Viral Complexity: Amino acid co-evolution in viral genomes as a possible metric

C. K. Sruthi[1] **and Meher K. Prakash**[1,*]

[1]Jawaharlal Nehru Centre for Advanced Scientific Research, Theoretical Sciences Unit, Bangalore, 560064, India
[*]meher@jncasr.ac.in

ABSTRACT

Viruses are simultaneously simple and complex. Simple because they have barely around ten types of proteins compared to tens of thousands of proteins in bacteria. Complex because amino acid mutation rates are very high, challenging host immune system and drugs. In this work we use the co-evolution of amino acids and the network characteristics that arise out of it to describe the complexity hidden in the multitude of variations in a viral genome. Using large-scale genomic data, the complexity in several viruses was compared. Interestingly, the co-evolutionary relations were primarily intra-protein in avian influenza and inter-protein in HIV-1. The network degree distributions showed two universality classes: a power-law with exponent -1 in HIV-1 and avian-influenza, random co-evolutionary behavior in human flu and dengue, suggesting the co-evolution as one way to statistically classify the complexity in viruses. The observed correlation between the network densities and the strengths on virus Richter scale raises interesting questions on whether it is possible to define the complexity of viruses using their evolutionary networks.

Introduction

The genome size and complexities in different organisms vary widely. While bacteria have genes encoding several thousand types of proteins, most viruses have barely around ten types of proteins. This is true for viruses as benign as common flu to the lethal ones like ebola. Interestingly, as the number of base-pairs encoding these genes varies from hundreds of millions to tens of thousands, the mutation rate which is the chance of making an error over a generation increases by many orders of magnitude.[1,2] Despite this high rate of mutations or errors in the amino acids of viral proteins, many viruses remain functional and infect the hosts possibly because many deleterious mutations are compensated by other simultaneous mutations. Continuously evolving viruses thus become much more unpredictable both for the immune system as well as the drugs developed against them. Characterizing the evolutionary behavior of viruses will thus be an important step towards understanding the complexity of viruses. Yet, to date there is no informatics way of describing the complexity of viruses and their evolution.

One way of describing the biological systems-level complexity involved in healthy and diseased cells is by studying interaction networks. Biological networks can be formed out of transient molecular interactions such as in proteins interacting with other proteins or from persistent physical interactions such as in neural networks. Metabolic[3] and gene regulatory networks,[4] protein-protein interaction networks,[5,6] and neural networks are examples of functional cellular networks. Disease networks on the other hand try to connect genotypes with phenotypes.[7,8] Protein-protein interaction networks have been used to describe the complexity of the different systems from *E. Coli* to humans[9]. Protein interactions became fine grained as the *C. elegans* interactomes initially identified and mapped at protein level[10], subsequently focused at the domain level.[11] Since viruses have only around ten types of proteins, but high mutation rates, further fine-graining with a focus on amino acid interactions is statistically more meaningful.[12–15] Studying the co-evolutionary relations among the amino acids is an important step towards describing and eventually deciphering the complexity of the viruses.

Amino acid level co-evolutionary interactions can arise either from structural constraints between proximal amino acids or because of functional constraints from amino acids at distal sites or other proteins. Several studies focused on amino acid interaction networks, starting from the three dimensional structural data of the proteins.[12,13,16] The utility of structure based methods is limited because of the limited structural information available, as well as because it more likely highlights the proximal relations. Conversely, using amino acid co-evolutionary couplings from abundant homologous sequence data of multiple species,[17] bioinformatic approaches such as Statistical Coupling Analysis (SCA)[18] and Direct Coupling Analysis[19] could predict hotspots of proteins, active centers of enzymes, to make *de novo* three dimensional structure prediction of proteins[20,21], to identify functionally related clusters of amino acids[22] and predict the vulnerability of viruses.[23] In studying viruses where there could be coupled relations between multiple proteins, it is thus useful to explore this functional coupling. In

this work, we use large-scale complete genome data to build and analyze amino acid co-evolutionary networks. The data is further analyzed to identify patterns or randomness in this co-evolution in intra- and inter-protein amino acids. The complexity of the different viral co-evolutions is also compared by studying the robustness of the network to a targeted or random removal of nodes from the network.

Results and Discussion

Data inclusion: Complete genome sequence data were obtained from the NCBI servers. The analysis was performed when large scale genomic data, from at least 1000 patients, was available. With the current publicly available data, only five viruses were chosen for analysis: HIV-1, hepatitis B, dengue, avian and human influenza. Our analysis was performed on data sets from minimum of 1,784 patient data (HIV-1) to about 8,689 (human influenza). However, the availability of such data is increasing, and in this work we focus on questions that can be posed with such large scale genomic data. Multiple Sequence Alignment (MSA) of the whole genome data from all patients was performed. Using a consensus sequence as a reference, the entire MSA was converted into a binary representation, 1 if the amino acid at a given position in a sequence is the same as that in the consensus sequence, 0 otherwise. Using the Statistical Coupling Analysis protocol,[18] weighted co-evolutionary matrix **C** that quantifies the relations among the different amino acids was created (**Methods** section). The data on pairwise co-evolutionary couplings was represented using a network for better visualization and analysis. The network representation translates the co-evolutionary information into *nodes* (amino acids) and *network edges* (connections between the amino acids) if the co-evolution matrix element C_{ij} relating amino acids i and j, is more than a threshold C, $C_{ij} > C^{th}$.

Clustering: Using the complete genome data from different patients, the co-evolution networks for different viruses were constructed. Clustering of nodes was performed using correlation as a weight (**Figure 1**) with the goal of observing patterns which are more general than those seen in pair-wise relations. About 3 to 4 significant clusters can be seen in each of the viruses, and no significant differences in the number of clusters were found when we performed Principal Component Analysis and used Cattell's criterion. However, there is a noticeable difference in the composition of each of the clusters in different viruses. Each of the clusters in the network of HIV-1 co-evolution network have a mixed representation from multiple proteins, suggesting a strong evolutionary relation across the genome, while avian influenza clusters are mostly from intra-protein relations. The inter-protein co-evolutionary relations are much stronger in HIV-1 (**Supplementary Tables 1a, 1b**) .

