19th International Computer Science and Engineering Conference (ICSEC)
Chiang Mai, Thailand, 23-26 November, 2015

Analysis of Amino Acid Pairs Relationships Based on Protein-Protein Interactions

Kittirat Thepsutum and Sudsanguan Ngamsuriyaroj
Faculty of Information and Communication Technology
Mahidol University
Thailand
{kittirat.the@student.mahidol.ac.th, sudsanguan.nga@mahidol.ac.th}

***Abstract*—A protein-protein interaction is important for all living cells since it performs important biological functions to form cells as well as control their mechanisms. Identifying PPIs is always a challenge for biology researchers. Many computational methods have been developed to predict PPIs using different data types including gene neighborhood and genetic profiles. However, these methods cannot be implemented if prior knowledge about proteins is not available. Furthermore, most methods have focused on the prediction accuracy. In this paper, we propose a novel method to analyze a well-known protein-protein interaction network with their strongest amino acid pairs using only protein sequences. Our work uses the Principal Component Analysis (PCA) technique to find outstanding amino acid pairs for each protein. We also use the Pearson's Correlation to find the strongest amino acid pairs of interactions.**

Keywords— Protein-Protein interaction; Principal component analysis; Amino acid; Correlation

I. Introduction

Proteins are large biological molecules, consisting of one or more long chains of amino acid residues. A linear chain of amino acid residues is called a polypeptide. Proteins perform functions within a living organism including catalyzing metabolic reactions, replicating DNA, responding to stimuli, and transporting molecules from one location to another [4]. Therefore, proteins are quintessential for all living organisms. One protein has its own function but whenever it interacts with each other, it also provides a new biological function called protein-protein interaction (PPI). PPIs occur when physical contacts established between two or more proteins as a result of biochemical events [5]. Understanding how proteins interact with each other and identifying biological networks is important to comprehend how proteins work within a cell [6]. Up to now, there are many studies on proteins and PPIs, a collection of PPI databases, protein functions, and PPI prediction [5-16]. Identifying a protein interaction would help researchers to find a way to determine biological functions.

In early studies, the biologists found an interaction from experiments but it has been a time consuming process and costly. Alternatively, having prior study on PPIs using computational approaches would help speed up the discovery. In addition, several databases have been developed to collect proteins and their interaction information from experiments and literatures for years [7-9]. A number of studies had proposed computational methods to predict PPIs [10], and most of them used the knowledge of biological data as found in GO (gene ontology) [11] or PPI prediction via functional regions [12]. Moreover, PPI studies are not limited to the same organism or species but cross organisms are investigated as well. For instance, a host-pathogen study focuses on the disease infection across different species, and the PPI study in host-pathogen is highly complicated due to more than one domain is involved [13].

Studying protein-protein interactions aims to find protein functional systems for a specific purpose. Since amino acids are composed to a protein, studying the relationships among amino acids would help predict protein-protein interaction indirectly. Recently, the amino acid residue association model, so called ARA model, had been proposed for large scale protein-protein interaction prediction [15]. There were six PPI datasets studied, and the prediction had significantly improved.

In our work, we further investigate which pairs of amino acids are outstanding among others to essentially influence an interaction. The concept is similar to human interaction. For example, among a group of close friends, each person has different habits but some common habits are strong or fit best in bringing those persons together. Thus, in one protein interaction, there might have some important attributes which make some pairs of proteins to interact. Since there are so many interactions in one pair of proteins, we will use the principal component analysis method to identify outstanding amino acid pairs, and use the Pearson's correlation coefficient to identify the strength of the interaction.

The remainder of this paper is organized as follows. Section 2 gives some background. Section 3 presents our proposed work. Section 4 describes the experimental results. Section 5 explains related work, and Section 6 concludes our work.

II. Background

In this section, we give background information on amino acids, proteins, protein interation and protein databases. The principal component analysis and Pearson's correlation methods are also explained.

A. Amino acids, Proteins and Protein Interaction

Proteins are large biological molecules consisting of one or more long chains of amino acid residues, as shown in Figure 1, which are the results of the protein synthesis process. There are well known 20 amino acids, and each amino acid is in a chemical structure formed and combined together with peptide bond.

