Chemical Characterization of Interacting Genes in Few Subnetworks of Alzheimer's Disease

Antara Sengupta
Department of Master of Computer Applications
MCKV Institute of Engineering
Liluah,India
antara.sngpta@gmail.com

Pabitra Pal Choudhury
Applied Statistical Unit
Indian Statistical Institute
Kolkata, India
pabitra@isical.ac.in

Hazel Nicolette Manners
Department of Information Technology
North-Eastern Hill University
Shillong, India
hazelmanners@nehu.ac.in

Swarup Roy
Department of Computer Applications
Sikkim University
Gangtok, India
sroy01@cus.ac.in

ABSTRACT

A number of genes have been identified as a key player in Alzheimer's disease (AD). Topological analysis of co-expression network reveals that key genes are mostly central or hub genes. The association between a hub gene and its neighbour genes can be derived easily using relative abundance of their expression levels. However, it is still unexplored fact that whether any hub and its neighbour genes within a subnetwork exhibits any kind of proximity with respect to their chemical properties of the DNA sequences or not, that code for a sequence of amino acids.

In this work, we try to make a quantitative investigation of the underlying biological facts in DNA sequential and primary protein level in mathematical paradigm. It may gives a holistic view of the interrelationships existing between hub genes and neighbour genes in few selective AD subnetworks. We define a mapping model from physicochemical properties of DNA sequence to chemical characterization of amino acid sequences. We use distribution of chemical groups present in a sequence after decoding into corresponding amino acids to investigate the fact that whether any hub genes are associated closely with its neighbour genes chemically in the subnetworks. Interestingly, our preliminary results confirm the fact the dependent genes that are coexpressed with its hub gene are also having proximity with respect to their amino acid chemical group distributions.

CCS Concepts

•**Applied computing** → **Computational genomics;**

Keywords

Alzheimer's Disease, Hub genes, Physicochemical, DNA sequence, Amino acid, Chemical properties

1. INTRODUCTION

Alzheimer's disease is a neuro-degenerative disease. AD is a type of dementia that causes problems with memory, thinking and behavior. Identifying any key genes and its underlying interaction network may put light on the disease mechanism. In turn it may help design effective therapeutic drug molecules for AD. It has been observed that such key genes or possible regulators are central genes [2], also called hub genes. Studies [1] reveals few genes namely, Amyloid Precursor Protein (APP), PreSenilin1 (PSEN1) and PreSenilin2 (PS2), Apolipoprotein E (ApoE) are some of the key genes in AD. Instead of single gene, generally a set of interacting genes forming a sub-network or module [17, 16] act towards disease abnormalities inside cell. The common way to identify such key genes and its subnetworks is to infer network computationally from microarray gene expression data [15, 3]. Drawing a relationship among pairs of genes in a co-expression network based on certain correlation measures is not always a conclusive step. We feel that similarity based on chemical properties between interacting genes also plays important role during interaction. We try to investigate the chemical proximity of any hub gene and its immediate neighbours within the subnetwork of the hub gene. As a case study, we consider AD responsible few key genes based on their centrality in the co-expression network. We first quantify DNA sequence of a hub gene and its neighbours with respect to their physicochemical properties. Finally, we compute proximity of a hub gene with its neighbours to show that hub genes are closely associated with their neighbours bio-chemically.

We organize the rest of the paper as follows. At first in Section 2 we describe the process of extracting hub genes and sub-networks. In Section 3, we develop a mathematical model to map and arbitrary DNA sequence to its corresponding chemical properties. We compute the proximity of each neighbour genes with all candidate hub genes to see

how they are associated with each other (Section 4). Few experimental results are reported in Section 5. Finally, we summarize our work in Section 6 with concluding remarks.

2. EXTRACTION OF ALZHEIMER'S SUB-NETWORKS

To perform our experiment we use the gene expression data of Alzheimer's disease (GSE1297) from the NCBI[1] data repository. In this datatset, 31 microarrays are used to analyse 9 control and 22 AD subjects of differing severity in the disease and test their correlation. In order to extract sub-networks from microarray gene expression data, we first perform network construction. Any two genes that have similar expression patterns will have lesser distance (higher similarity) and are more likely to interact with each other. Soft thresholding is then applied where interactions having a distance score above a certain threshold are removed. We use a simple parametric distance measure to compute the proximity between two gene expressions.