Scale-free vs. random networks: The complexity of the networks is analyzed by studying its node-degree distribution, $n(k)$ - the number of times a node with a certain number of edges k appears in the network.[24] Two commonly seen universality classes in these distributions - power-law and Poissonian, suggesting systematic or random underlying basis, occur in the amino acid degree distributions as well. In HIV-1, as well as in avian influenza, a power-law $n(k) \sim k^{-1}$, while dengue, human influenza show a Poissonian distribution (**Figure 2**).[24] Consistent with the observation in the clustering, using only the inter-protein co-evolution from HIV-1 did not change the observed powerlaw. Hepatitis B on the other hand showed a mixed behavior including both power law and Poissonian behaviors (**Figures 4a-4f**). We further analyzed the role of the threshold by varying C^{th} in the analysis of Hepatits B. As shown in **Figure 4**, as the C^{th} increases from 1.0 to 3.0, the power-law component becomes more pronounced (similar data for other viruses is shown in **Supplementary Figures 1** to **4**). The data shows a clear separation of network connections arising from two different origins, an organized network of co-evolution above a certain threshold and random network connections at lower thresholds of co-evolution. Within this power-law regime a further change in cutoff did not result in a change in the exponent significantly. The analysis presented so far is the statistical description of data collected from patients and is averaged over all the years of sample collection. In order to study the temporal evolution patterns, we performed time analysis on the data set which is most abundant, human influenza. We divided the complete genome data from human influenza into periods where the number of data sets is similar ($\sim$ 2000 complete genomes each). A node-distribution analysis shows that over this period, there is no significant change in the co-evolutionary complexity of viruses (**Supplementary Figure 5**).

Complexity Measure: It is difficult to describe complexity, and even more to quantify it with one single measure. The lack of a simple and precise metric for complexity is a problem both in biology and in network science. For biological complexity of viruses, here we use the strength on virus Richter scale[25] as a measure of their complexity. While it is understood that the Richter scale indicates mortality from viruses, which includes several factors from how fast the virus mutates to how poor the public health provisions are, for lack of a better way to compare the strengths of viruses or difficulty of developing vaccines against them, we use Richter scale. Figure 3 shows a plot between the virus strength and the network characteristic - network density. Avian influenza data from avian host was not part of this analysis as the Richter scale definition is irrelevant. Interestingly Figure 3 shows a correlation between the network metric and the biological metric. Clearly this correlation is not

conclusive, as they are based on studies of just four viruses. However, it raises the possibility that the complexity of the biology and the pathogenicity of the virus may be reflected in the amino acid co-evolution networks.

Power law: Random networks (Erdos-Renyi model), small world networks (Watts-Strogatz model[26]) and self-similar networks (Barabasi-Albert model[27,28]) arising in diverse contexts such as WWW, protein-protein interactions, co-authorship networks, etc have been well studied. Some of the mechanisms that explain the observed phenomena are preferential attachment model where newer edges are added to a node depending on its current degree, or based on its pre-defined fitness or a potential for a degree. The power-law with $\gamma \sim 1$ observed in the co-evolution network is different from the typical power-laws γ varying from 2 to 3 and is closer to the behavior in co-authorship networks. Unlike a citation network, there is no reason to believe that the co-evolutionary network evolves with a continuous increase in the number of nodes and edges. Considering amino acid conservation (ϕ) as a surrogate for their fitness, we developed a fitness based model[29]. The model uses two distributions derived from the whole genome data: (a) the distribution of the conservation among the amino acids, $p(\phi)$ (**Supplementary Figure 6**) (b) the co-evolutionary fitness potential of the node $\eta(\phi)$ corresponding to a given conservation of the amino acids. The latter can be modeled as a gaussian distribution, with minimal co-evolutionary fitness for amino acids with very high and very low conservation, a peak in between at ϕ_m and standard deviation σ. Considering a pair of amino acid nodes i and j, and two random numbers r_1 and r_2 drawn from a uniform distribution, edge i —j is created if $r_1 * r_2 \leq \eta(\phi_i) * \eta(\phi_j)$. This algorithm generates a node-degree distribution with $\gamma \sim 1$ (**Supplementary Figure 7**). For example, for HIV-1, the conclusion is relatively invariant for a gaussian with $\phi_m = 0.6 - 0.7$ and $\sigma = 0.02 - 0.07$. As the parameters go out of this range, node degree distribution eventually transforms to a random network model. While the model captures the observed power-law and poissonian distributions with minimal assumptions, the assumptions need to be related to the evolutionary stages of the viruses to see if the statistical complexity of viral co-evolution can be related to their biological complexity.

Robustness of networks: The co-evolution networks were checked for their robustness by removing different fractions of nodes and all the edges connecting to them,[30] the spirit being that the critical amino acids or groups of them can be a potential drug target. The nodes to be removed were chosen either randomly or by picking those with the highest degree, to simulate a random error or a targeted attack, **Figure 5** shows how the effective diameter - a metric of network connectivity - is affected by the targeted or random removal. Interestingly, random removal has the highest impact on human influenza network, and the least affected is HIV-1. Further, the impact of targeted removal is highest on HIV-1. The overall characteristics of robustness can be intuitively expected from the the power-law distribution of nodes. We also used another measure of robustness, which is the number of clusters it breaks into. The conclusions from these calculations, shown in **Figure 5b** are the same as from node removal.

Conclusions

By using a network representation of amino acid co-evolution we have seen two different characteristics in the large scale complete genome data - clustering with mostly intra-protein or inter-protein couplings and node degrees which have a structured power-law or random origins. When genomic data from more viruses will be available, it will be interesting to see if these two different measures of statistical complexity of genomes can be used to classify viruses into different categories, with a possible mapping to their biological or pathogenic complexity. Further it will be interesting to see if the inter-protein or intra-protein couplings are related to the host adaptation (HIV-1) or the host being a neutral carrier (avian influenza) and how such patterns evolve with time as the viruses adapt from being pandemics to epidemics.

