Comparative analysis of protein interaction networks reveals that conserved pathways are susceptible to HIV-1 interception

Xiaoning Qian*[1], Byung-Jun Yoon*[2]

[1]Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA
[2]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA

Email: Xiaoning Qian*- xqian@cse.usf.edu; Byung-Jun Yoon*- bjyoon@ece.tamu.edu;

*Corresponding author

Abstract

Background: Human immunodeficiency virus type one (HIV-1) is the major pathogen that causes the acquired immune deficiency syndrome (AIDS). With the availability of large-scale protein-protein interaction (PPI) measurements, comparative network analysis can provide a promising way to study the host-virus interactions and their functional significance in the pathogenesis of AIDS. Until now, there have been a large number of HIV studies based on various animal models. In this paper, we present a novel framework for studying the host-HIV interactions through comparative network analysis across different species.

Results: Based on the proposed framework, we test our hypothesis that HIV-1 attacks essential biological pathways that are conserved across species. We selected the *Homo sapiens* and *Mus musculus* PPI networks with the largest coverage among the PPI networks that are available from public databases. By using a local network alignment algorithm based on hidden Markov models (HMMs), we first identified the pathways that are conserved in both networks. Next, we analyzed the HIV-1 susceptibility of these pathways, in comparison with random pathways in the human PPI network. Our analysis shows that the conserved pathways have a significantly higher probability of being intercepted by HIV-1. Furthermore, Gene Ontology (GO) enrichment analysis shows that most of the enriched GO terms are related to signal transduction, which has been conjectured to be one of the major mechanisms targeted by HIV-1 for the takeover of the host cell.

Conclusions: This proof-of-concept study clearly shows that the comparative analysis of PPI networks across different species can provide important insights into the host-HIV interactions and the detailed mechanisms of HIV-1. We expect that comparative multiple network analysis of various species that have different levels of susceptibility to similar lentiviruses may provide a very effective framework for generating novel, and experimentally verifiable, hypotheses on the mechanisms of HIV-1. We believe that the proposed framework has the potential to expedite the elucidation of the important mechanisms of HIV-1, and ultimately, the discovery of novel anti-HIV drugs.

Background

Acquired immune deficiency syndrome (AIDS), one of the most destructive pandemics in recorded history according to the statistics from the World Health Organization (WHO) [1], has killed more than 25 million people since it was first recognized in 1981. Human immunodeficiency virus type one (HIV-1) has been found to be the causative pathogen of AIDS [2, 3]. HIV-1 is a lentivirus, a slow retrovirus that is responsible for long-duration illness with a long incubation period. HIV-1 has 9 genes which encode up to 19 proteins due to post-translational cleavage [4]. By reverse transcription from viral RNA to host-integrable DNA, the virus can become active and replicate to cause rapid T cell depletion, immune system collapse, and opportunistic infections that mark the advent of AIDS [5].

Although advances in antiviral therapy and management of opportunistic infection for AIDS have remarkably improved the general health, the expensive cost and adverse effects of the available drugs have motivated many researchers to explore novel avenues to anti-HIV-1 drug discovery. With the increasing coverage of HIV-1 and human protein interactions in the literature [6–11], a human/HIV-1 interactome has been created [12], which can play a critical role in better understanding the virology and pathology of this infectious disease and developing new therapeutics. In addition to this, the availability of large-scale biological networks, including protein-protein interaction (PPI) networks, has led to the introduction of systems biology approaches for novel HIV-1 drug discovery [13, 14]. In [13], Balakrishnan et al. proposed to find alternative pathways to circumvent the HIV-1 intercepted pathways based on the efficiency and robustness of biological processes. The main goal was to generate new hypotheses regarding HIV-1

targeted pathways and their effects on various molecular functions, which will help us better understand the mechanisms of HIV-1 takeover of the host cell and find ways to circumvent it. The study was based on curated signal transduction pathways obtained from multiple pathway databases. One practical problem of this pathway-based approach is that the currently known pathways cover only a limited number of human proteins, hence it may exclude important HIV-1 targets from the analysis. Moreover, many curated pathways in public databases overlap with each other, which may introduce bias in the analysis. On the other hand, Lin et al. [14] proposed comparative studies of host-virus protein interactions across human (*Homo sapiens*) and other animal models that may be invaded by similar lentiviruses that cause immunosuppression or immunoproliferation, including three mammalian species: chimpanzee (*Pan troglodyte*), rhesus macaque (*Macaca mulatta*), and mouse (*Mus musculus*). All these animal models have been extensively studied to understand the HIV-1 host-virus interplay [15, 16]. Comparative studies of host-virus interactions may provide new insights into why different species have different susceptibility to HIV-1, which may lead to the development of potential therapeutics in the long run.

Motivated by these works, we propose a novel framework for studying human/HIV-1 interactions, based on comparative analysis of the human PPI network and the PPI network of other species that are susceptible to lentivirus invasion. It has been shown that the comparative analysis of PPI networks of different species can identify conserved pathways that carry essential cellular functionalities [17–36]. Furthermore, HIV-1 has to be a "minimalist" in order to survive, and for this reason, it has been believed to target these essential pathways that are conserved across species [13, 37]. As a result, the comparative analysis of PPI networks may be used to generate new hypotheses that will be useful in improving our understanding of the mechanisms of HIV-1 takeover of the host cell, and ultimately, for developing effective therapeutics for AIDS.

Results and Discussion

Conserved pathways are susceptible to HIV-1 attacks

Our main goal in this paper is to validate the following hypothesis:

"Essential biological pathways that are conserved across different species that are susceptible to lentiviruses, have a high probability of being intercepted by HIV-1."

