

### Insight into redox-regulated gene networks in vascular cells

Charles D Johnson<sup>1</sup>, Yoganand Balagurunathan<sup>2</sup>, Edward R. Dougherty<sup>2</sup>, Cynthia A Afshari<sup>3</sup>, Qiang He<sup>1</sup> and Kenneth S. Ramos<sup>1\*</sup>

<sup>1</sup>Department of Biochemistry and Molecular Biology, University of Louisville, Louisville, Kentucky - 40292; <sup>2</sup>Department of Electrical Engineering, Texas A&M University, College Station, Texas - 77843; <sup>3</sup>Amgen Corporation, Thousand Oaks, CA - 91320; Kenneth S. Ramos\* - Email: kenneth.ramos@louisville.edu; Phone: +502 852 5217; Fax: +502 852 6221; \* Corresponding author received November 06, 2006; accepted January 21, 2007; published online April 10, 2007

#### Abstract:

To understand the complex nature of the atherogenic response initiated by oxidative stress in vascular smooth muscle cells (vSMCs), computational prediction methodology was employed to define putative gene-gene and gene-environment interactions in vSMCs subjected to oxidative chemical stress. Computational relationships were derived from the global gene expression profiles of murine cells challenged with a chemical pro-oxidant to cause oxidative stress or cells treated with anti-oxidant prior to oxidative injury. Target clones were chosen based on their biological relevance within the context of the atherogenic response and included lysyl oxidase, matrix metalloproteinase 2, insulin like growth factor binding protein 5, and lymphocyte antigen 6c. Established biological relationships were derived computationally confirming the usefulness of the algorithm in uncovering novel biological relationships worthy of future investigation. Thus, the predictive algorithm can be a useful tool to advance the frontiers of biological discovery.

**Keywords:** gene networks; vascular cells; oxidative stress; redox

#### Background:

The formation of atherosclerotic lesions involves the activation of vSMCs leading to modulation from a normally quiescent, contractile phenotype to a lesser differentiated and proliferative (i.e. atherogenic) phenotype. [1] This phenotypic modulation process is manifested by migration of vSMCs from the tunica media to the vessel lumen where they proliferate uncontrollably and give rise to occluding lesions that accumulate large amounts fat, undergo cellular necrosis and recruit clotting factors. To date, the interactive gene networks responsible for induction of atherogenic vSMC phenotypes have not been identified with certainty. We have previously established that oxidative chemical injury of vSMCs in vivo or in vitro mediate the phenotypic modulation of vSMCs to atherogenic phenotypes. [1]

To understand the complex nature of the atherogenic process, a computational approach was used to examine global patterns of gene expression and to define putative gene-gene interactions predictive of critical biological relationships during the course of atherogenesis. Several genes were chosen as targets for prediction using a method first described by Kim et al. [2] The target genes selected for analysis were lysyl oxidase, matrix metalloproteinase 2, insulin like growth factor binding protein 5, and lymphocyte antigen 6c. These genes encode for proteins known to be involved in the regulation of cellular growth and differentiation.

The experimental system employed involved acute challenge of vSMCs with benzo(a)pyrene (BaP), an aromatic hydrocarbon that causes oxidative stress in vSMCs [3] and initiates a cascade of genomic changes that culminates in induction of atherogenic phenotypes. [4] Our goal was to identify small sets of genes whose transcriptional states were predictive of the chosen targets, whether lying upstream or downstream within the gene interaction network, or based on chains of interaction among various intermediates. There was no assumption of causality in the prediction method and its sole focus was to identify sets of genes that may be associated with the target gene, and that could constitute the basis of hypothesis-driven biological investigations.

