An Application of Computer Networks toward Complex Biological Data Analysis

Christian K. Thomas* & Steve Smith**

*Lecturer, Dept. of Computer Science, Amsterdam University of Applied Sciences, Netherlands. E-Mail: Thomas_kchristian@gmail.com
** Lecturer, Dept. of Computer Science, Amsterdam University of Applied Sciences, Netherlands. E-Mail: smithsteveaps@yahoo.com

***Abstract*---** The practice of complex biological data analysis involves the collection, classification, storage and analysis of biological and biomedical information ─ using computers. Some of the specific areas in which these processes are applied include genomics and genetics. Also, an example of the specific data targeted by complex biological data analysis involves genetic codes. This paper has examined the concept of network analysis and its application in complex biological data analysis. Indeed, the process of network analysis has evolved with the aim of addressing the problem of using prior biological knowledge and expression data to establish differentially expressed gene or pathway groups. In particular, gene interaction networks and their related analysis processes have aided in searching for high-scoring sub-networks. From the findings reported in most of the previous scholarly studies, it can be inferred that there is increasing improvement in the network analysis process; including algorithms and scoring functions (such as the incorporation of averaging gene expression). When these improvements are achieved, network analysis exhibits higher accuracy and speed, as well as easier biological interpretations. Examples of human microarray datasets on which the network analysis idea has been applied include serum data, regulatory T-cell data, and lymphocyte data; especially that which is related to the immune system and cancer. In summary, the study has established that network analysis continues to play an important role in complex biological data analysis in such a way that improved algorithms and scoring functions have led to improvements in the speed and accuracy of identifying biological process- and disease-related pathways or genes.

***Keywords*---** Human Microarray, Biological Data Analysis, Biochemical Information, Network Analysis.

I. INTRODUCTION

AN example of complex biological data is the case of genetic codes [1]. Other scholarly reports suggest that complex biological data analysis is a process of collecting, classifying, storing and analyzing biological and biochemical information via computers; often applied to genomics and genetics [2]. From these definitions, it can be inferred that complex biological data analysis seeks to establish software tools and methods through which biological data could be understood. In biology, one of the central problems entails the identification of pathways or genes involved in biological processes and diseases [3]. With the evolution of microarray technology and related high-throughput approaches, there have evolved massively parallel approaches to the perceived problem [4]. This paper investigates the practice of applying network analysis in complex biological data analysis.

II. METHODOLOGY

The standard process of network analysis begins with data filtering and normalization before allowing for the computation of test statistics linked to the respective genes. In turn, various phenotypes' expression levels are compared [4, 5]. To account for many genes that are mostly tested, various techniques have been proposed. Also, gene sorting is conducted based on their adjusted *p*-values' increasing order; eventually using the most significant genes towards the generation of biological hypotheses (as well as subjecting them to experimental validation) [6]. As the application of network analysis in complex biological data analysis commences, data representation occurs in the form of undirected graphs whereby the respective nodes represent genes. Also, edges connect two nodes; especially in situations where the genes interact [7]. Nodes that connect to themselves form loops, and the network analysis process requires that the loops are eliminated; resulting in a graph with the desired number of nodes and edges [8]. With the highest degree expected to be 180, most of the previous scholarly affirmations hold that the most common degree is 1 [6-8].

Upon establishing the gene interaction network, the next step entails filtering. This process allows for comparisons between network-based approaches and the single-gene approaches by focusing only on genes associated with at least one interaction [9]. It is also worth noting that the number of false positives can be reduced through the selection and application of multiple testing only for genes linked to

adequate variability across the selected arrays; an example being the use of 0.5 as a threshold for M-value standard deviation [10]. Scoring functions follow the filtering procedure and operate in such a way that when a set of genes or single genes are available, a score is computed and is responsible for discerning the extent to which the gene is likely to be expressed differentially [1].