To validate the above hypothesis, we first aligned the *Homo sapiens* PPI network with the *Mus musculus* PPI network to find conserved pathways (Figure 1). The *Mus musculus* network was chosen as it is the

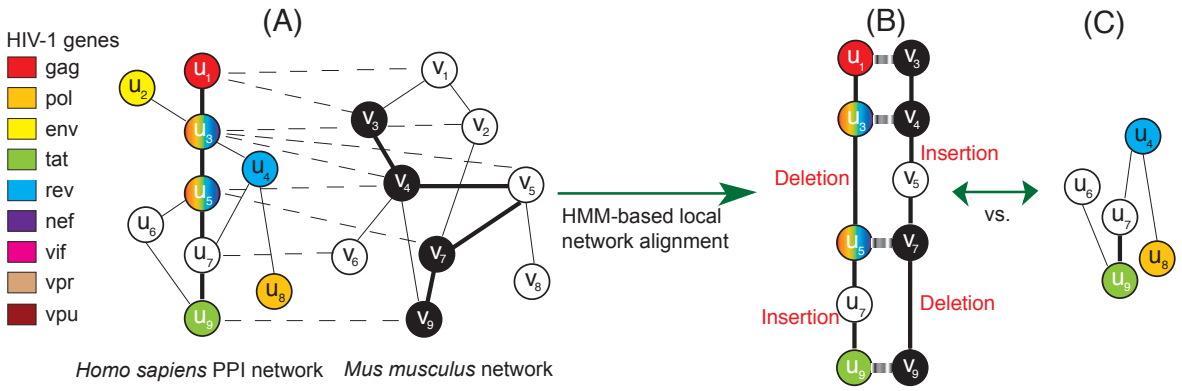

Figure 1: Overview of the proposed approach: (A) Illustration of the *Homo sapiens* and *Mus musculus* PPI networks along with HIV-1 interactions. The dashed line that connects two nodes u_i and v_j indicate that the corresponding proteins are orthologous. The solid lines represent protein-protein interactions. In the *Homo sapiens* network, the proteins are colored based on the HIV-1 proteins that can bind to them. Proteins with multiple colors are susceptible to multiple HIV-1 proteins, while proteins with no color have no known interactions with HIV-1 proteins. Note that the *Mus musculus* network is not colored. (B) The top-scoring alignment between two similar paths $\mathbf{u}$ and $\mathbf{v}$. Colored nodes represent matched proteins. (C) An example of a randomly extracted pathway in the *Homo sapiens* network.

largest animal network under lentivirus study that is available from public databases. This will allow us to reduce the bias that may arise from using smaller networks. For identifying conserved pathways, we used a local network alignment method based on hidden Markov models (HMMs), which we recently proposed in [36]. The HMM framework can naturally integrate both the "sequence similarity" of the proteins across different species and the "interaction reliability" between the proteins within the same PPI network into the scoring scheme for finding conserved pathways. The HMM-based local alignment method allows flexible number of consecutive insertions and/or deletions, and it can deal with a large class of path isomorphism. More importantly, the computational complexity for finding the best matching pathways grows linearly with respect to the size of each network, making it suitable for finding long conserved pathways in large PPI networks. We ran the HMM-based local alignment algorithm to find conserved pathways both with and without gaps. Since the *Mus musculus* network is still quite sparse (see **Methods**), the number of conserved pathways depends on whether we allow gaps and how long the pathways can be. Typically, we have fewer conserved pathways when we search for long pathways with no gaps allowed.

Next, we extracted $3,000$ random pathways of different sizes ($L = 16$, 32, and 64) by performing a random walk on the *Homo sapiens* network (see **Methods**). Then we compared the HIV-1 susceptibility of the conserved pathways with that of the random pathways in the *Homo sapiens* PPI network, by using the

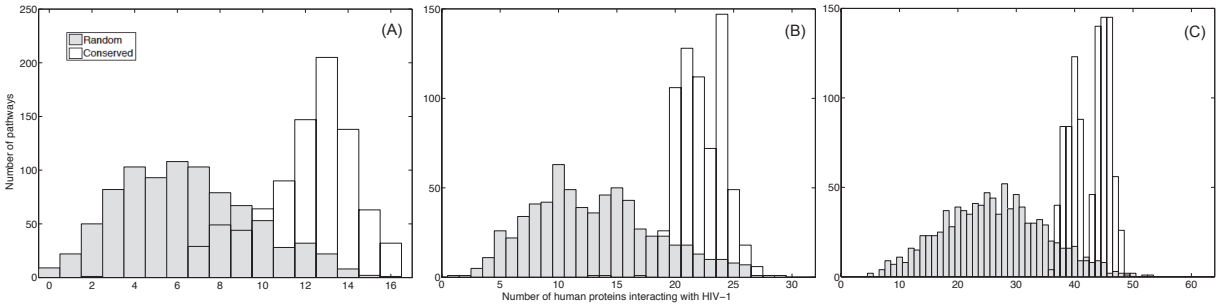

Figure 2: The number of proteins that interact with HIV-1 proteins in conserved pathways (with no gaps) and randomly extracted pathways. (A) The histograms for pathways of size $L = 16$. (B) The histograms for pathways of size $L = 32$. (C) The histograms for pathways of size $L = 64$.

predicted human/HIV-1 interactome map in [12].

Number of proteins in the pathways that are intercepted by HIV-1

Based on the identified conserved pathways and the randomly extracted pathways, we first computed how many proteins within each pathway can be intercepted by HIV-1, by mapping the predicted human/HIV-1 interactions in [12] onto these pathways. Figure 2 shows the histogram of the number of HIV-1 interacting proteins in conserved pathways as well as the histogram for random pathways. From the figure, there is a clear distinction between the two types of histograms. Typically, the separation between the two histograms increases with the length of the pathways. We can clearly see that highly conserved pathways are more susceptible to HIV-1 interception, in general.

Next, we considered conserved pathways that have been identified by the HMM-based local network alignment algorithm by allowing gaps. We compared the histograms for these pathways to the histograms for random pathways. These results are shown in Fig. 3. We can see that the histograms show similar trends as in Fig. 2, but the separation between the two types of histograms is smaller in this case. This might be due to the fact that the HMM-based algorithm may find less conserved pathways when we allow gaps.

