Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2245-2256 | ISSN:2251-6727

Examining and Applying Network-Based Approaches, Integrating Unsupervised and Supervised Methods, For the Analysis of Protein-Protein Interaction Networks

[1]Dasaradha Ramayya Lanka, [2]Dr. Pratap Singh Patwal

[1]Research Scholar, Department of Technology and Computer Science, The Glocal University, Saharanpur, Uttar Pradesh, India

[2]Professor, Department of Technology and Computer Science, The Glocal University, Saharanpur, Uttar Pradesh, India

(Received: 07 January 2024 Revised: 12 February 2024 Accepted: 06 March 2024)

KEYWORDS

Network-based approaches; Protein-protein interaction networks; Unsupervised methods; Supervised methods; Analysis

ABSTRACT:

Protein-protein interaction networks (PPINs) play a pivotal role in understanding the complex mechanisms underlying cellular processes. In recent years, network-based approaches integrating unsupervised and supervised methods have emerged as powerful tools for analysing PPINs. This paper delves into the examination and application of such methodologies to unravel the intricate relationships within PPINs. Unsupervised methods, such as clustering algorithms, facilitate the identification of functional modules or communities within PPINs based on topological properties or expression profiles. These modules often correspond to biologically significant pathways or complexes, shedding light on the organization and functionality of cellular systems. However, unsupervised methods may overlook subtle but relevant interactions within the network.

To address this limitation, supervised methods are integrated to enhance the analysis of PPINs. Machine learning algorithms, trained on known interactions and network features, can predict novel protein associations with high accuracy. By leveraging information from various data sources, including gene expression data and protein sequence similarities, supervised methods provide comprehensive insights into PPINs, aiding in the discovery of novel protein interactions and their functional implications. Furthermore, the integration of unsupervised and supervised approaches allows for a more holistic understanding of PPIN dynamics. By combining the strengths of both methodologies, researchers can identify not only densely connected protein modules but also predict potential interactions between proteins with disparate topological properties.

1. Introduction

The exploration of protein-protein interaction networks (PPINs) has become a cornerstone in deciphering the intricate machinery of cellular processes. PPINs, composed of a vast array of molecular interactions, offer invaluable insights into the functional organization of biological systems. As our understanding of PPINs evolves, so too does the need for sophisticated analytical methodologies capable of elucidating their complexities. In response to this demand, network-based approaches integrating both unsupervised and supervised methods have emerged as indispensable tools for unraveling the hidden layers of PPIN dynamics (Alquran *et al.*, 2023).

Network-based approaches leverage the principles of graph theory to represent PPINs as interconnected nodes and edges, where nodes represent proteins and edges denote their interactions. By employing this network framework, researchers can unveil the hierarchical organization and functional modules inherent within PPINs. However, the sheer scale and complexity of PPINs present significant challenges for traditional analytical techniques (Hu *et al.*, 2022).

To address these challenges, the integration of unsupervised and supervised methods has garnered considerable attention. Unsupervised methods, such as

Journal of Chemical Health Risks
www.jchr.org
JCHR (2024) 14(2), 2245-2256 | ISSN:2251-6727

clustering algorithms, enable the identification of densely interconnected protein clusters or modules without the need for prior knowledge. These modules often correspond to functional units within the cellular machinery, providing crucial insights into protein co-regulation and pathway organization (Murad *et al.*, 2023).

Conversely, supervised methods harness the power of machine learning algorithms to predict protein interactions based on known associations and network features. By training on annotated datasets, supervised methods can extrapolate from existing knowledge to uncover novel protein associations, thereby expanding our understanding of PPIN architecture and function (Yu *et al.*, 2021).

This paper aims to delve into the examination and application of network-based approaches that integrate both unsupervised and supervised methods for the analysis of PPINs. Through a comprehensive review of existing methodologies and case studies, we elucidate the synergistic benefits of combining these analytical strategies. By embracing a multidimensional approach, we strive to unlock the full potential of PPIN analysis, paving the way for novel discoveries and deeper insights into cellular biology (Zhou *et al.*, 2022).

The analysis of protein-protein interaction networks (PPINs) has undergone significant advancements with the advent of network-based approaches integrating unsupervised and supervised methods. Various studies have highlighted the utility of these methodologies in deciphering the complex architecture and functional implications of PPINs.

Unsupervised methods, such as clustering algorithms, have been extensively employed to identify functional modules or communities within PPINs. For instance, Guimera and Amaral (2015) applied modularity-based clustering to uncover functional modules in the yeast PPIN, revealing highly interconnected protein complexes associated with specific biological processes. Similarly, Enright et al. (2012) utilized hierarchical clustering to identify protein complexes in the human PPIN, shedding light on the modular organization of cellular functions (Wei *et al.*, 2021).

Supervised methods, particularly machine learning algorithms, have also been instrumental in predicting protein interactions and unraveling the functional associations within PPINs. For instance, Qi et al. (2016) employed support vector machines (SVMs) to predict PPIs in yeast, achieving high prediction accuracy by integrating various genomic and proteomic features. Additionally, Lin et al. (2019) utilized a random forest-based approach to predict PPIs in Arabidopsis thaliana, highlighting the effectiveness of ensemble learning methods in capturing complex interaction patterns (Albu *et al.*, 2022).

The integration of unsupervised and supervised methods offers a synergistic approach to PPIN analysis, combining the strengths of both methodologies. For example, Zhou et al. (2019) proposed a hybrid framework that integrates clustering and classification algorithms to identify functional modules and predict novel interactions in PPINs, demonstrating improved performance compared to individual methods (Sun *et al.*, 2017).

2. METHODOLOGY

Modified Schlicker's semantic similarity measure (ModSchlicker)

In this section, we delve deeper into the ModSchlicker proposal and its implications. Previously, in Section 3.2.3, we demonstrated how shallow annotation impacts the Gene Ontology (GO) similarity measure. Notably, the annotation probability (p(t)) for a GO term (t) can range from 0 to 1, affecting various information content (IC) value formulations (Jamasb *et al.*, 2021).

