Network modeling of Indian Traditional Medicine

S R Nikitha * **Mukhesh Pugalendhi Sudha** **

* *Department of Biotechnology, Indian Institute of Technology, Madras.(e-mail: be18b011@smail.iitm.ac.in)*
** *Electrical engineering department, Indian Institute of Technology, Madras.(e-mail: ee18b114@smail.iitm.ac.in)*

Abstract: Traditional Indian Medicine has been in effective practise for thousands of years for treating human diseases. The uniqueness about TIM comparing with the modern medicine is the herb formulations which are combinations of therapeutic herbs. The rationale of these herb combinations is yet unclear.In this study, we hypothesise that the combinations are based on the minimal set of plants that can interact with all the known genetic disease proteins. Using, network based connectivity and optimisation, we try to prove if the hypothesis is indeed correct. This will shed light on the rationale behind the herb formulations in TIM.

Keywords: Traditional Indian Medicine(TIM), Network modeling, p-median optimisation

1. INTRODUCTION

Diseases are a major problem to human health. Every single human out there suffers a disease or two in his/her lifetime. While it gently reminds us of our mortality, it inflicts serious pain and suffering. From the day born, we take a multitude of drugs (synthetic, biologics, etc) to fight various diseases. The design of treatments for diseases is a complex ideation and history tells us there has been a variety of approaches in treating diseases. One such methodology widely accepted by the modern medicine is the "one drug one target" principle of drug design. While such an approach helps identify a single target behind disease causation and identifying a drug to combat it, various complex diseases have multiple genes involved in the cause. Their malfunction is either triggered together or as a chain effect. Hence better drug therapies require combinatorial drug identifications.

Understanding combinatorial drug action requires the exploration of combinatorial drug space and finding potential drug combinations through high throughput phenotypic assays. This is a very tedious work and intractable in time. Hence there is a need to build computational models that can help understand such therapies and identify potential therapies. Interestingly, Indian traditional medicine treats diseases with combinatorial plant therapies.

2. TRADITIONAL INDIAN MEDICINE

Indian traditional medicine(TIM) is one of the world's first established medicinal system dating back to 250 BC. There are six different Indian traditional systems: Ayurveda, Siddha and Unani, Yoga, Naturopathy and Homeopathy. Ayurveda is the most popular among them. About 70 percent of the Indian rural rely on Ayurveda for immediate medical care. Treatments in Ayurveda include dietary interventions and herbs from natural plants.

Hence natural medicinal plants are a major source of drugs. These medicinal plants are indigeneous to the subcontinent. Treatments usually are extracts from these plants mixed together. Combinations of various herbs indicate that multiple phytochemicals (the small molecules from plants) act synergistically to to perform the healing action. It might be of great significance to understand the scientific rationale behind these plant combinations as it might shed light on combinatorial drug action. Such an understanding can help us device new methods of synthetic drug combinations for treating complex diseases.

Fig. 1. Natural herbs in India

One issue with understanding herb combinations is that each plant has multiple phytochemicals. Using herb combinations thus means that a plethora of phytochemicals act synergistically to provide enhanced therapeutic effects. Phytochemicals within a plant can also act synergistically adding to the complexity of understanding them.

2.1 Network Pharmocology

With the advent of high-throughput biological data, computational biology techniques have been developed to anal-

yse and explore the data. To understand drug action and interaction with multiple targets, network pharmacology was developed. It uses computational power to systematically catalogue the molecular interactions of a drug molecule in a living cell. This is an important tool in understanding complex interactions between botanical formulae and living body. Multi-phytochemical multi-target interactions can be elucidated with the help of NP. Based on molecular identities synthetic drugs similar to these phytochemicals can be put to action based on these understandings. NP also attempts to discover new drug leads and targets and to repurpose existing drug molecules for different therapeutic conditions by allowing an unbiased investigation of potential target spaces. Hence various complex diseases and their interactome can be analysed with the help of network pharmacological approaches.

3. METHODS

The relation between diseases and their potential traditional herb formulations are established through the proteins whose malfunction are causative in the disease and the plant phytochemicals that have a binding action upon these proteins. Disease-plant-phytochemical information was obtained from Indian Medicinal Plants, Phytochemistry and Therapeutics (IMPPAT) database. The database houses information on >1700 medicinal plants, >9500 phytochemicals, >1100 therapeutics and >900 traditional formulations. An example network of a plant (Acacia concinna) and it's various phytochemicals are shown in figure 2.

