A Dual Controllability Analysis of Influenza Virus-Host Protein-Protein Interaction Networks for Antiviral Drug Target Discovery

Ackerman, Emily E.[1]; Alcorn, John F.[2]; Hase, Takeshi[3,4]; Shoemaker, Jason E.[1,5,6]*

[1] Department of Chemical and Petroleum Engineering, University of Pittsburgh, Pittsburgh, PA USA

[2] Division of Pulmonary Medicine, Allergy, and Immunology, Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA USA

[3] The Systems Biology Institute, Saisei Ikedayama Bldg. 5-10-25 Higashi Gotanda, Shinagawa, Tokyo 141-0022, Japan

[4] Medical Data Sciences Office, Tokyo Medical and Dental University, M&D Tower 20F, 1-5-45 Yushima, Bunkyo, Tokyo 113-8510, Japan.

[5] The McGowan Institute for Regenerative Medicine (MIRM), University of Pittsburgh, Pittsburgh, PA USA

[6] Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA USA

* Corresponding author

ABSTRACT

Motivation: Host factors of influenza virus replication often are often found in key topological positions within protein-protein interaction networks. This work explores how protein states can be manipulated through controllability analysis: the determination of the minimum manipulation needed to drive the cell system to any desired state. Here we complete a two-part controllability analysis of two protein networks: a host network representing the healthy cell state and an influenza A virus-host network representing the infected cell state. This knowledge can be utilized to understand disease dynamics and isolate proteins for study as drug target candidates.

Results: Both topological and controllability analyses provide evidence of wide-reaching network effects stemming from the addition of viral-host protein interactions. Virus interacting and driver host proteins are significant both topologically and in controllability, therefore playing important roles in cell behavior during infection. 24 proteins are identified as holding regulatory roles specific to the infected cell.

Availability: None.

Contact: jason.shoemaker@pitt.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

A major step in the ability to predict novel anti-viral drug targets lies in the ability to understand how viruses interrupt and take control of healthy host cell functions (Rask-Andersen *et al.*, 2011). Traditional systems biology approaches for intercellular signaling pathways create detailed kinetic models which require experimentally-derived parameters or parameters estimated by simulation (Klipp and Liebermeister, 2006; Schoeberl *et al.*, 2002; Aldridge *et al.*, 2006). Without these key values and extensive training data, complications quickly arise and the model can become unattainable. Modeling studies are often focused on specific pathways, limiting their reach for analysis to particular cascades of reactions instead of considering the total cellular environment.

Network approaches in systems biology use protein-protein interaction (PPI) data to model cell-wide systemic changes associated with disease, changes in cell function, or cell fate (Cho *et al.*, 2012). This provides a holistic understanding of cell behavior by viewing protein as interdependent states, regardless of specific interaction mechanisms, allowing for the exploration of cell level relationships. The field of network theory is well established and basic network metrics like degree and betweenness (Freeman, 1977) have repeatedly been used to reveal the importance of specific proteins within biological processes that cannot be found from traditional modeling approaches (Zhu *et al.*, 2009; Vinayagam *et al.*, 2014; He and Zhang, 2006; Lopes *et al.*, 2015; Barabási *et al.*, 2011). Disease networks have been used to identify genes involved with cancer (Jonsson and Bioinformatics, 2006; Hase *et al.*, 2009; Mani *et al.*, 2008; Mine *et al.*, 2013), demonstrate that the genes that are responsible for similar diseases are likely to interact with each other (Mitchell *et al.*, 2013; Gandhi *et al.*, 2006), and predict novel drug targets (Arrell and Terzic, 2010; Pujol *et al.*, 2010).

Network studies have been performed for many common viruses including hepatitis C (Germain *et al.*, 2014; Chassey *et al.*, 2008), SARS (Mitchell *et al.*, 2013; Moni and bioinformatics, 2014), HIV (Moni and bioinformatics, 2014; Murali *et al.*, 2011; Ptak *et al.*, 2008; Shityakov *et al.*, 2015), and influenza virus (Schaefer *et al.*, 2013; Shoemaker *et al.*, 2012; Mitchell *et al.*, 2013; Korth *et al.*, 2013; Tripathi *et al.*, 2015). Past work studying the effects of influenza virus in the context of a PPI network have focused on identifying host factors involved in virus replication and improving the prediction of drug targets. Most analysis to date ends with basic topological measurements which only provide a general overview of the state of the network, and do not fully demonstrate performance under pressure (for example: network behavior in a diseased cellular state). The next logical step for analyzing biological networks lies in understanding how the network can be manipulated and exploited to manage specific biological properties.

In classic control theory, controllability is the idea that a deterministic system can be driven to any final state in finite time given an external input (Lin, 1974). This is commonly applied to linear, time invariant dynamic systems,

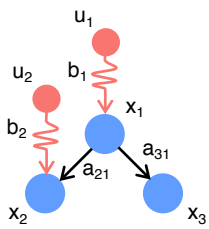
$$\frac{dx(t)}{dt} = Ax(t) + Bu(t)$$

where A is an NxN matrix of state coefficients that describes how N molecule states, $x(t)$, interact within the system and B is a matrix of input weights describing how external influences, $u(t)$, impact the system. In general, a system is controllable if the controllability matrix,

$$C = [B, AB, A^2B, \dots, A^{N-1}B]$$

is full rank, N. This means that the system can be manipulated into any desired combination of states within all of state space by the defined input, B. A controllability analysis identifies the key components of a system that can be manipulated to drive desired system outcomes (Wuchty, 2014).

a) Network representation



b) State space representation

$$A = \begin{bmatrix} 0 & 0 & 0 \\ a_{21} & 0 & 0 \\ a_{31} & 0 & 0 \end{bmatrix}, B = \begin{bmatrix} b_1 & 0 \\ 0 & b_2 \\ 0 & 0 \end{bmatrix}$$

where:

A = Interaction weight matrix
x = Protein concentation state matrix
B = Input process weight matrix
u = Protein translation process input matrix

$$\dot{x} = Ax + Bu$$

$$\dot{x} = \begin{bmatrix} 0 & 0 & 0 \\ a_{21} & 0 & 0 \\ a_{31} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} b_1 & 0 \\ 0 & b_2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

c) Controllability matrix

$$C = [B \quad AB \quad A^2B]$$

$$C = \begin{bmatrix} b_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & b_2 & a_{21}b_1 & 0 & 0 & 0 \\ 0 & 0 & a_{31}b_1 & 0 & 0 & 0 \end{bmatrix}$$

Rank (C) = 3
Full rank
∴ *System is fully controlled*

Figure 1: (a) An example protein-protein interaction network with three proteins and two protein translation process inputs. The state space representation (b) of the same network demonstrates that the change in state of a protein's concentration is a function of its current state and an input process. A classic controllability analysis (c) demonstrates that this system is fully controllable and could, therefore, be driven to any possible state change in every protein.