Imperative to note is that the procedure above occurs at the expense of introducing more false positives. From the documentation and functions above, it is worth indicating that the set *S* (such as size) determines the distribution of the score. Thus, there is a need to normalize the distribution before using it for purposes of comparison among sets [2]. In the normalization of the scores' distribution, a nonparametric normalization technique has been proposed in the majority of previous scholarly investigations. In this normalization method, the phenotypes' permutations are used in estimating the scores' null distribution. In turn, scores are adjusted in such a way that their quantiles are used to replace them [3]. Imperative to document is that the advantage of this technique lies in its nonparametric nature [4]. However, the central drawback is that the null distribution's estimates are only reliable if the scoring process focuses on enough permutations [4]; an aspect that is unlikely to hold if fewer phenotypes are used. Other scholarly assertions hold that this technique proves to be computationally intensive. To counter these drawbacks, maxTscaled has been proposed as an ideal variant [5]. The latter algorithm's assumption is that for all objects — such as graph scores arising from various root genes, the null distribution is the same.

III. Results

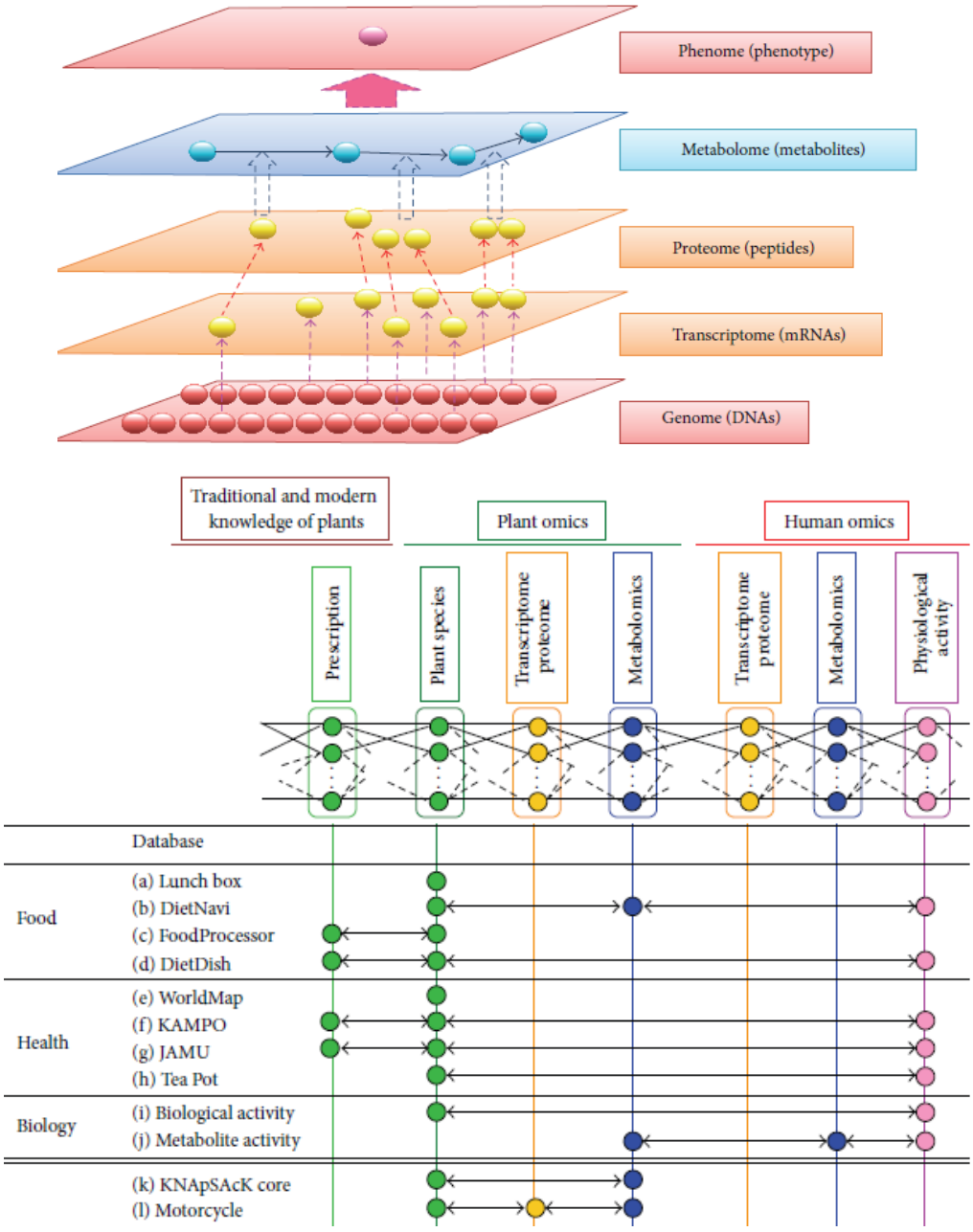

Figure 1 - An illustration of Computer Network Interaction in Complex Biological Data

Based on the results linked to the algorithm that relies on interaction networks to obtain results that the single-gene analysis technique is unlikely to obtain, promising results have been documented in the field of complex biological data analysis [7]. Some of the datasets on which the proposed algorithm (averaging gene expression scores) has been applied include serum data, regulatory T-cell data, and lymphocyte data [8]. Findings reported in most of the current literature hold that when network analysis is applied, it highlights groups of interacting and high-scoring genes [9].

IV. Conclusion