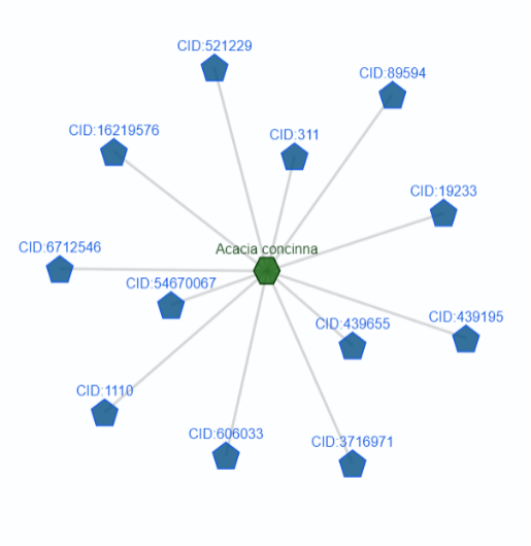

Fig. 2. Acacia concinna and it's phytochemicals

Information on binding activity of plant phytochemicals to human proteins can be obtained from databases on chemical-protein interactions. Here, we have obtained the binding score of the phytochemicals to proteins that constitute human proteome from the STITCH database. Human protein-protein interaction network was obtained from the STRING database. This way, three different networks were connected to relate plants and their target human proteins.

To identify the principle behind the traditional formulations, a minimal set of plants that constitute all the phytochemicals that affect the target proteins were hypothesised to be the formulation. Hence the aim was to develop a method/model that can identify these minimal set of plants given a set of target proteins causative of a particular disease.
Such a model, if proven, can in theory be used to identify synthetic alternatives for ayurvedic drugs and vice versa.

3.1 Network connectivity based

A network connectivity based method was developed to identify the minimal set of plants of therapeutic effect for a given disease. The method constitutes two steps: i) Predicting a preliminary set of all plants that are connected to the given list of disease target proteins. ii) Eliminating the redundant plants that have common phytochemicals with a less effectiveness. Effectiveness of a plant is defined as the number of proteins a plant can target from the given list. A plant with less effectiveness and constituting redundant proteins which are already targeted by plants of higher effectiveness were eliminated from the minimal list.

The mappings between plants, phytochemicals and proteins were represented by matrices. Two different matrices connected plants to phytochemicals (M1) and phytochemicals to proteins (M2). The matrices were binary matrices where the presence of a phytochemical in a plant and binding ability of a phytochemical with a protein is given a value of 1, else 0. The two matrices can also be stacked together to directly connect plants to proteins (Figure 3).

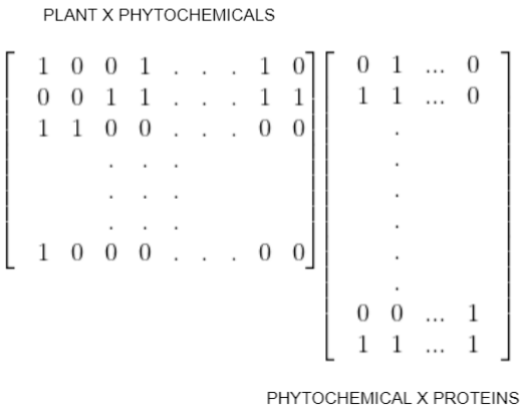

Fig. 3. Matrix representations

For a given set of proteins, a binary vector was generated. Multiplication between the input vector and the matrix M2 (row wise) would yield a binary matrix with only the necessary proteins along rows. Following this, the effectiveness of each row (phytochemical) can be calculated. The phytochemicals can then be sorted based on the effectiveness in descending order. Rows with the least effectiveness were then eliminated one at a time and the span (proteins covered) of the remaining rows were determined. This sequential elimination of phytochemicals were performed till all redundant phytochemical rows were removed. The final set of phytochemicals were then the minimal set required

to bind to all the target proteins.

The unique set of phytochemicals again form a binary vector for the plant-phytochemical (M1) matrix. The above process was again repeated to identify the minimal list of unique plants that can encompass the phytochemicals. This way we obtain the minimal set of plants required in the medicinal formulation for a specific disease.

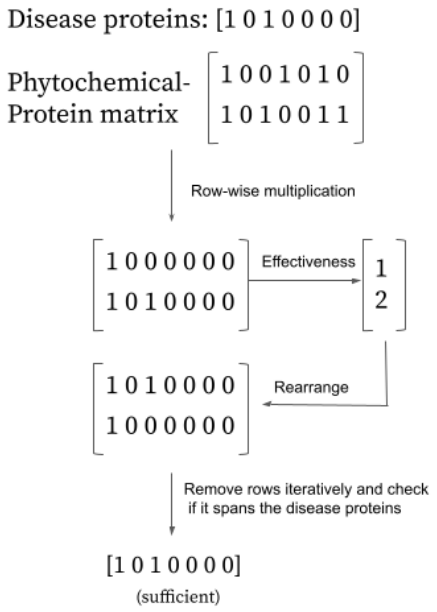

Fig. 4. Example of the method

3.2 p-median Optimisation problem

The identification of the minimal plant set for the given set of proteins can also be posed as an optimisation problem. This optimisation is similar to the warehouse location optimisation problem (p-median optimisation). The decision variables are the binary vector of plants represented by y and the binary matrix of plant-protein interactions x. A binary matrix (B) representing the plant-protein interactions was obtained by multiplying the matrices in Figure 3. The problem can be mathematically stated as follows,