DEFINITION 2.1 (PROXIMITY). *Given two expression vectors* $x =< x_1, x_2, \cdots x_n >$ *and* $y =< y_1, y_2, \cdots y_n >$ *for the genes* x *and* y*, the distance,* $\delta(x, y)$*, between two genes can be calculated by taking the normalized difference of standard deviations between the two expression profiles.*

$$\delta(x,y) = \frac{\sum_{i=1}^{n} \mid (x_i - \overline{x}) - (y_i - \overline{y}) \mid}{\sum_{i=1}^{n} (\mid (x_i - \overline{x}) \mid + \mid (y_i - \overline{y}) \mid)} \quad (1)$$

After the network construction, we then extract sub-networks using a density-based clustering approach [14]. These sub-networks are then analysed biologically as well as topologically. To determine how much each of the sub-networks may contribute to Alzheimer's disease we validate using KEGG pathway [8] analysis. We select sub-networks where the percentage of genes that participate in the AD pathway are considerably high along with low p-value. From these selected sub-networks, identification of hub genes is then done based on Maximum Clique Centrality(MCC) score[4]. MCC score considers the degree of connectivity of a node as well as the size of the branches that it connects to the rest of the network. Genes having high ranking MCC scores in the sub-networks are considered hub genes. MCC score of a node v is defined as follows.

$$MCC(v) = \sum_{C \in S(v)} (|C| - 1)! \quad (2)$$

where, $S(v)$ is the collection of maximal cliques which contain v, and $(|C| - 1)!$ is the product of all positive integers less than $|C|$. If no edge is present between v's neighbours then MCC(v) is equal to it degree.

3. CHEMICAL CHARACTERIZATION OF DNA SEQUENCE: A MAPPING

DNA sequences are combinations of four nucleotides A, T, C, G. Depending upon the chemical structures the four bases can form 16 combinations of dual nucleotides along with repeat bases (AG,GA,CT,TC,AC,CA,GT,TG,AT,TA,GC,CG,AA,TT,CC,GG).

[1]https://www.ncbi.nlm.nih.gov/geo/

Table 1: Classification of Dual Nucleotides

Physicochemical Group	Symbol	Dual Nucleotides
Purine	R	AG/GA
Pyrimidine	Y	CT/TC
Amino	M	AC/CA
Keto	K	GT/TG
Strong H bond	S	GC/CG
Weak H bond	W	AT/TA
Repeat Group	E	AA/TT/CC/GG

Table 2: Classification of 20 amino acids according to their chemical properties

Group Name	Amino Acids included
Acidic	Aspartate, Glutamate
Basic	Arginine , Histidine, Lysine
Aromatic side chain	Tyrosine, Phenylalanine, Tryptophan
Aliphatic side chain	Isoleucine, Leucine , Valine, Alanine, Glycine
Cyclic	Proline
Sulfur containing	Methionine, Cysteine
Hydroxyl containing	Serine , Threonine
Acidic amide	Glutamine , Asparagine

Depending upon the chemical structure along with repeating DN groups, the dual nucleotides are classified into seven (07) possible classes [10] given in Table 1. According to the definition of central dogma, transcription and translation are the two steps, through which the information in genes flows into proteins. A protein structure is dependent upon the chemical properties of amino acid by which it is formed. The chemical features of 20 amino acids are basically depending on eight chemical properties, which are shown in Table 2.

Quantitative understanding of a gene can be made with the help of chemical characterization without any interventions of wet lab experiments. Till date several researchers have tried booms and bursts to make quantitative analysis of a gene family. Some papers are reported where the authors have adopted different graph theoretical approaches to compare several DNA and primary protein sequences [19, 7, 18]. Some authors have tried to make mathematical models on various chemical properties of protein families to understand genes and genome[5, 6].

Analysis of DNA sequences using its underlying biological information hidden within dual nucleotides (DNs) is a good approach and has been reported in some previousresearches. Randic [13] introduced a qualitative approach to make quantitative comparisons of DNA sequences, whereas, Qi and Fan [11] proposed 3D graphical representation of DNA sequence based on dual nucleotides. They have proposed PN-curve for it. Wu et al. [12] tried to deal with neighbouring nucleotides of DNA sequence.

