A Network-Based Approach for Identifying Cancer Causing Pathogens

Joseph Hannigan
Department of Electrical Engineering & Computer Science
United States Military Academy
Joseph.Hannigan@usma.edu

Suzanne J. Matthews
Department of Electrical Engineering & Computer Science
United States Military Academy
Suzanne.Matthews@usma.edu

John K. Wickiser
Department of Chemistry and Life Science
United States Military Academy
John.Wickiser@usma.edu

Paulo Shakarian
Department of Electrical Engineering & Computer Science
United States Military Academy
Paulo.Shakarian@usma.edu

ABSTRACT

We present a new method to identify malignant cancer-causing pathogens by analyzing their interactions with the host protein interaction network. We introduce two new measurements, *core score* and *moment score* that is based on topological characteristics of the network of host proteins that interact with the pathogen. We applied these measurement to a data set consisting of the interactions of 135 pathogens and a human protein-interaction network. We show a strong linear relationship ($R^2 = 0.90$) between the core score and the probability that a pathogen leads to malignant cancer in humans and demonstrate, using a decision tree classifier, that both measurements can be used to correctly identify pathogens that lead to malignant cancer in humans with an accuracy of 97%.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics; G.2.2 [**Graph Theory**]: Network problems; H.2.8 [**Database Applications**]: Data Mining

General Terms

Theory, Measurement, Experimentation

Keywords

Protein Interaction Networks, Network topology, Pathogens, Cancer

1. INTRODUCTION

Pathogens (such as viruses and other microbes) are estimated to cause 20% of all fatal cancers in humans [15].

http://dx.doi.org/10.1145/2638404.2735459.

In recent years, much work [1, 9] has been done to study the role of viruses in causing tumors and malignant cancers. Most studies involve inducing tumors in animal models, as it is very difficult to observe viral tumor growth in human hosts, given the long latency periods for expression.

However, animal models have been limited in their ability to mimic the pathogenesis of cancer-causing viruses in humans [1]. There is great potential in leveraging information gained from human-pathogen protein interaction networks (PIN) for tracking the lethality of viruses. *In silico* detection of cancer causing pathogens can facilitate the process of creating vaccines. If these diseases were prevented, it would reduce cancer cases in developing and developed countries by 26.3%, and 7.7% respectively [15].

Recent analyses of host-pathogen interaction suggest that pathogens associate with proteins based on the topological features of the host protein interaction network [13, 14]. We hypothesize that pathogens interact best with host proteins containing specific features that enable optimal virulence. To test this hypothesis, we developed a new quantitative measurement of pathogens, *core score*, which is based on the topological characteristics of the proteins from the host organism. Based on a study of 135 pathogens interacting with a human protein interaction network, we not only show that this measurement has a strong linear relationship with the probability that the pathogen is cancer-causing in humans ($R^2 = 0.90$), but also that it can successfully be used to classify pathogens as cancer-causing (with an accuracy of 97%). Further, we introduce a second measurement, *moment score*, that when considered together with the *core score* can correctly classify pathogens as directly cancer-causing (leading to cancer in humans in a short time span without requiring additional pathogenic interaction) with an accuracy of 97%.

Our experimental results support the hypothesis that many pathogens may have evolved to achieve optimal virulence by interacting with portions of the host protein network thought to be critical for intra-cellular communication [10]. These results also suggest that the core score may be a useful metric to rate the lethality of a pathogen.

2. RELATED WORK

There have been many recent studies investigating the sig-

Cancer causing pathogens	
Name	S_C
*Hepatitis B virus	9.075
*Hepatitis C virus	45.170
Human herpesvirus 4	15.430
*Human herpesvirus 5	19.153
Human herpesvirus 8	1.403
*Human immunodeficiency virus 1	61.449
Human papillomavirus type 18	9.520
Human papillomavirus type 16	23.194
Human papillomavirus type 31	1.209
Human papillomavirus type 5	15.431
Human papillomavirus type 58	0.154
Human T-lymphotropic virus 1	9.408

Table 1: Cancer causing pathogens with core scores. Pathogens that indirectly lead to cancer are denoted with a star.