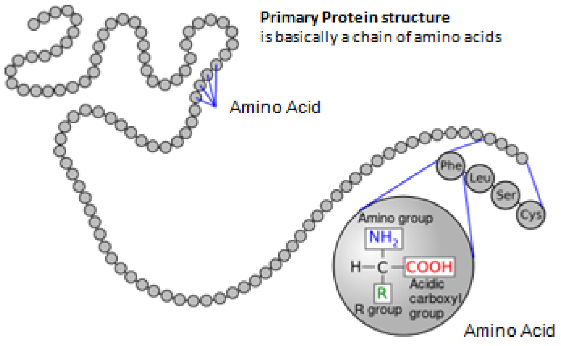


Fig. 1. An Amino Acid Chain (Protein) [22].

The main functions of protein include producing enzymes, transporting chemical tasks, and controlling hormones. A protein consisting of amino acids will have its own function but when two or more proteins bind together, they will generate another function; it is called protein interaction. Protein interactions are normally displayed in a form of a network of interaction as shown in Figure 2.

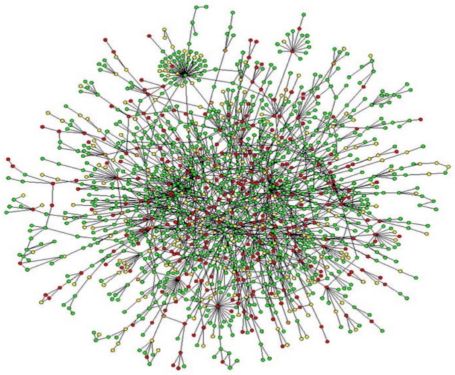


Fig. 2. An Amino Acid Chain (Protein) In Yeast [23].

B. Proteins and Protein Interaction Databases

One of the protein interaction databases includes DIP (Database of Interacting Proteins) [7] and IntAct [9]. In addition to being a large database for protein interactions, DIP provides a research tool for studying cellular networks of protein interactions. It also gives comprehensive and integrated tools for browsing and extracting information about protein-protein interactions and interaction networks in biological processes. Protein related information on DIP can be also downloaded for studying. IntAct [9] contains similar protein information. Another well-known site is Uniprot which is a freely accessible resource of protein sequence and functional information [2]. It provides protein knowledge base information including protein unique IDs, protein names, gene names, organisms, sequences and etc.

C. PCA (Principal Component Analysis)

Principal Component Analysis is a statistical technique used to examine the interrelation between variables among data in a data set. It is also a data reduction method used to express multivariable data with fewer dimensions. Figure 3 displays a data plot between variable X and Y. The regression line determines the best fit to a data set but PCA determines several orthogonal lines of the best fit to the data set [17, 18]. In this example, only two variables are present; therefore, there are two orthogonal lines. Each line shows different variation. The longest line explains 70% of the variation and the short line explains 30% of the variation. Typically, one data set could many variables or dimensions.

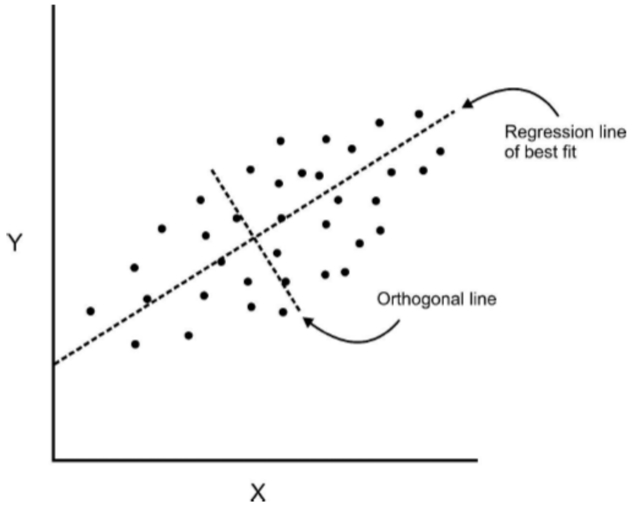


Fig. 3. A 2D Linear Graph with Two Orthogonal Lines [18]

Mathematically, the lines are called eigenvectors, and the length of the lines is called eigenvalue. Each line is also called as a component. Thus, the longest component is the principal component, which has the greatest variance of the data set. Since there are many components with different variance, the highest components will be selected as the main components. The component scores are based on the observation's component loading and their original variables among the data set. The highest scores will be the representatives of the data.