Our aim is to find out the nature of distribution of physicochemical properties of a DNA sequence in hub genes and its linked genes from the subnetworks of Alzheimer's. In this regard, it is worth enough to state that, 64 codons which code for 20 amino acids, consists of three nucleotides, where first two of them carry the properties of dual nucleotides.

In this work we try to find out the distribution of physicochemical properties of a DNA sequence from its primary protein sequence in hub and linked genes. Next, we investigate the similarities between linked genes with hub genes which are functionally dependent on hub genes. These investigations may put lights on the basic physical and chemical characteristics of the linked genes of a particular hub genes that may be responsible for Alzheimer's disease.

We try to investigate physicochemical properties of DNA

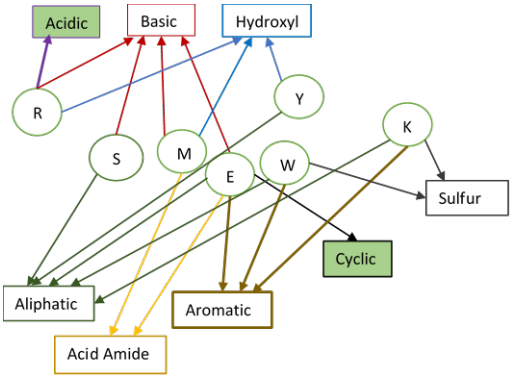

Figure 1: Mapping between physicochemical properties of DNA sequence and chemical properties of amino acid. Dual nucleotide classes are shown in circle and rectangle represents chemical groups

sequence that keep their signatures in protein synthesis. So, a mapping can be defined between physicochemical properties of any arbitrary DNA sequence and chemical properties of its amino acid sequence. As it is well known that all 16 possible combinations of dual nucleotides built from 4 bases (A,T,C,G) of a DNA sequence exhibits seven (07) different physicochemical properties (R,Y,S,W,A,K,E), which finally converges to the eight (08) chemical properties of amino acid sequences such as Aliphatic, Aromatic, Acidic, Basic, Cyclic, Acid Amide, Hydroxyl, Sulfur, after transcription and translation take place. A possible mapping from DN chemical properties to amino acid chemical properties is shown in Figure 1.

Detail view of the mapping, shown in Figure 2 may give a better understanding of the chemical structures underlying within the three nucleotide positions of the codon.

3.1 Formal Representation of the Mapping

It is possible to navigate the pathway of chemical properties which is transferred from physicochemical properties of DN sequence to chemical properties of primary protein sequence of itself. Formally, we can define the mapping as follows.

Let $\mathcal{A} = \{A_1, A_2, A_3, \cdots, A_k\}$ be an arbitrary k length nucleotide sequence, where A_i is any nucleotide base i.e. $A_i \in \{A, T, G, C\}$. Assume, $\{X_i, X_j, X_k\} \subset \mathcal{A}$ are the triplet nucleotides (codon) derived from given $\mathcal{A}$, where position of the nucleotides are represented by the lexicographical order of the variables such that $i < j < k$. Let $\mathcal{B} = \{B_1, B_2, \cdots B_n\}$ set of seven (07) possible physicochemical properties (R, Y, A, K, S, W, E) of any dual nucleotide. An intermediate mapping function $f_1 : \{X_i, X_j\} \rightarrow \mathcal{B}$ can convert every (X_i, X_j) from any triplet into corresponding physicochemical properties as denoted by $\mathcal{B}$.

On completion of this intermediate mapping, while we proceed to traverse to the final mapping, it is important to mention to here that the ordering of (X_i, X_j) matters biologically. That is, $(X_i, X_j) \neq (X_j, X_i)$ although (X_i, X_j) and (X_j, X_i) belong to the group with same physicochemical property from $\mathcal{B}$. Next, on considering the third nucleotide X_k with the dual nucleotide group, it maps to one of the eight (08)chemical groups of amino acids. Let $\mathcal{C} = \{C_1, C_2, \cdots, C_8\}$ be the set of this eight (08) chemical groups.