#### Methodology:

To define genomic profiles during the early phase of the atherogenic response, G0 synchronized cultures of vSMCs from C57BL/6J (6 wk old) mice at passage 12 and 75% confluence were released into growth by addition of fetal bovine serum (10%) in the presence of benzo (a)pyrene (BaP) (3  $\mu$ M; Sigma-Aldrich) or dimethyl sulfoxide (DMSO, 0.0075%; Sigma-Aldrich) for 24 h. A separate set of cultures was pretreated for 1 h with 0.5 mM *N*-acetylcysteine (NAC) (Sigma-Aldrich), a water-soluble antioxidant and precursor of cellular glutathione, dissolved in culture medium prior to BaP treatment to enhance antioxidant activity. Cultures were allowed to recover for 1 wk before mRNA isolation. Mouse cDNA arrays developed at National Institute of Environmental Health Sciences (NIEHS) were used for gene expression profiling. A complete listing of the 8,976 transcripts represented on the chip is available at <http://dir.niehs.nih.gov/microarray/chips.htm>. Comparisons between three treatment groups and one control were duplicated four times for a total of 12 independent hybridizations. Poly(A)RNA samples (2–4  $\mu$ g) were labeled with cyanine-3 (Cy3) or cyanine-5 (Cy5)-conjugated dUTP (Amersham) by reverse transcription using SuperScript (Invitrogen) and oligo-dT (Amersham). A subset of 200 differentially expressed genes was selected based on ANOVA *p* values that had been derived from gene expression profiles in response to pro-oxidant and anti-oxidant treatment as defined by cDNA microarrays. [5]

Using a heuristic method to discretize the data into ternary states that describe their behavior, the algorithm started by categorizing transcript levels into ternary expression data: -1 for down-regulated, 0 for invariant, and +1 for up-regulated genes. Invariant genes were defined as genes whose expression was not changed by the treatment relative to control. The data were then divided into training and test sets. Based on the training data, the conditional probability that the target gene takes on one of the three transcriptional states was calculated for all possible patterns of the predictor genes, and the predicted target value defined as the state with the largest conditional probability. In considering a predictor set with two genes, the

relationships can be defined as:  $t_1, \dots, t_9$  equal  $-1, 0,$  or  $+1$ . The analysis then reverted to the test data to examine the performance of these predictors. The error for each of the predictor function is given by  $|T_{obs} - T_*|$ , where  $T_{obs}$  is the observed and  $T_*$  is the predicted transcriptional state, which could be the optimal predictor state  $T_\psi$  obtained by the designed filter or the reference predictor state,  $T_\mu$  obtained by the reference filter.

The above procedure was repeated by randomly splitting the data into training and test sets, in a fixed proportion. The test error was estimated by averaging the prediction error across all iterations, and this error was computed for all possible predictor combinations. The performance of a set of predictors was determined by a statistic known as the coefficient of determination (COD. [6] This coefficient measured the degree to which the transcriptional levels of a set of genes can be used to improve the prediction of the transcriptional state of a target gene relative to the best possible prediction in the absence of predictors. In this case, the mean of the target gene was used as the reference metric, (its transcriptional state represented by  $T_\mu$ ). The COD ( $\theta$ ) is defined as

$$\theta = \frac{\varepsilon_* - \varepsilon_\psi}{\varepsilon_*}$$

Where  $\varepsilon_*$  is the average error for the best predictor in absence of observation and  $\varepsilon_\psi$  is the average error due to the optimal predictor designed. The errors with respect to  $n$  observations is given by,

$$\varepsilon_* = \frac{1}{n} \sum_{i=1}^n |T_{obs,i} - T_{\mu,i}|$$

$$\varepsilon_\psi = \frac{1}{n} \sum_{i=1}^n |T_{obs,i} - T_{\psi,i}|$$

The higher the COD  $\theta$  (close to 1), the more accurate the prediction of the target's transcriptional state, i.e., the higher the degree of relationship between the target and predictor genes. All possible combinations of 1, 2 and 3 gene predictors for the chosen targets were studied with possible predictors runs in the order of millions for multiple gene combinations for each target. Predictors were ordered with respect to their errors and the COD's, and the analysis focused on COD's greater than 0.9 and a test error less than 0.05. Information obtained was suggestive of biological commonality between predictor genes and their specified targets.