To evaluate the statistical significance of the difference between conserved and random pathways, we estimated the p-value of the number of HIV-1 interacting proteins for every conserved pathway. The detailed process for computing the p-values is described in **Methods**. Basically, it computes the statistical significance of the number of HIV-1 interacting proteins in each conserved pathway with respect to the

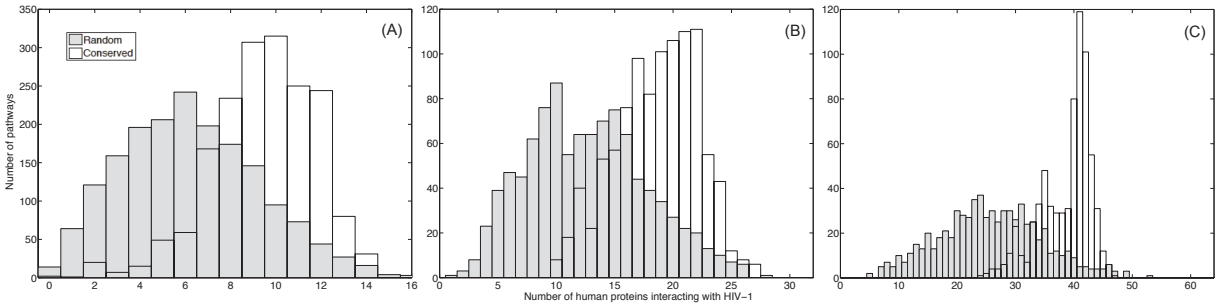

Figure 3: The number of proteins that interact with HIV-1 proteins in conserved pathways (with gaps) and randomly extracted pathways. (A) The histograms for pathways of size $L = 16$. (B) The histograms for pathways of size $L = 32$. (C) The histograms for pathways of size $L = 64$.

baseline distribution, which is estimated from the histogram of the numbers of HIV-1 interacting proteins in randomly extracted pathways. Figure 4(A) shows the plot of p-values for conserved pathways of various lengths and with no gaps allowed. Conserved pathways with gaps show a similar trend (results not shown). As we mentioned earlier, the number of conserved pathways decrease as the pathway size L gets larger. In the figure, the conserved pathways are sorted based on their alignment scores computed by the HMM-based local alignment method [36] (see **Methods**). The alignment score reflects the degree of conservation between the aligned pathways. We can see that highly conserved pathways are generally more susceptible to HIV-1 interception. In fact, such pathways typically contain more proteins that can be intercepted by HIV-1 proteins.

Total number of human/HIV-1 protein interactions within one pathway

To further validate our hypothesis, we checked the total number of human/HIV-1 interactions within each pathway. Again, we used the predicted human/HIV-1 interactome in [12] to count the total number of human/HIV-1 interactions within each pathway. Figure 4(B) shows the p-value of the total number of HIV-1 interactions within every conserved pathway (with no gaps). As before, the baseline distribution was estimated using the histogram of the total numbers of HIV-1 interactions in randomly extracted pathways. Note that, for long conserved pathways ($L = 64$), the p-values are always below 0.03. These results show that the difference in susceptibility to HIV-1 interception between the conserved and random pathways is statistically significant.

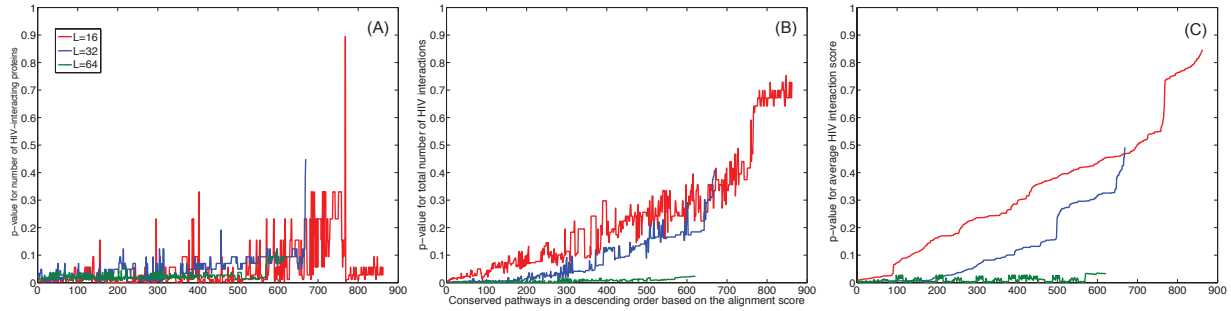

Figure 4: Statistical significance of the interactions between HIV-1 and the conserved pathways (with no gaps). (A) The p-values of the number of human proteins that interact with HIV-1 proteins within conserved pathways with different sizes ($L = 16$: red, $L = 32$: blue, $L = 64$: green). (B) The p-values of the total number of predicted human/HIV-1 interactions within conserved pathways. (C) The p-values for the average predicted interaction scores within conserved pathways.

HIV-1 interaction score of conserved pathways

Finally, we evaluated the HIV-1 interaction score for conserved pathways based on the scoring scheme in [12]. For this evaluation, we mapped the prediction scores of human/HIV-1 interactions onto the conserved pathways and computed their average. In a similar way, we computed the average HIV-1 interaction score for each random pathway and estimated the distribution of these average scores. Then we computed the p-values of the average prediction scores for all the conserved pathways. These p-values are shown in Fig. 4(C) for the conserved pathways with different lengths. By comparing the results in Fig. 4(C) and those shown in Fig. 4(A,B), we can clearly see that there exist considerable correlation between these results. This is especially interesting, if we consider the fact that our approach does not use any extra data except for the PPI networks, while the prediction algorithm in [12] is obtained by integrating the information from various sources, such as gene expression, domain and motif identification, tissue distribution, functional annotation, subcellular localization and human network features, and HIV-1's mimicry of human protein binding partners.