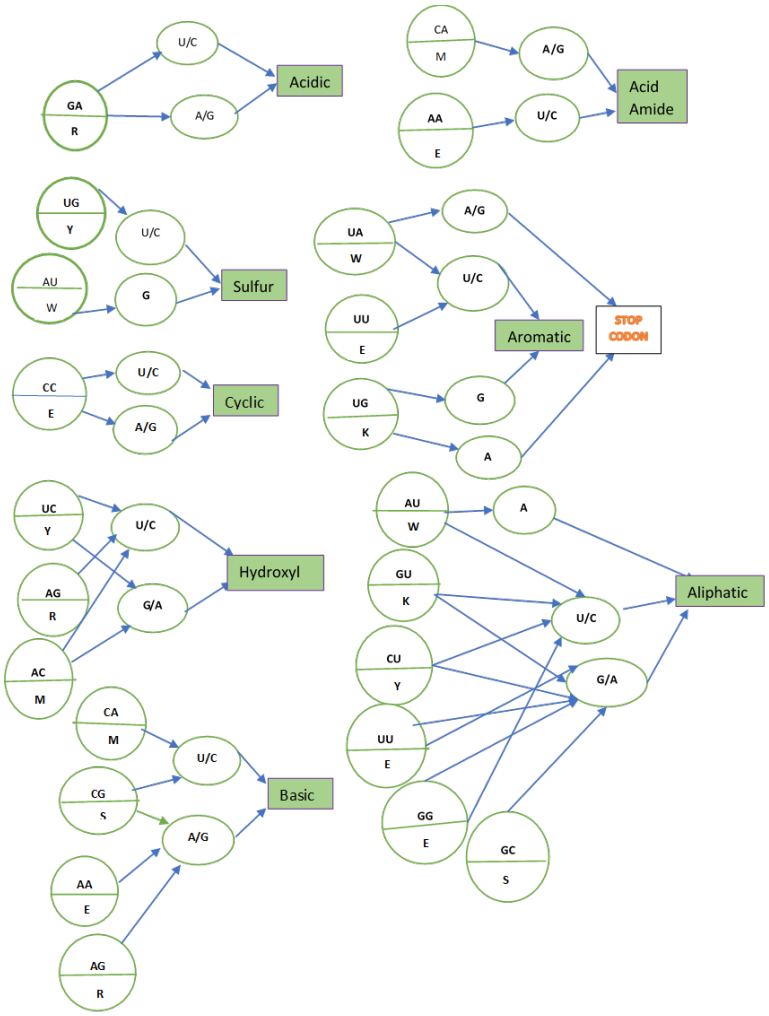

Figure 2: Microscopic view of the mapping between physicochemical properties of DNA sequence and chemical properties of amino acid. The first layer of the mapping represents dual nucleotides, second layer corresponds to nucleotide in third position of the codon and final layer shows the corresponding chemical group (shown inside rectangle) the amino acid belongs to. The corresponding physicochemical groups of DN are shown in half of the circle.

Thus, finally we get a mapping from triplet to its corresponding any chemical groups from $\mathcal{C}$ by using above intermediate mapping (f_1). We may represent the fact as follows.

$$\begin{aligned} & f_2 : \{X_i, X_j, X_k\} \rightarrow \mathcal{C} \\ = \; & f_2 : \{f_1 : \{X_i, X_j\} \rightarrow \mathcal{B}, X_k\} \rightarrow \mathcal{C} \\ = \; & f_2 : \{B_n \in \mathcal{B}, X_k\} \rightarrow \mathcal{C} \end{aligned}$$

We can now easily map the whole sequence $\mathcal{A}$ by extracting all sequence triplet $\{X_i, X_j, X_k\}$ and converting it to corresponding to chemical groups using above mapping functions.

3.2 Matrix Representation

If we consider the matrix representation of the above fact, then according to Figure 2, it is possible to make 8×33 vector representation of each DNA sequence. The vector representation of APP gene is shown in the Table 3 , where we can find that Glycine having chemical property of Aliphatic group and coming from repeating group E with relatively high abandance. As it is known that Aliphatic R groups are hydrophobic and nonpolar, those are one of the major driving forces for protein folding, give stability to globular or binding structures of protein [9]. It can be observed that DNA sequence keeps their signature even after translation.

4. CHEMICAL PROXIMITY BETWEEN SEQUENCES

In this section we try to compute association between DNA sequences of a pair of genes with respect to their distribution of chemical groups mapped from their sequences. The intension behind such proximity computation is to investigate how a hub gene is chemically associated closely with its neighbour or linked genes from its subnetwork.