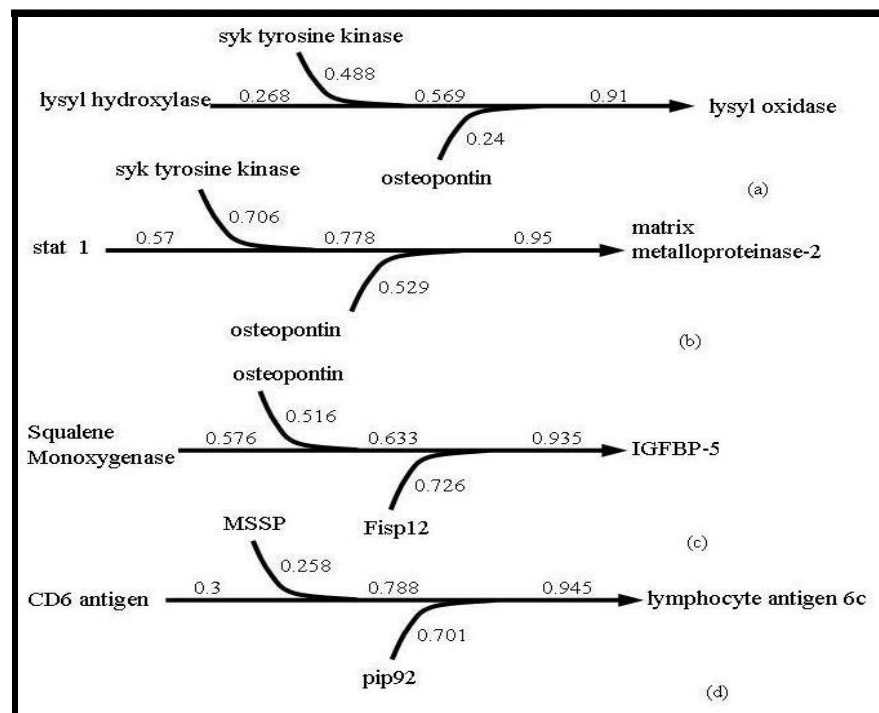
### Results and Discussion:

The present study was undertaken to understand the complex nature of the atherogenic process initiated by chemical atherogens present in

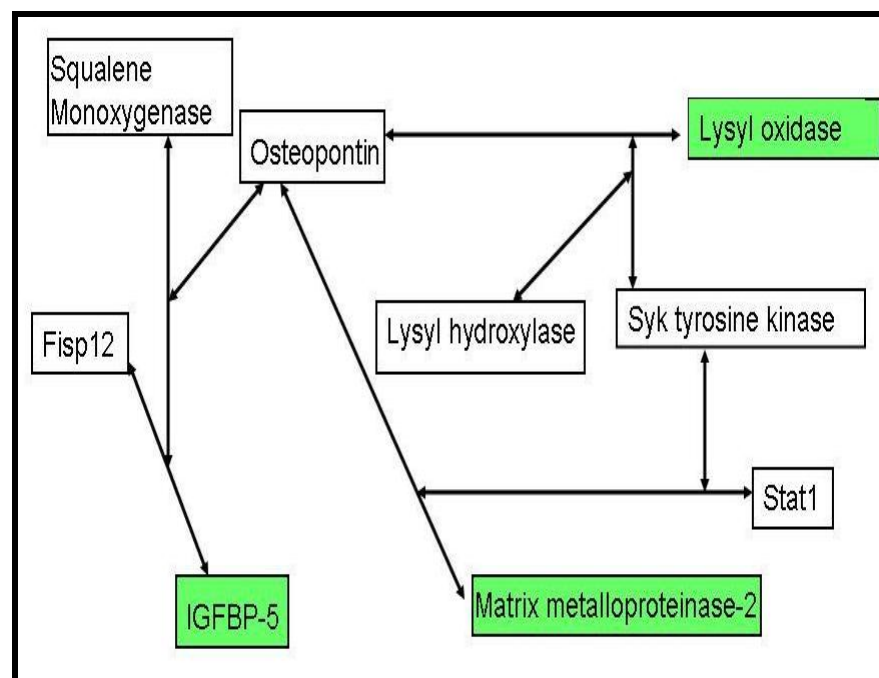
tobacco smoke using a novel computational approach. Based on ANOVA p-values  $\leq 0.01$  several clones were selected for further analysis using the computational target clone-predictor approach. This strategy selected for genes within the dataset that displayed a high probability to behave as superior singleton predictors. Target clones included lysyl oxidase, matrix metalloproteinase 2, insulin like growth factor binding protein 5, and lymphocyte antigen 6c. Multiple clone predictor combinations were ranked based on prediction error. Predictor combinations with CODs greater than 0.9 and errors less than 0.05 were selected for further analysis. A large number of three-clone combinations met these criteria for most targets, with one or two clones identified as predominant predictors within the sample pool.