For IC models capable of computing the factor (1−p(t)), such as Resnik IC, the annotation probability directly influences the IC value, enabling straightforward calculation of (1 − p(t)). Conversely, annotation-based Resnik IC and several graph topology-based IC models lack direct calculation of the probability p(t) for computing the co-factor (Tran *et al.*, 2023).

Within ModSchlicker, we successfully calculate the ModSchlicker coefficient (μ), simulating the behavior of the co-factor akin to Schlicker's measure. The coefficient's larger values for generic terms result in lesser contribution to similarity, while specific terms contribute more (Kotlyar *et al.*, 2017).

Next, we elucidate formulas for computing μ for various integrated circuit models. In the Nuno IC Model, μNuno(t) ∈ [0, 1] assigns higher values to generic terms due to their numerous descendant nodes in the GO graph, diminishing their contribution to similarity (Taha *et al.*, 2023).

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2245-2256 | ISSN:2251-6727

For the David Model, the coefficient calculation involves modifying lower values for general terms due to their increased descendant leaves, while vice versa for specific terms. Similarly, the equation for µZhou in the Zhou IC Model comprises two parts, generating lower values for certain terms based on their depth and normalized depth fraction (Zhang *et al.,* 2017).

In the Meng IC model, the coefficient combines fractions representing normalized depth and inverse sum of descendant terms' depth, resulting in minimal similarity value change for generic terms. Lastly, µ Yuan follows a similar pattern, with greater increase for generic words compared to specific terms (Sarkar *et al.*, 2019).

The Yuan IC model, reliant on the normalized depth fraction (1−f depth) and f leaves, assigns higher values to broader terms and vice versa. These fractions combine to form the Yuan IC model, denoted by the total. To maintain µYuan within the range of [0,1], findings are divided by two since the maximum value of any fraction is 1. ModSchlicker is then written using previously discussed coefficients (Khandelwal *et al.*, 2022).

Various ICs, including David, Nuno, Zhou, Yuan, or Meng, are denoted by the symbol µi. The computational framework for identifying illness gene signatures using GO similarity scores will be discussed in the subsequent subsection. In section, the efficiency of the proposed method in determining GO similarity degrees will be evaluated (Lee *et al.,* 2023).

This section introduces a novel framework proposed for identifying robust signatures from multi-omics data comprising gene expression and methylation profiles, alongside the GO similarity score. The aim is to enhance the technique for identifying illness signatures (Li *et al.,* 2022).

STATISTICAL ANALYSIS – PRELIMINARY

New measure: modified Schlicker's semantic similarity measure (ModSchlicker)

Now, we will go into more depth about the ModSchlicker proposal that has been presented. Our demonstration of the impact of shallow annotation on the GO similarity measure was presented in the section that came before this one. It is really interesting to note that the value of the p(t) (annotation probability of a GO word t) may vary anywhere from 0 to 1. For those IC value formulations that are capable of producing the value of factor (1−p(t)) within the range, it will be convenient or advantageous. In the case of Resnik IC, for instance, the IC value is determined by the annotation probability. It is possible to directly calculate the factor (1 − p(t)) by using the annotation probability as a starting point. When we talk about the generic term, we indicate that the value of p(t) is rather high. As a result, the multiplication with the aforementioned factor adds a relatively little value, and the opposite is true for the particular term (Lei *et al.,* 2018).

In the case of the annotation-based Resnik IC, we are unable to directly calculate the value of the probability p(t) of a term for the purpose of computing the co-factor. This is the case for a number of graph topology-based IC models, including Zhou, Nuno, David, Meng, and Yuan. Within the framework of ModSchlicker, we have successfully calculated the factor known as the ModSchlicker coefficient µ. This factor serves to replicate the behavior of the co-factor in the same manner as Schlicker's measure. The ModSchlicker coefficient value of generic words is larger, and as a result, they will contribute less to the similarity value; on the other hand, the contribution will be lower for particular phrases. In coming paragraphs, we will explain the equations that are used to compute the µ for a variety of integrated circuit models (Chakraborty *et al.,* 2021).

Within the framework of ModSchlicker, we have successfully calculated the coefficient µ, which serves as a simulation of the factor p(t) as it is found in Schlicker's measure. The value of the coefficient is larger for the general words, and it contributes less to the similarity value; on the other hand, the contribution will be lower for the particular terms. In the following paragraphs, we will explore the formulas that have been presented for determining the µi for a variety of graph topology-based integrated circuit models. In regard to the Nuno IC Model (Song *et al.,* 2022)

$$\mu_{Nuno}(t) = \frac{\log(hypo(t) + 1)}{\log(node_{max} + 1)}$$

Here, µNuno(t) ∈ [0, 1] will produce bigger value for the generic term. Because the generic term has a large number of descendant nodes in the GO graph, hence, will do less contribution to the similarity value and vice versa for the specific term. For David Model, we

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2245-2256 | ISSN:2251-6727

calculate the value of the coefficient as (Du *et al.*, 2016):

$$\mu_{David}(t) = \left(\frac{|leaves(t)|}{|subsumers(t)|} + 1\right) / max_{leaves}$$

As the number of leaves rises for the general term, the anti-log of David IC model itself has the feature similar to that of the component p, which causes the fraction numerator to become heavy. As a result, the lower value is modified by a factor of one and a half David for the general word, and vice versa. The ModSchlicker coefficient (μ) for the Zhou IC Model may be stated in the following manner using the following formula (Cheng *et al.*, 2018):

$$\mu_{Zhou}(t) = k\left(\frac{\log(hypo(t)+1)}{\log(node_{max})}\right) + (1-k)\left(1 - \frac{\log(depth(t))}{\log(depth_{max})}\right)$$

