Pacific Symposium on Biocomputing 11:354-366(2006)

DISCOVERING REGULATED NETWORKS DURING HIV-1 LATENCY AND REACTIVATION

SOURAV BANDYOPADHYAY, RYAN KELLEY, TREY IDEKER[1,2]

[1]Program in Bioinformatics, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA. [2]Department of Bioengineering, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
{sourav,rkelley,trey}@bioeng.ucsd.edu

Human immunodeficiency virus (HIV) affects millions of people across the globe. Despite the introduction of powerful anti-viral therapies, one factor confounding viral elimination is the ability of HIV to remain latent within the host genome. Here, we perform a network analysis of the viral reactivation process using human gene expression profiles and curated databases of both human-human and human-HIV protein interactions. Based on this analysis, we report the identification of active pathways in both the latent and early phases of reactivation. These active pathways suggest host functions that are altered and important for HIV pathogenesis.

1. Introduction

Human Immunodeficiency Virus (HIV-1) infects T lymphocytes and macrophages, resulting in the depletion of CD4+ T cells , which is the defining feature of acquired immune deficiency syndrome (AIDS). Recently, highly active anti-retroviral therapy (HAART) has led to a dramatic decrease in morbidity and mortality due to HIV and, when successful, results in undetectable levels of HIV-1 RNA in blood plasma. This stage of latency represents a period of proviral integration with little to no viral replication [1]. HIV-1 persists in a small reservoir of latently infected resting memory CD4+ T cells, which shows minimal decay even in patients on HAART and can persist for the lifetime of the patient [2, 3]. Latent HIV reservoirs are the principal barriers preventing the eradication of HIV infection and viral reactivation is necessary for targeting by antiviral drugs [2, 4]. Although HIV has been the subject of much study, the precise mechanisms by which the virus reactivates within the host cell remain unclear. Evidence points to the alteration of cellular transcription machinery in order to maximize the viral replication process [5].

Recently, microarray analysis has been employed to survey the changes of host cell transcription [6-8]. In particular, Krishnan and Zeichner have studied the changes in cellular gene expression associated with reactivation and

completion of the lytic viral cycle in cell lines chronically infected with HIV-1 [6]. The viral lytic cycle follows distinct mechanistic changes corresponding to stages of HIV synthesis. In the early stage, fully spliced mRNAs for Rev, Tat and Nef are exported from the nucleus for translation. In the late stage, the accumulation of Rev protein in the cytoplasm triggers the export of unspliced viral RNAs from the nucleus to form new viral particles. [9]. Results indicated that uninduced, latent cells had an altered gene expression program and that the host cell underwent specific and ordered changes in gene expression upon reactivation that corresponds to different stages of the viral life cycle.

The authors arrived at these conclusions through application of hierarchical clustering and functional categorization of differentially expressed genes. Clustering gene expression data allows similarly expressed genes to be grouped together, but does not provide any functional explanation for the mechanisms of regulation. Functional categorization identifies known pathways and functional categories that are enriched for differentially expressed genes. This type of analysis provides limited functional insights by constraining analysis to known pathways and reactions [10]. Ideally, we would like to integrate the clustering of similarly expressed genes with an incorporation of a wide variety pathway and biological protein interaction information in a coherent fashion [11, 12]. Network based analysis can improve upon this approach by identifying interesting groups of genes which have not been specifically delineated as a pathway in an ontological framework. One such approach, "Active modules", is a method for searching networks to find subnetworks of interactions with unexpectedly high levels of differential expression [13]. In this approach, gene expression data are mapped onto biological networks, a statistical measure is used to score sub-networks based on gene expression data, and a search algorithm is used to find sub-networks with high score.

Commonly, gene expression analysis using biological networks has been limited to lower organisms for which large scale experimentally derived protein-protein interactions networks are available. The generation of literature-based protein interaction networks from text mining has shown utility in the interpretation of gene expression [14]. However, manually curated protein interaction data is preferable to automated prediction of interactions through natural language processing. The Human Protein Reference Database (HPRD) includes information on protein–protein interactions, post-translational modifications, enzyme–substrate relationships and disease association of human genes which was derived manually by a critical reading of the published literature by expert biologists [15]. Another manually curated database, the HIV-1, human interaction database, provides protein-protein interaction data among 7

of the 9 (excluding *env* and *nef*) HIV-1 genes and human host cell genes[16]. These interactions are traceable to primary literature and are annotated by an ontology of terms describing the nature of the interaction. Together, these networks strive to encompass all of the information that has been published concerning protein inter-relationships between both HIV-1 and human proteins.