GO term enrichment analysis

We also performed a Gene Ontology (GO) term enrichment analysis [38] using GO::TermFinder [39]. We took the top 20 conserved pathways of size $L = 64$ and checked their GO terms. Table 1 shows some of the enriched GO terms, whose adjusted p-values are smaller than $2.0e - 7$.[1] Examples of highly enriched GO terms include: "signaling pathway", "signal transduction", "cell communication", "phosphate metabolic

[1]The complete enrichment analysis results can be found at `http://www.cse.usf.edu/~xqian/hiv/`.

process", "response to stimulus", "response to stress", "protein modification process", "regulation of immune system process", which are pathways that are widely known to be susceptible to HIV-1 interception [2, 3, 12, 13]. There are also more specific GO terms that are enriched, such as "hemopoietic or lymphoid organ development" and "lymphocyte proliferation". Interestingly, the pathogenesis of HIV-associated lymphomas has been conjectured to cause the complication of HIV infection as reported in [40]. From all the proteins covered by the top 20 conserved pathways, we have listed ten human proteins with the largest number of predicted HIV-1 interactions in Table 2. We also selected eight human proteins that are not known to be intercepted by HIV-1, which are shown in bold face in Table 2.. We have also listed their associated ontology keywords or GO terms based on the UniProt database.[2] We note that many of these proteins are related to the aforementioned biological processes and they might be unidentified targets of HIV-1. Further study on these proteins may lead to a better understanding of the biological mechanisms of HIV-1.

Results using curated human/HIV-1 protein interactions

In order to ensure that the obtained results are biologically meaningful, we further validate our hypothesis by testing the HIV-1 susceptibility using the curated human/HIV-1 protein interaction data in the Human Protein Interaction Database (HPID) [41]. These human/HIV-1 interactions were reported in the literature and curated by experts and can provide another supporting evidence that pathways which are conserved across species have a high probability of being attacked by HIV-1. As in our previous experiments, we counted the number of HIV-1 interacting proteins in conserved pathways (without allowing gaps) and compared it to the number of HIV-1 interacting proteins in random pathways. The resulting histograms are shown in Fig. 5. We can see that the histograms show similar trends as in Fig. 2, but the numbers of interacting human proteins are relatively smaller and the separation between the histograms of the conserved pathways and the random pathways is less significant. This is expected since the HPID curated interactions are sparser compared to the predicted interactions in [12] and smaller number of proteins in the *Homo sapiens* PPI network have been mapped with the HIV-1 interactions. We have also compared the total number of human/HIV-1 interactions in conserved pathways with that in random pathways. The resulting histograms are shown in Fig. 6. Figure 7 plots the computed p-values of both the number of the HIV-1 interacting proteins and the total number of human/HIV-1 interactions within every conserved

[2] `http://www.uniprot.org/`

Table 1: Selected GO terms enriched in the top 20 conserved pathways of size $L = 64$ with adjusted p-values.

Gene Ontology terms	Adjusted p-values
signaling pathway	1.39e-49
signaling	4.39e-47
signal transduction	1.97e-42
regulation of cellular process	1.14e-41
signal transmission	4.46e-41
signaling process	4.94e-41
regulation of biological process	1.46e-39
biological regulation	3.09e-37
intracellular signaling pathway	7.57e-37
intracellular signal transduction	2.56e-35
cell proliferation	8.71e-34
phosphate metabolic process	2.86e-33
system development	3.63e-31
enzyme linked receptor protein signaling pathway	9.49e-31
developmental process	6.77e-30
anatomical structure development	1.73e-29
organ development	2.16e-29
cell surface receptor linked signaling pathway	1.23e-28
response to endogenous stimulus	7.55e-27
cellular response to stimulus	3.96e-26
response to stimulus	4.21e-25
protein modification process	2.11e-24
regulation of metabolic process	1.12e-23
response to hormone stimulus	1.38e-21
cell communication	6.28e-21
regulation of biosynthetic process	1.55e-15
Ras protein signal transduction	2.85e-14
response to stress	1.65e-13
RNA biosynthetic process	5.10e-13
cellular macromolecule biosynthetic process	5.49e-13
regulation of transferase activity	1.15e-12
immune system development	2.09e-12
regulation of immune system process	2.26e-12
hemopoietic or lymphoid organ development	9.12e-12
hemopoiesis	4.55e-11
neurogenesis	5.75e-08
leukocyte differentiation	6.68e-08
lymphocyte proliferation	1.43e-07

Table 2: UniProt accession numbers of selected proteins in the top 20 conserved pathways of size $L = 64$ with protein names and the associated top ontology keywords and GO terms listed by `http://www.uniprot.org/`.

UniProt IDs	Protein names	Gene Ontology terms
P04637	Cellular tumor antigen p53	apoptosis; host-virus interaction; DNA damage response; protein tetramerization
P17612	cAMP-dependent protein kinase catalytic subunit α	hormone-mediated signaling pathway; intracellular protein kinase cascade
P28482	Mitogen-activated protein kinase 1	Ras protein signal transduction; cell cycle; transcription; interspecies interaction between organisms; chemotaxis; synaptic transmission
P27361	Mitogen-activated protein kinase 3	Ras protein signal transduction; interspecies interaction between organisms
P05412	Transcription factor AP-1	SMAD protein signal transduction; positive regulation by host of viral transcription; transforming growth factor (TGF) β receptor signaling pathway
P06241	Tyrosine-protein kinase Fyn	T cell receptor signaling pathway; interspecies interaction between organisms
P06493	Cell division protein kinase 1	anti-apoptosis; cell division; mitosis
Q15796	Mothers against decapentaplegic homolog 2	SMAD protein complex assembly; intracellular signaling pathway; palate development; transcription; TGF β receptor signaling pathway
P06400	Retinoblastoma-associated protein	Cell cycle; Host-virus interaction; androgen receptor signaling pathway; myoblast differentiation
P04049	RAF proto-oncogene serine/ threonine-protein kinase	Ras protein signal transduction; cell proliferation; protein amino acid phosphorylation
O14512	Suppressor of cytokine signaling 7	Ubl conjugation pathway; regulation of growth; negative regulation of signal transduction
P10721	Mast/stem cell growth factor receptor	male gonad development; transmembrane receptor protein tyrosine; kinase signaling pathway
P15884	Transcription factor 4	cerebral cortex development; regulation of smooth muscle cell proliferation; transcription
P16410	Cytotoxic T-lymphocyte protein 4	immune response; negative regulation of regulatory T cell differentiation
P29597	Non-receptor tyrosine-protein kinase TYK2	intracellular protein kinase cascade; peptidyl-tyrosine phosphorylation
Q13480	GRB2-associated-binding protein 1	cell proliferation; epidermal growth factor receptor signaling pathway; insulin receptor signaling pathway
Q15503	Son of sevenless homolog 2	apoptosis; regulation of Rho protein ; signal transduction; small GTPase mediated signal transduction
Q96TE0	Cdk inhibitor p27KIP1	cell cycle arrest