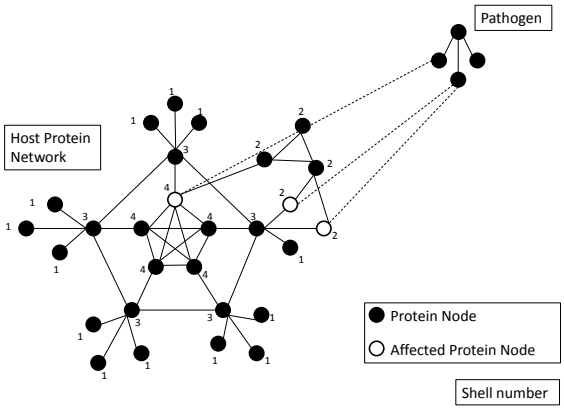

Figure 1: Example network. The first number represents the shell number of the node. See the methods section for a detailed explanation of shell decomposition.

nificance of PINs in various organisms; this work has shown that essential proteins can be determined by network structure [19], allowed for the modeling of protein interactions [6, 2], helped identify disease-causing proteins [5], and proved useful in identifying proteins in an organism associated with cancer [12].

Specifically related to this paper is the recent work examining the study of pathogenic interaction with host PINs and the study of the topological characteristics of proteins that interact with various pathogens [13, 11, 14, 17, 18]. All of these previous studies have focused on identifying topological properties of nodes (or sets of nodes) in the host PIN that are likely targets for pathogenic interaction. This work differs from these previous studies in that we characterize the *pathogen* through analysis of topological properties of host proteins with which it interacts. Specifically, we are concerned with identifying if a pathogen leads to malignant cancer in humans.

3. APPROACH

We utilize the data set of [14] which consists of a human PIN which consists of $63,099$ host interactions (not including self-interactions) over $10,057$ proteins as well as $2,099$ pathogen-host interactions with 416 pathogen proteins belonging to 135 pathogens. Using the available literature, we determined whether or not each pathogen leads to cancer in humans. These are listed in Table 1. We present a list of all supporting references for cancer-causing pathogens in the supplement. Note that some of the pathogens listed, such as HIV which are highly correlated with cancer, but either have a long term dormancy period or there is no reported evidence for direct causation. These are denoted with a star. The remaining pathogens on the list can be thought of as directly leading to cancer. In this study, we consider both types of cancer-causing pathogens. All other pathogens either supported by literature as non-carcinogenic or which lacked reports of carcinogenicity were considered not cancer causing.

To characterize the pathogens by the topology of the targeted host proteins, we created two new measurements that are introduced for the first time in this paper: core score (S_C) and moment score (S_M). These measures depend on the degree and shell number of the host proteins which with the pathogen interacts. The degree of a host protein is simply the number of other interacting host proteins. An S-core of a PIN is the maximum sub-graph where each protein is connected to at least S other proteins and the shell number of a protein is largest value S such that the protein is included in that S-core. The shell number can be easily determined using shell decomposition, described in [16] and proceeds as follows: remove nodes with less than or equal to degree 1 and assign them a shell number of 1, recalculating degree every time a node is removed. Repeat this process but for nodes with less than or equal to degree 2, assigning them shell number 2. Then this process is repeated for nodes with less than or equal to degree 2, assigning them shell number 2. This process is continued until all nodes have are removed and have a shell number. For a given shell S, we define its moment $\langle k_S \rangle$ be the moment (average degree) of nodes in that shell. Figure 1 shows an example shell protein network with a pathogen interaction.

Hence, the core score and moment score are calculated as follows: For a given pathogen P let $P_{interact}$ be the total number of proteins in the host that P interacts with. For given shell S let $P_{interact}(S)$ be the number of pathogen interactions with shell S. For shell S, let $size(S)$ be the size (number of nodes in) shell S. Core score and moment score are defined in equations 1 and 2 below:

$$S_C = \sum_S \frac{S \cdot P_{interact}(S)}{size(S)} \tag{1}$$

$$S_M = \sum_S \frac{\langle k_S \rangle \cdot P_{interact}(S)}{size(S)} \tag{2}$$

From Figure 1, S_C is calculated for the sample pathogen. Because the pathogen interacts with 3 different proteins in the host network, $\sum_S P_{interact}(S) = 3$. $size(1) = 13$, $size(2) = 5$, $size(3) = 5$, and $size(4) = 5$. $P_{interact}(1) = 0$, $P_{interact}(2) = 2$, $P_{interact}(3) = 0$, and $P_{interact}(4) = 1$. The summation yields $S_C = 1.60$.