We read a given DNA sequence in terms of triplets and classify them according to the chemical nature of the amino acid to which it codes. At first, we compute the distribution of chemical groups per sequence from the above matrix representation followed by distance calculation between every neighbour genes with all other hub genes irrespective of any particular subnetwork. We explain the steps below.

4.1 Calculating Distribution of Amino Acids

In nature, twenty (20) amino acids are distributed unevenly in a DNA sequence and hence their chemical groups. However, the percentage of abundance of any chemical groups definitely play certain roles in protein structure formation. Hence, it may be responsible for any functional dependencies between a pair of genes. We use the matrix (subsection 3.2) derived after mapping a sequence to corresponding chemical groupings. We calculate the occurrence percentage of each groups with respect to the sequence size. For a given DNA sequence $\mathcal{A} = \{A_1, A_2, \cdots, A_k\}$, the chemical group distribution can be represented as a vector $\mathcal{D} = \{D_1, D_2, \cdots, D_8\}$ of size eight (08) for eight different chemical groups. Each D_i indicates the percentage of occurrence of a particular chemical group in the given sequence.

4.2 Constructing Distance Matrix

Depending upon the distribution of chemical groups of amino acid in hub and linked genes, we calculate a distance between them to see their proximity with respect to their chemical distribution. The functional dependencies between two genes are directly proportional to the degree of similarity of them. Given two chemical group distribution vectors say, $\mathcal{D}_i$ and $\mathcal{D}_j$, for two target genes, we may calculate the proximity with respect to their chemical properties as follows.

$$\xi(\mathcal{D}_i, \mathcal{D}_j) = \sum_{l=1}^{8} |D_{il} - D_{jl}|. \quad (3)$$

We extend the pairwise distance calculation among all linked genes and all hub genes to get a final distance matrix.

5. EXPERIMENTAL RESULTS

We use few selective Alzheimer's subnetworks and based on their significance in the disease. In this study, six hub genes and their corresponding linked genes which are functionally dependent on them are taken into account to carry out the experiments. The set of hub genes and their neighbours are shown in Table 4.

During the analysis of those 6 hub genes and their associated genes it is observed that out of 20 amino acids, amino acids of aliphatic group contribute a large to construct primary protein sequence. Although DNA sequences have six physicochemical properties along with repeating group. It is observed that in most of the cases they come from repeating group (AA/TT/CC/GG) and code for corresponding amino acids. Out of all neighbours of APP, RPL1 shows high enrichment of Glycine which is Aliphatic and comes from GG (repeating group). MYOMB shows high abundance of Proline, which is from cyclic amino acid group but comes from CC (repeating group). APP, the hub gene also has Glycine the most and is from repeating group of physicochemical property and Aliphatic in nature. The chemical distribution of two hub genes PSEN1 and NDUFB2 and their neighbours are shown in the Figures 3 and 4. Interestingly both the figures reveals an interesting pattern of similar chemical group distribution between members of a subnetwork centered around hub genes.

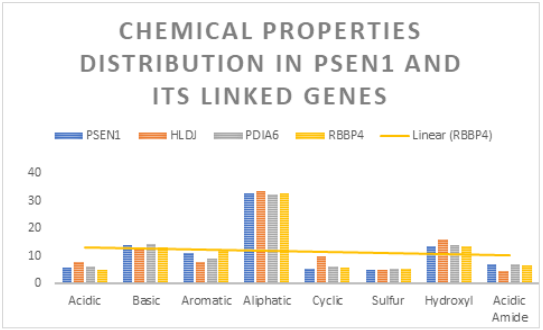

Figure 3: Chemical characteristics of PSEN1 subnetwork

To investigate the proximity of hub genes with its neighbours we calculate the distance matrix from chemical distributions as discussed in section 4.2. Distance matrix for all the neighbour genes of PSEN1 and NDUFB2 with all other hub genes are shown in Table 5 and 6 respectively. We consider 6 hub genes along with their linked genes to investigate the distribution of chemical properties of amino acid in them and to observe how much similarities do the linked

Table 3: Matrix representation of DNA sequence of APP showing the mapping between physicochemical properties of DN and chemical properties of its primary protein sequence.

