Comparison of Sequence Similarity Networks.

Prerana B Parthasarathy and Shruti S Patil, *Dept. of Computer Science, Washington State University*

***Abstract*—Sequence Similarity Networks are useful for identifying complex relations between proteins and find homologous proteins. SSNs can be constructed using threshold-based methods that use pairwise similarity calculation and hence is an expensive step. This is overcome by the DiWANN method of SSN construction. The goal of the project is to compare these methods by creating the networks of the proteins present in Chloroplast gene of rice and wheat crops. We draw conclusions about the crops by analyzing properties like density and centrality.**

***Index Terms*—Directed Weighted All Nearest Neighbors, Sequence Similarity Networks, Similarity Measures, Chloroplast gene.**

I. INTRODUCTION

In this era of biological research there has been a data explosion which has compelled the development of more data analysis approaches and tools. Network analysis is one such approach that has found application in bioinformatics. Researchers work on DNA data and the underlying proteins to come up with conclusions about the organism. The proteins in the gene play a crucial role in cellular function. And networks of proteins have found an application in finding homologous protein or in predicting the function of the proteins. One type of protein network is the sequence similarity network which are constructed on the basis of protein similarity. The nodes of these networks represent individual proteins and the edges between then is the distance between the sequences or the similarity of one sequence to another. The advantage of using SSNs are that it provides a graphical visualization of the protein interrelationships unlike the traditional methods like creating phylogenetic trees for analysis. The edges and nodes are overlaid with orthogonal information that makes networks a powerful tool for generating hypothesis.

The project examines three different models of SSNs they are exact threshold-based, inexact threshold based and directed weighted all nearest neighbors' method. In the exact similarity threshold-based method we use BLAST similarity scores as the measure. While the other two methods use edit distance as the similarity measure. We see how the methods differ in the construction of the network and how this affects the structure of the sequence similarity network. These models are applied to two types of data in out project. One is the Anaplasma Marginale msp1α gene and the other is the chloroplast protein of different crops.
We have chosen 2 different crops that is the rice and wheat. Wheat is most widely grown of all crops and the cereal with the highest monetary yield. It is a mid-tall annual or winter annual grass with flat leaf blades and a terminal floral spike consisting of perfect flowers. *T. aestivum* is a cereal for temperate climates. There are four common classes of wheat: hard red winter, hard red spring, soft red winter, and white. Oryza sativa (rice) is one of the most important crops in the world and it provides the main resource of energy for more than half of the world population. Rice production represents 30% of the world cereal production today. There are about 120,000 varieties known to exist. Two of the common types are Indica and Japonica. *Indica* varieties are generally adapted to tropical lowland cultivation, whereas most *japonica* varieties are adapted to more temperate climates.

The results of comparing the models show that DiWANN works better and faster than the other two models. We will then compare the DiWANN networks of the chloroplast protein sequence for rice and wheat crop to find how different this protein is in the different species of plants and this can be used to predict how it affects the growth and yield.

II. MEASURES

The backbone of a sequence similarity network as the name suggests is the similarity between the sequences. There are many similarity metrics that can be used to compute the similarity between numeric data like the cosine similarity and Euclidian distance. But genome data are sequence of strings. They represent either a sequence of amino acids or of DNA nucleotides. The measure that is most commonly used for such data is Edit Distance. It is the minimum value at which one sequence can be transformed into another sequence. It handles insertion, deletion and substitution. There are some other alignment algorithms that are used to achieve a similarity score. Alignment is an important issue when it comes to protein sequences as any slight change in the sequence could mean it is a different protein. Needleman-Wunsch and Smith Waterman are some frequently used alignment algorithms. Because of the ease of use and implementation we have selected Edit distance as the similarity measure for this project.

One challenge faced by using these methods be it an alignment algorithm or the edit distance is the fact that it computes pairwise distances which is an expensive step if there are more number of sequences, long sequences and in the worst case scenario both. A faster approach is using approximate distances like BLAST which is used in the inexact threshold-based method of SSN generation.

III. Implementation

In this section we will discuss the different datasets and models and their implementation.

A. Dataset

The first dataset is the short sequence repeats (SSRs) of msp1α gene of Anaplasma Marginale. SSRs are sequences in which a pattern occurs two or more times. These proteins sequences have roughly 28 amino acids each and we have 284 such sequences. These sequences have been extracted from the different strains of A. marginale found in different countries. The diversity of this strain places a challenge for taxonomic classification.

The idea of our project is to compare the chloroplast protein of crops like rice and wheat. Chloroplasts are a type of plastid in plants and algae that are required for photosynthetic energy production. It is responsible for photosynthesis in plants and for reactions between enzymes and other biosynthetic reactions. The chloroplast genome is believed to be clonal and has its own replication and DNA repair systems. Chloroplast genomes in plant leaf cells are not absolutely homogenous but are similar as the mutation rate of the chloroplast sequences is very low. These sequences are shorter and as the protein sequence for chloroplast is separately available, it is convenient to work with the Chloroplast protein sequence.