In other words, S_C is the summation of the ratio of interactions a virus has with each shell of a host v.s. the total interactions, weighted with the shell number itself. S_M is the same, except it is weighted with the moment of each

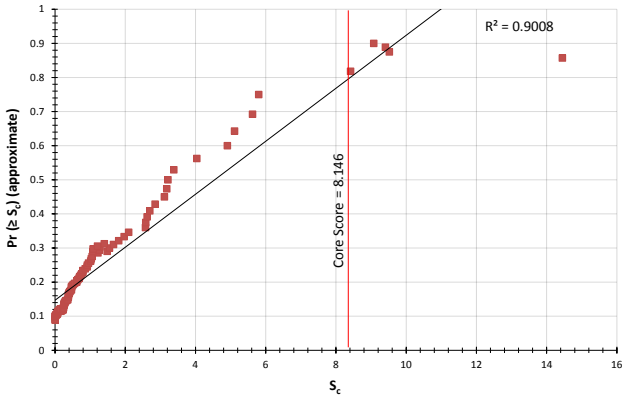

Figure 2: Direct and Indirect Cancer Causing Pathogens represented as the core score versus the fraction of cancer causing pathogens that are greater than or equal to the corresponding core score on the x-axis. This graph omits the data points at the end of the x axis where y = 1.

shell. Therefore, S_M approaches S_C from above as the network becomes more connected with fewer, higher shells. Our intuition behind these measures is derived from our previous work in [17] where we showed that shells with higher moments had a greater number of pathogenic interactions.

The Weka J-48 decision tree classifier [7] was used to ascertain if our measures can be used to correctly classify pathogens as being cancer causing or not. Weka is a commonly used software package for data mining analysis, and the J-48 decision tree classifier is typically used for decision tree analysis.

4. RESULTS

Let $Pr(\geq S_C)$ denote the probability that a pathogen is directly or indirectly cancer-causing in humans given that its core score is greater than or equal to S_C. We approximate this probability by taking the fraction of the pathogens examined in this study that have a certain core score or greater (using the observed core scores from the population of pathogens examined). We found a linear relationship between the core score and this probability (Figure 2, $R^2 = 0.90$). This strong correlation suggests that the core score can be used to identify cancer causing pathogens. To test this hypothesis, we used a decision tree to identify cancer causing pathogens based on this measure. We found, based on our population, pathogens with a core score greater than 8.416 were often correctly identified as cancer causing (97% accuracy, 3 false negatives, 1 false positive).

Table 2 summarizes the incorrectly classified pathogens. The false negatives were Human herpes virus 8 (HHV 8), Human papillomavirus 31 (HPV 31), and Human papillomavirus 58 (HPV 58). These three false negative pathogens have abnormally low core scores (1.403, 1.209, and 0.154 respectively). However, we note that infection with HHV 8 (also known as Kaposi's Sarcoma virus, or KSV) while known to directly cause tumorgenesis in humans, is reported to be insufficient to produce disease alone [9]. Due to the pathogen's requirement for another virus such as HIV to cause cancer, we hypothesize that the virus does not need

Type I Error(Rejected True)		
Name	S_C	S_M
Human herpesvirus 8	1.403	6.280
Human papillomavirus type 31	1.209	4.863
Human papillomavirus type 58	0.154	0.373
Type II Error(Failed to Reject False)		
Name	S_C	S_M
Vaccinia virus	14.447	108.04

Table 2: Results of decision tree analysis on indirect and directly cancer causing pathogens. Method yielded 3 false negatives (Type I error) and 1 false positive (Type II error).

Type I Error(Rejected True)		
Name	S_C	S_M
Human herpesvirus 8	1.403	6.280
Human papillomavirus type 31	1.209	4.863
Human papillomavirus type 58	0.154	0.373

Table 3: Results of decision tree analysis on indirect and directly cancer causing pathogens. Method yielded 3 false negatives (Type I error) and 1 false positive (Type II error).

to infiltrate the core of the host PIN in order to achieve optimal virulence. Further, HPV 58 has a strong geographic component. For instance, it is highly prevalent in cervical cancer of East Asian women but is rare in cancer among North American women. Further exploration is needed to establish a link between low core score and these two viruses. As for Vaccinia, our sole false-positive, there may be some similarities between its host interaction and that of a carcinogenic pathogen as Vaccinia has recently been noted to interfere with cancer [3, 4, 8]. This may suggest that false-positives found with this measurement may be used to repress cancer growth in certain cases - this is may be an important direction for future work.