The development and validation of analytical tools that detect multivariate influences on cellular decision-making within complex genetic networks is essential. COD methodology provides an advantage over linear correlations because gene associations are measured based on categorization of discrete variables into a finite numbers of subgroups that enhance the accuracy of prediction. This is in contrast to Pearson's correlation where a pair of continuous variables is examined in the absence of criteria that examine putative interactions among multiple genes. CoD can in fact be used for nonlinear filtering of small datasets such as those often encountered in DNA microarray experiments as CoD is based on error estimation of patterns of gene expression. The determination coefficient permits biologists to focus on particular connections in the genome and coefficient estimates are useful even if they are biased and not overly precise, because at least the estimated coefficients provide a practical means of discrimination among potential predictor sets.

A complete listing of target-predictor clones is presented as Appendix 1. Biologically relevant three gene combinations for each selected target are presented in Figures 1 and 2. The combination of lysyl hydroxylase, syk tyrosine kinase, and osteopontin was shown to predict the behavior of lysyl oxidase (COD 0.91). Lysyl oxidase functions in the maturation of collagen and elastin and is a putative tumor suppressor through a Ras related mechanism. [7] The two matrix related targets, lysyl oxidase (LO) and matrix metalloproteinase-2 (mmp-2) shared two common predictors syk tyrosine kinase (Syk) and osteopontin (OPN). The substitution of stat1 for lysyl hydroxylase and the combination of syk tyrosine kinase, and osteopontin were shown to predict the behavior of matrix metalloproteinase-2 (COD 0.95). This is significant given the role of these two targets in matrix remodeling during atherogenesis. The prediction of genes related to insulin like growth factor binding protein 5 included squalene monooxygenase, osteopontin, and connective tissue growth factor (fisp12) (COD 0.935). Lastly, the best predictors of lymphocyte antigen 6c included MSSP, pip92 and CD6 antigen (COD 0.945).



**Figure 1:** Three gene combinations to predict the behavior of selected target genes. The values are the COD values for each predictor and the effect that addition of each predictor has on overall model prediction potential



**Figure 2:** Network inference based on overlapping edges in the predictor-target relationships resolved using CoD values

MMP-2 functions in the digestion collagens of and elastin to promote cell migration and vascular remodeling. Syk plays a critical role in signaling of various receptors of the adaptive immune system and functions as an inhibitor of breast cancer cell growth and metastasis.

[8] Syk participates in the integrin signaling pathway in monocytic cells leading to activation of NF- $\kappa$ B and increased levels of cytokine mRNAs. [9] Studies have shown that OPN induces activation of mmp-2 through the I $\kappa$ B alpha/IKK signaling pathways [10], and that

mmp-2 plays a direct role in OPN-induced cell migration, invasion, and tumor growth. Both LO and OPN are activated by platelet-derived growth factor in vSMC. [11]