Our project uses three other datasets. One is the Chloroplast protein sequence of the Chinese Spring strain of Triticum aestivum, also known as wheat, which consists of 77 protein sequences This data was obtained from UniProt (https://www.uniprot.org/proteomes/UP000019116) and the genome accession number for this data is AB042240. The data collected for the rice chloroplast protein has 83 protein sequences for both the sub-species of rice. The Chloroplast protein for Rice sub species Japonica was obtained from UniProt (https://www.uniprot.org/proteomes/UP000059680) and the genome accession number for this data is AY522330 and the Chloroplast protein for Rice sub species Indica whose accession number is AY522329 was obtained from UniProt (https://www.uniprot.org/proteomes/UP000007015#). For each of the three datasets we have implemented all the three models explained above and compared the structures based on density, degree distribution and centrality. We have then compared the DiWANN graphs of each rice and wheat chloroplast protein to make a few conclusions about it.

Two applications have been implemented using the Chloroplast protein sequences: genotype analysis and inter-species same protein analysis. For genotype analysis, we have constructed SSNs for the Chloroplast protein sequences of wheat and rice and for inter-species same protein analysis, we have compared the Chloroplast sequences of two sub-species of rice: Indica and Japonica.

B. Models

The three models of sequence similarity networks chosen for the project are two threshold-based methods and the DiWANN method.

1) Inexact Threshold method

The implementation of approximate threshold or the inexact threshold-based method makes use of BLAST. BLAST is a basic local alignment search tool. It finds an alignment between two sequences and gives it a score. It provides a bit-score which is the factor used as the measure in the project. The clear disadvantage in threshold-based methods is the selection of the threshold which is mostly based on trial and error and so changes based on the needs of the user. This is not a very definitive way for creating a model. As the bit scores represent the similarity, the greater the value the more similar the sequences are. Once the threshold is selected then for all the sequences where the bit-score is greater than the threshold an edge is drawn.

2) Exact Threshold method

The implementation of the exact threshold method makes use of the bounded edit distance as the similarity measure. The bounded edit distance take 3 parameters, two sequences and a threshold. When the distance between the two sequences is less than the threshold then the distance is returned otherwise it returns maxsize. Again, we face the same challenge of selecting a good threshold which will capture the structure of the network well. The problem is that if the threshold is too small then there could be loss of some relationships between proteins and if the threshold is large then the network may end up being too dense to analyze. A favorable threshold is selected on trial and error basis.

3) DiWANN

This network model overcomes the challenges of the other two models. A major issue faced by the model discussed above is how it is expensive to compute pairwise similarity when there are a large number of sequences. This makes the runtime for such computations as $O(n^2m^2)$ where n is the length of the sequences and m is the number of sequences present. The second challenge is the selection of the threshold. DiWANN algorithm for generating sequence similarity networks mitigates these shortcomings. It reduces the number of sequence similarity computations by using the property of triangle inequality.

The algorithm works as follows – consider we have m sequences the algorithm creates an m X m matrix which stores the similarity. The matrix is filled in row by row but only the first row is computed using edit distance. All the other rows are filled by pruning the distance calculation. The triangle inequality principle works by creating bounds for the distance in question. Say we have 3 sequences A,B and C the row for A is filled completely(the first row). The distance between B and C is bounded below by the difference of distances between A, B and A,C. It is bounded above by the sum of the distances between A,B and A,C. These lower and upper bounds are stored in vectors called the minED and maxED and is computed for each row. If the lower bound for the cell is lesser than the current minimum that has been computed in the row, then this distance is calculated using the bounded threshold method explained before. If the lower bound is greater that the row minimum, then that cell is filled with MAXINT. This means that this cell does not play a role in any of the edges and is hence skipped. Skipping computation of these cells is what makes this process of creating the similarity matrix faster and more advantageous than the pairwise computation. The result of this is presented in figure 1

The next part of the algorithm involves the construction of the graph based on the similarity matrix. This also works row-wise where an edge is drawn to every node that has the row minimum value in the similarity matrix. This works better than k nearest neighbors where there is a tie breaker. In DiWANN if there is a tie then an edge is drawn to all the nodes this keeping the integrity of the structure. Figure 2 represents the graph creation for the example data considered.

	A	B	C	D.	E	F	H	I	J	AustralianType
	0	1	2	3	4	5	6	7	8	9
0	92233720368...	4	4	2	1	3	4	2	5	6
1	4	92233720368...	1	4	92233720368...	3	4	3	1	5
2	4	1	92233720368...	4	3	2	3	2	2	4
3	2	4	4	92233720368...	1	4	5	4	92233720368...	6
4	1	92233720368...	3	1	92233720368...	3	4	3	92233720368...	92233720368...
5	3	3	2	4	3	92233720368...	1	1	92233720368...	6
6	4	4	3	5	4	1	92233720368...	2	5	5
7	2	3	2	4	3	1	2	92233720368...	92233720368...	6
8	5	1	2	92233720368...	92233720368...	92233720368...	5	92233720368...	92233720368...	6
9	6	5	4	6	92233720368...	6	5	6	6	92233720368...