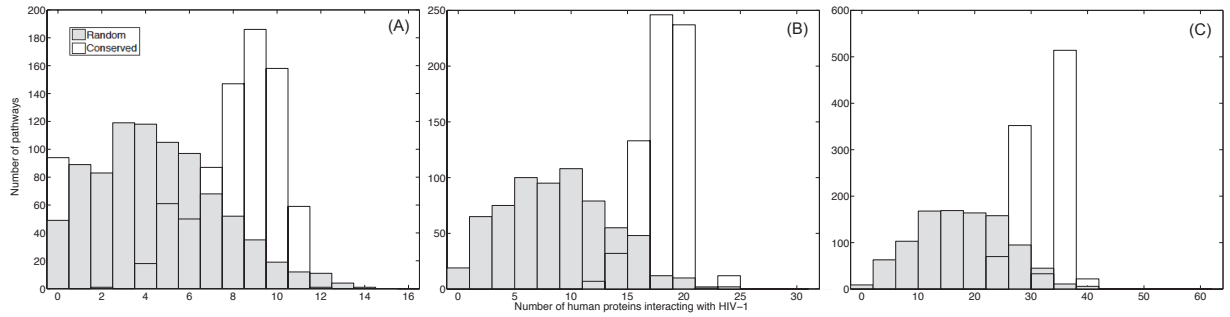

Figure 5: The number of proteins that interact with HIV-1 proteins based on the HPID interaction data in conserved pathways (with no gaps) and randomly extracted pathways. (A) The histograms for pathways of size $L = 16$. (B) The histograms for pathways of size $L = 32$. (C) The histograms for pathways of size $L = 64$.

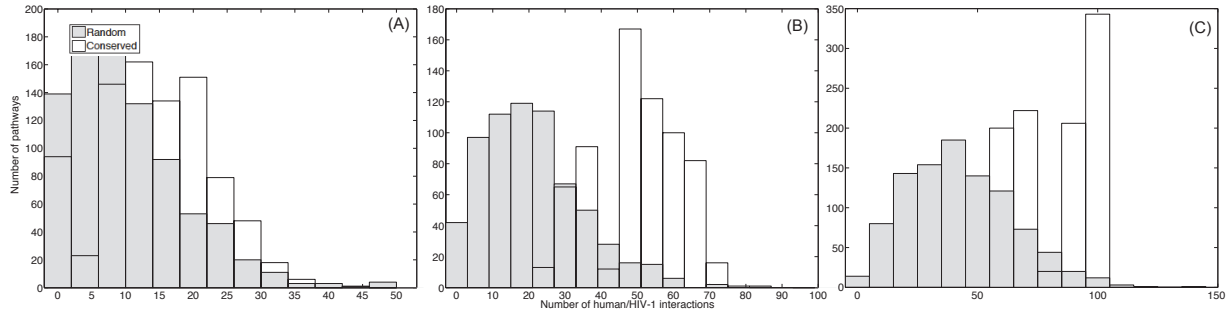

Figure 6: The total number of human/HIV-1 protein interactions based on the HPID interaction data in conserved pathways (with no gaps) and randomly extracted pathways. (A) The histograms for pathways of size $L = 16$. (B) The histograms for pathways of size $L = 32$. (C) The histograms for pathways of size $L = 64$.

pathway (with no gaps). Both results show that the predicted susceptibility of conserved pathways in the *Homo sapiens* PPI network and the *Mus musculus* PPI network to HIV-1 interception is statistically significant.

Conclusions

Local network alignment can effectively identify conserved pathways that are biologically meaningful [36]. If HIV-1 is a minimalist in order to survive and therefore targets essential pathways [13], as other viruses do, it is natural to expect that essential pathways that are conserved across different species should be highly vulnerable to HIV-1 attacks. Our analysis based on comparing the *Homo sapiens* PPI network and the *Mus musculus* PPI network indicates that our conjecture is indeed true. This proof-of-concept study

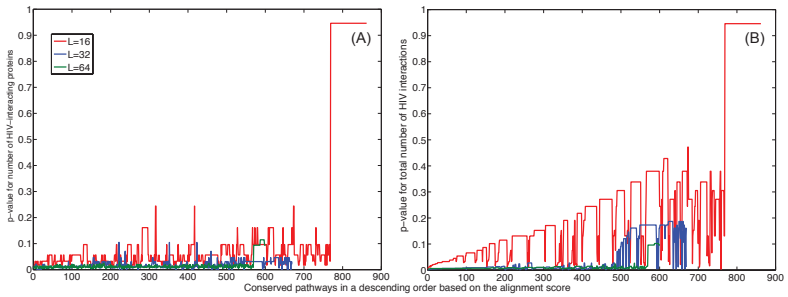

Figure 7: Statistical significance of the interactions between HIV-1 and the conserved pathways (with no gaps) according to the curated interactions in HPID. (A) The p-values of the number of human proteins that interact with HIV-1 proteins within conserved pathways with different sizes ($L = 16$: red, $L = 32$: blue, $L = 64$: green). (B) The p-values of the total number of predicted human/HIV-1 interactions within conserved pathways.

that we present clearly shows that the comparative network analysis of different species can provide important insights into the mechanisms of human/HIV-1 interactions. We believe that further studies based on aligning the networks of various species that are susceptible to similar lentiviruses will lead to breakthroughs in HIV research. For example, although chimpanzees are the human's closest relative in nature, AIDS is rarely life-threatening to them [14]. Identifying the main reasons for this difference in HIV-1 susceptibility may lead to the development of novel therapeutics for this highly destructive disease. Balakrishnan et al. [13] proposed a heuristic way to search for alternative pathways that can circumvent HIV-intercepted pathways, whose ultimate goal is to identify potential drug targets. In a similar way, comparative network analysis may also be used to identify alternative pathways in the PPI network, by querying known HIV-intercepted pathways in the human PPI network. Although comparative network analysis is still at an early stage and is not yet as mature as comparative sequence analysis, it can take direct advantage of the large-scale interaction measurements that have become available these days and it has the potential to generate experimentally verifiable hypotheses on the biological mechanisms of HIV-1, which may lead to the identification of better drug targets and innovative AIDS therapeutics in the future.