Methods

Undirected co-evolution networks: The chance of co-evolution C_{ij} between a pair of amino acids i and j is calculated by averaging the columns i and j of the boolean sequences using either an unweighted or weighted protocol following the Statistical Coupling Analysis protocol.[22] Unweighted and normalized co-evolution is defined as $C_{ij}^{unweighted} = (\langle x_i x_j \rangle_s - \langle x_i \rangle_s \langle x_j \rangle_s) \Big/ \left(\sqrt{\langle x_i^2 \rangle_s - \langle x_i \rangle_s^2} \sqrt{\langle x_j^2 \rangle_s - \langle x_j \rangle_s^2} \right)$, where x_i is the i^{th} column in the boolean sequence and $\langle\rangle_s$ denotes the average over sequences. Weighted co-evolution is defined as $C_{ij}^{weighted} = \phi_i \phi_j \, |\langle x_i x_j \rangle_s - \langle x_i \rangle_s \langle x_j \rangle_s|$, where $\phi_i = ln\left((\langle x_i \rangle_s (1 - q^{a_i})) \Big/ (q^{a_i} (1 - \langle x_i^s \rangle_s)) \right)$, and q^{a_i} is the probability with which the amino acid a_i at position i in the consensus sequence occurs among all proteins. background probability of the most frequent amino acid a_i at position i frequency of occurs among all proteins. One could work with either $C_{ij}^{unweighted}$ or $C_{ij}^{weighted}$, and in the present work on networks we use $C_{ij}^{weighted}$. If the chosen C_{ij} exceeds a chosen cutoff c, we consider an undirected network link i — j to be present. The sensitivity of the analysis to c is discussed in the article. The analysis reported in the article is based on $C_{ij}^{weighted}$. However, changing the

$C_{ij}^{unweighted}$ the power-law distribution in HIV-1 was still around 1, changed from 0.91 to 1.37. Thus, we believe the general conclusions do not change with the weighting.

Data Availability: The datasets generated and analysed during the current study are available from the corresponding author on request.

References

1. DRAKE, J. A constant rate of spontaneous mutation in DNA-based microbes . *Proceedings of the National Academy of Sciences, USA* **88**, 7160–7164 (1991). DOI 10.1073/pnas.88.16.7160.
2. Sanjuan, R. & Domingo-Calap, P. Mechanisms of viral mutation. *Cell.* **73**, 4433–4448 (2016). DOI 10.1007/s00018-016-2299-6.
3. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of escherichia coli. *Nat. genetics* **31**, 64–68 (2002).
4. Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* **9**, 770–780 (2008).
5. Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nat.* **444**, 364–368 (2006).
6. Kar, G., Gursoy, A. & Keskin, O. Human cancer protein-protein interaction network: a structural perspective. *PLoS Comput. Biol* **5**, e1000601 (2009).
7. Goh, K.-I. *et al.* The human disease network. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA* **104**, 8685–8690 (2007). DOI 10.1073/pnas.0701361104.
8. Barabasi, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *NATURE REVIEWS GENETICS* **12**, 56–68 (2011). DOI 10.1038/nrg2918.
9. Stelzl, U. *et al.* A human protein-protein interaction network: A resource for annotating the proteome. *CELL* **122**, 957–968 (2005). DOI 10.1016/j.cell.2005.08.029.
10. Li, S. *et al.* A map of the interactome network of the metazoan C-elegans. *SCIENCE* **303**, 540–543 (2004). DOI 10.1126/science.1091403.
11. Boxem, M. *et al.* A protein domain-based interactome network for C-elegans early embryogenesis. *CELL* **134**, 534–545 (2008). DOI 10.1016/j.cell.2008.07.009.
12. Amitai, G. *et al.* Network analysis of protein structures identifies functional residues. *J. molecular biology* **344**, 1135–1146 (2004).
13. Brinda, K. & Vishveshwara, S. A network representation of protein structures: implications for protein stability. *Biophys. journal* **89**, 4159–4170 (2005).
14. Aurora, R., Donlin, M. J., Cannon, N. A., Tavis, J. E. & Grp, V.-C. S. Genome-wide hepatitis C virus amino acid covariance networks can predict response to antiviral therapy in humans. *Journal of Clinical Investigation* **119**, 225–236 (2009). DOI 10.1172/JCI37085.
15. Donlin, M. J., Szeto, B., Gohara, D. W., Aurora, R. & Tavis, J. E. Genome-Wide Networks of Amino Acid Covariances Are Common among Viruses. *Journal of Virology* **86**, 3050–3063 (2012). DOI 10.1128/JVI.06857-11.
16. Estrada, E. *The structure of complex networks: Theory and Applications* (Oxford University Press, 2011).
17. Shendure, J. & Ji, H. Next-generation dna sequencing. *Nat. biotechnology* **26**, 1135–1145 (2008).
18. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Sci.* **286**, 295–299 (1999).
19. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci.* **106**, 67–72 (2009). URL http://www.pnas.org/content/106/1/67.abstract. DOI 10.1073/pnas.0805923106. http://www.pnas.org/content/106/1/67.full.pdf.
20. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *NATURE BIOTECHNOLOGY* **30**, 1072–1080 (2012). DOI 10.1038/nbt.2419.
21. Ovchinnikov, S. *et al.* Large-scale determination of previously unsolved protein structures using evolutionary information. *ELIFE* **4** (2015). DOI 10.7554/eLife.09248.

22. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774–786 (2009).

23. Dahirel, V. *et al.* Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA* **108**, 11530–11535 (2011). DOI 10.1073/pnas.1105315108.

24. Barabasi, A.-L. *Network Science* (Cambridge University Press, 2016).

25. Weiss, R. & McLean, A. What have we learnt from SARS? *PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B-BIOLOGICAL SCIENCES* **359**, 1137–1140 (2004). DOI 10.1098/rstb.2004.1487. Discussion Meeting on Emerging Infections - What Have We Learnt from SARS, London, ENGLAND, JAN 13, 2004.

26. Watts, D. & Strogatz, S. Collective dynamics of 'small-world' networks. *NATURE* **393**, 440–442 (1998). DOI 10.1038/30918.

27. Barabasi, A., Albert, R. & Jeong, H. Mean-field theory for scale-free random networks. *PHYSICA A* **272**, 173–187 (1999). DOI 10.1016/S0378-4371(99)00291-5.

28. Albert, R. & Barabasi, A. Statistical mechanics of complex networks. *REVIEWS OF MODERN PHYSICS* **74**, 47–97 (2002). DOI 10.1103/RevModPhys.74.47.

29. Nguyen, K. & Tran, D. A. *Handbook of Optimization in Complex Networks* (Springer, 2012).

30. Albert, R., Jeong, H. & Barabasi, A. Error and attack tolerance of complex networks. *NATURE* **406**, 378–382 (2000). DOI 10.1038/35019019.

31. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498–2504 (2003).

32. Newman, M. E., Strogatz, S. H. & Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. review E* **64**, 026118 (2001).

Acknowledgements

We thank Prof. Reka Albert and Prof. Hemalatha Balaram for insightful discussions.

Author contributions statement

C.K.S. developed Python scripts, performed calculations, data analysis, and literature review; M.K.P. conceived the study, interpreted results and wrote the manuscript.

Additional information

Competing financial interests statement

The authors declare no competing financial interests.

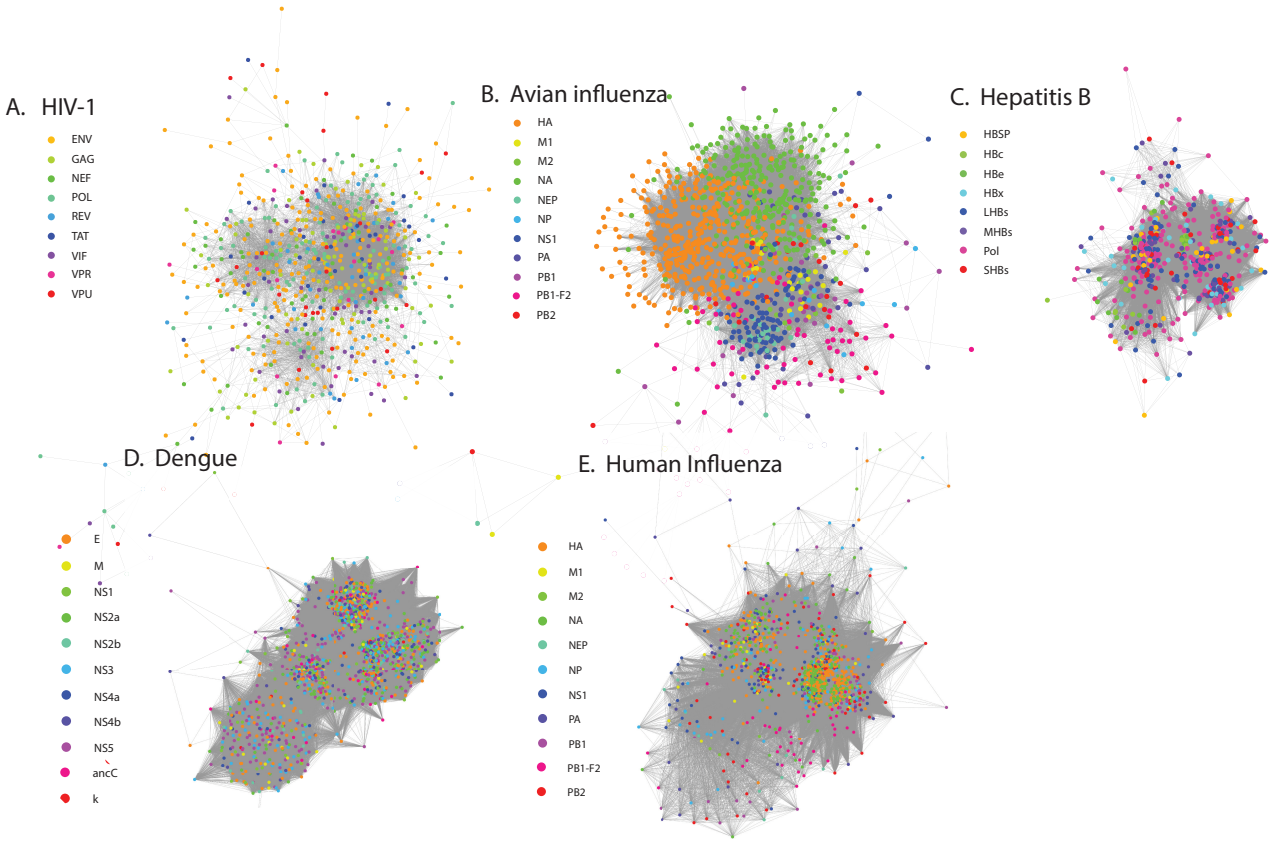


Figure 1. Co-evolutionary network from complete genome analysis of different viruses: (a) HIV-1, (b) avian influenza and (c) Hepatitis B (d) Dengue (e) Human influenza. The networks are generated using co-evolution strength as a weight. The side bar indicates the different types of proteins found in these viruses, as well as the coloring notation used. The networks show three to four major clusters. While in HIV-1, each cluster has a mixed representation from all the proteins, avian influenza clusters are mainly from intraprotein co-evolutionary relations. Network representations were generated using Cytoscape[31]