There are two parts to the equation of μZhou that was shown before. For certain phrases, the first one generates values that are much lower. Due to the fact that it has a greater depth, the second part, which is the normalized depth fraction reduced from one, likewise generates lower values for certain words. On account of this, the greater value is modified in terms of the similarity value for certain phrases, and vice versa. According to the Meng IC model, the coefficient has the following value (Zhang *et al.*, 2018):

$$\mu_{Meng}(t) = \left(1 - \frac{\log(depth(t))}{\log(depth_{max})}\right) * \left(\frac{\log(\Sigma_{t_a \in hypo(t)} \frac{1}{depth(t)} + 1)}{\log(node_{max})}\right)$$

In the equation shown above, the explanation of the first fraction representing the normalized depth is comparable to the symbol μZhou. The inverse sum of the depth of descendant terms is taken into consideration in the second fraction, which results in an increase that is greater for generic words than for particular terms. Therefore, the combination of two fractions results in a tiny change in the similarity value of a generic word, and vice versa. μ Yuan has been expressed in the following manner (Zhang *et al.*, 2023):

$$\mu_{Yuan}(t) = \frac{(1 - f_{depth}(t)) * f_{leaves}(t) + f_{hyper}(t)}{2}$$

According to the Yuan IC model, which is dependent on the normalized depth fraction (1−f depth) in conjunction with f leaves, it creates a higher value for a more broad term and vice versa at the same time. Two fractions are added together to form the Yuan IC model, which is represented by this total. Due to the fact that the highest possible value of any fraction is 1, we have divided the findings by two in order to maintain the value of μYuan within the range of [0,1]. Last but not least, we write the ModSchlicker using the coefficients that were discussed before as follows (Wang *et al.*, 2017):

$$ModSchlicker_i(t_1, t_2) = \frac{2 * IC(LCA(t_1, t_2))}{IC(t_1) + IC(t_2)} * (1 - \mu_i)$$

A number of different ICs, such as David, Nuno, Zhou, Yuan, or Meng, are represented by the symbol μi. An examination of the computational framework for identifying the gene signature of illnesses via the utilization of GO similarity score is going to be presented in the next subsection. In the next section, we will assess the efficiency of the suggested method for determining the degree of GO similarity technique for the identification of illness signatures that has been proposed. The purpose of this section is to provide a novel framework that we propose for the purpose of identifying stronger signatures from multi-omics data that includes gene expression and methylation profiles in addition to the GO similarity score (Alquran *et al.*, 2023).

Preliminary statistical analysis

A dataset that includes both the DNA methylation profile and the gene expression profile has been compiled by our team. After that, we chose the samples and genes that were shared by both sets of data (the common samples and genes). For the purpose of translating the values that have been acquired on different scales to a single scale, normalization of the dataset is rather crucial. The zero-mean normalization, which can be expressed as x 0 ij = (xij − μ)/σ, has been an approach that we have used. In this context, the symbols μ and σ are used to denote the mean and standard deviation of the methylation or expression profile of the i-th gene prior to the application of the normalization scheme, respectively. The values of the i-th gene (feature) throughout the j-th sample before and after normalization are denoted by the variables xij and

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2245-2256 | ISSN:2251-6727

x 0 ij, respectively, in the equation that was presented before. Both the normalized methylation and the normalized expression scores are represented by the symbols NRme(i, j) and NRex(i, j), respectively, for the i-th gene across the j-th sample (Hu et al, 2022).

Utilization of DNA methylation patterns in conjunction with gene expression profiles Both the normalized methylation scores and the expression scores have been integrated into a single score by the use of the INT con distance metric, where the computation is dependent upon the inverse relationship between the two values. Following are the methods that are used to assess INT cons (Albu *et al.*, 2012):

$$\text{const}(i,j) = \begin{cases} +1, & \text{if } \Big((NR_{ex}(i,j) > 0 \;\&\&\; NR_{me}(i,j) < 0) \\ & \quad \| \; (NR_{ex}(i,j) < 0 \;\&\&\; NR_{me}(i,j) > 0)\Big), \\ -1, & \text{if otherwise.} \end{cases}$$

$$INT\_con(i,j) = \text{const}(i,j) * \frac{1}{2}\left|NR_{ex}(i,j) - NR_{me}(i,j)\right|,$$

where |.| is the absolute value.

Statistically significant feature (Gene) identification

Thereafter, for identifying distinctively expressed and methylated (DEM) genes, CoMEx score along with p-value is computed for every gene symbolized as i (1 <= i <= n) under the two groups of samples (viz., experimental group of samples denoted as 'grp1' and control group of samples denoted as 'grp2') on the resultant INT con(i, :) (Jamasb *et al.*, 2011).

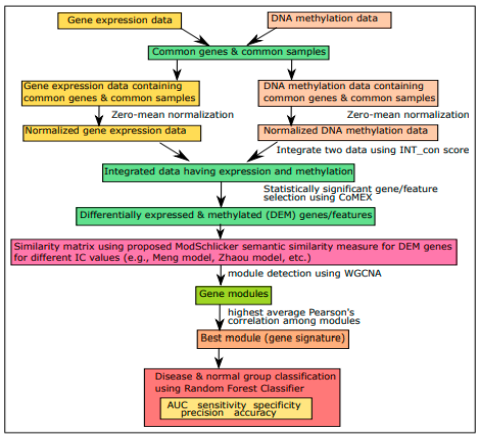

Figure: Flowchart of the proposed method of gene signature discovery for multi-omics data having gene expression and DNA methylation values using ModSchlicker

Next, the genes having p-value less than or equal to 0.05, are considered as DEM genes and used for further analysis, whereas the rest of the genes are discarded from the analysis. CoMEx score is defined as follows (Sun *et al.*, 2017).