Here, we report the identification of protein interaction modules that are significantly activated or repressed across different stages of the latent HIV-1 replication cycle. Our results indicate multiple significant clusters of gene expression in which genes are linked together through established interactions in the literature. This computational analysis allows for the evaluation of a mechanism of the observed changes in gene expression. Analysis of the observed differences in active networks between HIV-1 life cycle stages suggest that these differences are associated with the movement from latent to actively replicating HIV *in vivo*.

2. METHODS

2.1. *Microarray Data*

Microarray data was taken from Krishnan et. al. [6] Each array was normalized and the fold changes reported. 131 genes showed altered expression before induction and 1,740 spots showed significant altered gene expression at some point though the lytic replication cycle. Data were averaged over time points corresponding to specific stages in the HIV-1 life cycle which were determined by RT-PCR analysis of specific HIV-1 mRNA fragments. The early stage gene expression of the lytic cycle was taken as the mean over the 0.5,3,6 and 8 hour time points. The intermediate stage gene expression was taken as the mean over 12, 18 and 24 hours post induction. The late stage was the mean over 48, 72 and 96 hours post induction. Each stage included at least 18 arrays including replicates. In total, 1,334 genes were differentially expressed during the early time points, 756 during the intermediate stage, and 566 during the late stage ($P<0.001$). P-values were assigned to each probe for each of the four stages using the t-test using log ratios for all arrays for a specific stage, including replicates against a mean of 0 (no differential expression).

2.2. *Network Generation*

Human – human protein interaction data was culled from HPRD [15] (June 2005 download) and protein fragments were BLAST matched to Entrez Gene and unigene identifiers. HIV-human interactions were taken from the HIV-1, human

interaction database[16]. Each protein –protein link can be traced to a specific literature citation which assists in any hypothesis generation.

2.3. *Algorithm Implementation*

Briefly, the ActiveModules algorithm attempts to identify connected regions of a network, which have an unexpectedly high occurrence of genes with significant changes in expression. These network regions represent putative "active modules" in response to a particular test condition. The score of a subgraph is defined as the sum of expression Z-values divided by the square root of the number of nodes in the subgraph.

$$\text{Score}(V) = \frac{1}{\sqrt{|V|}} \sum_{v \in V} z_v$$

Here, V is a set of nodes which define a subgraph, while Z_v refers to the Z score for node v. Individual Z scores are determined by application of an approximation of the inverse normal CDF to the individual expression p-values. This scoring system ensures that if the original Z scores are distributed according to the normal distribution, the expected mean and variance of the subgraph scores are independent of subgraph size. In order to find high scoring regions according to this criterion, a greedy search is initiated from each protein in the network. At each step of the search, all adjacent proteins are considered for inclusion in the result network. The search is executed with a search depth of one node and a maximum diameter of three nodes (corresponding to a local search of "depth"=2 and "max depth"=2 with the jActiveModules plugin available for the Cytoscape Network Modeling package at http://www.cytoscape.org). In order to reduce the influence of network topology on the significance of our final result, we employed a "neighborhood scoring" method [17]. In this method, the search procedure is required to add either all or no node neighbors at each step in the search process. This prevents the selection of a few highly scoring adjacent nodes in a large neighborhood. In order to assess the significance of our result, the search was repeated one-hundred times with the assignment of expression significance values to proteins randomly permuted in each trial. The top scoring result from each of these trials was retained. Those networks which scored higher than 95% of these retained networks were considered significant. To produce smaller subnetworks for visualization in Figures 1 and 2 we repeated the Active Modules search within each original subnetwork with a local search of "depth"=1 and "max depth"=1 to identify singleton nodes which had a significant number of neighbors with differential expression.