Figures

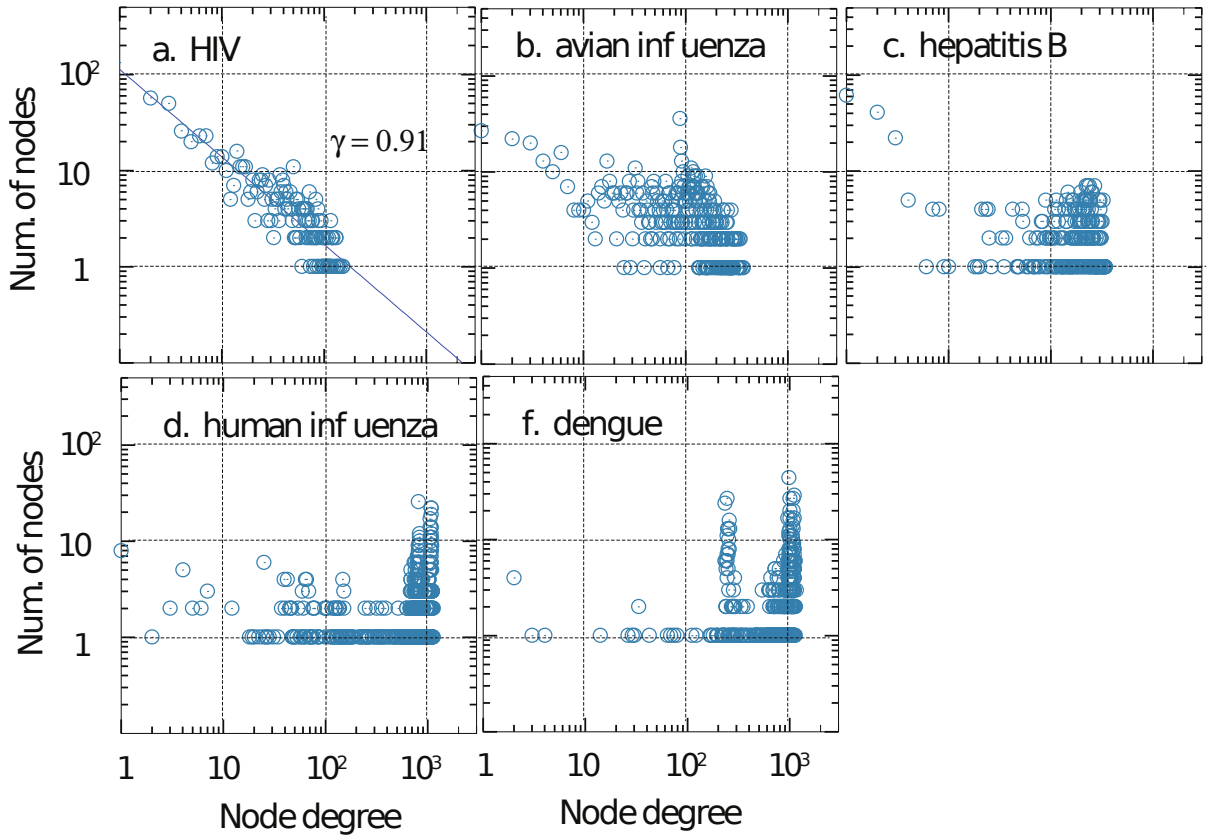


Figure 2. Node degree distribution from the complete genome data in different viruses showing a range of behavior from a pure power-law (HIV-1 and avian influenza) to a pure-random network behavior (dengue and human influenza). A cutoff $C^{th} = 0.85$ was used as a threshold for establishing network edge connections. The effect of changing the cutoff is discussed separately.

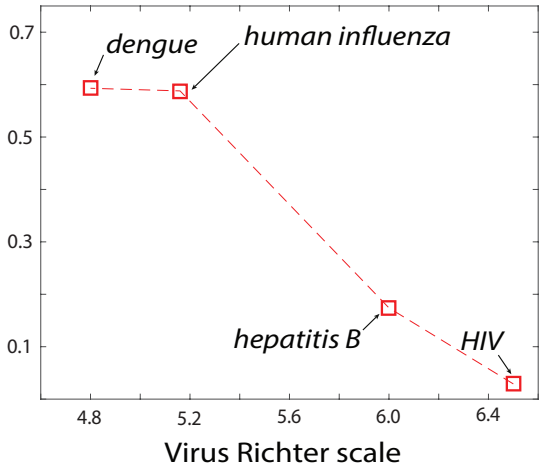


Figure 3. The relation between the complexity of the virus, as described by Virus Richter scale, and network characteristics - node density (□)

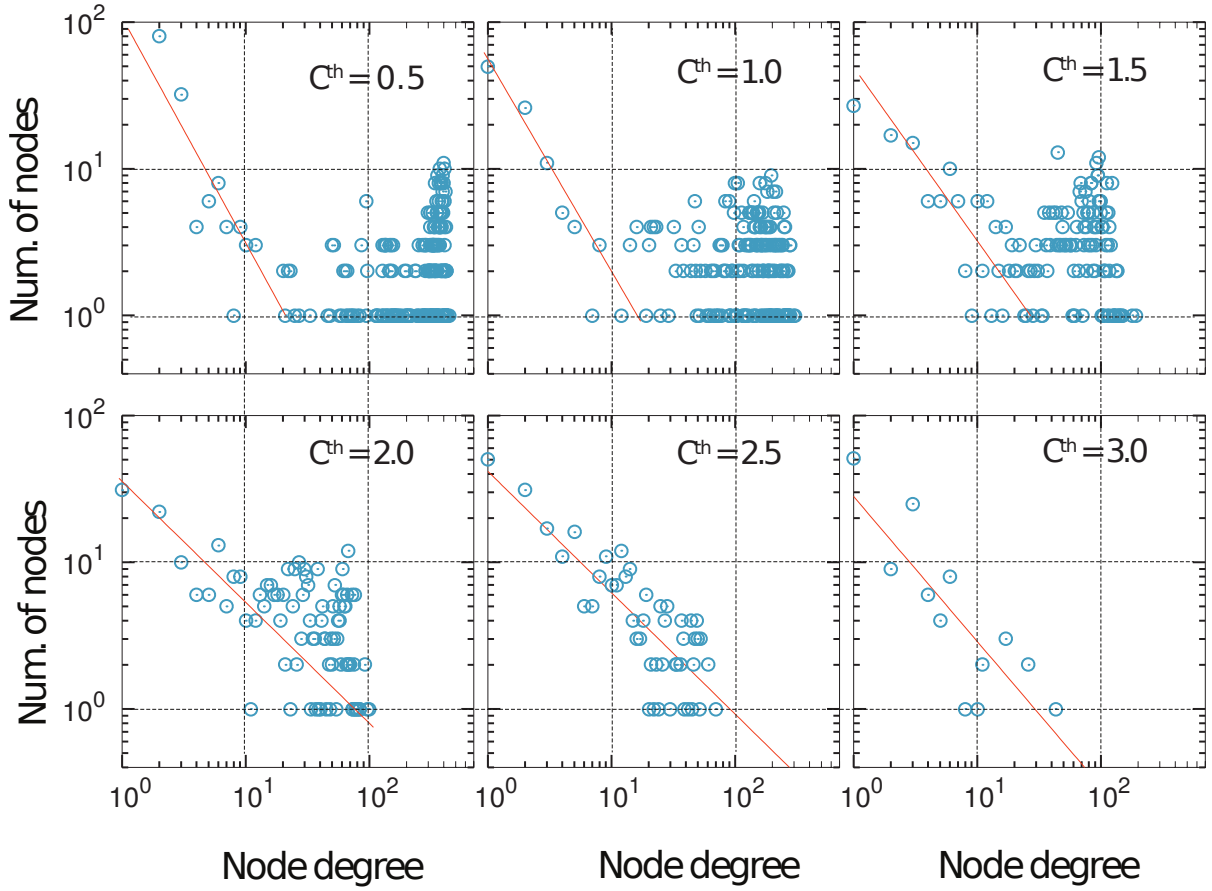


Figure 4. Node degree distribution sensitivity was studied in Hepatitis-B network by changing the cut-off value used for defining edge connectivity between the nodes. At a very low cut-off there is a mixed behaviour in the node degree distribution, with both power-law as well as a random component. As the cutoff is increased, the random component is selectively removed, while preserving the power-law component. This suggests a clear separation of network connections from random and systematic origins. By choosing a threshold value, one can filter and study just the systematic component.