The difference between LO and mmp-2 in our predictor model is the addition of lysyl hydroxylase in the case of LO and stat1 in the case of mmp-2. Lysyl hydroxylase catalyzes hydroxylation of lysyl residues in collagens and other proteins with collagenous domains. [12] Both LO and lysyl hydroxylase are involved in post-translational modifications of collagen as part of the cross-linking pathway and would thus be expected to behave in a similar manner. The regulation of mmp-2 by stat1 has been demonstrated in tumor cells. [13] Stat1 is a signal transducer and activator of early transcription factor by adherence, and a modulator of ICAM gene expression. Studies have shown the involvement of Stat1 in IFN and growth factor-dependent signaling and in positive, negative and constitutive regulation of gene expression. [14]

Insulin like growth factor binding protein 5 (IGFBP-5) was predicted by OPN, squalene monooxygenase (SMO) and connective tissue growth factor (Fisp12). IGFBP-5 is the most conserved IGFBP across species and as an essential regulator in bone, kidney and mammary gland. In addition, IGFBP-5 plays a decisive role in the control of proliferation of specific tumor cell types. [15] In vSMCs, IGFBP-5 and OPN promote IGF-I effects and OPN binds to IGFBP-5 with high affinity. [16] These interactions are important for concentrating intact IGFBP-5 in the extracellular matrix and modulation of the cooperative interaction between the IGF-I receptor and integrin  $\alpha_v\beta_3$  signaling pathways in atherosclerotic lesion. [17] Fisp12 mediates cell adhesion and migration through integrin  $\alpha_v\beta_3$ , and promotes cell survival, and induces angiogenesis *in vivo*. [18] Fisp12 is also known as insulin-like growth factor binding protein related proteins (IGFBP-rPs). [19] Therefore, it is likely that Fisp12 is regulated in a like fashion to IGFBP-5.

It is unclear how squalene monooxygenase (squalene epoxidase) is related to IGFBP-5, but squalene monooxygenase catalyzes the second committed step in cholesterol biosynthesis from farnesyl pyrophosphate to squalene. [20] Studies have shown that squalene monooxygenase is bound to the endoplasmic reticulum of cells in association with NADPH-cytochrome P450 reductase, its electron transfer partner. [21] Squalene monooxygenase is regulated at the transcriptional level in response to sterol levels and may compete with HMG-CoA reductase as the regulated step in cholesterol synthesis. Studies have identified a link between farnesyl pyrophosphate and post-translational processing of Ras and Ras-related proteins. [20] The Ramos laboratory has demonstrated that Ras is a key factor in atherogenesis [22], and others have revealed that OPN is also critical for Ras expression. [23] Thus, LO, MMP-2 and IGFBP-5 share OPN and integrin signaling as common factors, a relationship identified by computational methodology.

The last target examined was lymphocyte antigen 6c, which plays a role in the T cell activation cascade and is modified by atherogenic challenge in vSMC. [5] This was predicted by the combination of CD6 antigen, pip92 and MSSP. CD6 belongs to the scavenger receptor cysteine-rich protein super family that triggers co-activating signaling of T cells. Its regulation during T cell ontogeny and activation has been extensively investigated. [24] MSSP promotes

ras/myc cooperative cell transforming activity by binding to c-Myc [25], while Pip92 is an early response gene, activated by growth factors. The activation of pip92 is mediated by JNK and p38 kinase, but not ERK. [26] Type I interferon is the primary regulator of inducible Ly-6C expression on T cells [27], and studies have shown that interferon-alpha down regulates c-myc.[28] c-myc single-strand binding protein (MSSP) may function in a similar fashion, a pattern that fits well with our current understanding of chemical-induced atherosclerosis. [29]

### Conclusion:

The greatest challenge in computational prediction is biological validation. Clearly, biological relevance is required to establish the true significance of findings derived computationally. Based on established biological relationships, we demonstrate that predictor gene combinations derived computationally using the COD algorithm can be related at the biological level and consistent with established biological contexts. We have applied a similar strategy to define gene network structure and function of LINE-1 and found that unknown predicted relationships could be validated using functional genomics approaches. [30] Thus, the fidelity of computational relationships lends confidence that COD methodology can in fact predict undiscovered relationships. These findings may yield new areas of exploration for those interested in atherogenesis and computational biology.