2.4. *GO ontology analysis*

We utilized the BiNGO plugin for Cytoscape to determine which Gene Ontology (GO) Molecular Function and Biological Process categories are statistically over-represented in a set of genes [18]. We applied a hypergeometric test to determine which categories were significantly represented (p-value cutoff of 0.01). This significance value was adjusted for multiple hypothesis testing using the Bonferroni Family-wise error rate correction. Only those over-represented terms present at the 8^{th} level of the GO hierarchy were reported.

3. Results

3.1. *An integrated network of protein – protein interactions.*

We sought to elucidate mechanisms of HIV-1 latency and reactivation through integration of biological networks based on literature with measurements of cellular gene expression. Recently, the curation efforts of the Human Protein Reference Database (HPRD) have produced high confidence protein-protein interaction networks derived from literature on a scale such that systems biology based modeling is possible [15]. Additionally, the NCBI has produced a HIV-1-human protein interaction database, providing a summary of known interactions of HIV-1 proteins with those of the host cell [18]. The human protein-protein interaction data consisted of 17,558 interactions among 6,050 genes. HIV-human interactions consisted of 2,420 total interactions over 796 human genes. Representing HIV-1 and human proteins as nodes and interactions between those nodes as edges, we constructed an integrated network which summarizes the corpus of literature-based knowledge about human and HIV interactions.

3.2. *Discovering regulated subnetworks*

The study of Krishnan and Zeichner [6] assayed cellular gene expression of human cell lines chronically infected with HIV, before and during activation of the lytic viral replication cycle. In this study, latently infected ACH-2 cells (derived from a human T-cell line) were treated with phorbol myristyl acetate (PMA), which induces the lytic replication cycle. The changes in cellular gene expression were assayed and compared to uninfected cells exposed to equal amounts of PMA. The authors defined several time points which correspond to different stages of the lytic replication cycle (early, intermediate, late) by comparative analysis of spliced to unspliced mRNAs and cell viability. Their analysis of these stages indicates significant changes of host

gene expression in latently infected cells as compared to uninfected cells, as well as systematic, synchronous changes of gene expression in reactivated cells compared to uninfected controls.

To further characterize expression changes at various stages of the viral reactivation cycle, we used an integrated approach of expression clustering and network analysis to find "activated modules" of connected proteins with significant levels of differential activity. We identified highly significant subnetworks for both the latent (uninduced) and early (up to 8 hour post-induction) stages. Both the intermediate and late stages did not produce significant networks ($P<0.05$), which may be due to any combination of factors, notably loss of synchronization, the broad effects of cytopathicity, lack of adequate interaction data or this particular grouping of time points into phases.

In the uninduced stage of the HIV latent infection we found a single active network of 116 Tat-interacting proteins with a score of 10.9 ($P<0.01$). The overview in Figure 1 shows the active subnetwork with differentially expressed neighbors ($P<0.05$) and all HIV interactions removed for clarity. The network was significantly enriched for proteins associated with various aspects of HIV replication, including genes involved in apoptosis and cell death regulation [19] (see Table 1). To condense the network and facilitate interpretation, we ran the algorithm again on the significant subnetwork to find local regions of significant differential expression. The top five modules from this analysis are shown in Figure 1 (a-d). In this stage there is significant down regulation of collagen and fibronectin associated genes (Figure 1c). These are mostly upregulated by Tat [20] during the intermediate stage of the lytic cycle (data not shown) corresponding to their roles in cell-cell adhesion. In the tubulin associated network (Figure 1a), TubA3 interacts with multiple differentially expressed genes. TubA3 expression levels were not available, but due to its in interactions with both Rev and Tat (Rev acts to depolymerize microtubules that are formed by tubulin [21] and Tat binds tubulin[22]) and other differentially expressed neighbors, one can infer its role in the maintenance of the HIV latent phase.

In the early stages of HIV reactivation we found a single active network with a score of 10.7 ($P<0.01$) composed of 79 proteins which all interact with Tat. The subnetwork was enriched for proteins involved in transcription. The apparent importance of this process is consistent with the considerable transcription of integrated viral genes which occurs at this state of viral reactivation (Table 1). The overview Figure 2 shows the active subnetwork with differentially expressed neighbors ($P<0.01$) and all HIV links removed for clarity. We ran the active modules algorithm again and returned the top five proteins within the module which had significant numbers of differentially