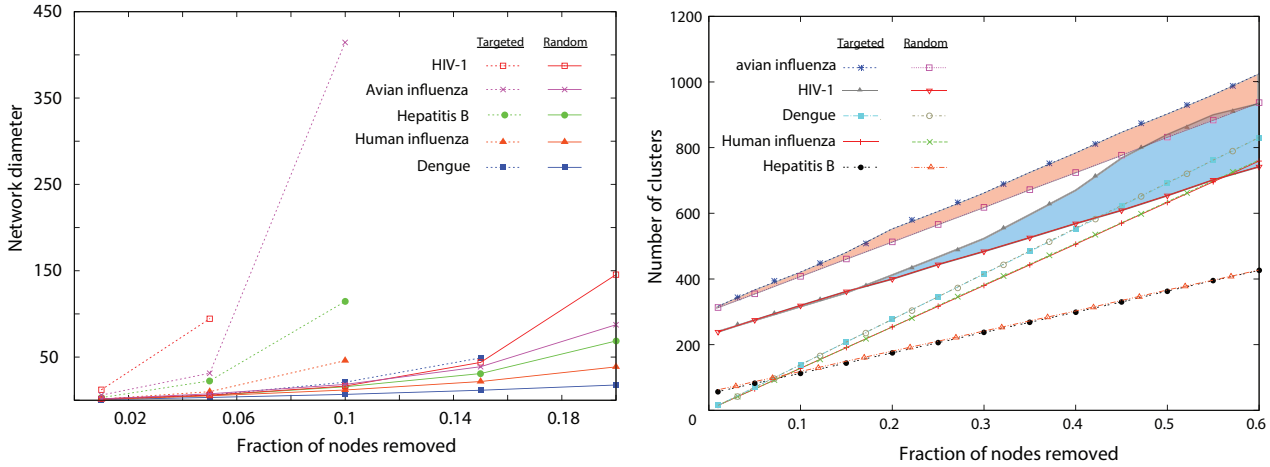


Figure 5. The robustness of the networks is studied using a targeted and random removal of a fraction of nodes. Two different measures were used to estimate how robust the networks are to random and targeted removal: (a) diameter of the network after the removal of the node (b) number of clusters the network breaks into. In HIV-1 and avian influenza there is a clear difference between targeted and random removal, while in others there is not. The avian-influenza and HIV-1 data were shifted up along y-axis by 300 and 200 units for clarity of representation. Network diameter was calculated following the procedure in Ref.[32]

Supplementary Material

Viral Complexity: Amino acid co-evolution in viral genomes as a possible metric

C. K. Sruthi and Meher K. Prakash*

Theoretical Sciences Unit

Jawaharlal Nehru Centre for Advanced Scientific Research,

Bangalore-560064, India

Corresponding author: meher@jncasr.ac.in

October 14, 2017

Methods

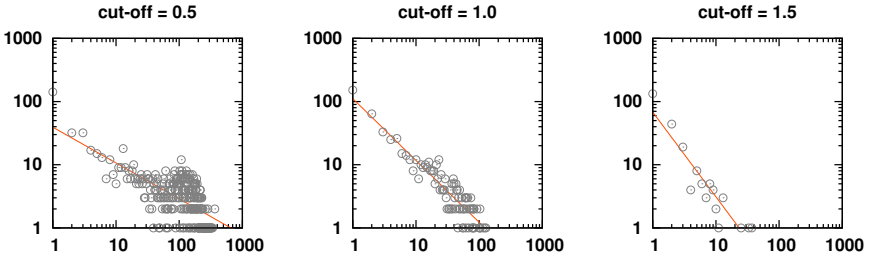
Undirected co-evolution networks: The chance of co-evolution C_{ij} between a pair of amino acids i and j is calculated by averaging the columns i and j of the boolean sequences using either an unweighted or weighted protocol following the Statistical Coupling Analysis protocol. Unweighted and normalized co-evolution is defined as $C_{ij}^{unweighted} = (\langle x_i x_j \rangle_s - \langle x_i \rangle_s \langle x_j \rangle_s) \,/\, \left(\sqrt{\langle x_i^2 \rangle_s - \langle x_i \rangle_s^2} \sqrt{\langle x_j^2 \rangle_s - \langle x_j \rangle_s^2} \right)$, where x_i is the i^{th} column in the boolean sequence and $\langle \rangle_s$ denotes the average over sequences. Weighted co-evolution is defined as $C_{ij}^{weighted} = \phi_i \phi_j \; |\langle x_i x_j \rangle_s - \langle x_i \rangle_s \langle x_j \rangle_s|$, where $\phi_i = ln\left((\langle x_i \rangle_s (1 - q^{a_i})) \,/\, (q^{a_i} (1 - \langle x_i^s \rangle_s)) \right)$, and q^{a_i} is the probability with which the amino acid a_i at position i in the consensus sequence occurs among all proteins. One could work with either $C_{ij}^{unweighted}$ or $C_{ij}^{weighted}$, and in the present work on networks we use $C_{ij}^{weighted}$. If the chosen C_{ij} exceeds a chosen cutoff c, we consider an undirected network link i — j to be present. The sensitivity of the analysis to c is discussed in the article. The analysis reported in the article is based on $C_{ij}^{weighted}$. However, changing the $C_{ij}^{unweighted}$ the power-law distribution in HIV-1 was still around 1, changed from 0.96 to 1.35. Thus, we believe the general conclusions do not change with the weighting.