An example PPI network in Fig. 1a, has been transformed into its state space matrix representation in Fig. 1b. With the inclusion of two independent inputs (u_1 and u_2), the controllability matrix in Fig. 1c is full rank. Therefore, the system is fully controllable and it is possible to drive the protein concentrations to any desired state. Applying the idea of controllability to a cell at the onset of viral infection, a virus seeks to control of cellular functions (the system of proteins), promote virus replication tasks, and reach a final infected cell state. While it would be advantageous to interpret the infection from this control perspective, mathematical limits in large system dimensions prevent the direct application of traditional controllability methods to PPI networks.

Advances in network theory have created alternative methods of network controllability evaluation which survey each node's (protein's) importance in the ability of an external set of inputs to fully control the network. Controllability classification is founded in "driver node" calculations: identifying the network components which must be manipulated for the system to be fully controlled (analogous to determining the non-zero elements of the B matrix in classic controllability). Without manipulation, driver nodes will remain unaffected by changes to the rest of the system, rendering

the total system uncontrollable. A set of driver nodes (size N_D) that is capable of controlling the total network is called a minimum input set (MIS). The MIS is not unique and the number of possible MISs scales exponentially with the size of the network (Jia and Barabási, 2013). After a primary MIS is calculated, two methods of controllability node classification can be used.

In the first method by Liu et al. [34], the MIS is re-calculated (size $N_D{}'$) after removing each node from the network. The node is then classified by its effect on the manipulation required to control the network, represented by the MIS (Liu *et al.*, 2011). The absence of: an *indispensable node* increases the number of driver nodes ($N_D{}' > N_D$), a *dispensable node* decreases the number of driver nodes ($N_D{}' < N_D$), and a *neutral node* has no effect on the number of driver nodes ($N_D{}' = N_D$). This method has previously been applied to many network types such as gene regulatory networks, food webs, citation networks, and PPI networks to better understand what drives the dynamics of each system (Liu *et al.*, 2011; Vinayagam *et al.*, 2016). While it is useful to observe the structural changes to the network in the absence of singular nodes, this method only considers one possible MIS. In a second classification method by Jia et al. (Jia *et al.*, 2013), a node is classified by its role across all possible MISs. A *critical node* is included in all possible MISs, an *intermittent node* is included in some possible MISs, and a *redundant node* is not included in any possible MISs. This method places each node in the broader context of all possible control configurations.

In total, this study aims to determine key host factors in the progression of influenza infection for the prediction of novel antiviral targets. We have completed a two-part controllability analysis of a host PPI network (HIN) and a hybrid network of human host PPI data combined with influenza A virus-host protein interaction data (VIN). The controllability characteristics of influenza virus interacting host proteins and driver proteins are compared to the characteristics of the total

network. A set of host factors which hold value, both topologically and in controllability, are identified as candidates for further study based on their specialized behavior during influenza infection.

RESULTS

Topology of the Host Interaction Network and Virus Integrated Network

The directed PPI network from Vinayagam et al (Vinayagam *et al.*, 2011) was restricted to confident interactions (see Methods), creating a network containing 6,281 proteins and 31,079 interactions. This network will be referred to as the "Host Interaction Network" (HIN). Influenza A virus-host interactions from Watanabe et al (Watanabe *et al.*, 2014) were narrowed to 2,592 interactions between 11 influenza A virus (IAV) proteins (HA, M1, M2, NA, NP, NS1, NS2, PA, PB1, PB2, and PB1-F2 proteins) and 752 "IAV interacting proteins" found in the HIN. After the integration into the HIN, the network contains 6,292 proteins and 33,671 interactions. This network will be referred to as the "Virus Integrated Network" (VIN).

Degree and betweenness calculations were completed for the HIN and VIN. As expected, the only proteins which display a shift in degree between the two networks are the 752 IAV interacting proteins (Marked in blue in Fig. 2a). The degree shift of the group of IAV interacting proteins is significant as compared to all proteins in both the VIN (log scaled median of IAV interacting proteins: 1.04; log scaled median of all proteins: 0.70; student t-test of log scaled data p-value $< 2.20\text{x}10^{-16}$) and the HIN (log scaled median of IAV interacting proteins: 0.85; log scaled median of all proteins: 0.70; Student t-test of log scaled data p-value: $5.97\text{x}10^{-12}$). The degree distributions of both networks are proven to be scale free (Fig. S1).

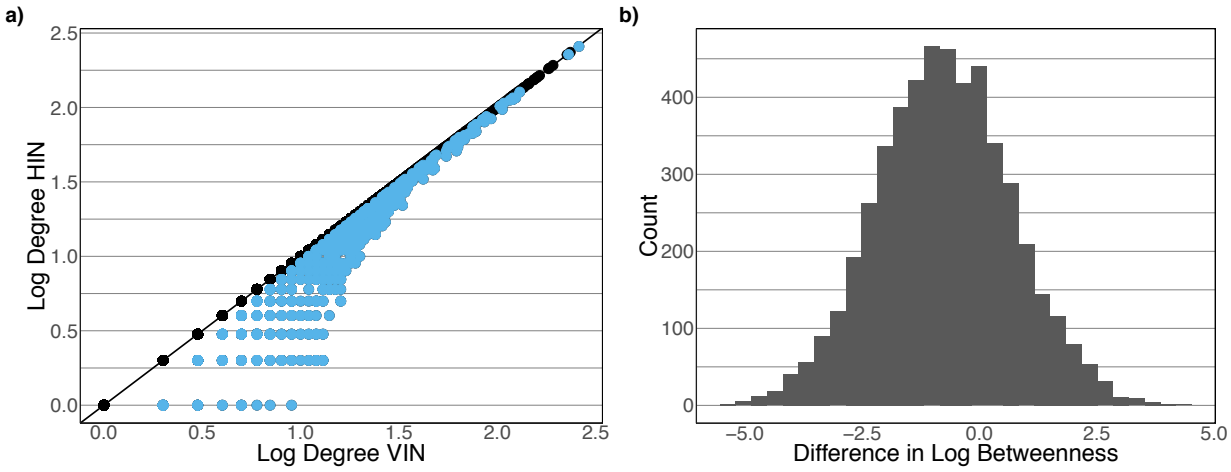


Figure 2: (a) Degree of the VIN vs degree of the HIN where the IAV interacting proteins are marked in blue. The degree distributions of the networks are scale free. (b) Difference in betweenness between the VIN and HIN for proteins which exhibit a difference greater than one.

Because betweenness is sensitive to the information flow through all proteins instead of only neighboring proteins, 2,735 proteins exhibit an increase in betweenness after the addition of IAV interactions. Of these proteins, 207 proteins' log betweenness exhibits an increase of 2 or more in the VIN compared to the HIN (Fig. 2b). This suggests that the addition of IAV interactions has an effect on network topology that reaches over 3.5 times the number of host proteins that are directly interacting with IAV proteins. The betweenness shift in the group of IAV interacting host proteins is significant as compared to all proteins in both the VIN (Log scaled median of IAV interacting proteins: 3.23; Log scaled median of all proteins: 2.82; Student t-test of log scaled data p-value: 2.20×10^{-16}) and the HIN (Log scaled median of IAV interacting proteins 3.22; Log scaled median of all proteins: 2.82; Student t-test of log scaled data p-value: 2.13×10^{-15}). This is a result of being the limited protein set responsible for information flow from the viral proteins to the rest of the network.