Methods

Protein-protein interaction (PPI) networks

We have obtained both the *Homo sapiens* and *Mus musculus* protein-protein interaction (PPI) networks from the open platform `NATALIE`[3], managed by the Knowledge Management in Bioinformatics group of the Humboldt-Universität Berlin. Both networks were obtained from several open databases [42–47] as

[3] `https://www.mi.fu-berlin.de/wiki/pub/LiSA/Natalie/natalie-0.9.tgz`

described in [24, 48]. The *Homo sapiens* network has 34,979 interactions among 9,695 proteins, and the *Mus musculus* network has 3,116 interactions among 3,247 proteins. The similarity between proteins in the two networks were determined based on protein sequences, protein domain information (InterPro domains) and functional annotations (GO annotations) [48–51]. Pairs of similar proteins in the two networks were identified based on a minimum protein identity threshold of $\alpha = 0.4$ as in [24, 48].

Human-HIV interaction data

As mentioned in [13], there are several types of interactions between HIV-1 proteins and human proteins, including direct physical interactions that are reported in the literature, indirect interactions reported in the literature, and interactions that have been manually annotated by experts [12, 41]. However, many HIV-1 virologists do not agree upon the majority of these interactions [13]. For this reason, we focus on the human/HIV-1 interactome in [12] in our analysis, which has been computationally predicted by integrating various types of protein features. The human/HIV-1 interactome data can be obtained from the authors' website.[4] For further validation, we also performed similar analysis based on the curated human/HIV-1 protein interactions in HPID [41].

Identification of conserved pathways through local network alignment

To align the *Homo sapiens* and *Mus musculus* PPI networks, we used the local network alignment algorithm based on hidden Markov models (HMMs) [35, 36]. Let us represent the *Homo sapiens* and *Mus musculus* PPI networks as $\mathcal{G}_h = (\mathcal{U}, \mathcal{D})$ and $\mathcal{G}_m = (\mathcal{V}, \mathcal{E})$, respectively. $\mathcal{U} = \{u_i\}$ and $\mathcal{V} = \{v_i\}$ represent the sets of nodes, where each node represents a protein in a given network. $\mathcal{D} = \{d_{ij}\}$ and $\mathcal{E} = \{e_{ij}\}$ are the sets of edges, where each edge represents the interaction between the connected proteins. For the orthologous protein pairs in the *Homo sapiens* and *Mus musculus* PPI network that have been predicted based on the method in [48], we define the node similarity score $s(u, v)$ between a pair of proteins $u \in \mathcal{U}$ and $v \in \mathcal{V}$ as follows:

$$s(u,v) = \begin{cases} 1, & \text{if } u \text{ and } v \text{ are orthologous} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The interaction reliability scores between two nodes are binary for both PPI networks. For example, the interaction score $w_h(u_i, u_j)$ between u_i and u_j in the *Homo sapiens* network is defined as:

$$w_h(u_i,u_j) = \begin{cases} 0, & \text{if interaction between } u_i \text{ and } u_j \text{ exists;} \\ -\infty, & \text{otherwise.} \end{cases} \tag{2}$$

[4] `http://www.cs.cmu.edu/~oznur/hiv/hivPPI.html`

The interaction reliability score $w_m(v_i, v_j)$ between v_i and v_j in the *Mus musculus* network is defined in a similar way.

In order to find the pathways that are conserved in both PPI networks, we first search for the best matching pair of paths $\mathbf{u} = u_1 u_2 \dots u_L$ ($u_i \in \mathcal{U}$) and $\mathbf{v} = v_1 v_2 \dots v_L$ ($v_i \in \mathcal{V}$) in the respective networks (Fig. 1) that maximizes a predefined pathway alignment score $S(\mathbf{u}, \mathbf{v})$. The pathway alignment score $S(\mathbf{u}, \mathbf{v})$ integrates the *similarity score* $s(u_i, v_i)$ between the aligned nodes u_i and v_i ($1 \leq i \leq L$), the *interaction reliability score* $w_h(u_i, u_{i+1})$ between u_i and u_{i+1} ($1 \leq i \leq L-1$), the interaction reliability score $w_m(v_j, v_{j+1})$ between v_j and v_{j+1} ($1 \leq j \leq L-1$), and the penalty for potential gaps in the alignment. We denote the number of nodes in a pathway as L. In this paper, we search for conserved pathways of size $L = 16$, 32, 64. Based on the HMM framework [35, 36], we transform the problem of "finding the best matching pair of paths" to a problem of "finding the optimal pair of state sequences in the two HMMs" that jointly maximize the observation probability of a virtual path (see Fig. 1). We use two different types of settings for finding conserved pathways, where we do not allow gaps in the pathway alignment in one setting while we allow gaps in the other setting. In general, the two setting will yield different predictions.
We can find the best matching pair of paths in the given networks using dynamic programming. For this purpose, we first define the score for the most probable pair of a subsequence paths of length $t\,(\leq L)$ as follows:

$$\gamma(t, j, \ell) = \max_{i,k} \Big[\gamma(t-1, i, k) + w_h(u_i, u_j) + w_m(v_k, v_\ell) + s(u_j, v_\ell)\Big]. \tag{3}$$

Next, we find the optimal pair of paths $(\mathbf{u}^*, \mathbf{v}^*)$

$$S(\mathbf{u}^*, \mathbf{v}^*) \quad = \quad \max_{\mathbf{u},\mathbf{v}} \Big[S(\mathbf{u}, \mathbf{v})\Big] = \max_{j,\ell} \gamma(L, j, \ell), \tag{4}$$

by computing the score (3) iteratively. As discussed in [35, 36], we can add auxiliary states to the HMMs that represent the PPI networks to find gapped path alignments. Instead of finding only the best matching pair of paths, we can also search for the top k path pairs by replacing the max operator in (4) and (3) by an operator that finds the k largest scores. The computational complexity of the described dynamic programming algorithm is only $O(kLM_1M_2)$ for finding the top k pairs of matching paths, where M_1 is the number of edges (i.e., interactions) in $\mathcal{G}_h$, and M_2 is the number of edges in $\mathcal{G}_m$. Note that the computational complexity is linear with respect to each parameter k, L, M_1, and M_2. To avoid multiple occurrences of the same protein in the conserved pathways that are predicted by the algorithm, we incorporate a "look-back" step into each iteration of the dynamic programming algorithm [36].