### References:

- [01] K. S. Ramos, *Ann Rev Pharmacol Toxicol.*, 29:243 (1999) [PMID: 10331084]
- [02] S. Kim, *et al.*, *Genomics*, 67:201 (2000) [PMID: 10903845]
- [03] K. K. Kerzee & K.S. Ramos, *Mol. Pharmacol.*, 58:152 (2000) [PMID: 10860937]
- [04] K. P. Lu, *et al.*, *Atherosclerosis*, 160:273 (2002) [PMID: 11849648]
- [05] C. D. Johnson, *et al.*, *Physiol Genomics*, 13:263 (2003) [PMID: 12657712]
- [06] E. R. Dougherty, *et al.*, *Signal Processing*, 80:2219 (2000)
- [07] S. Contente, *et al.*, *Science*, 249:796 (1990) [PMID: 1697103]
- [08] P. J. Coopman, *et al.*, *Nature*, 406:742 (2000) [PMID: 10963601]
- [09] T. H. Lin, *et al.*, *J Biol Chem.*, 270:16189 (1995) [PMID: 7541794]
- [10] S. Philip *et al.*, *J. Biol. Chem.*, 276:44926 (2001) [PMID: 11564733]
- [11] R. S. Green, *et al.*, *Lab Invest.*, 73:476 (1995) [PMID: 7474918]
- [12] C. M. Kielty, *et al.*, *J Cell Sci.*, 103:445 (1992) [PMID: 1478946]
- [13] S. Huang, *et al.*, *Oncogene*, 21:2504 (2002) [PMID: 11971185]
- [14] L. Malmgaard, *J Virol.*, 76:4520 (2002) [PMID: 11932417]
- [15] M. R. Schneider, *et al.*, *J. Endo.*, 172:423 (2002) [PMID: 11874691]
- [16] T. J. Nam, *et al.*, *Endocrinology*, 141:1100 (2000) [PMID: 10698186]
- [17] T. C. Nichols, *et al.*, *Circ Res.*, 85:1040 (1999) [PMID: 10571535]
- [18] M. Babic, *et al.*, *Mol Cell Biol.*, 19:2958 (1999) [PMID: 10082563]
- [19] V. Horn, *et al.*, *Int J Mol Med.*, 9:645 (2002) [PMID: 12011983]

- [20] S. A. Holstein, *et al.*, *Biochemistry*, 41:13698 (2002) [PMID: 12427032] [26] K. C. Chung, *et al.*, *Eur J Biochem.*, 267:4676 (2000) [PMID: 10903500]
- [21] T. Ono & K. Bloch, *J Biol Chem.*, 250:1571 (1975) [PMID: 234459] [27] A. J. Schlueter, *et al.*, *J Interferon Cytokine Res.*, 21:621 (2001) [PMID: 11559440]
- [22] K. S. Ramos, *et al.*, *Arch. Biochem. Biophys.*, 332:213 (1996) [PMID: 8644998] [28] S. Radaeva, *et al.*, *Gastroenterology*, 122:1020 (2002) [PMID: 11910354]
- [23] Y. Wu, *et al.*, *Br J Cancer*, 83:156 (2000) [PMID: 10901364] [29] M. Fujimoto, *et al.*, *Int J Oncol.*, 16:245 (2000) [PMID: 10639566]
- [24] J. E. Skonier, *et al.*, *Protein Eng.*, 10:943 (1997) [PMID: 9415444] [30] K. S. Ramos, *et al.*, *Genomics*, (accepted) (2007)
- [25] T. Niki, *et al.*, *Genes Cells*, 5:127 (2000) [PMID: 10672043]

Edited by Susmita Datta

Citation: Johnson *et al.*, *Bioinformatics* 1(10): 379-383 (2007)

**License statement:** This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.