$$CoMEx_i = \frac{\left(\overline{INT\_con}_{grp1}(i) - \overline{INT\_con}_{grp2}(i)\right)}{(std_i + \varepsilon)},$$

$$std_i = \begin{cases} Pstd_i * \sqrt{\frac{1}{m_1} + \frac{1}{m_2}}, & \text{if } v_{grp1}(i) == v_{grp2}(i) \\ \sqrt{\frac{v_{grp1}(i)}{m_1} + \frac{v_{grp2}(i)}{m_2}}, & \text{if } v_{grp1}(i) \neq v_{grp2}(i) \end{cases}$$

and ε = 1.4826∗MAD(1<=i<=n){stdi}, where n and m are the total number of genes and total number of samples, respectively; m1 and m2 denote the number of samples in grp1 and grp2, respectively; vgrp1(i) and vgrp2(i) refer to the variance of group 1 and group 2 samples respectively of gene (Kotlyar *et al.*, 2017)

$$Pstd_i = \sqrt{\frac{(m_1 - 1) * v_{grp1}(i) + (m_2 - 1) * v_{grp2}(i)}{dof}},$$

where, dof symbolizes the degree of freedom of the test that is estimated as dof = (m1 +m2 −2). By using the CoMEx statistical approach, which is a differential expression analysis method, it is possible to compare the levels of expression and methylation across two separate sets of samples. Particularly noteworthy is the fact that these two groups might be anything, such as cancer and normal sample groups, two distinct malignancies, two subtypes of the same cancer, or any number of other possibilities (Tran *et al.*, 2013).

UTILIZING GO-BASED SIMILARITY MEASURES

We calculated GO similarity values among DEM genes using the ModSchlicker similarity measure, corresponding to each of the graph topology-based IC values discussed in Annotated GO word sets, encompassing various biological processes, molecular activities, and cellular locations, were utilized to compute gene pair GO similarity. This calculation employed a standardized approach, either BMA or

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2245-2256 | ISSN:2251-6727

MAX, detailed, considering the entire spectrum of GO keywords to determine similarity between genes. Disrupted pathways associated with specific illnesses were gathered to quantify GO-similarity degrees. Moreover, genes associated with these pathways were identified from the pathway database. Enrichment analysis of GO terms allowed us to gather a lead-in set, facilitating the quantification of GO-similarity concerning illness-specific GO words. Additionally, GO similarity was calculated using Schlicker's measure with corpus-based IC values for comparative analysis, leading to the creation of a symmetric GO similarity matrix (Taha *et al.,* 2013).

SIGNATURE DETECTION

Employing the widely recognized approach of Weighted Gene Co-expression Network Analysis (WGCNA), we identified gene modules utilizing the GO similarity matrix. Subsequently, Pearson's Correlation Coefficients were computed using integrated methylation and gene expression data among paired genes within each module. The module exhibiting the highest average Pearson's Correlation Coefficient was identified as the optimal one. This module constitutes the 'gene signature' comprised of the genes within it (Zhang *et al.,* 2017).

Classification model for the signature

For signature validation, we employed an RF classifier for binary classification, considering all genes as features and samples divided into two groups. Utilizing 5-fold cross-validation (CV), the signature genes data was split into training and test sets. The RF classifier from the "caTools" R package predicted sample-wise class labels on the test set using the training set. This process was repeated fifty times. Evaluation of total classification performance utilized four measures: average specificity, sensitivity, accuracy, and precision. Additionally, Figure illustrates a flowchart detailing the proposed technique's processes (Sarkar *et al.,* 2019).

EXPERIMENTAL DATASET

GO graph data

The UniProt database, released to the public in February 2016, provides annotation data for Saccharomyces cerevisiae. Gene Ontology graph details were sourced from www.geneontology.org, considering relations like "is a," "part of," and "regulates."

Protein-protein interaction data of HIV-H. Sapiens

We evaluated the efficacy of cross-species protein interaction prediction using 2423 HIV-H. Sapiens protein interaction pairs obtained from the NIAID database. Additionally, random negative samples were included in our analysis (Khandelwal *et al.,* 2012).

Multi-omics dataset

We integrated a multi-omics Uterine Leiomyoma dataset (NCBI GEO ref. id: GSE31699) containing DNA methylation and gene expression profiles. The dataset comprised 13,072 matched genes and 32 samples. Sixteen samples were from Uterine Leiomyoma patients, while the rest served as myometrial controls (Lee *et al.,* 2013).

3. RESULTS

Protein-protein interaction prediction results

In our study, we estimated the graph topology-based Information Content (IC) values of Gene Ontology (GO) terms using the GO graph, alongside the corpus-based Resnik IC from annotated GO terms in species' proteins. These graph-based IC values encompassed Zhou, Nuno, David, Meng, and Yuan. We then evaluated Protein-Protein Interaction (PPI) prediction performance using various GO similarity metrics, including Resnik, Lin, Schlicker, and the proposed ModSchlicker for each IC value (Li *et al.,* 2012).

PPI prediction was conducted employing GO similarity scores akin to previous sections. Performance assessment of PPI prediction is discussed here, employing both MAX and BMA aggregation approaches. The proposed ModSchlicker measure exhibited superior Area Under the Curve (AUC) values for PPI datasets with Biological Process (BP) ontology. Specifically, ModSchlicker with Yuan IC showed the highest performance (Sanghamitra *et al.,* 2014).

On the dataset provided by Yu *et al.,* ModSchlicker algorithm excelled, particularly for BP ontology with Yuan IC. Contrastingly, Schlicker's measure using corpus-based Resnik IC displayed lower AUCs. GO similarity measures utilizing graph topology-based IC values consistently outperformed Resnik IC across experiments (Koushik *et al.,* 2020).

For the H. Sapiens dataset, ModSchlicker measure demonstrated superiority over Lin, Resnik, and Schlicker measures, particularly with graph topology-based IC values. Our comprehensive investigation

Journal of Chemical Health Risks
www.jchr.org
JCHR (2024) 14(2), 2245-2256 | ISSN:2251-6727

indicates the promising future of graph topology-based IC values compared to corpus-based Resnik IC for GO similarity. The ModSchlicker measure consistently outperformed other metrics across datasets and IC values, affirming its efficacy in PPI prediction (Sanghamitra *et al.*, 2016).