D. Pearson's Correlation

Pearson's correlation is widely used to measure the degree of linear dependence between two variables

X and Y, giving a value between +1 and −1, where 1 indicates the total positive correlation, 0 indicates no correlation, and −1 indicates the total negative correlation. The data should be expressed an interval or as ratio scales [20,21]. The correlation coefficient can be calculated using the Equation (1) given as follows.

$$r_{xy} = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} \quad (1)$$

r_{xy} :Correlation coefficient value between X and Y

$\sum X$:Summation of X

$\sum Y$:Summation of Y

$\sum XY$:Summation of multiplication between X and Y

$\sum X^2$:Summation of square of X

$\sum Y^2$:Summation of square of Y

n :the number of samples

III. Proposed Method

Figure 4 presents our proposed system overview. In the first step, we do data collection and data cleansing. We use the protein sequences from UniProt [2] and PPIs of Saccharomyces cerevisiae (Yeast) from DIP as of Jan 17, 2014. In addition, we need to do data cleansing before inserting them into our database for further analysis. During the data cleansing, we would make sure that all proteins obtained have sequences and they are all unique. In other words, proteins have no sequence and do not appear in UniProt database will be removed. Thus, for all 22,637 PPIs obtained from DIP, only 22,055 PPIs are cleansed, and only 4,929 proteins are unique in our data collection as all duplicated proteins are eliminated.

In the second step, we compute the feature encoding of each unique protein. The feature is computed as a proportion of all 20 amino acids appeared in one protein. In other words, we compute the frequency of each amino acid in one protein sequence divided by the length of its sequence. We call this feature as amino acid ratio. Thus, one protein sequence will have 20 features described.

In the third step, the outstanding amino acids are estimated via applying PCA on the features obtained for each protein sequence. First, the variance of all 20 amino acids are calculated. Second, the eigenvector and the eigenvalue of the covariance matrix are calculated, so called a component. Since there are only 20 attributes, 20 components are computed. Eigen vectors having high eigenvalue are principal components, and only the components having the cumulative proportion variance of more than 50% will be considered and selected. The chosen components are the representatives of the data set. Third, we need to find amino acids belonging to each selected component by calculating component loading values. The loading values tell how much variation of a component is. Next, based on the loading values, only the top three amino acids for each component are selected. Finally, the component score is calculated to determine the outstanding amino acid of each protein.

In the fourth step, the absolute correlation value of all outstanding amino acids for each protein is computed using Pearson's correlation. The result is an amino acid correlation matrix. Finally, the PPIs with their amino acid correlation are displayed as a matrix diagram, and they will be grouped into Strong, Weak and Very weak correlation.

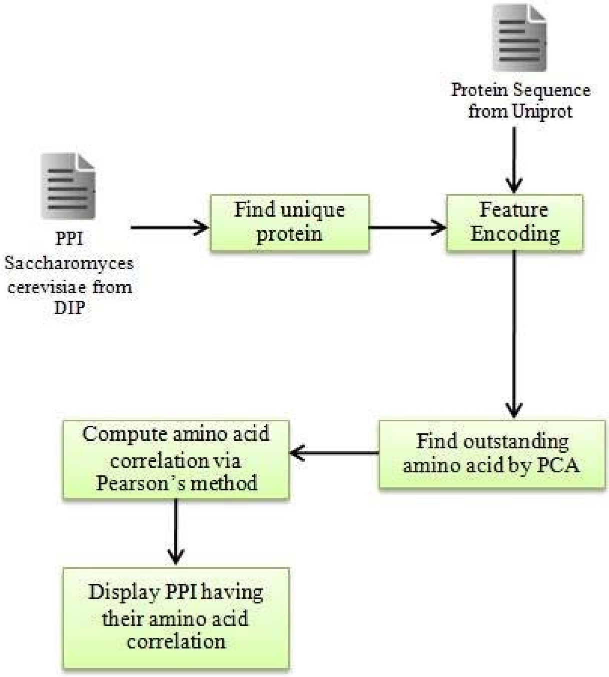

Protein Sequence from Uniprot
PPI Saccharomyces cerevisiae from DIP
Find unique protein
Feature Encoding
Find outstanding amino acid by PCA
Compute amino acid correlation via Pearson's method
Display PPI having their amino acid correlation

Fig. 4. System Overview of the Proposed Work

IV. Experimental Results

Due to high computation required, our experiment only randomly selected only 50 PPIs from 4,929 unique proteins in Yeast protein dataset were, and only 64 unique proteins from the PPIs were used. From each unique protein, we calculated the amino acid ratio as features for PCA and Pearson's correlation as shown in Figure 5. Table I shows the variance of the components from PCA and their cumulative proportion. The first three components, namely C1, C2 and C3, having the cumulative proportion more than 50% are selected as the representatives of all the data set since they are considered as the majority of the components.