Figure 1: Similarity matrix construction

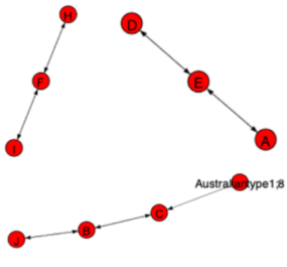


Figure 2: Graph construction

C. Application

The first application of the given network models is genotype analysis. Genotype analysis refers to mapping variations in genome across individual organisms. We study the Chloroplast proteins of Triticum aestivum (Wheat) and Oryza sativa subsp. Japonica (Rice) using the above network models. The structure of the SSN of the protein sequences can be studied to find similarity and differences in the Chloroplast protein of the two crops. Sequence similarity can identify "homologous" proteins or genes by detecting excess similarity, that is, statistically significant similarity that reflects common ancestry. DIWANN makes it easier to study the protein sequences, thus comparing the two DIWANN plots, inferences about the similarity between the two proteins can be deduced.

Inter-species same protein analysis involves comparing the sequences for the same protein in different species. Chloroplast indirectly affects the yield of a crop. The resources produced by photosynthesis process are used up for the development of the seeds. Thus, at early stages, photosynthesis rates and efficiency are very important. The protein sequences can alter depending on the outside environment changes, for instance, the protein can alter during droughts, trying to sustain it. Thus, these features can be studied in the future to decide which crops perform better and which would be the most suitable crop to be grown during that period. This can be highly beneficial to farmers. Here, the same protein, that is, the Chloroplast protein of two species Wheat and Rice are compared using their SSNs. The structural similarities and differences of the two are studied. Homology is inferred when two sequences or structures share more similarity than would be expected by chance. Homologous proteins in both the species are found using different centrality measures.

For intra-species same protein analysis, we have compared the Chloroplast protein of two sub-species of rice: Indica and Japonica. These two varieties differ clearly in morphological and agronomic traits, in physiological and biochemical characteristics and in their genomic structure. They even have differences in yield, quality and stress resistance. Studies showed that some unique proteins in the two sub-species displayed functional specificity and were involved only in functions such as electron carrier, structural molecule, biological regulation and pigmentation. Thus, studying the Chloroplast would help us identify some dissimilarities. The SSNs of the two protein sequences are produced using DIWANN and the networks are compared. The protein sequences are also compared based on their centrality and clusters.

IV. ANALYSIS

The implementation of the three models on A. Marginale dataset resulted in three networks. We looked at the degree distribution of the networks and observed the important property of DiWANN networks. All nodes in the DiWANN network have an outgoing edge although it may or may not have an incoming edge. Which means that there are no edges with degree zero. This makes the network connected.

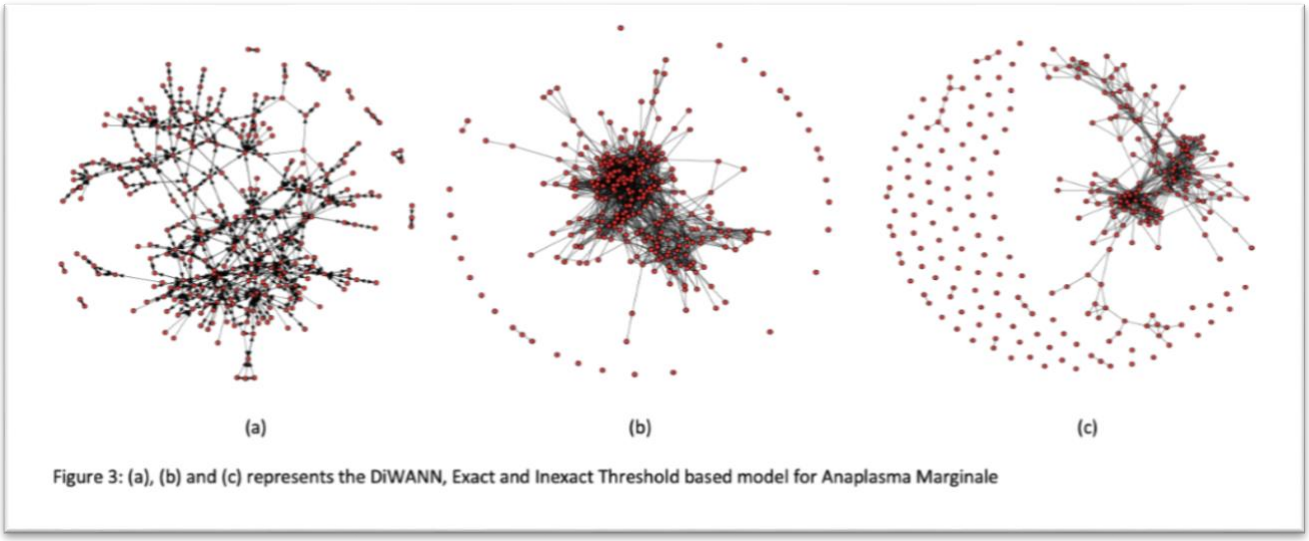


Figure 3: (a), (b) and (c) represents the DiWANN, Exact and Inexact Threshold based model for Anaplasma Marginale

Comparing this to the other two model of threshold-based networks. It is observed to have many singleton nodes in the network and therefore is not connected. This means that there is a definite loss of stricture in the threshold-based networks compared to the directed weighted all nearest neighbor network. An observation can also be made about the density of the networks. The threshold-based networks are denser than the DiWANN network even though they are not as connected as the DiWANN. All this is represented in Figure 3. Figure 4 represents the degree distribution of the three networks in figure 3. Notice that at degree zero there are no edges in the DiWANN network.