Extraction of random pathways

In order to extract random pathways from the *Homo sapiens* PPI network, we performed random walks on the network starting from a randomly selected node in network G_h. We randomly walk on the network to choose a random pathway, until the size of the pathway reaches a pre-specified size L. During this random walk, we avoid visiting a node that has been previously visited, so that the extracted random pathway contains only distinct nodes.

Comparison between conserved pathways and random pathways

To compare the HIV-1 susceptibility of conserved pathways with that of random pathways, we computed the following values: (1) The *number of proteins* within these pathways that have been predicted to be intercepted by at least one of the HIV-1 proteins according to the human/HIV-1 interactome in [12]. (2) The *total number of predicted human/HIV-1 protein interactions* within these pathways; (3) The *average HIV interaction scores* within pathways. We also computed the p-values of the estimated results for conserved pathways, according to the process described in the next subsection. Finally, for GO term enrichment analysis, we used an open source software called the GO::TermFinder [39].

Computing p-values

In order to evaluate the statistical significance of the estimated results in conserved pathways, we first extract a large number of random pathways ($3,000$) from the *Homo sapiens* PPI network, based on random walk. For each random pathway, we also estimate the indices of HIV-1 susceptibility (i.e., the number of human proteins intercepted by HIV-1; the total number of human/HIV-1 interactions, the average interaction scores among the proteins in each pathway). Baseline distributions of different indices are estimated from these results. We can either model the baseline distributions using Gumbel distributions [52] or simply use histograms. The latter approach was adopted in this paper. Once we have estimated the baseline distributions, we can compute the p-values of the estimated results in conserved pathways according to the estimated distributions.

Authors contributions

Conceived and designed the experiments: XQ, BJY. Performed the experiments: XQ. Analyzed the results: XQ, BJY. Wrote the paper: XQ, BJY.

Acknowledgements

XQ was supported in part by the University of South Florida Internal Awards Program under Grant No. 78068. BJY was supported in part by the Texas A&M Faculty start-up fund.

References

1. **Joint United Nations Programme on HIV/AIDS: Overview of the global AIDS epidemic**. *2006 Report on the global AIDS epidemic. ISBN 9291734799* 2006.
2. Weiss R: **How does HIV cause AIDS?** *Science* 1993, **260**(5112):1273–1279.
3. Douek D, Roederer M, Koup R: **Emerging concepts in the immunopathogenesis of AIDS**. *Annu Rev Med* 2009, **60**:471–484.
4. **HIV Sequence Compendium 2008 Introduction**. *Various* 2008, [http://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/2008/frontmatter.pdf].
5. Moore J: **Coreceptors: implications for HIV pathogenesis and therapy**. *Science* 1997, **276**(5309):51–52.
6. Fu W, Sanders-Beer B, Katz K, Maglott D, KD KP, Ptak R: **Human immunodeficiency virus type 1, human protein interaction database at NCBI**. *Nucleic Acids Res* 2009, **37**(Database):D417–422.
7. Ptak R, Fu W, Sanders-Beer B, Dickerson J, Pinney J, Robertson D, Rozanov M, Katz K, Maglott D, Pruitt K, et al: **Cataloguing the HIV type 1 human protein interaction network**. *AIDS Res Hum Retroviruses* 2008, **24**(12):1497–1502.
8. Brass A, Dykxhoorn D, Benita Y, Yan N, Engelman A, Xavier R, Lieberman J, Elledge S: **Identification of host proteins required for HIV infection through a functional genomic screen**. *Science* 2008, **319**(5865):921–926.
9. Konig R, Zhou Y, Elleder D, Diamond T, Bonamy G, Irelan J, Chiang C, Tu B, Jesus PD, Lilley C, et al: **Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication**. *Cell* 2008, **135**:49–60.
10. Zhou H, Xu M, Huang Q, Gates A, Zhang X, Castle J, Stec E, Ferrer M, Strulovici B, Hazuda D, et al: **Genome-scale RNAi screen for host factors required for HIV replication**. *Cell Host Microbe* 2008, **4**(5):495–504.
11. Pinney J, Dickerson J, Fu W, Sanders-Beer B, Ptak R, Robertson D: **HIV-host interactions: a map of viral perturbation of the host system**. *AIDS* 2009.
12. Tastan O, Qi Y, Carbonell J, J JKS: **Prediction of interactions between HIV-1 and human proteins by information integration**. In *Pac Symp Biocomput, Volume 14* 2009:516–527.
13. Balakrishnan S, Tastan O, Carbonell J, Klein-Seetharaman J: **Alternative paths in HIV-1 targeted human signal transduction pathways**. *BMC Genomics* 2009, **10**(Suppl 3):S30.
14. Lin F, Pan C, Yang J, Chuang T, Chen F: **CAPIH: A web interface for comparative analyses and visualization of host-HIV protein-protein interactions**. *BMC Microbiology* 2009, **9**(164):10 pages.
15. McCune J: **AIDS RESEARCH: Animal models of HIV-1 disease**. *Science* 1997, **278**(5346):2141–2142.
16. Zink M, Laast V, Helke K, Brice A, Barber S, Clements J, Mankowski J: **From mice to macaques–animal models of HIV nervous system disease**. *Curr HIV Res* 2006, **4**(3):293–305.
17. Kelley B, Sharan R, Karp R, Sittler T, Root D, Stockwell B, Ideker T: **Conserved pathways within bacteria and yeast as revealed by global protein network alignment**. *Proc Natl Acad Sci USA* 2003, **100**(20):11394–11399.
18. Sharan R, Ideker T: **Modeling cellular machinery through biological network comparison**. *Nat Biotechnol* 2006, **24**:427–433.
19. Ideker T, Sharan R: **Protein networks in disease**. *Genome Research* 2008, **18**:644–652.
20. Flannick J, Novak A, Srinivasan B, McAdams H, Batzoglou S: **Græmlin: general and robust alignment of multiple large interaction networks**. *Genome Res* 2006, **16**(9):1169–1181.