$$\begin{aligned}
\underset{y}{\text{minimize}} \quad & \sum_{j=1}^{n} y_j \\
\text{subject to} \quad & \sum_{j=1}^{n} x_{ij} \geq 1,\ i = 1, \ldots, m. \\
& \sum_{i=1}^{m} x_{ij} \leq C_j,\ j = 1, \ldots, n. \\
& x_{ij} = y_j * B_{ij} \\
& x_{ij} \in 0, 1 \\
& y_j \in 0, 1
\end{aligned}$$

where, n is the number of plants present in the matrix and m is the number of target proteins for a given disease.

$$x_{ij} = \begin{cases} 1, & \text{if } \textit{protein } i \textit{ can bind to plant } j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$y_j = \begin{cases} 1, & \text{if } \textit{plant } j \textit{ was selected} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$$C_j = \textit{capacity of a plant } (\textit{max proteins it can bind to}) \quad (3)$$

C vector was calculated by adding the column of B matrix to obtain the maximum proteins with which each plant can bind. This way, a minimal set of plants can be obtained in vector y, which was finally cross checked with the original set of therapeutic plants for a given disease.

4. RESULTS

A set of five different diseases which are known to be genetic in nature were considered for the analysis. These diseases were,

(1) Alzheimer's disease
(2) Fibrosarcoma
(3) Obesity
(4) Liver cirrhosis
(5) Endometriosis

The proteins which are genetically associated with the chosen diseases were obtained from the database, DisGeNET. The database integrates information on disease-gene associations from various repositories by means of disease and gene vocabulary mapping. Based on the literature survey, the database allocates a score for the association between a specific disease and a gene. Since each disease had 1000s of genes associated with them but most had a very low score, we applied a threshold score of above 0.5 out of 1 to be the condition to filter the disease genes.

The protein list vectors generated by this were used as inputs for both the methods explained above. The plant list generated by the methods were compared with the original plant list associated with the disease in IMPPAT database. On applying method I, it was observed that a dozen plants, which had extensive connections with the proteins owing to their large repository of phytochemicals, kept appearing for all the diseases. This led to poor correlation between the original plant vector and the predicted plant vector. Removal of these plants resulted in a better correlation of the predicted plants and the original plant lists. The results are recorded in Table 1.

The same set of diseases were analysed by applying method II. The optimisation was performed with CPLEX solver. The predicted plants were even lesser in count compared to the plants predicted by method I. The results on false positives and true negatives are reported in Table 2.

A detailed record of the predicted plants can be found in the link: RESULTS

5. DISCUSSION

We see that in both the methods the matching plant proportions are low. There is a high incidence of false positives and true negatives. Comparison with the predicted plant list from method I show that only 40% of the predicted plants were present in the original plant list.

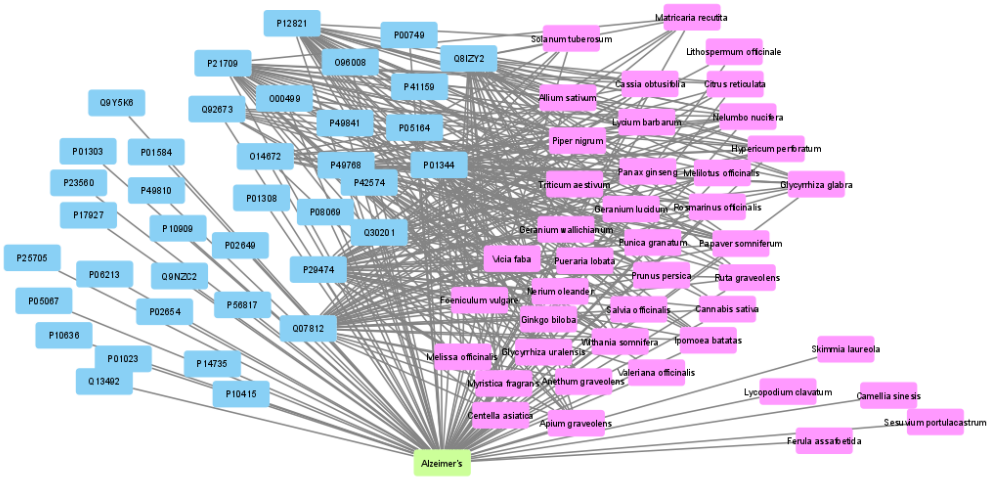

Fig. 5. Network mapping of plants and proteins of Alzheimer's disease constructed using Cytoscape.