Table II presents the outstanding amino acid of some sample proteins based on the component scores. From the sample results, we can see that the amino acids: L, K and S are outstanding among 3 proteins.

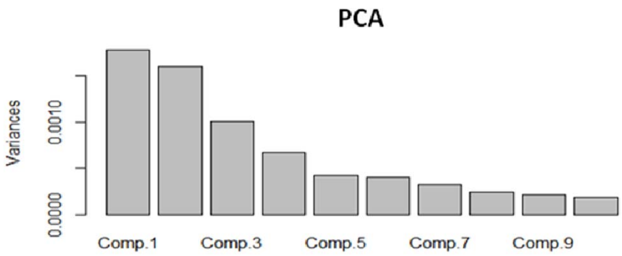


Fig. 5. Sample Component Variance for PCA

TABLE I. SUMMARY OF COMPONENTS FROM PCA

	C1	C2	C3
Standard deviation	0.0422	0.0401	0.0317
Variance	0.0018	0.0016	0.0010
Proportion of variance	0.2333	0.2106	0.1321
Cumulative Proportion	0.2333	0.4439	0.5760

TABLE II. OUTSTANDING AMINO ACIDS OF SAMPLE PROTEINS

Proteins	Selected Component	Amino acids
P42073	C3	L , K , S
P32578	C2	N , S , E
P22219	C3	L , K , S
P11978	C3	L , K , S

From our selected proteins, the correlation coefficients of outstanding amino acids, which are E, G, K, L, N, S, are presented in Table III as the correlation matrix. We also categorize those correlation values into three groups and labeled them as Strong, Weak and Very Weak to indicate the level of correlation among those amino acids. In the table, the Strong correlation is highlighted in green, the Weak one is highlighted in yellow, and the Very Weak one is highlighted in red.

Table IV shows the values of sample amino acid correlation values in all three categories. We can see that the Strong correlation has high values greater than 0.2, the Weak correlation has the values ranging between 0.1 and 0.2, and the Very week correlation has the values lower than 0.1. Note that the abbreviated letters in the table are amino acids. They are E (Glutamic Acid), G (Glycine), K (Lysine), L (Leucine), N (Asparagine) and S (Serine).

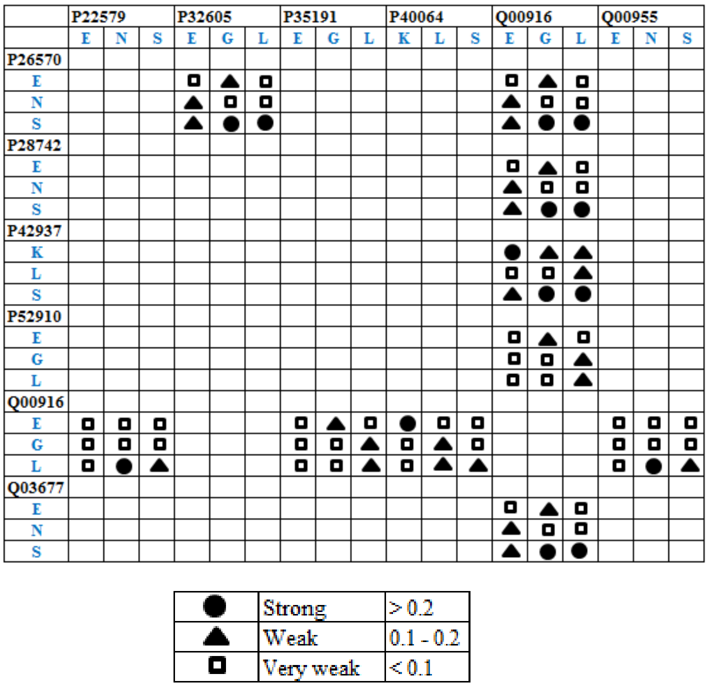
TABLE III. AMINO ACID CORRELATION MATRIX

	E	G	K	L	N	S
E	0.067	0.111	0.239	0.007	0.015	0.010
G	0.010	0.029	0.036	0.117	0.005	0.093
K	0.246	0.130	0.070	0.126	0.011	0.011
L	0.060	0.053	0.039	0.173	0.212	0.113
N	0.132	0.029	0.052	0.006	0.059	0.027
S	0.155	0.294	0.117	0.336	0.367	0.017