Driver proteins

Driver proteins (nodes) are the foundation of both types of controllability calculations, representing the protein set which must be manipulated for the system to be fully controlled. The

proteins are identified through maximum matching algorithms (Hopcroft and Karp, 1973). The HIN and VIN both require N_D = 2,463 driver proteins to achieve controllability, suggesting that the magnitude of network control is unchanged by the influence of the IAV interactions. However, the identity of driver proteins shifts slightly as the 11 viral proteins replace 11 host proteins within the primary MIS as drivers in the VIN. Table 1 lists their identities along with the shortest distance to an IAV protein in the network, and protein degree and betweenness). Of these 11 host proteins, only five are directly interacting with IAV proteins. One of the remaining proteins is two steps (two interactions and one connecting protein) from any IAV protein, and the remaining five proteins are three steps from any IAV protein. The number of paths between viral proteins and the group of shifting proteins are reflective of the number of paths between viral proteins and all host proteins (Fisher test p-value: 0.99). This supports the idea that viral interactions have lasting effects on the system's control structure, affecting proteins which are multiple paths away.

Table 1: Identities of proteins that are drivers in the HIN but not the VIN with the shortest number of paths to an Influenza A viral protein. Degree and betweenness of the proteins of the VIN is provided (with the values from the HIN in parenthesis). Only 45% of these proteins are directly interacting with the viral proteins, demonstrating the cascade effect caused by the inclusion of viral interactions.

Entrez ID	**Gene Name**	**Shortest Distance to IAV Protein**	**Degree**	**Betweenness**
10658	CUGBP, Elav-Like Family Member 1 (CELF1)	1	4 (4)	81 (81)
1969	EPH Receptor A2 (EPHA2)	1	14 (13)	93 (0)
6733	SRSF Protein Kinase 2 (SRPK2)	1	6 (2)	6023 (6023)
10318	TNFAIP3 Interacting Protein 1 (TNIP1)	1	7 (7)	115 (115)
2997	Glycogen Synthase 1 (GYS1)	3	4 (4)	384 (384)
10949	Heterogeneous Nuclear Ribonucleoprotein A0 (HNRNPA0)	2	9 (2)	5 (0)
64112	Modulator of Apoptosis 1 (MOAP1)	1	8 (8)	6942 (6931)
10419	Protein Arginine Methyltransferase 5 (PRMT5)	3	26 (17)	6996 (4743)
10262	Splicing Factor 3b Subunit 4 (SF3B4)	3	13 (7)	82 (44)
23321	Tripartite Motif Containing 2 (TRIM2)	3	2 (2)	15 (15)
81603	Tripartite Motif Containing 8 (TRIM8)	3	3 (3)	0 (0)

Lastly, an analysis was performed to observe the network metrics of proteins that are both IAV interacting and driver proteins, totaling 8.9% of all driver proteins. There is a significant increase in the betweenness of driver proteins depending on their status as IAV interacting or IAV non-interacting proteins (Fisher test p-value: 2.20×10^{-16}) where there is no significant difference in degree of the same groups (Fisher test p-value: 0.7161). This is further evidence that the addition of virus interactions to the network magnifies information flow through the proteins most involved in controlling network behavior.

Liu Controllability

Liu controllability was calculated (see Methods) for all proteins of the HIN and VIN (as shown in Table 2 with and without parentheses, respectively). The addition of IAV interactions to the network has no effect on the distribution of Liu classification of host proteins, meaning that the Liu classification of the "IAV Interacting proteins" also does not change with the addition of viral interactions. Upon entry to the VIN, the 11 IAV proteins are classified as neutral, meaning their removal does not alter the number of driver proteins required to control the VIN ($N_D = N_D'$). This reveals that singular viral proteins do not make integral changes to the control structure of the VIN.

Table 2: Liu types of all proteins, driver proteins, and virus interacting proteins in the VIN (HIN in parenthesis)

	All Proteins	**Driver Proteins**	**IAV Interacting Proteins**
Indispensable	1,169 (1,169)	0 (0)	186 (186)
Neutral	2,669 (2,658)	803 (799)	312 (312)
Dispensable	2,454 (2,454)	1,660 (1,664)	254 (254)

While none of the proteins change Liu classification between networks, the replacement of host protein driver proteins with viral proteins creates a small change in Liu type distribution for driver proteins. Of the 11 displaced host proteins (found in Table 1), seven are neutral in the HIN

(meaning that their removal from the network does not change the number of driver proteins) and four are redundant in the HIN (meaning their removal reduces the number of driver proteins needed). Of the five host proteins that are both drivers and IAV interacting, four are neutral and one is redundant. The most notable change in degree and betweenness between the HIN and VIN is *PRMT5*, with an increase of 9 and 2,250, respectively. Overall, Liu classification results suggest that the HIN is stable against potential changes in the control structure that could be caused by the addition of IAV interactions.

We next developed an analysis to test if IAV is selectively targeting host proteins based on controllability characteristics. 10,000 random sets of 752 proteins (the number of IAV interacting proteins) were pulled from the host proteins of the VIN. Their Liu type distributions were plotted against the percent classification of IAV interacting proteins, driver proteins, and all proteins in the VIN (Fig. 3a-c). The randomly sampled sets closely resemble all proteins of the network, suggesting that the true IAV interacting protein set's Liu control behavior is not a coincidence of network construction (one-sided p-values of 0.50, 0.52, and 0.49 for indispensable, neutral, and dispensable, respectively). IAV interacting proteins tend to be indispensable when compared to the percentage of all proteins that are indispensable (Fig. 3a). This suggests that viruses prefer to interact with proteins that are vital to cellular control. Additionally, driver proteins are very likely to be dispensable proteins compared to the percent of all proteins that are dispensable (Fig. 3c). Further, the mean and median log degree and betweenness of the randomly sampled protein sets is lower than the same measurements of the true IAV interacting set (Fig. 4), signifying that virus interacting proteins are in positions of network significance. Overall, the Liu controllability results of IAV interacting proteins suggest that the virus may be selectively targeting host proteins based on controllability characteristics.

