Journal of Carcinogenesis & Mutagenesis

Carson et al., J Carcinogene Mutagene 2012, 3:1
DOI: 10.4172/2157-2518.1000e103

Editorial Open Access

Knowledge Mining of Disease Network can Provide New Insights in Cancer Research through Analysis of Other Diseases

Matthew B. Carson, Cong Liu and Hui Lu*

Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL, 60612-7340, USA

Cancer

The word itself seems to hold a dark power over modern humankind. Nearly every one of us has come in close proximity to this life-threatening disease. In recent years, researchers have produced a body of work that has given us a clearer (albeit more complicated) picture of how cancer comes to be, how it develops, and how it can be treated. The roles of genetics (in the form of single nucleotide polymorphisms or SNPs) [1], epigenetics [2], miRNA [3], copy number variation [4], chromatin structure [5], and protein biomarkers [6] in cancer have been shown. While great scientific advances have been made in the understanding and treatment of this disease in the last 50 years, we still do not have a clear understanding of the 'how' and 'why'. Given a set of initial conditions in the body defined by genetics, lifestyle, environmental exposure, etc., cancer begins and proceeds to develop through an evolutionary process. This results in all cancers having unique characteristics [7]. Clearly, cancer is a multi-dimensional problem for which we have an enormous amount of data now. Gaining knowledge from the existing data, however, is a non-trivial task.

In recent years, bioinformatics and computational biology have made a variety of contributions to disease analysis using existing data in an attempt to increase our understanding of many diseases. Popular topics include the discovery, prediction, and analysis of genes related to disease [8], statistical analysis of SNPs and disease [9], the prediction and discovery of new drug targets [10], the development of the disease ontology and its application to the human genome [11,12], the analysis of protein-protein interaction networks as they relate to disease [13], and many others. Of particular interest is the development of 'disease networks' [14,15], which are in most cases bipartite graphs describing disease-disease as well as disease-gene relationships. In the projection of the disease-gene network that describes disease-disease relationships (Figure 1), nodes indicate diseases and the edge between two nodes represents how these diseases are related. These edges may signify one or more shared genes, metabolic pathways, miRNAs, or a number of other data types. The disease network reveals the interconnected nature of various diseases, which begs the question; can we gain new knowledge of a disease such as cancer by studying 'connected', non-cancer diseases? Many diseases including obesity [16,17], various infections [18], diabetes [19], and possibly even psychological stress [20] have been reported some relationship to cancer. Often the relationship type is unknown or partially known, which indicates that a deeper understanding of these relationships is needed. However, those relationships have not been explored as a whole, but rather as individual links.

Due to the complicated nature of many diseases, which may involve the failure of multiple levels of biological function including DNA repair, gene regulation, epigenetic and histone modifications, metabolic pathways etc., elucidation of disease relationships requires a systematic and computational solution. Though there may be a plethora of data available to quantify this disease problem, the data itself does nothing for us if we cannot turn that data into knowledge (a similar problem arose after the sequencing of the human genome). Merely combining sources of data is not sufficient. We must identify patterns within the data, which is manually infeasible when the number of data points and characteristics to be compared is large. Clearer understanding could be gained by finding, among all attributes of a relationship, those that characterize it most accurately. Several existing machine learning algorithms can help achieve this including multiple instance learning [21], positive/unlabeled (PU) learning [22], Bayesian inference [23], the alternating decision tree, or ADTree [24], and others. In the past we have used the ADTree algorithm to analyze methylation patterns on DNA [25] and to predict DNA-binding proteins [26]. In both cases, this algorithm helped us to understand what characteristics have the most influence on determining the class to which the examples belonged. A similar method of 'rule discovery' is needed in the case of the disease

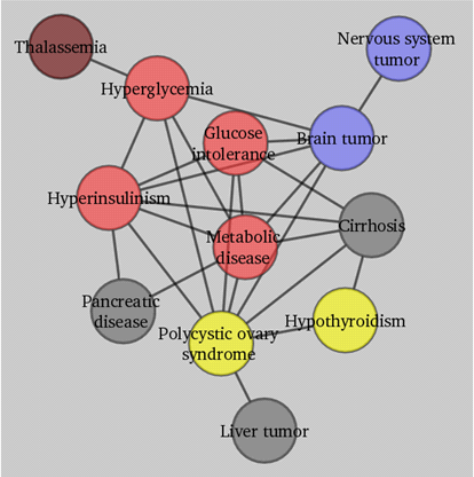

Figure 1: A small example of a projection of the disease-gene bipartite graph that describes disease-disease relationships. Nodes indicate diseases; edges between nodes represent disease relationships. Edges may signify one or more genes, metabolic pathways, miRNAs, or a number of other data types.

***Corresponding author:** Hui Lu, Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL, 60612-7340, USA, E-mail: huilu.bioinfo@gmail.com

Received March 28, 2012; **Accepted** March 29, 2012; **Published** March 31, 2012

Citation: Carson MB, Liu C, Lu H (2012) Knowledge Mining of Disease Network can Provide New Insights in Cancer Research through Analysis of Other Diseases. J Carcinogene Mutagene 3:e103. doi:10.4172/2157-2518.1000e103

network. Of course, the rules may be heavily dependent upon the types of disease in question (i.e. metabolic, infectious, autoimmune and genetic). By analyzing a combination of available genetic, epigenetic, and proteomic data, one will be able to use these algorithms to enrich the edges between cancer and other diseases in the disease network, as well as to predict new edges within disease clusters.