TABLE IV. GROUPING OF AMINO ACID PAIRS INTO THREE LEVELS

Strong (> 0.2)		Weak (0.1 - 0.2)		Very weak (< 0.1)	
S-N	0.367	S-E	0.155	G-E	0.010
S-L	0.336	N-E	0.132	L-E	0.060
S-G	0.294	E-G	0.111	G-G	0.029
K-E	0.246	K-G	0.130	L-G	0.053
E-K	0.239	S-K	0.117	N-G	0.029
L-N	0.212	G-L	0.117	G-K	0.036

Figure 6 shows which pairs of PPIs having the outstanding amino acid pairs as they are categorized into three levels according to the correlation values of those amino acids.



	P22579			P32605			P35191			P40064			Q00916			Q00955		
	E	N	S	E	G	L	E	G	L	K	L	S	E	G	L	E	N	S
P26570																		
E				□	▲	□							□	▲	□			
N				▲	□	□							▲	□	□			
S				▲	●	●							▲	●	●			
P28742																		
E													□	▲	□			
N													▲	□	□			
S													▲	●	●			
P42937																		
K													●	▲	▲			
L													□	□	▲			
S													▲	●	●			
P52910																		
E													□	▲	□			
G													□	□	▲			
L													□	□	▲			
Q00916																		
E	□	□	□				□	▲	□	●	□	□				□	□	□
G	□	□	□				□	□	▲	□	▲	□				□	□	□
L	□	●	▲				□	□	▲	□	▲	▲				□	●	▲
Q03677																		
E													□	▲	□			
N													▲	□	□			
S													▲	●	●			

●	Strong	> 0.2
▲	Weak	0.1 - 0.2
□	Very weak	< 0.1

Fig.6. Sample Interacting PPIs with Their Amino Acid Pairs

V. RELATED WORK

Protein interactions have been studied for years by biological experiments. But, the experiments are costly and require lots of resources. Recently, using computational methods gains high attention from both biologists and computer scientists. To find good methods to predict PPIs would not be easy. Juwen Shen and et.al. The article in [14] proposed a method to predict protein-protein interactions based only on sequence information since it was based on the concept that amino acid sequence alone might be sufficient to predict the interaction. In the experiments, more than diverse 16,000 PPIs are used and the features of amino acids are also extracted. Using the prediction model via the support vector machine, the prediction results gave high accuracy.

Zhu Hong You and et.al. proposed the method to predict PPIs which uses less time taken by using PCA and another prediction called Ensemble Extreme Learning Machine (EELM) but gives high accuracy [16]. The results gave good performance in prediction and fast learning speed. Similarly, the PCA is used to reduce unused data dimensions and thus, the computation time had been improved.

A few studies on large scale of proteins have been conducted. But there is only one work using information from amino acids to predict PPIs [15]. The PPIs were collected from six species: E.coli, H.pylori, Yeast, C.elegans, Fruit fly, Human. Moreover, amino acid ratio of each protein is computed as a feature for the dataset. The correlation between amino acids of interacting proteins is also calculated using Pearson correlation. The results are the correlation matrix of amino acids from both positive PPIs and negative PPIs. Furthermore, the logistic regression model is used for prediction. But, the prediction accuracy is not so high, just above 50%.

In summary, most work focus on PPI prediction, but our work focuses on the computation of amino acid correlation matrix between known pairs of PPIs in order to identify significant pairs of amino acids which in turn would mainly influence the mechanism of the protein interaction. We do not aim to propose any prediction model for protein-protein interactions.

VI. Conclusions

In this paper, we propose a novel method to study the correlation of outstanding amino acids among known protein-protein interactions. We used the statistical and computational techniques including the PCA which helps identify the outstanding amino acids and Pearson's correlation which computes the strength of amino acid pairs. Our experimental results present the PPIs correlation matrix illustrating the correlation values of the most correlated amino acid pairs. To exemplify, the amino acid S has strong relationship with the amino acid G and L, for all pairs of PPIs having S as the outstanding amino acid. This observation should be further investigated with other PPI pairs, and could be confirmed via a real biological experiment. Thus, we hope that the results of this work could be useful for biologists to further investigate which pairs of amino acids significantly influence the interaction of the known PPIs that they are interested in.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Anuphap Prachumwat at the Center of Excellence for Shrimp Molecular Biology and Biotechnology (CENTEX Shrimp), Faculty of Science, and Dr. Apirak Hoonlor, Faculty of ICT, Mahidol University for their valuable comments and suggestions during the proposal session.