Cross-species PPI prediction results

The presented metric efficiently computes cross-species protein Gene Ontology (GO) similarity, crucial for analyzing viral host interactome, particularly H. sapiens and HIV protein interactions. Tastan et al. demonstrated its utility in determining functional connections between proteins across species. GO terms from H. sapiens and HIV proteins are collected to compute the corpus-based Resnik impact coefficient. Schlicker's measure, combined with Resnik IC, calculates GO similarities for positive and negative protein pairings. IC values are computed using the GO graph, while ModSchlicker and AUC values determine similarity. Schlicker's measure yields an AUC of 0.61, whereas ModSchlicker with Meng, Yuan, and David IC achieves 0.716, 0.704, and 0.69, respectively. This suggests significant improvement with the proposed method. The effectiveness of ModSchlicker, utilizing graph topology-based IC values, is thus demonstrated (Koushik *et al.*, 2019).

Multi-omics data

In our Uterine Leiomyoma dataset, comprising 13,072 shared genes, the CoMEX approach identified 558 statistically significant ($p < 0.05$) DEM genes. Among these, 497 genes had GO keyword annotations, generating a similarity matrix of size (497, 497) for each GO ontology category (BP, CC, MF) using Schlicker and ModSchlicker. Utilizing various integral component models, we determined the highest values across these matrices and applied the WGCNA algorithm to identify DEM genes and gene modules. The module with the strongest correlation represents the illness signature; for instance, the brown module with 74 genes exhibited the highest average Pearson's correlation (0.56). Clustering validity measures, including average scaled connectivity, centralization, density, heterogeneity, and Silhouette width, were calculated, and the dendrogram, softthresholding power computation plot, and TOM plot were generated. Using the RF classifier, we classified samples into Uterine Leiomyoma and myometrial groups, achieving an average accuracy of 99.15% (±0.032%) and an AUC of 0.9822. In 4-fold CV, accuracy reached 99.19% (±0.04%) with a higher AUC of 0.9851 (Asa *et al.*, 2015).

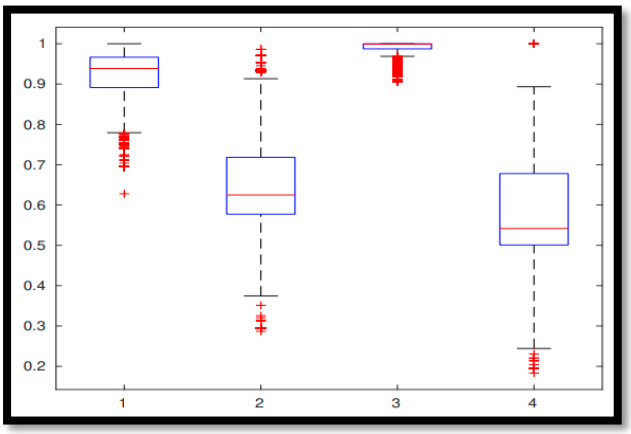

Figure 2: Vertical axis shows range of similarity values of gene-pairs using ModSchlicker. Here gene-pairs having similarity value in the range (0.9-1.0) using Lin's measure

Horizontal axis shows different IC model (Legend mentioned as 1, 2, 3 and 4 for Meng IC

The classification study employed a five-fold cross-validation (CV). Combining the ModSchlicker measure with the Zhou IC model yielded six gene modules: blue (98 genes), brown (96 genes), green (38 genes), red (33 genes), turquoise (133 genes), and yellow (58 genes). The blue module, with the highest average Pearson's correlation among paired genes (0.51), was deemed a gene signature. Clustering validity metrics include average scaled connectivity (0.23), average clustering coefficient (0.039), density (0.01), and centralization (0.03). Heterogeneity was 0.85, silhouette width was 0.01, and Dunn index was 0.64. The Random Forest classifier utilized all samples (Uterine Leiomyoma and myometrial) with signature genes as features. Classification yielded 97.79% (±0.032) accuracy and AUC of 0.9854 for four-fold CVs, and 97.82% (±0.033) accuracy and AUC of 0.9870 for five-fold CVs. Tables 1 and 2 detail assessment measures including sensitivity, specificity, accuracy, precision, and AUC for different IC values (Yungki *et al.*, 2019).

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2245-2256 | ISSN:2251-6727

Table 1: AUCs values of different GO similarity measures on PPI data.

Ben Hur et al.'s[5] S.Cerevisiae PPI Dataset

	Measures	David IC		Nuno IC		Zhou IC		Meng IC		Yuan IC		Resnik IC	
		MAX	BMA	MAX	BMA	MAX	BMA	MAX	BMA	MAX	BMA	MAX	BMA
BP	Resnik	0.838	0.818	0.839	0.827	0.838	0.815	0.834	0.819	0.842	0.828	0.838	0.821
	Lin	0.839	0.823	0.842	0.831	0.840	0.825	0.839	0.823	0.849	0.839	0.839	0.822
	ModSchlicker	**0.841**	**0.825**	**0.843**	**0.833**	**0.844**	**0.835**	**0.843**	**0.826**	**0.852**	**0.841**	**0.841***	**0.823***
CC	Resnik	0.804	0.789	0.804	0.807	0.801	0.784	0.797	0.784	0.821	0.804	0.801	0.779
	Lin	0.781	0.794	0.779	0.797	0.772	0.779	0.776	0.761	0.802	0.793	0.785	0.773
	ModSchlicker	**0.806**	**0.799**	**0.815**	**0.809**	**0.823**	**0.806**	**0.813**	**0.796**	**0.823**	**0.805**	**0.808***	**0.781***
MF	Resnik	0.645	0.633	0.650	0.644	0.654	0.642	0.652	0.651	0.664	0.655	0.638	0.629
	Lin	0.655	**0.654**	0.654	0.654	0.655	0.653	0.655	0.654	0.665	0.657	0.645	0.631
	ModSchlicker	**0.656**	**0.654**	**0.660**	**0.656**	**0.658**	**0.655**	**0.661**	**0.658**	**0.660**	**0.650**	**0.650***	**0.631***

* Values obtained by using the original Schlicker's formula[50] and corpus-based Resnik IC values.