In summary, it is evident that network analysis has evolved in the field of complex biological data analysis for purposes of addressing the problem of the use of prior biological knowledge and expression data for purposes of identifying differentially expressed groups of genes or pathways. Particularly, the use of the gene interaction network (and its associated analysis) is seen to play the role of searching for high-scoring sub-networks. Indeed, most of the results suggest that when several improvements are made to the network analysis process (in terms of algorithms and scoring functions (such as the incorporation of averaging gene expression)), network analysis exhibits higher accuracy and speed, as well as easier biological interpretations. Some of the human microarray datasets on which the network analysis idea has been applied include serum data, regulatory T-cell data, and lymphocyte data; with specific target sets involving those related to conditions or parameters such as the immune system and cancer. Overall, this study has established that network analysis plays a crucial role in the field of complex biological data analysis whereby improved algorithms and scoring functions yield improvements in the speed and accuracy of biological process and disease-related pathway (or gene) identification.

References

[1] Leiserson, M.D., Eldridge, J.V., Ramachandran, S., Raphael, B.J. (2013). Network analysis of GWAS data. *Curr Opin Genet Dev.* 23(6), 602-610.

[2] Bolouri, H. (2014). Modeling genomic regulatory networks with big data. *Trends in Genetics*, 30(5), 182-191.

[3] Halldórsson, B.V., & Sharan, R. (2013). Network-based interpretation of genomic variation data. *Journal of molecular biology*, 425(21), 3964-3969.

[4] Vandin, F., Upfal, E., & Raphael, B.J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18(3), 507-522.

[5] Jia, P., Zheng, S., Long, J., Zheng, W., & Zhao, Z. (2010). dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics*, 27(1), 95-102.

[6] Liu, J.Z., Mcrae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., & Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, 87(1), 139-145.

[7] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., & Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3), 559-575.

[8] Khurana, E., Fu, Y., Chen, J., & Gerstein, M. (2013). Interpretation of genomic variants using a unified biological network approach. *PLoS computational biology*, 9(3), e1002886.

[9] Korcsmáros T, Farkas IJ, Szalay MS, et al. (2010). Uniformly curated signaling pathways reveal tissue-specific cross-talks and support drug target discovery. *Complex biological data analysis.* 2010; 26(16), 2042-2050.

[10] Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., & Maitland, A. (2010). The Gene MANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(suppl_2), W214-W220.