We then studied the problem of identifying pathogens that only directly cause cancer. As stated earlier, we label a pathogen as directly causing cancer if there was clear evidence in the literature establishing a direct link between infection and development of cancer. These pathogens are denoted with an star in Table 1. Using this more narrow definition of "cancer causing" we find that our linear relationship between S_C and $Pr(\geq S_C)$ is significantly weakened, noting that this probability monotonically increases for core scores under 9.075 and reverses after this point.

However, if we consider both core score and moment score together, we can correctly identify pathogens that directly cause cancer with a 97% accuracy based on the results of our decision tree (Figure 3). Using this two-variable classifier, we obtained no false positives and only three false negatives 3. We hypothesize this arises due to moment score being associated with a virus' ability to penetrate deep into the core of a host. However, based on our previous findings [17] where we showed that the core proteins are often under-targeted, we suspect that high penetration of the core may cause apitosis, or programmed cell death. As most cancer causing viruses are known to have an ability to disable or delay the host's mechanism for apitosis, this could explain the cut off for very high moment scores.

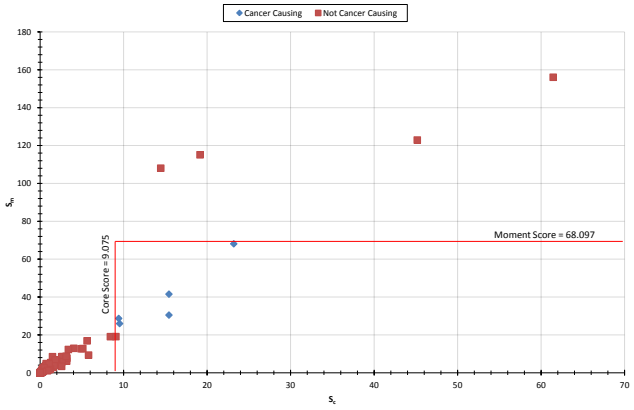

Figure 3: Direct Cancer Causing Pathogens: Moment Score vs. Core Score. Each data point represents a data point with a given core and moment score. The legend specifies which pathogens are/are not cancer causing. The box represents the area that the decision tree determines as cancer causing.

Overall, our results suggest that the core and moment scores of host-pathogen protein interactomes show promise in helping facilitate the classification of unknown pathogens as cancer causing. Our method could be used as a "pre-check" prior to conducting more expensive wet lab testing, and assist in the rapid identification of newly discovered pathogens as being carcinogenic.

5. CONCLUSIONS

In this paper, we characterized pathogens by network-topological characteristics of the host proteins they interact with using two new measurements that we call the core score and the moment socre. Using "ground truth" data based on a literature review, we showed that these measurements can be used to identify if the pathogens lead to malignant cancer in humans with 97% accuracy. Further, the linear relationship we found between the core score and the probability of a pathogen leading to cancer indicates that the techniques presented here could be potentially used to measure the level of carcinogenicity for a given pathogen.

One potential shortcoming of our results is that we relied entirely on existing data. We note that many of these data sources were derived from previous studies that focused on cancer-causing pathogens, which may bias some our results. An important direction for future research in this topic is to create a new, less-biased dataset of pathogens and their host interactions. That said, we feel this work is a useful "first step" toward identifying cancer-causing pathogens based on host protein network interaction.

There are other important avenues for future work as well. These include how to best characterize these measurements on networks with erroneous, missing or uncertain interaction, determining the relevance of the topological characteristics of the proteins in the pathogen, and exploring the possibility of these measurements as useful predictors for other types of diseases.

Acknowledgements

The authors are supported under by the Army Research Office (project 2GDATXR042). The opinions in this paper are those of the authors and do not necessarily reflect the opinions of the funders, the U.S. Military Academy, or the U.S. Army.