APP		R				Y				W					S				M				K				E							
		GA	GA	AG	AG	TC	TC	CT	CT	AT	AT	AT	TA	TA	CG	CG	GC	GC	AC	AC	CA	CA	GT	GT	TG	TG	AA	AA	GG	GG	CC	CC	TT	TT
		T/C	A/G	T/C	A/G	T/C	A/G	/T/C	A/G	U/C	A	G	T/C	A/G	T/C	A/G	T/C	A/G	T/C	A/G	T/C	A/G	T/C	A/G	T/C	A/G	T/C	A/G	T/C	A/G	T/C	A/G	T/C	A/G
ACIDIC	Asp	6																																
	Glu		11																															
Basic	Arg				3										5	3																		
	His																				2													
	Lys																											7						
Aromatic	Trp																								4									
	Phe																																8	
	Tyr												2																					
Aliphatic	Ile									8	1																							
	Leu																																	5
	Val																						8	2										
	Ala	6															6	3																
	Gly																												18	46				
Cyclic	Pro																														2	5		
Sulphur	Met											5													7									
	Cys																								7									
Hydroxyl	Ser					6	5						2																					
	Thr																		4	6														
Acidic amide	Gln																				5													
	Asn																										3							

Table 4: Hub Genes and Linked Genes

Hub Genes	Linked Genes	Hub Genes	Linked Genes	Hub Genes	Linked Genes
APP	MYO9B	NDUF1	ATP5H	NDUFB2	NDUFC2
	PLA2G6		ATP5O		NDUFS32
	RNGTT		ATP6V1G2		UCQRQ
	RPL18		COX8A		UQCRC2
	SCRAB2		NDUFA5		ATPMC3
	SEMA6A				ATP5J2
	SSH3	UQCR10	ATP5B		ATP5J
PPP3R1	CN1H3		ATP5C1	PSEN1	COX7B
	GRIA1		ATP6V1G2		HLA-J
	GRIN2A		COX6C		PDIA6
	GSPT2		COX7C		QK1
	PDE10A		NDUFA4		RBBP4
	TUBA3D		UQCR10		
	VDAC1				

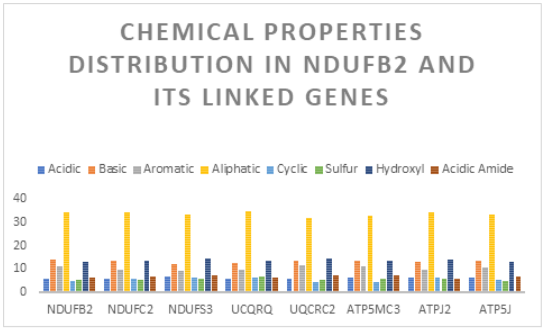

Figure 4: Chemical group distribution of NDUFB2 and its neighbour genes

genes have with their hub genes along with their neighbouring hub genes too. We can easily observe from the table that the linked genes of a hub genes PSEN1 and NDUFB2 respectively have smaller distances in comparison to other hub genes. This establish a new and interesting fact that any interacting genes based on their co-expressions are also chemically close to each others. Among 6 networks, hub gene PSEN1 and its three linked genes have strong similarities in distribution of chemical properties of amino acids. But it is also observed that the neighbours of APP such as MYO9B, PLA2G6 and RPL18 showing high proxity with PSEN1 and NDUFB2 instead of APP. So PSEN1 and NBUFB2 may have enough strength to be a hub gene compared to APP for those neighbour genes. It may indicate a fact that association derive from expression data may not always be true from biological point of view.

6. CONCLUSION

Table 5: Distance matrix for neighbours of PSEN1 with all other hub genes.

	HLDJ	PDIA6	RBBP4
PSEN1	4.393	4.9389	16.4018
NDUFB2	6.5219	7.3552	17.7422
NDUFA1	42.0535	39.2301	45.0277
PPP3R1	88.5546	9.2677	21.2383
UQCR10	141.523	7.5615	17.1946
APP	30.5543	25.3935	25.8261

In silico inference of gene interaction networks from experimental data is a challenging and important task in computational biology. Majority of the works use expression data as an as a major source of input to infer the network. We believe that during interaction the chemical properties of the amino acids and their distributions in a gene plays a vital role. In this work we demonstrated the justification of the fact by applying our idea in a selective subnetworks of Alzheimer's disease networks which is derived from gene expression data. We observed that chemical properties are distributed very similar way between a hub gene and its neighbours. We proposed a new mapping technique for converting a DNA sequence to corresponding amino acids chemical groups. It is our firm belief that our investigation will give a new dimension in the future research on computational network inference.