The same analysis has been carried out with the chloroplast protein sequence of wheat and rice. The plots for this are in figure 5 and 6. The DIWANN network was more efficient in terms of constructing the SSNs and also for studying the proteins for the two species. The SSNs of wheat and rice do have some noticeable differences in structure. The complexity of the rice DIWANN plot is more as the protein sequences are more connected together and are densely distributed, whereas the DIWANN plot for wheat is not so dense. The main reason being wheat has a longer genome sequence and thus, the proteins are more spread apart in terms of sequence similarity. Another observation is that the wheat SSN plot consists of 4 weakly connected components whereas rice SSN consists of only 2 weak components. This indicated the components being more well connected in rice.

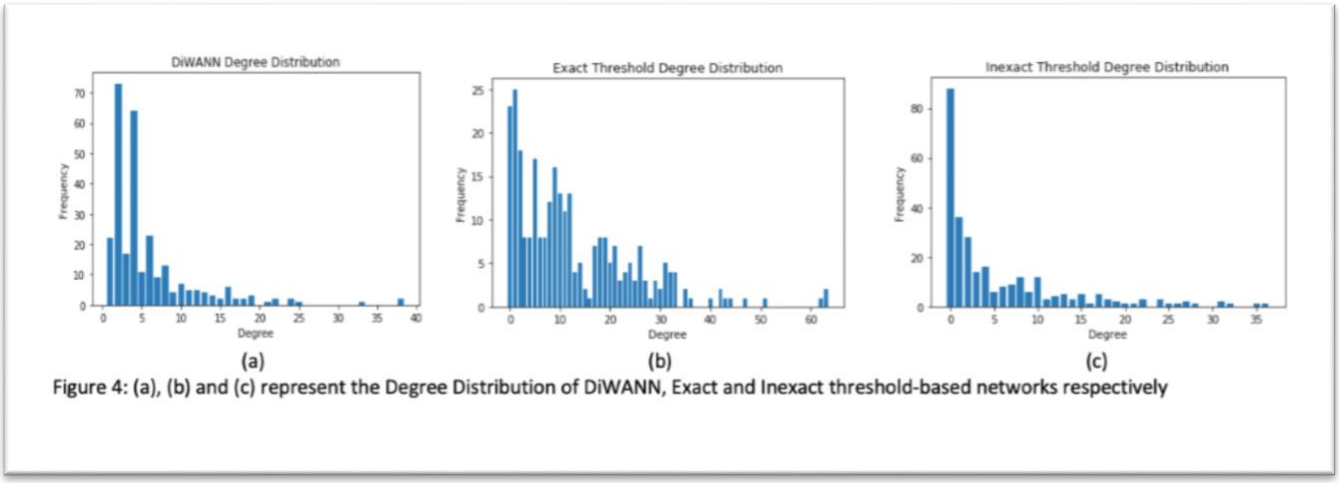


Figure 4: (a), (b) and (c) represent the Degree Distribution of DiWANN, Exact and Inexact threshold-based networks respectively

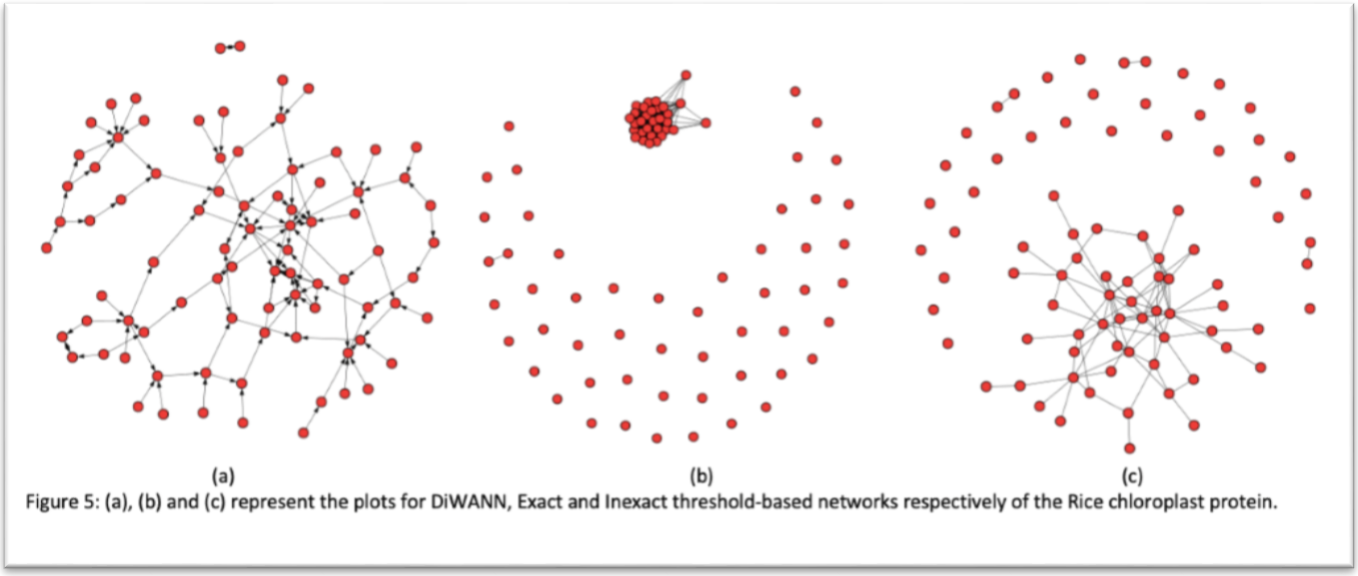


Figure 5: (a), (b) and (c) represent the plots for DiWANN, Exact and Inexact threshold-based networks respectively of the Rice chloroplast protein.