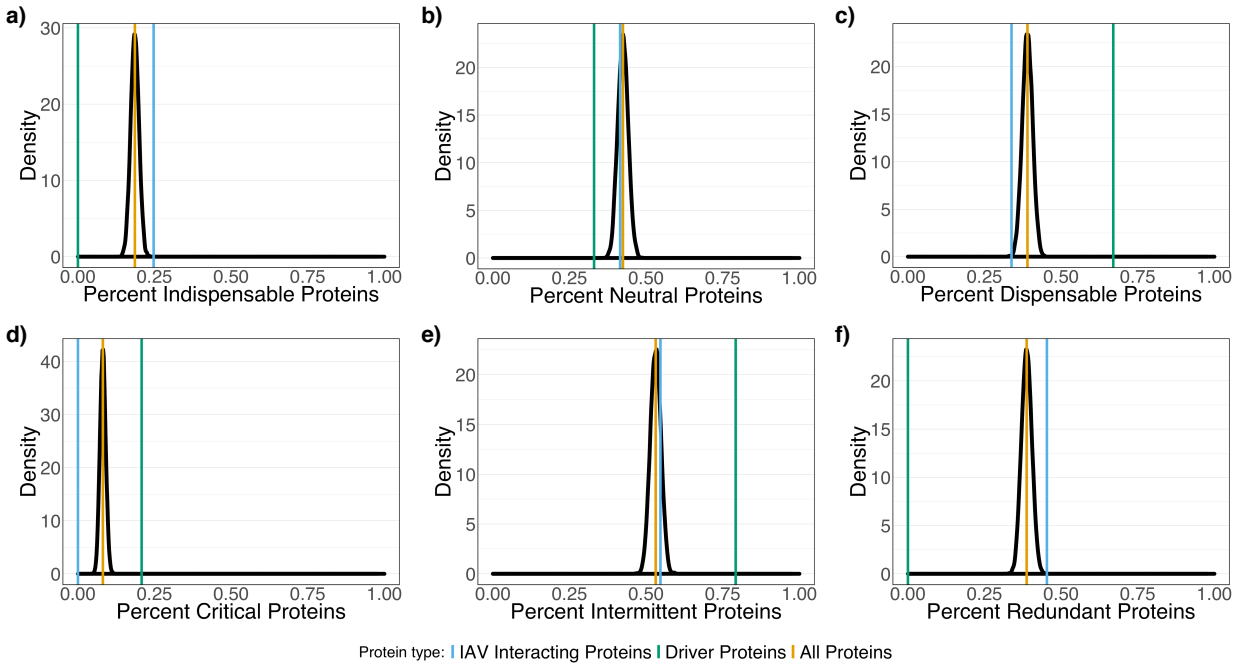


Figure 3: a-c) Density plots of distribution of Liu type for 10,000 random pulls of 752 proteins (number of virus interacting proteins in network). d-f) Density plots of distribution of Jia type for 10,000 random pulls of 752 proteins (number of virus interacting proteins in network). Values for IAV interacting proteins (red), driver proteins (green), and all proteins (blue) are pictured for all figures.

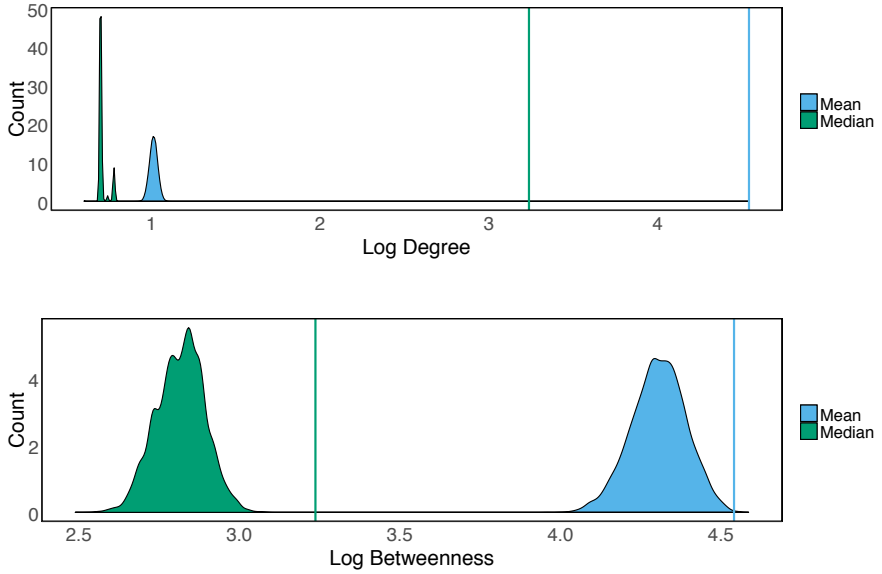


Figure 4: Density plots of a) mean (blue) and median (green) log degree of random IAV interacting protein sets and b) mean (blue) and median (green) log betweenness of random IAV interacting protein. Values for the true IAV interaction set shown as vertical lines, evidence that host proteins that directly interact with viral proteins are in positions of network significance.

Jia Controllability

Jia controllability was calculated (see Methods) for all proteins of the HIN and VIN (as shown in Table 3 with and without parentheses, respectively). Unlike in Liu controllability, there is a small disturbance to Jia type distributions of host proteins after the addition of IAV-host interactions. There are 24 host proteins that shift from being classified as critical (a member of all MISs) to intermittent (a member of some MISs) proteins. Identities of these proteins can be found in Table 4 along with the shortest distance to an IAV protein in the network and protein degree and betweenness. The two most notable changes in degree and betweenness between the HIN and VIN is EPH receptor A2 with an increase of 1 and 93, respectively, and transferrin receptor, with an increase of 3 and 164, respectively. All 24 proteins are drivers and IAV interacting. There are only two proteins (*EPHA2* and *HNRNPA0*) that are also members of the set of 11 proteins that are removed from the MIS after the addition of IAV interactions to the network. 45% of IAV interacting proteins are never drivers, suggesting that they are always manipulated by neighboring proteins. IAV interacting proteins are not enriched for driver proteins (Fisher test p-value: 0.14).

Table 3: Jia types of all proteins, driver proteins, and virus interacting proteins in the VIN (HIN in parenthesis)

	All Proteins	**Driver Proteins**	**IAV Interacting Proteins**
Critical	512 (525)	512 (525)	0 (24)
Intermittent	3,342 (3,318)	1,951 (1,938)	411 (387)
Redundant	2,438 (2,438)	0 (0)	341 (341)

Again, a randomized protein set was created to test if IAV may be selectively interacting with host proteins based on their controllability. 10,000 random sets of 752 proteins (the number of IAV interacting proteins) were sampled from the host proteins of the VIN. Their Jia type distributions were plotted against the percent classification of IAV interacting proteins, driver

proteins, and all proteins in the VIN (Fig. 3d-f). As with the Liu classification, the random sets closely resemble all proteins of the network (one-sided p-values of 0.48, 0.50, and 0.49 for indispensable, dispensable, and redundant, respectively). While there are no redundant driver proteins by definition, driver proteins are more likely to be intermittent proteins than critical proteins (Fig. 3d-e), showing more than 75% of all driver proteins are missing from at least one possible MIS. This means the majority of possible driver proteins are able to be controlled by a neighboring protein in at least one MIS. IAV interacting proteins tend to be redundant compared to the number of all proteins that are redundant (Fig. 3f). This suggests that viruses prefer to interact with proteins that are part of existing control structures, receiving input from neighboring proteins.