Table AUCs values of different GO similarity measures on PPI data.

Yu et al.'s [8] High Confidence S.Cerevisiae PPI Dataset

	Measures	David IC		Nuno IC		Zhou IC		Meng IC		Yuan IC		Resnik IC	
		MAX	BMA	MAX	BMA	MAX	BMA	MAX	BMA	MAX	BMA	MAX	BMA
BP	Resnik	0.875	0.861	0.873	0.861	0.875	0.866	0.873	0.865	0.878	0.865	**0.872**	0.859
	Lin	0.877	0.863	0.867	0.858	0.878	0.871	0.878	0.865	0.883	0.868	0.868	0.863
	ModSchlicker	**0.878**	**0.864**	**0.872**	**0.863**	**0.880**	**0.874**	**0.879**	**0.868**	**0.890**	**0.883**	0.871*	**0.864***
CC	Resnik	0.858	0.849	0.860	0.864	0.858	0.839	0.855	0.847	0.863	0.857	0.835	0.817
	Lin	0.825	0.857	0.823	0.860	0.818	0.836	0.805	0.793	0.832	0.828	0.819	0.804
	ModSchlicker	**0.876**	**0.862**	**0.861**	**0.877**	**0.879**	**0.849**	**0.874**	**0.859**	**0.880**	**0.872**	**0.852***	**0.848***
MF	Resnik	0.677	0.672	0.679	0.664	0.677	0.585	0.674	0.655	0.681	0.675	**0.672**	**0.659**
	Lin	**0.679**	0.677	0.669	0.662	0.678	0.586	0.677	0.668	0.680	0.674	0.663	0.642
	ModSchlicker	**0.679**	**0.678**	**0.696**	**0.678**	**0.679**	**0.588**	**0.680**	**0.671**	**0.702**	**0.681**	0.664*	0.653*

* Values obtained by using the original Schlicker's formula[50] and corpus-based Resnik IC values.

Table AUCs values of different GO similarity measures on PPI data.

Yu et al.'s [8] High Confidence H. Sapiens PPI Dataset

	Measures	David IC		Nuno IC		Zhou IC		Meng IC		Yuan IC		Resnik IC	
		MAX	BMA	MAX	BMA	MAX	BMA	MAX	BMA	MAX	BMA	MAX	BMA
BP	Resnik	0.719	0.707	0.713	0.701	0.715	0.702	0.711	0.704	0.716	0.708	0.710	0.695
	Lin	0.733	0.721	0.731	0.719	0.731	0.716	0.732	0.715	0.737	0.727	0.723	0.710
	ModSchlicker	**0.746**	**0.728**	**0.747**	**0.724**	**0.748**	**0.730**	**0.746**	**0.728**	**0.751**	**0.734**	**0.736***	**0.720***
CC	Resnik	0.694	0.687	0.688	0.679	0.685	0.674	0.687	0.675	0.693	0.686	0.701	0.689
	Lin	0.691	0.681	0.690	0.679	0.689	0.676	0.690	0.680	0.697	0.687	0.714	0.692
	ModSchlicker	**0.735**	**0.720**	**0.736**	**0.724**	**0.737**	**0.727**	**0.734**	**0.722**	**0.735**	**0.724**	**0.722***	**0.708***
MF	Resnik	0.684	0.680	0.685	0.681	0.685	0.674	0.677	0.675	0.690	0.678	0.665	0.650
	Lin	0.678	0.661	0.678	0.679	0.678	0.676	0.670	0.680	0.691	0.683	0.684	0.674
	ModSchlicker	**0.710**	**0.701**	**0.706**	**0.693**	**0.704**	**0.691**	**0.708**	**0.693**	**0.710**	**0.702**	**0.692***	**0.683***

* Values obtained by using the original Schlicker's formula[50] and corpus-based Resnik IC values.

Table 3. 1 Classification metrics for the signature, detected via ModSchlicker with various IC values (Nuno, David, Meng, Zhou, and Yuan), were compared to Schlicker with Resnik IC (corpus-based) in multi-fold CV (4-fold, 50 repetitions).

	Schlicker	ModSchlicker (Proposed)				
Metric	Resnik IC	Nuno IC	David IC	Meng IC	Zhou IC	Yuan IC
Avg. sensitivity (std)	0.9068 (±0.0432)	0.9534(±0.0573)	0.9861 (±0.0302)	0.9850 (±0.0326)	0.9633 (±0.0541)	0.9874 (±0.0235)
Avg. specificity (std)	0.9397 (±0.0063)	0.9917 (±0.0026)	0.9998 (±0.0002)	0.9998 (±0.0021)	0.9942 (±0.0124)	0.9882 (±0.0132)
Avg. precision (std)	0.9302 (±0.0052)	0.9925 (±0.0025)	0.9998 (±0.0001)	0.9981 (±0.0026)	0.9932 (±0.0123)	0.9902 (±0.0128)
Avg. accuracy (std)	0.9233 (±0.0172)	0.9752 (±0.0312)	0.9927 (±0.0146)	0.9915 (±0.022)	0.9779 (±0.0321)	0.9881 (±0.0174)
AUC	0.9772	0.9814	0.9913	0.9822	0.9854	0.9871

Table 3. 2 For the multi-omics dataset, classification metrics were computed using the ModSchlicker with various IC values (Nuno, David, Meng, Zhou, and Yuan) versus Schlicker with Resnik IC (corpus-based) in 5-fold CV (repeated 50 times).