Protein (Num of amino acids)	GAG (500)	POL (1003)	VIF (192)	VPR (96)	TAT (100)	REV (116)	VPU (82)	ENV (856)	NEF (205)
GAG	343	735	287	105	178	248	130	847	277
POL		530	398	141	241	319	174	1112	400
VIF			93	66	79	126	60	432	143
VPR				6	31	43	16	137	50
TAT					25	99	37	297	89
REV						65	55	410	126
VPU							16	208	66
ENV								714	462
NEF									126

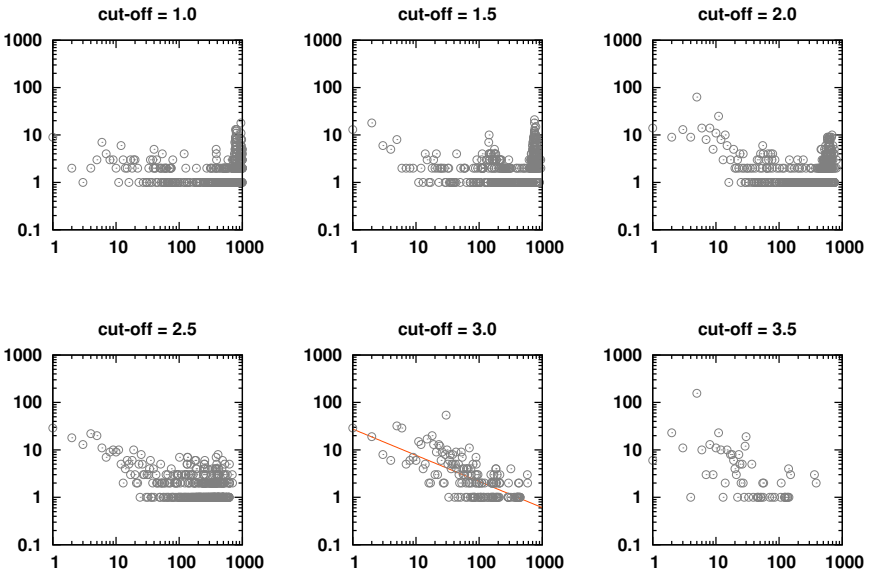
Supplementary Table 1a. Table showing the number of inter-protein and intra-protein amino acid co-evolutionary couplings from HIV-1 data, with a $C^{th} = 0.85$

Protein (No. of amino acids)	NP (498)	PB2 (759)	HA (566)	M1 (252)	M2 (97)	NA (469)	NS1 (230)	NEP (121)	PA (716)	PB1-F2 (90)	PB1 (757)
NP	49	101	219	210	99	77	291	59	137	156	150
PB2		56	304	185	78	134	267	49	145	127	130
HA			20000	900	192	5362	852	106	184	377	242
M1				171	172	233	570	94	296	268	252
M2					38	33	233	29	130	103	116
NA						12005	352	48	30	93	92
NS1							2832	1436	400	352	376
NEP								187	53	66	56
PA									153	168	197
PB1-F2										637	211
PB1											86

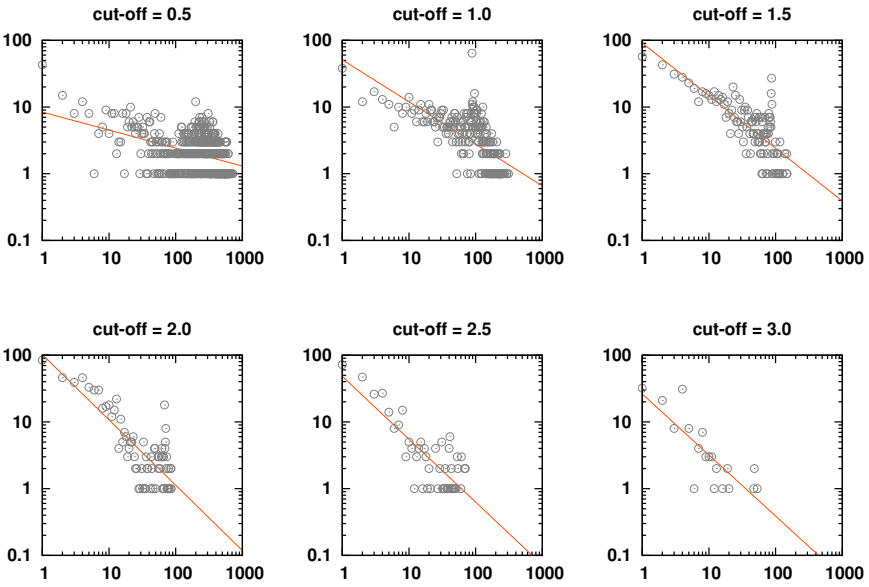
Supplementary Table 1b. Table showing the number of inter-protein and intra-protein amino acid co-evolutionary couplings from avian influenza data, with a $C^{th} = 0.85$



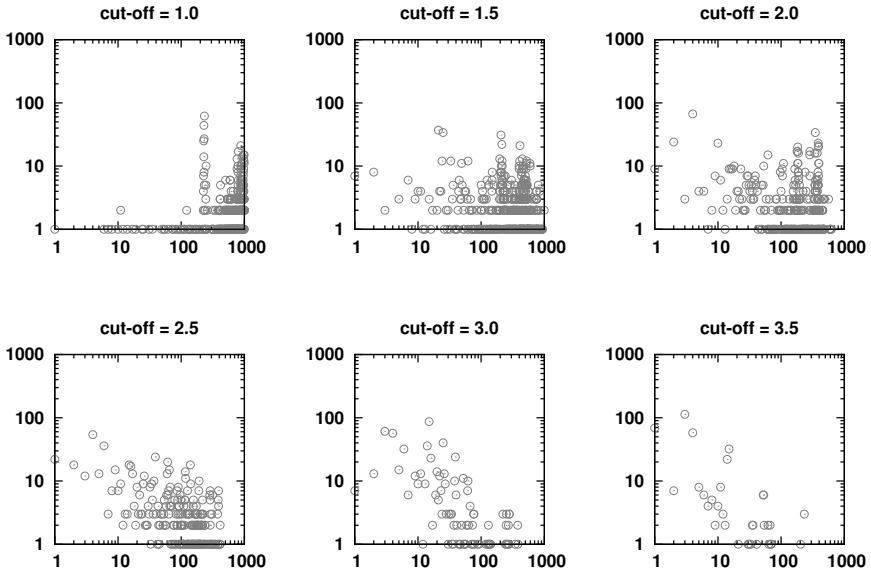
Supplementary Figure 1. Variation in node distribution of HIV-1 co-evolution network as the cutoff C^{th} is changed.