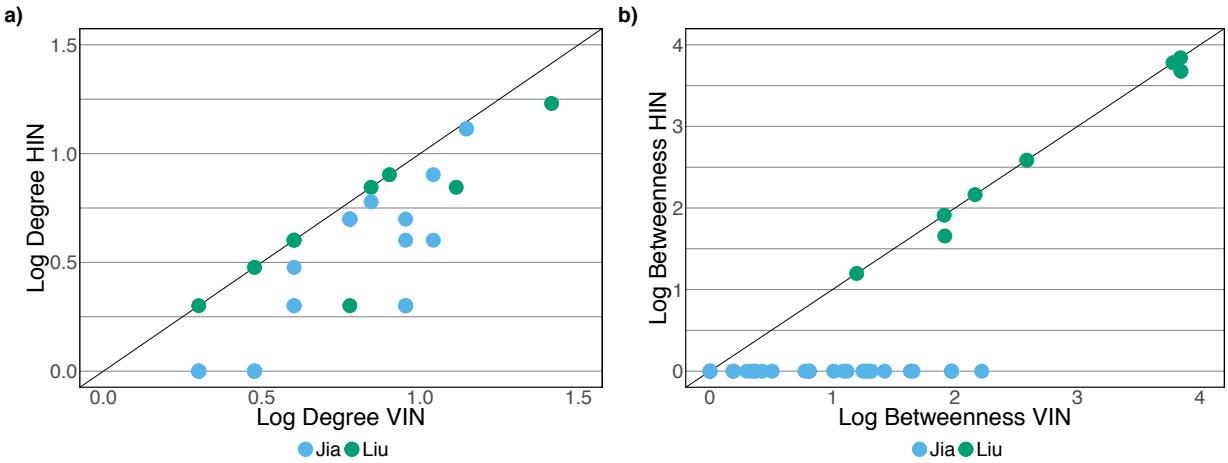


Figure 5: a) Degree and b) betweenness of proteins that change Liu and Jia classification between the HIN and VIN.While proteins identified in the Liu controllability analysis do not show significant deviation in degree or betweenness, proteins identified in the Jia controllability analysis show a shift in both measures after the addition of viral interactions.

Overall, Jia calculations identify a set of proteins for consideration that act differently within the VIN. This is demonstrated through a comparison of degree and betweenness for the identified driver proteins that are shifting controllability classification in the HIN and VIN (Fig. 5). Proteins identified in the Liu analysis show little deviation in both degree (Fig. 5a) and betweenness (Fig. 5b) measures between the HIN and VIN. In contrast, proteins identified in the

Jia analysis show much larger deviations in degree (Fig. 5a) and betweenness (Fig. 5b) with all proteins having a betweenness of 0 in the HIN (Table 4). This means that the identified proteins were not responsible for information flow until the addition of IAV interactions.

Table 4: Identities of proteins that shift Jia classification between the HIN and VIN. All identified proteins are directly interacting with viral proteins. Degree and betweenness of the proteins of the VIN is provided (with the values from the HIN in parenthesis).

Entrez ID	Gene Name	Shortest Distance to IAV Protein	Degree	Betweenness
56655	DNA Polymerase Epsilon 4, Accessory Subunit (POLE4)	1	2 (1)	1 (0)
30846	EH Domain Containing 2 (EHD2)	1	3 (1)	1 (0)
1969	EPH Receptor A2 (EPHA2)	1	14 (13)	93 (0)
2665	GDP Dissociation Inhibitor 2 (GDI2)	1	3 (1)	2 (0)
51552	RAB14, Member RAS Oncogene Family (RAB14)	1	2 (1)	1 (0)
2091	Fibrillarin (FBL)	1	9 (4)	19 (0)
10949	Heterogeneous Nuclear Ribonucleoprotein A0 (HNRNPA0)	1	9 (2)	5 (0)
3032	Hydroxyacyl-Coa Dehydrogenase/3-Ketoacyl-Coa Thiolase/Enoyl-Coa Hydratase (Trifunctional Protein), Beta Subunit (HADHB)	1	9 (5)	26 (0)
3419	Isocitrate Dehydrogenase 3 (NAD(+)) Alpha (IDH3A)	1	3 (1)	2 (0)
4191	Malate Dehydrogenase 2 (MDH2)	1	3(1)	1 (0)
64949	Mitochondrial Ribosomal Protein S26 (MRPS26)	1	2 (1)	0 (0)
9180	Oncostatin M Receptor (OSMR)	1	6 (5)	18 (0)
5052	Peroxiredoxin 1 (PRDX1)	1	11 (4)	44 (0)
5213	Phosphofructokinase, Muscle (PFKM)	1	6 (5)	17 (0)
26227	Phosphoglycerate Dehydrogenase (PHGDH)	1	4 (2)	9 (0)
5817	Poliovirus Receptor (PVR)	1	7 (6)	42 (0)
5686	Proteasome Subunit Alpha 5 (PSMA5)	1	6 (5)	11 (0)
5464	Pyrophosphatase (Inorganic) 1 (PPA1)	1	6 (5)	5 (0)
113174	Serum Amyloid A Like 1 (SAAL1)	1	2 (1)	1 (0)
6745	Signal Sequence Receptor Subunit 1 (SSR1)	1	4 (2)	12 (0)
7037	Transferrin Receptor (TFRC)	1	11 (8)	164 (0)
8834	Transmembrane Protein 11 (TMEM11)	1	4 (3)	20 (0)
30000	Transportin 2 (TNPO2)	1	2 (1)	1 (0)
7407	Valyl-Trna Synthetase (VARS)	1	3 (1)	0 (0)

Validation of controllability significant host factors

All proteins were checked against 6 siRNA screens for host factors involved in influenza replication (Brass et al (Brass *et al.*, 2009), Hao et al (Hao *et al.*, 2008), Karlas et al (Karlas *et al.*, 2010), König et al (König *et al.*, 2010), Shapira et al (Shapira *et al.*, 2009), and Watanabe et al (Watanabe *et al.*, 2014)), grouped by both Liu and Jia controllability classifications. Less than 5% of all classifications of both types are validated by any of the 6 screens (Fig. 6), suggesting that no controllability classification is more enriched for host factors than others. This is likely due to the low agreement observed across siRNA studies (Hao *et al.*, 2013). However, the driver proteins that change Liu and Jia classification have higher hit rates in siRNA screens, with 1 of 11 changing MIS proteins (9% validation) and 2 of 24 Jia-identified proteins (8% validation), though neither are significant results (Fisher p-values of 1 and 0.5696, respectively). However, *HNRNPA0* is triple validated, which is highly unlikely (the odds of detecting the same protein from the proteome twice, 1:20,000) (Bushman *et al.*, 2009; Hao *et al.*, 2013).