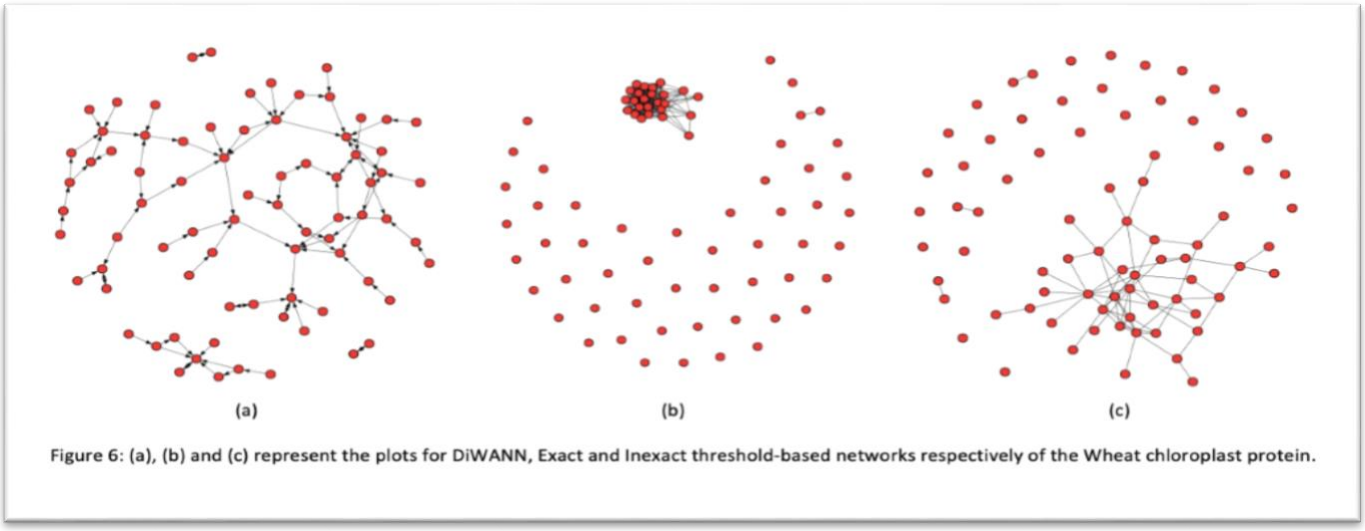


Figure 6: (a), (b) and (c) represent the plots for DiWANN, Exact and Inexact threshold-based networks respectively of the Wheat chloroplast protein.

We noticed a high level of homology among the Chloroplast proteins of wheat and rice. When similarity between the sequences was checked using BLAST, the sequences with the highest similarity score along with a large alignment length were considered. It was observed that the sequences with high similarities, had the same functionality in both the protein sequences. The figure 7 shows an instance where the highly similar protein sequences have similar functions. This could also indicate alternative splicing in which several proteins can be encoded by a single gene and this allows varied proteome and evolutionary flexibility. Additionally, these protein sequences are represented by nodes that have a similar structure in the SSN, that is they have the same degree and a similar orientation.