6. REFERENCES

[1] J. S. Butel. Viral carcinogenesis: revelation of molecular mechanisms and etiology of human disease. *Carcinogenesis*, 21(3):pp. 405–426, 2000.

[2] A. I. M. Consortium. Evidence for Network Evolution in an Arabidopsis Interactome Map. *Science*, 333(6042):601–607, July 2011.

[3] S. Gholami, A. A. Marano, N. G. Chen, A. Frentzen, C.-H. Chen, C. Eveno, E. Lou, L. Belin, A. A. Szalay, and Y. Fong. Enhanced therapeutic effects of a novel oncolytic and anti-angiogenic vaccinia virus against triple-negative breast cancer. *Journal of the American College of Surgeons*, 215(3), Sep. 2012.

[4] M. Gil, M. Seshadri, M. P. Komorowski, S. I. Abrams, and D. Kozbor. Targeting cxcl12/cxcr4 signaling with oncolytic virotherapy disrupts tumor vasculature and inhibits breast cancer metastases. *Proc Natl Acad Sci U S A*, 2013.

[5] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabasi. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[6] M. W. Hahn and A. D. Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. 22(4):803–806, 2005.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutmann, and I. H. Witten. The WEKA data mining software: An update. volume 11 of *SIGKDD Explorations*, 2009.

[8] J. Heo, C. Breitbach, A. Moon, C. Kim, R. Patt, M. Kim, Y. Lee, S. Oh, H. Woo, K. Parato, J. Rintoul, T. Falls, T. Hickman, B. Rhee, J. Bell, D. Kirn, and T. Hwang. Sequential therapy with jx-594, a targeted oncolytic poxvirus, followed by sorafenib in hepatocellular carcinoma: preclinical and clinical demonstration of combination efficacy. *Mol Ther.*, 19(6):1170, Jun. 2011.

[9] R. T. Javier and J. S. Butel. The history of tumor virology. *Cancer Research*, 68(7693), October 2008.

[10] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse. Identification of influential spreaders in complex networks. *Nat Phys*, 6(11):888–893, November 2010.

[11] T. Milenkovíc, V. Memisević, A. Bonato, and N. Przulj. Dominating biological networks. *PLOS One*, 6(8), 2011.

[12] T. Milenkovíc, V. Memisević, A. K. Genesan, and N. Przulj. Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related interaction networks. *Journal of the Royal Society Interface*, 7(44):pp. 423–437, 2010.

[13] M. S. Mukhtar, A.-R. Carvunis, M. Dreze, P. Epple, J. Steinbrenner, J. Moore, M. Tasan, M. Galli, T. Hao, M. T. Nishimura, S. J. Pevzner, S. E.

Donovan, L. Ghamsari, B. Santhanam, V. Romero, M. M. Poulin, F. Gebreab, B. J. Gutierrez, S. Tam, D. Monachello, M. Boxem, C. J. Harbort, N. McDonald, L. Gai, H. Chen, Y. He, E. U. E. Consortium, J. Vandenhaute, F. P. Roth, D. E. Hill, J. R. Ecker, M. Vidal, J. Beynon, P. Braun, and J. L. Dangl. Independently Evolved Virulence Effectors Converge onto Hubs in a Plant Immune System Network. *Science*, 333(6042):596–601, July 2011.

[14] V. Navratil, B. de Chassey, C. R. R. Combe, and V. Lotteau. When the human viral infectome and diseasome networks collide: towards a systems biology platform for the aetiology of human diseases. *BMC systems biology*, 5(1):13+, 2011.

[15] D. M. Parkin. The global health burden of infection-associated cancers in the year 2002. *International Journal of Cancer*, 118(12):pp. 3030–3044, June 2006.

[16] S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269 – 287, 1983.

[17] P. Shakarian and J. K. Wickiser. Similar pathogen targets in arabidopsis thaliana and homo sapiens protein networks. *PLoS ONE*, 7, 09 2012.

[18] R. W. Solava, R. P. Michaels, and T. Milenkovic. Graphlet-based edge clustering reveals pathogen-interacting proteins. *Bioinformatics*, 28(18):480–486, 2012.

[19] H. Yu, D. Greenbaum, H. X. Lu, X. Zhu, and M. Gerstein. Genomic analysis of essentiality within protein networks. *Trends Genet*, 20(6):227–231, June 2004.