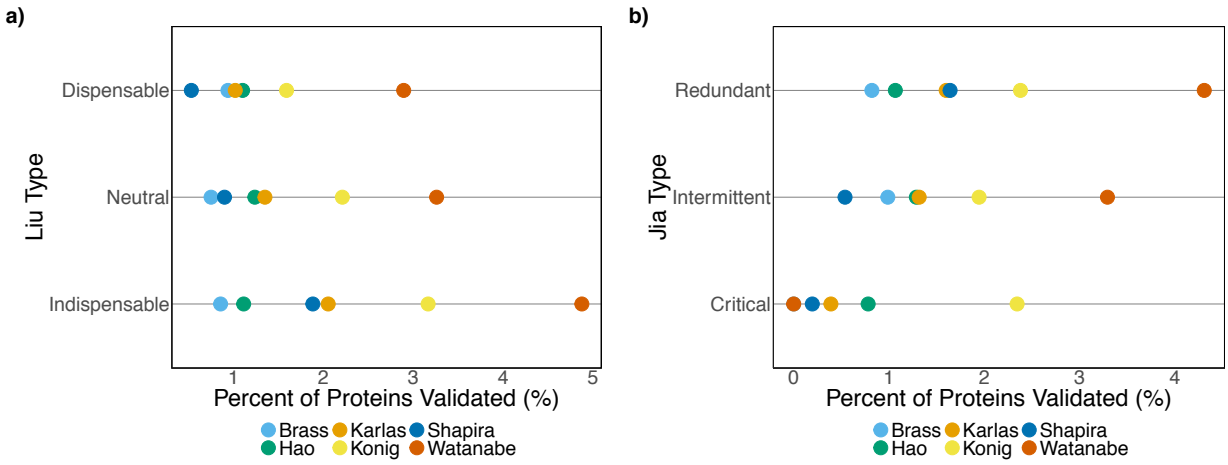


Figure 6: Percent of each a) Liu protein type and b) Jia protein type confirmed in 6 siRNA screens (Brass, Karlas, Shapira, Hao, Konig, Watanabe). None of the 6 possible classifications are more than 5% validated in the screenings, suggesting that experimental findings do not favor certain protein controllability types.

An analysis of both protein sets of interest was performed using Ingenuity Pathway Analysis (IPA) (Krämer *et al.*, 2013). The network created for the 11 changing MIS proteins identified cellular compromise, cell death, and cell cycle functions. The network created for the 24 Jia-identified proteins identified protein synthesis functions, all centered around NF-kB. The Jia network notably recognizes six proteins (*EPHA2, FBL, PFKM, PSMA5, SSR1,* and *TFRC*) for their involvement in the infection of cells (p-value: 9.58×10^{-4}). Four proteins (*CELF1, HNRNPA0, SF384,* and *SRPK2*) were identified for their involvement in mRNA processing (p-value: 3.33×10^{-6}) in both networks.

Lastly, Interferome (Samarajiwa *et al.*, 2008) was used to determine if the 11 changing MIS proteins and 24 Jia-identified proteins are interferon regulated genes (IRGs). When treated with interferon, all 11 changing MIS proteins and 12 of 24 Jia-identified proteins exhibit a 2-fold change in expression in at least one experimental dataset. In particular, *HNRNPA0* is significantly down regulated in 13 separate studies.

DISCUSSION

In total, this study has completed a two-part network controllability analysis on both a host protein-protein interaction network (HIN) and an integrated influenza virus-host protein-protein interaction network (VIN) to enhance the prediction of antiviral drug targets for influenza A virus. The unique construction of the VIN includes experimentally-derived virus-host interaction data (Watanabe *et al.*, 2014) which represents opportunities for the virus to manipulate host intracellular machinery using protein-protein interactions. While Liu controllability methods have been previously applied to PPI networks (Liu *et al.*, 2011; A. Vinayagam *et al.*, 2016), a Jia controllability analysis has never been applied to PPI networks to our knowledge. Our workflow observes both the effect of structural changes to the network in the case of potential protein

knock outs, as well as each protein's role in all MISs, representing all possible ways of controlling the system. In combination, the results of these methods provide deeper understanding of how changes to cell behavior at the onset of infection are able to occur through the work of a small set of viral proteins. This offers new possibilities to "outsmart" viral attack by dismantling the control structure which allows the viral infection to take hold.

Network representations of the cellular environment demonstrate that the effects of infection (represented by the addition of virus-host interactions) cascade through the system, altering basic topology measures. The betweenness shift between the two networks, particularly in IAV interacting proteins, supplies evidence that the topological effect of viral infection is wide reaching (Tables 1 and 4). Further, a comparison of the betweenness of driver proteins that are and are not also IAV interacting proteins shows a significant difference. By definition, driver proteins that are IAV interacting are not receiving control influence from viral proteins and require additional external influence to achieve network control. However, the increased betweenness of proteins that are both driver and IAV interacting proteins suggests that this group is still of great importance to information flow through the network. This is one example where differences in network topology measures can emphasize the importance of select proteins that are overlooked by controllability principles.

Controllability analysis shows that IAV interacting proteins are in positions of significance for both types of classification. The increased population of indispensable IAV interacting proteins (Liu controllability: $N_D{}' > N_D$, Fig. 3a) compared to what would be expected by random chance suggests that it would be more difficult for an outside influence (such as viral infection) to control the network in the absence of IAV interacting proteins opposed to a randomly selected protein. This is logical as IAV interacting proteins act as the connection between viral proteins

and the host network where control is initiated. The increased population of redundant IAV interacting proteins (Jia controllability: never a driver protein, Fig. 3f) when compared to the random expectation indicates that more IAV interacting proteins are always being manipulated internally than would be expected by chance. This means that they are fully incorporated into the control structure of the VIN. From these two results, one can conclude that IAV interacting proteins contribute to both the "gate" (the ease of entering the system) and the "heart" (the proteins responsible for propagating control through the system) of the network control structure during infection. These findings support the idea that viruses are likely to interact with proteins which offer an advantage to total network control.

Similarly, both sets of controllability results demonstrate that driver proteins play interesting roles in the network control structure. The large population of dispensable driver proteins (Liu controllability: $N_D{}' < N_D$, Table 2) signifies that the majority of driver proteins are making it more difficult to control the network by requiring more external inputs to control system behavior. In their absence, the number of driver proteins would decrease and it would theoretically be easier for a viral attack to compromise the network control structure. As such, one possible strategy for drug development is to protect these proteins from drastic changes to abundance. Over 75% of driver proteins are classified as intermittent (Jia controllability: sometimes a driver protein, Table 3), meaning there is at least one MIS where these driver proteins are not drivers, and receive control influence through internal propagation. This lends itself to the idea of viral escape routes: under pressure, virus proteins could utilize alternative pathways to maintain system control and reach the goal of hijacking cellular function.

The two network controllability methods identify protein sets of interest through changes to classification in both controllability types between the HIN and VIN. Unfortunately, the results

of Liu classification do not detect a change between the two networks in this study. As it is a measure of the robustness of the network to structural changes in the absence of each protein, this suggests that the HIN upholds its typical control structure during IAV infection. This could be a consequence of the interaction data being used being or it may be that the strategy applied here cannot distinguish between the behavior of healthy and diseased states. Knowing the extent of changes to cell behavior such as immune response (Koyama *et al.*, 2008; Thompson *et al.*, 2011; Iwasaki and immunology, 2004), apoptosis signaling (death and differentiation, 2001; journal of experimental pathology, 2001), and transcriptional processes (Gale *et al.*, 2000; Sonenberg and Cell, 2009; Walsh *et al.*, 2013) during infection, the IAV infected cell can be interpreted as a different system. The failure to see this distinction may be a short coming of the Liu controllability calculation, especially knowing that the 11 changing MIS proteins are not unique due to the method's use of a single MIS. Overall, the Liu analysis should be applied to additional virus-host networks to further evaluate the method.