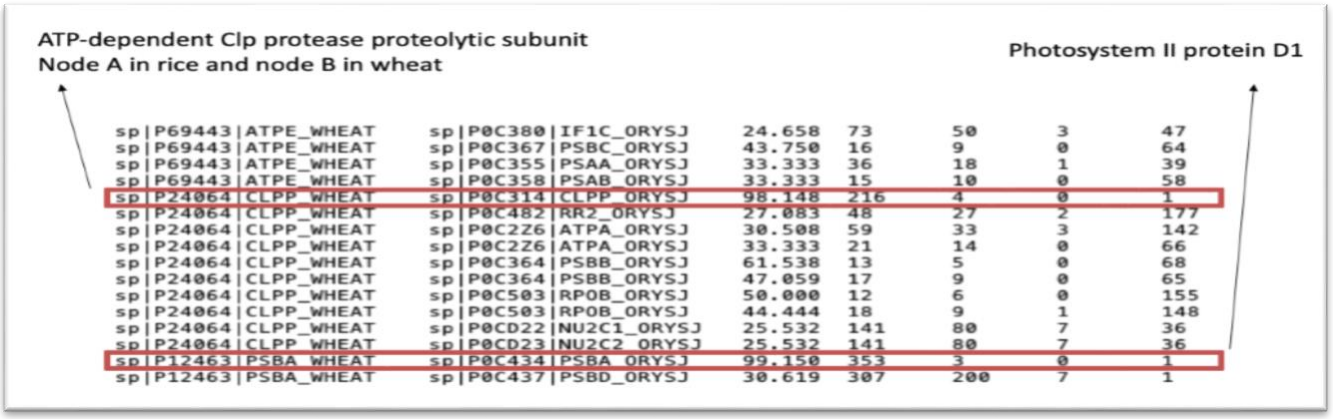
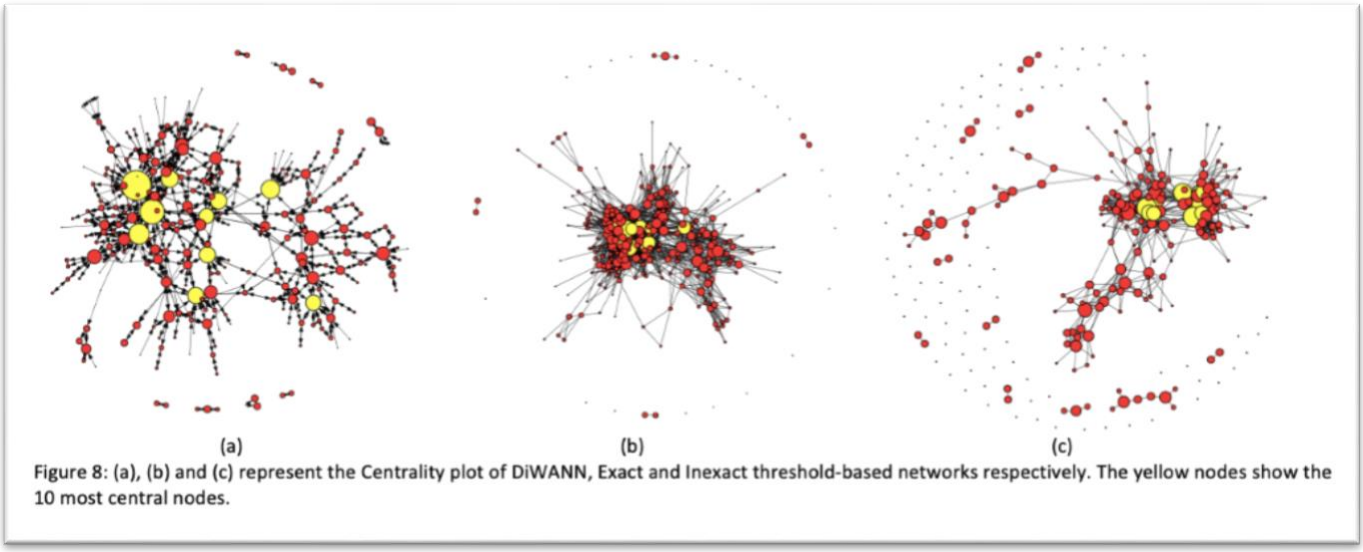


Figure 7: BLAST instances of similarity between Wheat and Rice protein sequence

We then implemented page rank centrality on the networks created by the three models. This analyses the most central nodes in the network. We have compared the corresponding central nodes in the network and how they relate to the sequence. In A. Marginal it is seen that the central nodes in DiWANN and exact threshold network correspond to each other while it is different for inexact. The figure 8 represents the centrality plots of the network variants. The nodes in yellow are the most central nodes and are sized according to the centrality score. The most central nodes for the three networks are shown in table 1. This signifies that some strains that are widely geographically distributed also have a higher centrality score. Hence centrality can be applied as a marker for the non-structural property like geographic distribution.



(a) (b) (c)

Figure 8: (a), (b) and (c) represent the Centrality plot of DiWANN, Exact and Inexact threshold-based networks respectively. The yellow nodes show the 10 most central nodes.

DiWANN	Exact Threshold	Inexact Threshold
M	UP33	β
61	F	2
γ:y	N	Z,ϕ
E	Q	15
F	M	11
4	B	F
B	13	P
27	4	5
3	27	6
13	3	7

Table 1: Central nodes

The same analysis was conducted for the rice and wheat variants of the chloroplast proteins. The plots for this are in Figure 9. Centrality of the two plots gave us the most important and fundamental proteins in both the cases. It can be observed that the most central nodes in both the species protein sequence are similar. The central nodes represent protein sequences which are responsible for fundamental functions of the Chloroplast such as Photosystem I, Photosystem II and Cytochrome functions. Other proteins seemed to have evolved after these proteins.

The DIWANN plots for the two sub-species of rice: indica and japonica are similar in structure. Similar to the DIWANN plot for Oryza sativa subsp. japonica (Rice), the DIWANN plot for indica sub-species of Rice, Oryza sativa subsp. Indica is complex. The proteins are tightly bound. The SSN for indica too has two weak components, showing high structural similarity.