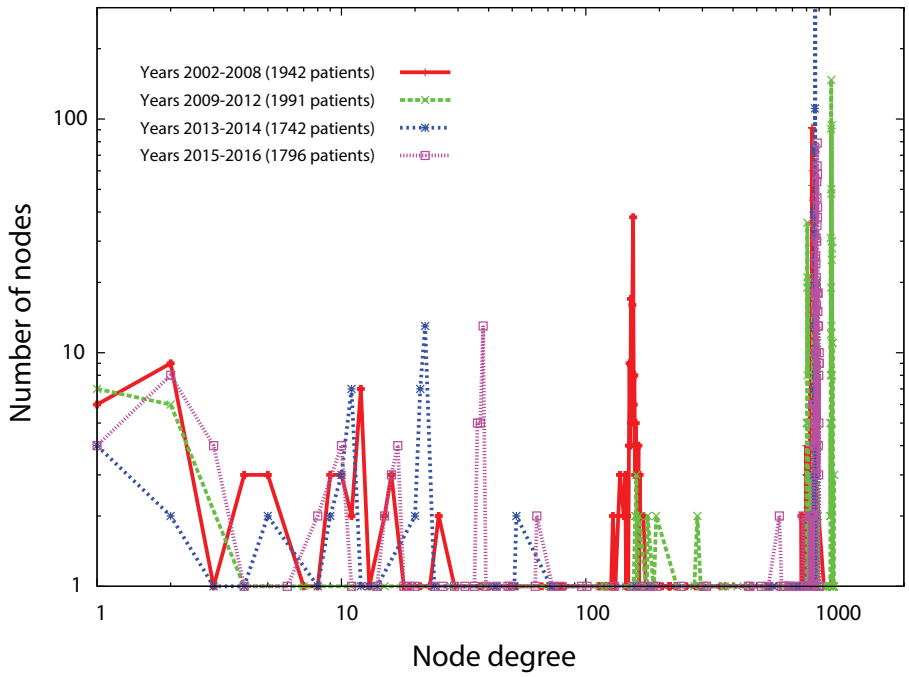
Supplementary Figure 2. Variation in node distribution of human influenza co-evolution network as the cutoff C^{th} is changed.



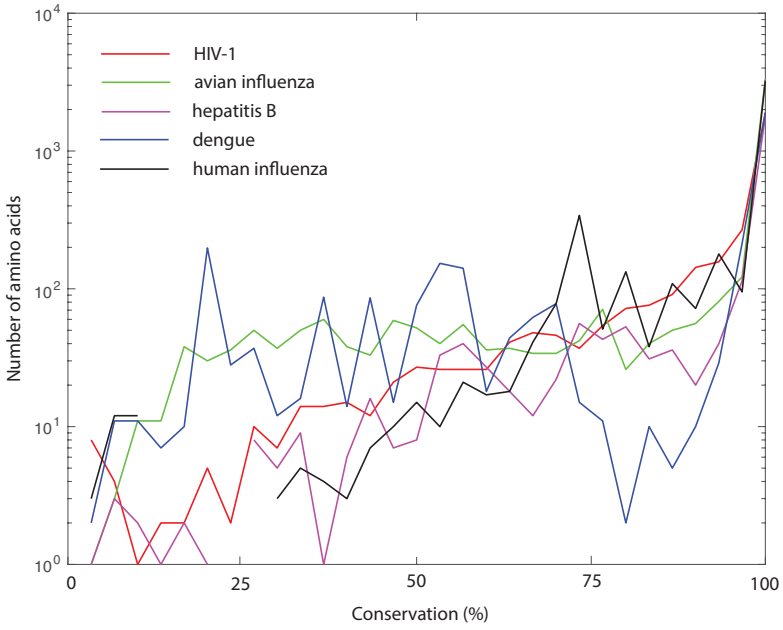
Supplementary Figure 3. Variation in node distribution of avian influenza co-evolution network as the cutoff C^{th} is changed.



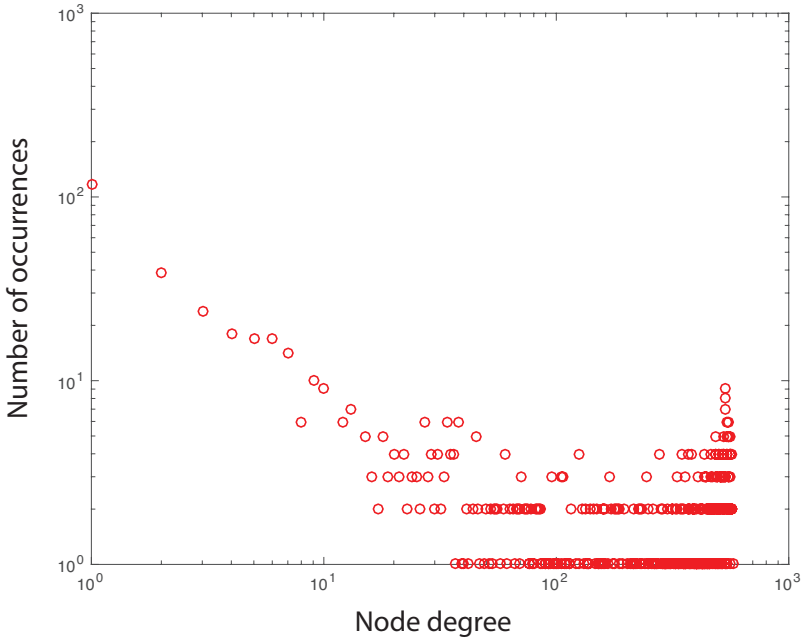
Supplementary Figure 4. Variation in node distribution of dengue co-evolution network as the cutoff C^{th} is changed.



Supplementary Figure 5. Variation of the node degree distribution over years in the human influenza. Human influenza data was abundant, so we sorted it according to the year of incidence, and made 4 groups of about 2000 patients each. No noticeable trend in the node degree distribution was observed in the data between 2002-2016.



Supplementary Figure 6. Distribution of the conservation of amino acids in different viruses.



Supplementary Figure 7. Model network generated using the amino acid conservation distribution from HIV-1, and $\eta(\phi)$ with parameters $\phi_m = 0.05$ and $\sigma = 0.7$