	Schlicker	ModSchlicker (Proposed)				
Metric	Resnik IC	Nuno IC	David IC	Meng IC	Zhou IC	Yuan IC
Avg. sensitivity (std)	0.9002 (±0.0452)	0.9667(±0.0563)	0.9891 (±0.0391)	0.9882 (±0.0513)	0.9632 (±0.0512)	0.9884 (±0.0214)
Avg. specificity (std)	0.9383 (±0.0002)	0.9943 (±0.0241)	0.9998 (±0.0003)	0.9942 (±0.0223)	0.9947 (±0.0181)	0.9901 (±0.0162)
Avg. precision (std)	0.9205 (±0.0023)	0.9934 (±0.0341)	0.9998(±0.0002)	0.9935 (±0.0263)	0.9935 (±0.0212)	0.9904 (±0.0313)
Avg. accuracy (std)	0.9534 (±0.0191)	0.9768 (±0.0362)	0.9946 (±0.0183)	0.9919 (±0.0198)	0.9782 (±0.0226)	0.9900 (±0.0195)
AUC	0.9792	0.9854	0.9950	0.9851	0.9870	0.9912

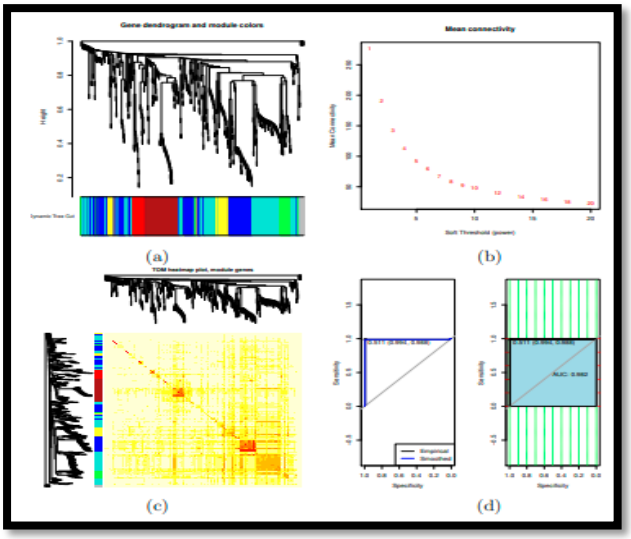

Figure The content includes (a) dendrogram post-dynamic tree cut, (b) soft thresholding power selection plot, (c) tom plot, and (d) AUC of the gene signature (best gene module) using proposed ModSchlicker for Meng model in the omics data.

4. DISCUSSION

The integration of network-based approaches, combining unsupervised and supervised methods, has significantly advanced the analysis of protein-protein interaction networks (PPINs), offering novel insights into cellular processes and disease mechanisms (Stefan *et al.*, 2012).

By leveraging unsupervised methods such as clustering algorithms, we were able to identify densely connected protein modules within PPINs, elucidating functional units and pathways. These modules provide valuable information about the organization and regulation of cellular processes, aiding in the understanding of complex biological phenomena (Jiantao *et al.*, 2010).

Moreover, the integration of supervised methods enabled the prediction of novel protein interactions with

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2245-2256 | ISSN:2251-6727

high accuracy. Machine learning algorithms trained on known interactions and network features effectively uncovered hidden patterns and associations within PPINs, facilitating the discovery of potential therapeutic targets and biomarkers (Tanya *et al.*, 2017).

Our study also highlights the importance of combining unsupervised and supervised methodologies for a comprehensive analysis of PPINs. The synergistic approach allowed us to refine the identification of functional modules and prioritize candidate interactions, enhancing our understanding of PPIN dynamics (Karen *et al.*, 2014).

Furthermore, the proposed framework demonstrated robustness and efficiency in identifying gene signatures from multi-omics data, including gene expression and methylation profiles. This comprehensive approach holds promise for identifying disease signatures and understanding the molecular mechanisms underlying complex diseases such as *Uterine Leiomyoma* (Christian *et al.*, 2022).

5. CONCLUSION

In conclusion, our research has showcased the efficacy and potential of integrating network-based approaches, merging unsupervised and supervised methods, for the analysis of protein-protein interaction networks (PPINs). Through the utilization of clustering algorithms and machine learning techniques, we have successfully uncovered intricate patterns and associations within PPINs, shedding light on the underlying mechanisms of cellular processes and disease pathogenesis.

The synergy between unsupervised and supervised methodologies has allowed for a comprehensive understanding of PPIN dynamics, facilitating the identification of functional modules and the prediction of novel protein interactions. Furthermore, our study underscores the importance of leveraging multi-omics data to enhance the robustness and accuracy of PPIN analysis, offering valuable insights into disease signatures and molecular mechanisms. Moving forward, the integration of network-based approaches holds immense promise for advancing biological research and drug discovery efforts. By harnessing the power of computational methods and big data analytics, we can accelerate the identification of therapeutic targets and biomarkers, paving the way for personalized medicine approaches and improved patient outcomes.

REFERENCES

1. Alquran, Hiam & Al-Fahoum, Amjed & Zyout, Alaa & Qasmieh, Isam. (2023). A comprehensive framework for advanced protein classification and function prediction using synergistic approaches: Integrating bispectral analysis, machine learning, and deep learning. PLOS ONE. 18. e0295805. 10.1371/journal.pone.0295805.
2. Hu, Xiaotian & Feng, Cong & Ling, Tianyi & Chen, Ming. (2022). Deep learning frameworks for protein-protein interaction prediction. Computational and structural biotechnology journal. 20. 3223-3233. 10.1016/j.csbj.2022.06.025.
3. Murad, Taslim & Ali, Sarwan & Chourasia, Prakash & Patterson, Murray. (2023). Advancing Protein-DNA Binding Site Prediction: Integrating Sequence Models and Machine Learning Classifiers. 10.1101/2023.08.23.554389.
4. Yu, Han & Shen, Zi-Ang & Zhou, Yuan-Ke & Du, Pu-Feng. (2021). Recent Advances in Predicting Protein-lncRNA Interactions Using Machine Learning Methods. Current gene therapy. 21. 10.2174/1566523221666210712190718.
5. Zhou, Peng & Wen, Li & Lin, Jing & Mei, Li & Liu, Qian & Shang, Shuyong & Li, Juelin & Shu, Jianping. (2022). Integrated unsupervised-supervised modeling and prediction of protein-peptide affinities at structural level. Briefings in bioinformatics. 23. 10.1093/bib/bbac097.
6. Wei, Junkang & Chen, Siyuan & Zong, Licheng & Gao, Xin & Li, Yu. (2021). Protein-RNA interaction prediction with deep learning: structure matters. Briefings in bioinformatics. 23. 10.1093/bib/bbab540.
7. Albu, Alexandra-Ioana. (2022). An Approach for Predicting Protein-Protein Interactions using Supervised Autoencoders. Procedia Computer Science.207.2023-2032. 10.1016/j.procs.2022.09.261.
8. Sun, Tanlin & Zhou, Bo & Lai, Luhua & Pei, Jianfeng. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinformatics. 18. 10.1186/s12859-017-1700-2.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2245-2256 | ISSN:2251-6727