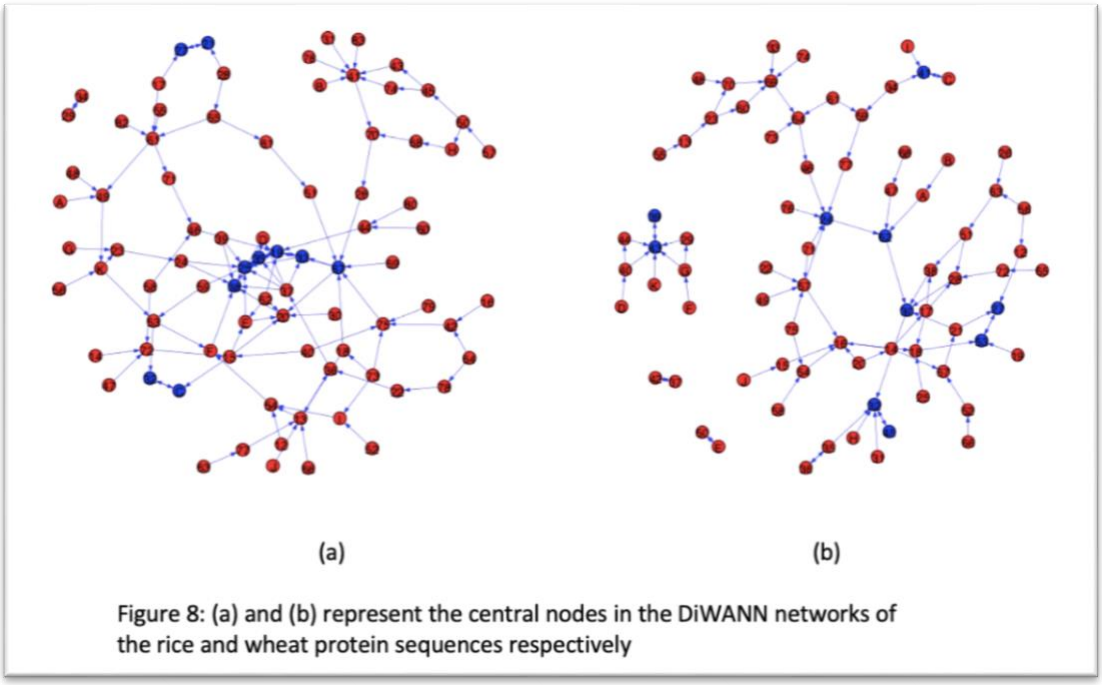


Figure 8: (a) and (b) represent the central nodes in the DiWANN networks of the rice and wheat protein sequences respectively

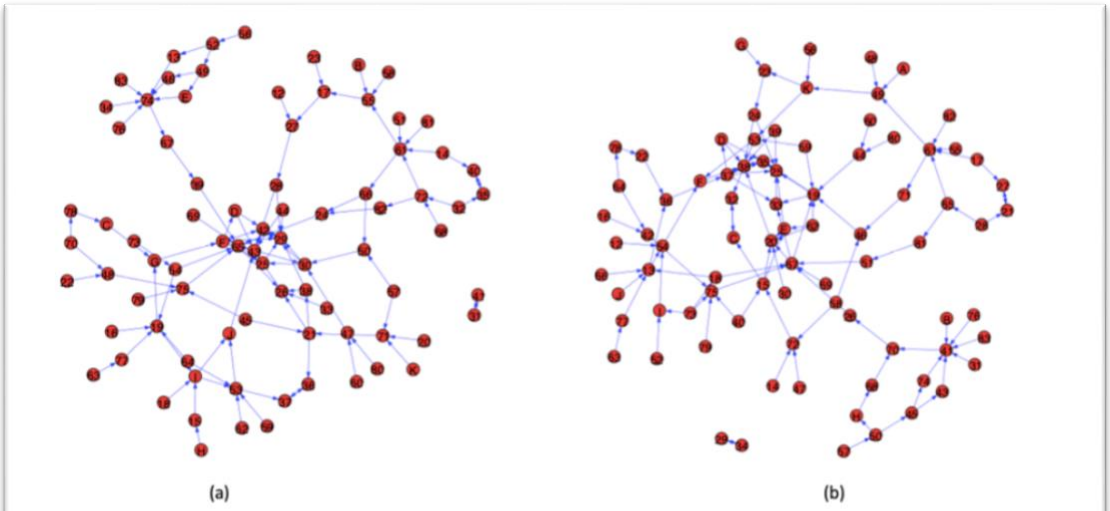
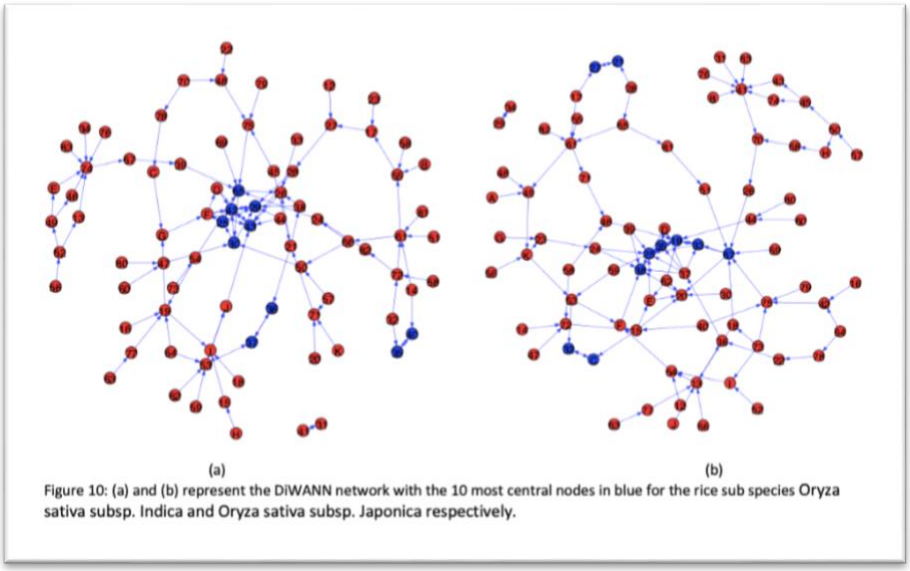


Figure 9: (a) and (b) represent the DiWANN networks of the rice sub species Oryza sativa subsp. Indica and Oryza sativa subsp. Japonica respectively.