The 24 proteins identified by the Jia controllability analysis show promise as indicators of regulatory roles specific to the infected state, particularly *PRMT5* whose betweenness is 250 times its HIN value after the addition of viral interactions to the network (Table 4). It is also noteworthy that *PRDX1* has been implicated in respiratory syncytial virus (RSV) (Pavia, 2011), a lower respiratory tract infection that is often associated with influenza virus (Jamaluddin *et al.*, 2010). Though a higher percentage of the proteins identified as changing MIS proteins and in the Jia controllability analysis are validated by existing siRNA screening data than any of the groups of 6 possible controllability classifications, this validation is still not statistically significant. Remarkably, one protein (*HNRNPA0*) is validated in 3 studies: a rare occurrence. The IPA analysis reveals that some of the identified proteins hold roles in mRNA processing, an integral

part of the influenza virus' ability to spread through processing its own RNA using host machinery (Dubois *et al.*, 2014). The Jia-identified network is centered around NF-kB, which is implicated in host immunity with evidence that the virus directly inhibits NF-kB activity (Kumar *et al.*, 2008; Ludwig and Planz, 2008). The interferon regulating roles of proteins in a high number of both identified sets (100% and 50% of changing MIS and Jia-identified proteins, respectively) speak to their responsibility in controlling infection. Again, *HNRNPA0* appears as downregulated in 13 studies when treated with interferon compared to a control. In total, this evidence suggests that controllability analyses hold power as predictors for important host factors in influenza infection and, therefore, hold power for drug target prediction.

Existing influenza virus studies using PPI networks require additional data such as differentially expressed gene information (Shoemaker and Fukuyama, 2012) or protein context (Schaefer *et al.*, 2013) to construct host response networks. Alternative methods such as DeltaNet (Noh and Bioinformatics, 2016; Noh *et al.*, 2016) and ProTINA (Noh *et al.*, 2018) utilize gene transcription profiles to infer protein drug targets, but rely on the accurate deduction of gene regulatory networks. More recent PPI studies have used network growing functions such as GeneMANIA, STRING, and IPA (Taye *et al.*, 2017) to predict IAV host factors and studied infected cell systems through the integration of screening data with network methods (Tripathi *et al.*, 2015; Heaton *et al.*, 2016). Approaches incorporating time course data into network analysis have also been explored (Jain *et al.*, 2016). While these methods (which include basic network metrics such as degree and betweenness of PPI networks) have been successful at identifying disease host factors and in drug target development in the existing body of work, this dual controllability study offers a novel, in-depth analysis of the role of individual proteins in the context of total system function and how possible changes to the system can be interpreted.

Lastly, though this study has used experimental data from Influenza A studies, this analysis can be used to improve the prediction of drug targets for any pathogen-host interaction given the availability protein interaction data because of the generality of the method. The limits of the work lie in limited availability of large-scale, dependable databases of protein-protein interactions. Additionally, foundational maximum matching algorithms for the calculation of driver proteins must be performed with directed networks. Therefore, the future of this field depends on continued establishment of large, confident, directed PPI networks.

METHODS

Protein-protein interaction network

The host protein-protein interaction network used was downloaded from Vinayagam et al (Vinayagam *et al.*, 2011). A confidence level cutoff of 0.7 was used. Influenza A virus-host interactions from Watanabe et al (Watanabe *et al.*, 2014) were narrowed to interactions which contained host proteins already found in host network interactions to avoid skewing degree and betweenness network metrics. These interactions were directly integrated into the host network. All analysis was completed in R 3.4.3 using the igraph package.

Liu classification

Calculations for Liu classification were adopted from Liu et al (Liu *et al.*, 2011). For a network of *n* nodes, a set of driver nodes for the bipartite representation of the network, N_D, is found using a maximum matching algorithm such as Hopcroft-Karp (Hopcroft and Karp, 1973). Each node of the network is iteratively removed ($N' = N - 1$) and maximum matching, N_D', is reevaluated. Nodes are classified as indispensable ($N_D' > N_D$), neutral ($N_D' = N_D$), or dispensable ($N_D' < N_D$).

Jia classification

Calculations for Jia classification were adopted from Jia et al (Jia *et al.*, 2013). For a network of n nodes, a set of driver nodes for the bipartite representation of the network, N_D, is found using a maximum matching algorithm such as Hopcroft-Karp (Hopcroft and Karp, 1973). For all N_D, control adjacent nodes were identified iteratively and an input graph was created as dictated in Zhang et al (Zhang *et al.*, 2016). The input graph was used to classify nodes as critical (in all minimum input sets), neutral (in some minimum input sets), or redundant (in no minimum input sets).

ACKNOWLEDGEMENTS

Thank you to the Department of Chemical and Petroleum Engineering at the University of Pittsburgh for funding this research. Thank you to the Systems Biology Institute, Tokyo, for expertise and computational training. Thank you to the Howard Hughes Medical Institute for supporting this project through the James H. Gilliam Fellowships for Advanced Study program.

AUTHOR CONTRIBUTIONS

Emily E. Ackerman assisted in conceptualization of the study, designed the study, performed all computational experiments, and wrote the manuscript. Jason E. Shoemaker conceptualized and funded the study. John F. Alcorn assisted in conceptualization of the study and advised on relevant virology and immunology. Takeshi Hase provided computational training.

COMPETING INTERESTS

The authors have no competing interests.

CITATIONS

Aldridge, BB *et al.* (2006) Physicochemical modelling of cell signalling pathways. *Nature cell*
Arrell, DK and Terzic, A (2010) Network systems biology for drug discovery. *Clinical Pharmacology & Therapeutics*.
Barabási, AL *et al.* (2011) Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*.
Brass,A.L. *et al.* (2009) The IFITM proteins mediate cellular resistance to influenza A H1N1

