaction with arsenic [2], we have developed a sequence-based method to generalize cell biological, molecular, biochemical and *in vivo* data and systematically predict direct molecular targets of arsenic in DNA repair pathways, structural subfamilies [3], and associated binding affinities [2]. Predictions are then compared against experimental results to further refine the classifier and inform experimental studies. A principal motivation for this work is to understand critical mechanisms of arsenic carcinogenesis and toxicity in the context of zinc finger structure and function in DNA repair pathways.

II. Methodology

A preliminary report of the methodology has been given in [4]. The string-based pattern-recognition step is implemented as a regular expression search for C_xH_y consensus sequences in the proteome. Results are correlated and filtered using the relational database ZincGrep, which enables the automated integration of online structural and text-based literature annotations (e.g. Entrez GeneRIFs, SCOP, PDB, GO) within a comprehensive local schema. ZincGrep merges information derived from trusted sources, manual search, and automated Web spider software. Each refinement of the regular expression pattern and database yields candidate zinc-finger proteins potentially implicated in DNA repair.

III. Results and Conclusion

We have developed a machine learning and bioinformatic framework for the identification and functional characterization of proteins associated with specific biological processes and pathways. Results from initial computations have led to the identification of 31 candidate zinc finger proteins associated with DNA repair, further annotated by SCOP domain and structural subclasses that extend beyond conventional zinc finger classifications. These are undergoing further evaluation using cell biology and biochemical approaches to establish sensitivities to arsenic and provide information on structural characteristics for iterative refinement of the framework.

Acknowledgements: This work was supported in part by NIH/NIEHS Grant 1R21 ES021499-01. We are grateful to the UNM Cancer Center Shared Resource for Biostatistics and Bioinformatics, supported by D.H.H.S NIH/NCI P30 Grant CA 118100, and the UNM Center for Advanced Research Computing for computational resources.

[1]Department of Physics and Astronomy and UNM Cancer Center, University of New Mexico, Albuquerque, NM 87131. E-mail: susie@sapphire.phys.unm.edu

[2]University of New Mexico School of Medicine, Albuquerque, NM 87131. E-mail: osmith14@salud.unm.edu

[3]University of New Mexico College of Pharmacy, Albuquerque, NM 87131. Email: lhudson@salud.unm.edu

References

[1] Ding, W., Liu, W, Cooper, KL, Qin, XJ, de Souza Bergo, PL, Hudson, LG, and Liu, KJ (2009). Inhibition of poly(ADP-ribose) polymerase-1 by arsenite interferes with repair of oxidative DNA damage. J Biol Chem 284:6809-6817.

[2] Zhou, X, Sun, X, Cooper, KL, Wang, F, Liu KJ, and Hudson, LG (2011). Arsenite interacts selectively with zinc finger proteins containing C3H1 or C4 motifs. J. Biol. Chem. 286:22855-22863.

[3] Krishna, S, Majumdar, I, and Grishin, NV (2002). Structural classification of zinc fingers. Nucl. Acids Res. 31:532-550.

[4] Smith, SH (2011). *A Novel Sequence-Based Approach for Identification of Zinc Finger Motifs Involved in Arsenic-Induced Zinc Release.* B.S. honors thesis, University of New Mexico.