21. Li Z, Zhang S, Wang Y, Zhang X, Chen L: **Alignment of molecular networks by integer quadratic programming**. *Bioinformatics* 2007, **23**(13):1631–1639.
22. Kalaev M, Bafna V, Sharan R: **Fast and accurate alignment of multiple protein networks**. In *Proc of the 10th Annu Int Conf Res Comput Mol Bio (RECOMB 2008)* 2008.
23. Singh R, Xu J, Berger B: **Global alignment of multiple protein interaction networks with application to functional orthology detection**. *Proc Natl Acad Sci USA* 2008, **105**(35):12763–12768.
24. Klau G: **A new graph-based method for pairwise global network alignment**. *BMC Bioinformatics* 2009, **10**(Suppl 1):S59.
25. Zaslavskiy M, Bach F, Vert J: **Global alignment of protein-protein interaction networks by graph matching methods**. *Bioinformatics* 2009, **25**:259–267.
26. Liao C, Lu K, Baym M, Singh R, Berger B: **IsoRankN: spectral methods for global alignment of multiple protein networks**. *Bioinformatics* 2009, **25**:253–258.
27. Tian W, Samatova N: **Pairwise alignment of interaction networks by fast identification of maximal conserved patterns**. In *Pac Symp Biocomput, Volume 14* 2009:99–110.
28. Steffen M, Petti A, Aach J, D'haeseleer P, Church G: **Automated modelling of signal transduction networks**. *BMC Bioinformatics* 2002, **3**:34.
29. Koyut*ü*rk M, Grama A, Szpankowski W: **An efficient algorithm for detecting frequent subgraphs in biological networks**. *Bioinformatics* 2004, **20**:SI200–207.
30. Pinter R, Rokhlenko O, Yeger-Lotem E, Ziv-Ukelson M: **Alignment of metabolic pathways**. *Bioinformatics* 2005, **21**(16):3401–3408.
31. Sharan R, Suthram S, Kelley R, Kuhn T, McCuine S, Uetz P, Sittler T, Karp R, Ideker T: **Conserved patterns of protein interaction in multiple species**. *Proc Natl Acad Sci USA* 2005, **102**(6):1974–1979.
32. Scott J, Ideker T, Karp R, Sharan R: **Efficient algorithms for detecting signaling pathways in protein interaction networks**. *J Comput Biol* 2006, **13**:133–144.
33. Shlomi T, Segal D, Ruppin E, Sharan R: **QPath: a method for querying pathways in a protein-protein interaction network**. *BMC Bioinformatics* 2006, **7**(199).
34. Yang Q, Sze S: **Path matching and graph matching in biological networks**. *J Comput Biol* 2007, **14**:56–67.
35. Qian X, Sze S, Yoon B: **Querying pathways in protein interaction networks based on hidden Markov models**. *J Comput Biol* 2009, **16**(2):145–157.
36. Qian X, Yoon B: **Effective identification of conserved pathways in biological networks using hidden Markov models**. *PLoS ONE* 2009, **4**(12):e8070.
37. Dyer M, Murali T, Sobral B: **The landscape of human proteins interacting with viruses and other pathogens**. *PLoS Pathog* 2008, **4**(2):e32.
38. Ashburner M, et al: **Gene ontology: tool for the unification of biology. The Gene ontology consortium**. *Nat Genet* 2000, **25**:25–29.
39. Boyle E, Weng S, Gollub J, Jin H, Botstein D, Cherry J, Sherlock G: **GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes**. *Bioinformatics* 2004, **20**(18):3710–3715.
40. Monroe J, Silberstein L: **HIV-mediated B-lymphocyte activation and lymphomagenesis**. *J Clin Immunol* 1995, **15**(2):61–68.
41. **HIV-1, Human Protein Interaction Database**[http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/].
42. Blake J, et al: **The Mouse Genome Database (MGD): genetic and genomic information about the laboratory mouse. The Mouse Genome Database Group**. *Nucleic Acids Research* 1999, **27**:95–98.
43. Bader G, Betel D, Hogue C: **BIND: the Biomolecular Interaction Network Database**. *Nucleic Acids Research* 2003, **31**:248–250.
44. Boeckmann B, et al: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic Acids Research* 2003, **31**:365–370.

45. Peri S, et al: **Development of human protein reference database as an initial platform for approaching systems biology in humans**. *Genome Res* 2003, **13**(10):2363–2371.

46. Salwinski L, et al: **The Database of Interacting Proteins: 2004 update**. *Nucleic Acids Research* 2004, **32**(Database-Issue):449–451.

47. Pagel P, et al: **The MIPS mammalian protein-protein interaction database**. *Bioinformatics* 2005, **21**(6):832–834.

48. Jaeger S, Leser U: **High-precision function prediction using conserved interactions**. In *Proc German Conference on Bioinformatics (GCB 07)*, *Volume 115*. Edited by Falter C, et al 2007:146–162.

49. Needleman S, Wunsch C: **A general method applicable to the search for similarity in the amino acid sequences of two proteins**. *J Mol Biol* 1970, **48**:443–453.

50. Hermjakob H, et al: **IntAct: an open source molecular interaction database**. *Nucleic Acids Research* 2004, **32**(Database-Issue):452–455.

51. O'Brien K, Remm M, Sonnhammer E: **Inparanoid: a comprehensive database of eukaryotic orthologs**. *Nucleic Acids Res* 2005, **33**(Database issue):D476–D480.

52. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK 1998.