virus, West Nile virus, and dengue virus. *Cell*, **139**, 1243–1254.
Bushman, FD *et al.* (2009) Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS ….*
Chassey,D.B. *et al.* (2008) Hepatitis C virus infection protein network. *Molecular systems ….*
Cho, DY *et al.* (2012) Network biology approach to complex diseases. *PLoS computational biology*.
Death and differentiation,B.-G. (2001) Host defense, viruses and apoptosis. *Cell death and differentiation*.
Dubois,J. *et al.* (2014) Influenza viruses and mRNA splicing: doing more with less. *MBio*, **5**, e00070–14.
Freeman, LC (1977) A set of measures of centrality based on betweenness. *Sociometry*.
Gale, M *et al.* (2000) Translational control of viral gene expression in eukaryotes. *Microbiology and molecular biology ….*
Gandhi *et al.* (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, **38**, 285–293.
Germain, MA *et al.* (2014) Elucidating novel hepatitis C virus–host interactions using combined mass spectrometry and functional genomics approaches. *Molecular & Cellular ….*
Hao, L *et al.* (2008) Drosophila RNAi screen identifies host genes important for influenza virus replication. *Nature*.
Hao,L. *et al.* (2013) Limited Agreement of Independent RNAi Screens for Virus-Required Host Genes Owes More to False-Negative than False-Positive Factors. *Plos Comput Biol*, **9**, e1003235.
Hase,T. *et al.* (2009) Structure of protein interaction networks and their implications on drug design. *PLoS computational biology*, **5**, e1000550.
He, X and Zhang, J (2006) Why do hubs tend to be essential in protein networks? *PLoS genetics*.
Heaton, NS *et al.* (2016) Targeting viral proteostasis limits influenza virus, HIV, and Dengue virus infection. *Immunity*.
Hopcroft,J.E. and Karp,R.M. (1973) An n^5/2 algorithm for maximum matchings in bipartite graphs. *SIAM Journal on computing*, **2**, 225–231.
Iwasaki, A and immunology,M.-R. (2004) Toll-like receptor control of the adaptive immune responses. *Nature immunology*.
Jain,S. *et al.* (2016) Reconstructing the temporal progression of HIV-1 immune response pathways. *Method Biochem Anal*, **32**, i253–i261.
Jamaluddin, M *et al.* (2010) Role of peroxiredoxin 1 and peroxiredoxin 4 in protection of respiratory syncytial virus-induced cysteinyl oxidation of nuclear cytoskeletal proteins. *Journal of ….*
Jia,T. *et al.* (2013) Emergence of bimodality in controlling complex networks. *Nat Commun*, **4**, 2002.
Jia,T. and Barabási,A.-L. (2013) Control Capacity and A Random Sampling Method in Exploring Controllability of Complex Networks. *Sci Reports*, **3**.
Jonsson, PF and Bioinformatics,B.-P. (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics*.
Karlas, A *et al.* (2010) Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature*.
Klipp, E and Liebermeister, W (2006) Mathematical modeling of intracellular signaling pathways. *BMC neuroscience*.

König, R *et al.* (2010) Human host factors required for influenza virus replication. *Nature*.
Korth,M.J. *et al.* (2013) Systems approaches to influenza-virus host interactions and the pathogenesis of highly virulent and pandemic viruses. **25**, 228–239.
Koyama, S *et al.* (2008) Innate immune response to viral infection. *Cytokine*.
Krämer, A *et al.* (2013) Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*.
Kumar, N *et al.* (2008) NF-κB signaling differentially regulates influenza virus RNA synthesis. *Journal of virology*.
Lin, CT (1974) Structural controllability. *IEEE Transactions on Automatic Control*.
Liu,Y.-Y. *et al.* (2011) Controllability of complex networks. *Nature*, **473**, 167–173.
Lopes,T.J. *et al.* (2015) Identifying problematic drugs based on the characteristics of their targets. *Front Pharmacol*, **6**, 186.
Ludwig,S. and Planz,O. (2008) Influenza viruses and the NF-κB signaling pathway – towards a novel concept of antiviral therapy. *Biol Chem*, **389**, 1307–1312.
Mani, KM *et al.* (2008) A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Molecular systems …*.
Mine,K.L. *et al.* (2013) Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer. *Nature communications*, **4**, 1806.
Mitchell, HD *et al.* (2013) A network integration approach to predict conserved regulators related to pathogenicity of influenza and SARS-CoV respiratory viruses. *PLoS …*.
Moni, MA and bioinformatics,L.-P. (2014) Network-based analysis of comorbidities risk during an infection: SARS and HIV case studies. *BMC bioinformatics*.
Murali, TM *et al.* (2011) Network-based prediction and analysis of HIV dependency factors. *PLoS computational …*.
Noh, H *et al.* (2016) Inferring Causal Gene Targets from Time Course Expression Data. *IFAC-PapersOnLine*.
Noh, H *et al.* (2018) Network perturbation analysis of gene transcriptional profiles reveals protein targets and mechanism of action of drugs and influenza A viral infection. *Nucleic Acids Research*.
Noh, H and Bioinformatics,G.-R. (2016) Inferring gene targets of drugs and chemical compounds from gene expression profiles. *Bioinformatics*.
Journal of experimental pathology,T.-B. (2001) Viruses and apoptosis. *International journal of experimental pathology*.
Pavia,A.T. (2011) Viral infections of the lower respiratory tract: old viruses, new viruses, and the role of diagnosis. *Clinical Infectious Diseases*, **52**, S284–S289.
Ptak, RG *et al.* (2008) Cataloguing the HIV type 1 human protein interaction network. *... and human …*.
Pujol, A *et al.* (2010) Unveiling the role of network and systems biology in drug discovery. *Trends in pharmacological sciences*.
Rask-Andersen,M. *et al.* (2011) Trends in the exploitation of novel drug targets. *Nature reviews. Drug discovery*, **10**, 579.
Samarajiwa, SA *et al.* (2008) INTERFEROME: the database of interferon regulated genes. *Nucleic acids …*.
Schaefer,M.H. *et al.* (2013) Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Comput. Biol.*, **9**, e1002860.
Schoeberl, B *et al.* (2002) Computational modeling of the dynamics of the MAP kinase cascade

activated by surface and internalized EGF receptors. *Nature* ….
Shapira, SD *et al.* (2009) A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell*.
Shityakov, S *et al.* (2015) Gene expression profiles and protein–protein interaction network analysis in AIDS patients with HIV-associated encephalitis and dementia. *HIV/AIDS (Auckland*.
Shoemaker, JE and Fukuyama, S (2012) Integrated network analysis reveals a novel role for the cell cycle in 2009 pandemic influenza virus-induced inflammation in macaque lungs. *BMC systems* ….
Shoemaker,J.E. *et al.* (2012) Integrated network analysis reveals a novel role for the cell cycle in 2009 pandemic influenza virus-induced inflammation in macaque lungs. *BMC Syst Biol*, **6**, 117.
Sonenberg, N and Cell,H.-A. (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*.
Taye,B. *et al.* (2017) Benchmarking selected computational gene network growing tools in context of virus-host interactions. *Sci Reports*, **7**, 5805.
Thompson, MR *et al.* (2011) Pattern recognition receptors and the innate immune response to viral infection. *Viruses*.
Tripathi,S. *et al.* (2015) Meta-and orthogonal integration of influenza OMICs data defines a role for UBR4 in virus budding. *Cell host & microbe*, **18**, 723–735.
Vinayagam *et al.* (2011) A Directed Protein Interaction Network for Investigating Intracellular Signal Transduction. *Sci Signal*, **4**, rs8–rs8.
Vinayagam, A *et al.* (2016) Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proceedings of the* ….
Vinayagam, A *et al.* (2014) Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. *Nature* ….
Vinayagam,A. *et al.* (2016) Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc National Acad Sci*, **113**, 4976–4981.
Walsh, D *et al.* (2013) Tinkering with translation: protein synthesis in virus-infected cells. *Cold Spring Harbor* ….
Watanabe,T. *et al.* (2014) Influenza virus-host interactome screen as a platform for antiviral drug development. *Cell Host Microbe*, **16**, 795–805.
Wuchty (2014) Controllability in protein interaction networks. *Proceedings of the National Academy of Sciences*, **111**, 7156–7160.
Zhang,X. *et al.* (2016) Input graph: the hidden geometry in controlling complex networks. *Sci Rep*, **6**, 38209.
Zhu,M. *et al.* (2009) The analysis of the drug–targets based on the topological properties in the human protein–protein interaction network. *J Drug Target*, **17**, 524–532.