Table 1. Results of method I with highly connected plants removed

Disease	No. of plants in formulations	No. of predicted plants	No. of plants correctly predicted	False positives	True negatives
Alzheimer's disease	42	14	2	12	40
Fibrosarcoma	9	1544	9	1535	0
Obesity	50	10	1	9	49
Liver cirrhosis	2	1582	2	1580	0
Endometriosis	1	1	0	1	1

Table 2. Results of method II (p-median optimisation)

Disease	No. of plants in formulations	No. of predicted plants	No. of plants correctly predicted	False positives	True negatives
Alzheimer's disease	42	5	0	5	42
Fibrosarcoma	9	0	0	0	9
Obesity	50	5	0	5	50
Liver cirrhosis	2	0	0	0	2
Endometriosis	1	1	0	1	1

Method II results were also very less correlative with the original plant lists. The minimal set of plants predicted were very low and did not match with the original list. An analysis on the actual plants corresponding to the formulations of the five diseases was performed to understand the reason behind the failure of the hypothesis. It was seen that the plants in the formulation do not encompass all the target proteins of that particular disease. For example, the plant combinations for Alzheimer's disease is mapped with it's target proteins. The figure 5 illustrates the network mapping.

It can be seen from this figure that the plants do not encompass all the target proteins. Hence the minimal list that tried to bind all the target proteins wasn't a necessary condition. Also, not all the plants used in the formulation need to bind to the target proteins. Some plants in the network do not bind to any protein at all and are still used in the formulations.

6. CONCLUSION

The low level of true positive values indicate that there is some error either in the method built or in the hypothesis itself. From figure 5 it is evident that not all plants should cover the entire set of disease proteins. Hence the hypothesis that TIM is indeed the minimal list might be wrong. Few other potential reasons for the low true positive plants could be,

(1) Insufficiency in the dataset. It is possible that interactions between phytochemicals-proteins or the presence of phytochemicals have not been discovered yet. This leads to loss of connections in the network. Since the data is reliant on the source database, they might have undergone other computational processes before constructing the interaction network. Some databases which are literature survey based (IMPPAT) might have missed collecting information as well.
(2) The idea of connecting plants and diseases just based on the known disease proteins might be wrong. We

are aware of disease modules which are a subset of proteins in the human protein-protein interaction network that are known to be causative of the disease. These modules contain additional proteins apart from just the direct disease proteins.

(3) The idea of obtaining the minimal set of plants might be the way in which actual TIM combinations were derived.

(4) The DisGeNET threshold score considered for shortlisting the disease proteins should be reconsidered.

But since these set of predicted plants have a binding affinity for the disease proteins they might be novel therapeutic combinations. Their effects could be experimentally verified! This way the obtained minimal sets have set new avenues for further research.

ACKNOWLEDGEMENTS

We would like to sincerely thank prof. Karthik Raman for guiding us to look at different ways to solve the problem statement with different paper and database suggestions. We would also like to thank our families who constantly battle with us about the lost knowledge of Indian traditional medicine and it's immense potential in curing diseases compared to the conventional pharmaceutical interventions. They nudged us to look into the Indian traditional medicine and made us think about computationally modeling it.

REFERENCES

[1] IMPPAT: A curated database of Indian Medicinal Plants, Phytochemistry And Therapeutics, Karthikeyan Mohanraj, Bagavathy Shanmugam Karthikeyan, R.P. Vivek-Ananth, R.P. Bharath Chand, S.R. Aparna, P. Mangalapandi and Areejit Samal*, Scientific Reports 8:4329 (2018).

[2] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, von Mering C. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019 Jan; 47:D607-613.

[3] Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res. 2008;36(Database issue):D684-D688. doi:10.1093/nar/gkm795

[4] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, Laura I Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucl. Acids Res. (2019) doi:10.1093/nar/gkz1021

[5] Zhao H, Shan Y, Ma Z, Yu M, Gong B. A network pharmacology approach to explore active compounds and pharmacological mechanisms of epimedium for treatment of premature ovarian insufficiency. Drug Des Devel Ther. 2019;13:2997-3007 https://doi.org/10.2147/DDDT.S207823

[6] G. B. Dantzig, D. R. Fulkerson, and S. Johnson. Solution of a large scale traveling salesman problem. Operations Research, pages 393–410, 1954.

[7] C. E. Miller, A. W. Tucker, and R. A. Zemlin. Integer programming formulation of traveling salesman problems. Journal of ACM, :326–329, 1969.

[8] Yinyin Wang, Hongbin Yang, Linxiao Chen, Mohieddin Jafari, Jing Tang, Network-based modeling of herb combinations in traditional Chinese medicine, Briefings in Bioinformatics, 2021;, bbab106, https://doi.org/10.1093/bib/bbab106