The key to understanding the disease network is to enrich the value of existing edges and to infer new ones based on this enriched value. There is a wealth of information concerning diseases, metabolism, gene ontology, drug targets, miRNA, protein-protein interaction, gene regulation, and gene expression. Unfortunately, there are large areas of missing and overlapping data as well as many false positives and even more false negatives. This makes it difficult to assemble the puzzle and gain knowledge. One can use algorithms such as ADTree which can filter through noisy data to find the most informative and conserved characteristics of a disease-disease relationship. Cancer A and non-cancer disease B, though they may not share a causal gene(s) according to OMIM, but may be related at some distance through a common metabolic pathway, co-regulating transcription factor, or negative regulation by one or more miRNAs. Any of these three could be a false positive association. When analyzed together along with other available data, however, a more complete biological process comes into focus and the noise problem can be mitigated. The ADTree allows us to easily visualize which biological processes contribute most to the disease relationship, eliminating the 'black box' effect of many machine learning algorithms.

Overall, we believe cancer is both unique and related to other diseases. Study of all diseases as a network system can generate many interesting results. For example; drug of related non-cancer diseases may help treat the side effects of cancer drugs; the complex relationship between bacteria and cancer: bacteria can be both beneficial and cancer-causing, can provide new ideas about cancer treatment; mechanisms and tissue-specificity of non-cancer diseases may prime the cellular environment for metastasis. We expect in the near future, with enormous genotype and phenotype data available for all diseases, there will be a novel view point for cancer research that will emerge from the disease network study.

References

1. Dutt A, Beroukhim R (2007) Single nucleotide polymorphism array analysis of cancer. Curr Opin Oncol 19: 43-49.
2. Sharma S, Kelly TK, Jones PA (2010) Epigenetics in cancer. Carcinogenesis 31: 27-36.
3. Visone R, Croce CM (2009) MiRNAs and cancer. Am J Pathol 174: 1131-1138.
4. Kuiper RP, Ligtenberg MJ, Hoogerbrugge N, Geurts van Kessel A (2010) Germline copy number variation and cancer risk. Curr Opin Genet Dev 20: 282-289.
5. Brock MV, Herman JG, Baylin SB (2007) Cancer as a manifestation of aberrant chromatin structure. Cancer J 13: 3-8.
6. Gagnon A, Ye B (2008) Discovery and application of protein biomarkers for ovarian cancer. Curr Opin Obstet Gynecol 20: 9-13.
7. Greaves M, Maley CC (2012) Clonal evolution in cancer. Nature 481: 306-313.
8. Wang E (2010) Cancer systems biology in Chapman & Hall/CRC mathematical and computational biology series. CRC Press: Boca Raton 191-212.
9. Li H, Lee Y, Chen JL, Rebman E, Li J, et al. (2012) Complex-disease networks of trait-associated single-nucleotide polymorphisms (SNPs) unveiled by information theory. J Am Med Inform Assoc 19: 295-305.
10. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, et al. (2010) Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput Biol 6: e1000662.
11. Osborne JD (2007) GeneRIF is a more comprehensive, current and computationally tractable source of gene-disease relationships than OMIM, in Technical Report: Bioinformatics Core. Northwestern University.
12. Osborne JD, Flatow J, Holko M, Lin SM, Kibbe WA, et al. (2009) Annotating the human genome with Disease Ontology. BMC Genomics 10 Suppl 1: S6.
13. Ideker T, Sharan R (2008) Protein networks in disease. Genome Res 18: 644-652.
14. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. Proc Natl Acad Sci U S A 104: 8685-8690.
15. Zhang M, Zhu C, Jacomy A, Lu LJ, Jegga AG (2011) The orphan disease networks. Am J Hum Genet 88: 755-766.
16. Kushi LH, Byers T, Doyle C, Bandera EV, McCullough M, et al. (2006) American Cancer Society Guidelines on Nutrition and Physical Activity for cancer prevention: reducing the risk of cancer with healthy food choices and physical activity. CA Cancer J Clin 56: 254-81.
17. Taubes G (2012) Cancer research. Unraveling the obesity-cancer connection. Science. 335: 28,30-32.
18. Anand P, Kunnumakkara AB, Sundaram C, Harikumar KB, Tharakan ST, et al. (2008) Cancer is a preventable disease that requires major lifestyle changes. Pharm Res 25: 2097-2116.
19. Wang F, Herrington M, Larsson J, Permert J (2003) The relationship between diabetes and pancreatic cancer. Mol Cancer 2: 4.
20. Garssen B (2004) Psychological factors and cancer development: evidence after 30 years of research. Clin Psychol Rev 24: 315-338.
21. Dietterich TG, Richard HL, Lozano PT (1997) Solving the multiple instance problem with axis-parallel rectangles. Artif Intell 89: 31-71.
22. Liu B (2007) Web data mining : exploring hyperlinks, contents, and usage data. Data-centric systems and applications. Berlin New York Springer 532.
23. Bickel PJ, Doksum KA (2001) Mathematical statistics: basic ideas and selected topics. 2nd ed. Upper Saddle River NJ Prentice Hall.
24. Freund Y, Mason L (1999) The Alternating Decision Tree Learning Algorithm, in Proceedings of the Sixteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 124-133.
25. Carson MB, Langlois R, Lu H (2008) Mining knowledge for the methylation status of CpG islands using alternating decision trees. Conf Proc IEEE Eng Med Biol Soc 2008: 3787-3790.
26. Langlois RE, Lu H (2010) Boosting the prediction and understanding of DNA-binding domains from sequence. Nucleic Acids Res 38: 3149-3158.