7. REFERENCES

[1] E. Bagyinszky, Y. C. Youn, S. S. A. An, and S. Kim. The genetics of Alzheimer's disease. *Clinical interventions in aging*, 9:535, 2014.

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

[3] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, pages 418–429. World Scientific, 1999.

[4] C.-H. Chin, S.-H. Chen, H.-H. Wu, C.-W. Ho, M.-T. Ko, and C.-Y. Lin. cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Systems Biology*, 8(4):S11, 2014.

[5] J. K. Das and P. P. Choudhury. Chemical property based sequence characterization of ppca and its

Table 6: Distance matrix for the neighbours of NDUFB2.

	NDUFC2	NDUFS3	UCQRQ	UQCRC2	ATP5MC3	ATPJ2	ATP5J
PSEN1	4.6205	8.6653	9.1508	3.5074	3.4736	7.5481	2.6911
NDUFB2	4.3518	10.7625	8.3721	6.097	5.1517	7.1289	4.042
NDUFA1	38.3227	40.5954	42.151	37.85	37.7692	40.0121	37.6762
PPP3R1	9.4344	12.378	12.9302	5.1694	4.3579	11.3445	6.9629
UQCR10	5.964	9.0659	7.4803	9.6024	9.3967	6.3885	8.577
APP	28.464	27.1783	28.2659	26.5347	26.9442	26.9848	28.792

homolog proteins ppcb-e: A mathematical approach. *PloS one*, 12(3):e0175031, 2017.

[6] J. K. Das, P. Das, K. K. Ray, P. P. Choudhury, and S. S. Jana. Mathematical characterization of protein sequences using patterns as chemical group combinations of amino acids. *PloS one*, 11(12):e0167651, 2016.

[7] N. Helal, R. A. Moneim, and M. Fathi. Mathematical modeling of p53 gene based on matlab code. *TC*, 2(1):3, 2012.

[8] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[9] H. Lomeli, R. Sprengel, D. J. Laurie, G. Köhr, A. Herb, P. H. Seeburg, and W. Wisden. The rat delta-1 and delta-2 subunits extend the excitatory amino acid receptor family. *FEBS letters*, 315(3):318–322, 1993.

[10] L. R. Nemzer. A binary representation of the genetic code. *Biosystems*, 155:10–19, 2017.

[11] X.-Q. Qi, J. Wen, and Z.-H. Qi. New 3d graphical representation of dna sequence based on dual nucleotides. *Journal of Theoretical Biology*, 249(4):681–690, 2007.

[12] Z. Qi and X. Qi. Novel 2d graphical representation of dna sequence based on dual nucleotides. *Chemical Physics Letters*, 440(1-3):139–144, 2007.

[13] M. Randić, M. Vracko, A. Nandy, and S. C. Basak. On 3-d graphical representation of dna primary sequences and their numerical characterization. *Journal of chemical information and computer sciences*, 40(5):1235–1244, 2000.

[14] S. Roy and D. Bhattacharyya. An approach to find embedded clusters using density based techniques. In *International Conference on Distributed Computing and Internet Technology*, pages 523–535. Springer, 2005.

[15] S. Roy, D. K. Bhattacharyya, and J. K. Kalita. Reconstruction of gene co-expression network from microarray data using local expression patterns. *BMC bioinformatics*, 15(7):S10, 2014.

[16] S. Roy, H. N. Manners, M. Jha, P. H. Guzzi, and J. K. Kalita. Soft computing approaches to extract biologically significant gene network modules. In *Soft Computing for Biological Systems*, pages 23–37. Springer, 2018.

[17] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34(2):166, 2003.

[18] A. Sengupta, J. K. Das, and P. P. Choudhury. Investigating evolutionary relationships between species through the light of graph theory based on the multiplet structure of the genetic code. In *Advance Computing Conference (IACC), 2017 IEEE 7th International*, pages 854–859. IEEE, 2017.

[19] H.-J. Yu. Similarity analysis of dna sequences based on three 2-d cumulative ratio curves. In *International Conference on Intelligent Computing*, pages 462–469. Springer, 2011.