In intra-species same protein analysis, centrality is measured for the DIWANN plots of both the sub-species of rice: indica and japonica as shown in figure 10. The proteins represented by the most central nodes are observed to be the same in both the networks. Table 2 shows a comparison of the central nodes in the DIWANN plots of the two sub-species. It can be observed that the most central nodes are the proteins representing fundamental and most vital protein in Chloroplast.



(a) (b)

Figure 10: (a) and (b) represent the DiWANN network with the 10 most central nodes in blue for the rice sub species Oryza sativa subsp. Indica and Oryza sativa subsp. Japonica respectively.

In Indica :	**In Japonica:**
25 Photosystem I reaction center subunit VIII	C Cytochrome b559 subunit beta
29 Photosystem II reaction center protein M	19 Photosystem II reaction center protein Z
30 Photosystem II reaction center protein Z	21 Photosystem II D2 protein
35 Photosystem II D2 protein	25 Photosystem II reaction center protein M
36 Cytochrome b559 subunit beta	27 Photosystem II protein D1
37 Photosystem II reaction center protein J	32 Photosystem II reaction center protein J
40 Photosystem II protein D1	33 Photosystem I reaction center subunit VIII
42 Cytochrome b6-f complex subunit 8	35 Cytochrome b6-f complex subunit 8
43 Cytochrome b6-f complex subunit 6	38 Cytochrome b6-f complex subunit 6
65 ATP synthase subunit c, chloroplastic	67 ATP synthase subunit c, chloroplastic

Table 2: Central nodes in DiWANN plots of the sub-species of Rice.

V. Conclusion

Three network models: DIWANN, exact threshold based and in-exact threshold based were implemented to construct sequence similarity networks. These models were applied on various datasets of two different types: Short sequence repeats (SSRs) and protein sequences. It was observed that DIWANN is the best model among the three to construct SSNs as it took the least amount of time and displayed the nodes which represented the sequences in a suitable manner to study. It proved to be more efficient as it avoided pairwise computation which was unnecessary in measuring the distance between sequences. Another drawback of the threshold-based methods was that there was ambiguity as a suitable threshold was selected based on trial and error. This led to loss in the structure of the network.

Genotype analysis, inter species same protein analysis and intra species same protein analysis was carried out. Major surface protein for Anaplasma Marginale and Chloroplast protein sequences were studied for two different species: Wheat and Rice. This study revealed the homology between the same protein in the two different species. Intra-species study of the chloroplast protein also showed a high level of similarity between the two sequences. The most central nodes represented proteins with the same functionality in both the applications. Chloroplast has a low mutation rate thus it was convenient to perform this study.

VI. Future work:

SSNs help us study the similarity of sequences based on the sequence of amino acids. It individually compared sequences for which we can check functionality and is thus more of a structural analysis. Using Protein-Protein Interaction (PPI) networks would give us a better idea for functionality analysis. This would also tell us which proteins interact with each other to produce something new of make variations in features and functions. Understanding PPIs is crucial for understanding cell physiology in normal and disease states. It is also essential in drug development, since drugs can affect PPIs. Drugs, enzymes and other external factors can have effects on these networks.

Rice proteomic studies can be done on the protein profiles of various organs, tissues and subcellular structures and the influences of a variety of environmental factors on gene expression. Apart from the Chloroplast protein, other important proteins can also be identified and studied. Inter-species different protein analysis can be conducted to study the interactions between various proteins and how different functionalities can be altered. A temporal network can be constructed to observe changes in the protein sequences over time. The evolution of a particular protein or the proteome can be studied using this network. One can study the leaf color and photosynthesis efficiency using such models, which would help in improving the crop.

VII. References:

[1] Catanese HN, Brayton KA, Gebremedhin AH. A nearest-neighbors network model for sequence data reveals new insight into genotype distribution of a pathogen. *BMC Bioinformatics*. 2018;19(1):475. Published 2018 Dec 12. doi:10.1186/s12859-018-2453-2

[2] Yang Y, Zhu K, Xia H, Chen L, Chen K. Comparative proteomic analysis of indica and japonica rice varieties. *Genet Mol Biol*. 2014;37(4):652–661. doi:10.1590/S1415-47572014005000015

[3] Tang J, Xia H, Cao M, et al. A comparison of rice chloroplast genomes. *Plant Physiol*. 2004;135(1):412–420. doi:10.1104/pp.103.031245

[4] A Johns, Mitrick & Mao, Long. (2007). Differentiation of the two rice subspecies indica and japonica: A Gene Ontology perspective. Functional & integrative genomics. 7. 135-51. 10.1007/s10142-006-0036-1.

[5] Alejandro Cabezas-Cruz, José de la Fuente,Anaplasma marginale major surface protein 1a: A marker of strain diversity with implications for control of bovine anaplasmosis,Ticks and Tick-borne Diseases,Volume 6, Issue 3, 2015,ISSN 1877-959X,https://doi.org/10.1016/j.ttbdis.2015.03.007.

GitHub Repository
https://github.com/PreranabParthasarathy/NetworkScienceProject