References

[1] DIP:Home.[Online].Available:http://dip.doe-mbi.ucla.edu/dip/ Main.cgi. [Accessed: 25- Jan- 2014].

[2] UniProt. [Online]. Available: http://www.uniprot.org/. [Accessed: 30- Jan- 2014].

[3] Protein Structure. [Online]. Available: http://www.chemguide.co.uk/organicprops/aminoacids/proteinstruct.html. [Accessed: 21- Jun- 2014].

[4] Protein. [Online]. Available: http://en.wikipedia.org/ [Accessed: 27- Sep- 2014].

[5] Protein–protein interaction. [Online]. Available: http://en.wikipedia.org/ [Accessed: 27- Sep- 2014].

[6] Overview of Protein-Protein Interaction Analysis | Life Technologies. [Online]. Available: http://www.piercenet.com/. [Accessed: 27- Sep- 2014].

[7] I. Xenarios, 'DIP: the Database of Interacting Proteins', Nucleic Acids Research, vol. 28, no. 1, pp. 289-291, 2000.

[8] A. Chatr-aryamontri, A. Ceol, L. Palazzi, G. Nardelli, M. Schneider, L. Castagnoli and G. Cesareni, 'MINT: the Molecular INTeraction database', Nucleic Acids Research, vol. 35, no., pp. D572-D574, 2007.

[9] H. Hermjakob, 'IntAct: an open source molecular interaction database', Nucleic Acids Research, vol. 32, no. 90001, pp. 452D-455, 2004.

[10] K. A. Theofilatos, C. M. Dimitrakopoulos, A. K. Tsakalidis, S. D. Likothanassis, S. T. Papadimitriou and S. P. Mavroudi, 'Computational Approaches for the Prediction of Protein-Protein Interactions: A Survey', CBIO, vol. 6, no. 4, pp. 398-414, 2011.

[11] P. Li, L. Heo, M. Li, K. Ho Ryu and G. Pok, 'Protein Function Prediction Using Frequent Patterns In Protein-Protein Interaction Networks', Proceedings of the Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 26-28 July 2011, Shanghai, China, pp. 1617-1620.

[12] H. Fei-Hung and C. Hung-Wen, 'Protein-Protein Interaction Prediction based on Association Rules of Protein Functional Regions', Proceedings of the Second International Conference on Innovative Computing, Information and Control (ICICI), 5-7 Sept. 2007, Kamamuto, Japan.

[13] R. Arnold, K. Boonen, M. Sun and P. Kim, 'Computational analysis of interactomes: Current and future perspectives for bioinformatics approaches to model the host–pathogen interaction space', Methods, vol. 57, no. 4, pp. 508-518, 2012.

[14] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li and H. Jiang, 'Predicting protein-protein interactions based only on sequences information', Proceedings of the National Academy of Sciences, vol. 104, no. 11, pp. 4337-4341, 2007.

[15] R. Rao, K. Tun, Y. Makita, S. Lakshminarayanan and P. K. Dhar, 'Amino-acid residue association models for large scale protein protein interaction prediction', 2009.

[16] Z. You, Y. Lei, L. Zhu, J. Xia and B. Wang, 'Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis', 2013.

[17] A tutorial on Principal Components Analysis,: Department of Computer Science - University of Otago, New Zealand, 2002, pp. 1-26.

[18] 'Principal Components Analysis'. [Online]. Available: http://webspace.ship/edu/pgmarr/Geo441/Lectures/Lec%2017%20-%20Principal%20Component%20Analysis.pdf. [Accessed: 27- Jun-2014].

[19] Correlation coefficient and p-values: what they are and why you need to be very wary of them, Queen Mary: Queen Mary University, 2012.

[20] Statistics How To, 'Pearson Correlation: Definition and Easy Steps for Use', 2013. [Online]. Available: http://www.statisticshowto.com/what-is-the-pearson-correlation-coefficient/. [Accessed: 27- Jun- 2014].

[21] Amino acid, [Online]. Available: http://en.wikipedia.org/ [Accessed: 22- May- 2015].

[22] M. Lima, 'visualcomplexity.com | Protein-Protein Network', Visualcomplexity.com, 2015. [Online]. Available: http://www.visualcomplexity.com/vc/project.cfm?id=184. [Accessed: 22- May- 2015].