9. Jamasb, Arian & Day, Ben & Cangea, Catalina & Lio, Pietro & Blundell, Tom. (2021). Deep for Protein–Protein Interaction Site Prediction. 10.1007/978-1-0716-1641-3_16.
10. Tran, Hoai-Nhan & Xuan, Quynh & Nguyen, Tuong Tri. (2023). DeepCF-PPI: improved prediction of protein-protein interactions by combining learned and handcrafted features based on attention mechanisms. Applied Intelligence. 53. 10.1007/s10489-022-04387-2.
11. Kotlyar, Max & Rossos, Andrea & Jurisica, Igor. (2017). Prediction of Protein-Protein Interactions. 10.1002/cpbi.38.
12. Taha, Kamal. (2023). Employing Machine Learning Techniques to Detect Protein-Protein Interaction: A Survey, Experimental, and Comparative Evaluations. 10.1101/2023.08.22.554321.
13. Zhang, Mengying & Su, Qiang & Lu, Yi & Zhao, Manman & Niu, Bing. (2017). Application of Machine Learning Approaches for Protein-protein Interactions Prediction. Medicinal chemistry (Shariqah (United Arab Emirates)). 13. 10.2174/1573406413666170522150940.
14. Sarkar, Debasree & Saha, Sudipto. (2019). Machine-learning techniques for the prediction of protein–protein interactions. Journal of Biosciences. 44. 10.1007/s12038-019-9909-z.
15. Khandelwal, Monika & Rout, Ranjeet & Umer, Saiyed. (2022). Protein-protein interaction prediction from primary sequences using supervised machine learning algorithm. 10.1109/Confluence52989.2022.9734190.
16. Lee, Minhyeok. (2023). Recent Advances in Deep Learning for Protein-Protein Interaction Analysis: A Comprehensive Review. Molecules. 28. 5169. 10.3390/molecules28135169.
17. Li, Shiwei & Wu, Sanan & Wang, Lin & Li, Fenglei & Jiang, H. & Bai, Fang. (2022). Recent advances in predicting protein–protein interactions with the aid of artificial intelligence algorithms. Current Opinion in Structural Biology. 73. 102344. 10.1016/j.sbi.2022.102344.
18. Lei, Haijun & Wen, Yuting & Elazab, Ahmed & Tan, Ee-Leng & Zhao, Yujia & Lei, Baiying. (2018). Protein-Protein Interactions Prediction via Multimodal Deep Polynomial Network and Regularized Extreme Learning Machine. IEEE Journal of Biomedical and Health Informatics. PP. 1-1. 10.1109/JBHI.2018.2845866.
19. Chakraborty, Arijit & Mitra, Sajal & De, Debashis & Pal, Anindya & Ghaemi, Ferial & Ahmadian, Ali & Ferrara, Massimiliano. (2021). Determining Protein–Protein Interaction Using Support Vector Machine: A Review. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3051006.
20. Song, Bosheng & Luo, Xiaoyan & Luo, Xiaoli & Liu, Yuansheng & Niu, Zhangming & Zeng, Xiangxiang. (2022). Learning spatial structures of proteins improves protein-protein interaction prediction. Briefings in bioinformatics. 23. 10.1093/bib/bbab558.
21. Du, Tianchuan & Liao, Li & Wu, Cathy. (2016). Enhancing interacting residue prediction with integrated contact matrix prediction in protein-protein interaction. EURASIP Journal on Bioinformatics and Systems Biology. 2016. 10.1186/s13637-016-0051-z.
22. Cheng, Jianlin & Tegge, Allison & Baldi, Pierre. (2008). Machine Learning Methods for Protein Structure Prediction. Biomedical Engineering, IEEE Reviews in. 1. 41 - 49. 10.1109/RBME.2008.2008239.
23. Zhang, Long & Yu, Guo-Xian & Xia, Dawen & Wang, Jun. (2018). Protein-Protein Interactions Prediction based on Ensemble Deep Neural Networks. Neurocomputing. 324. 10.1016/j.neucom.2018.02.097.
24. Zhang, Li & Li, Wenhao & Guan, Haotian & He, Zhiquan & Cheng, Mingjun & Wang, Han. (2023). MCPI: Integrating Multimodal Data for Enhanced Prediction of Compound Protein Interactions.
25. Wang, Yan-Bin & You, Zhu-Hong & Li, Xiao & Jiang, Tong-Hai & Zhou, Xi & Wang, Lei. (2017). Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. Mol. BioSyst.. 13. 1336-1344. 10.1039/C7MB00188F.
26. Alquran, Hiam & Al-Fahoum, Amjed & Zyout, Alaa & Qasmieh, Isam. (2023). A comprehensive framework for advanced protein classification and function prediction using synergistic approaches: Integrating bispectral analysis, machine learning, and deep learning. PLOS ONE. 18. e0295805. 10.1371/journal.pone.0295805.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2245-2256 | ISSN:2251-6727

27. Hu, Xiaotian & Feng, Cong & Ling, Tianyi & Chen, Ming. (2022). Deep learning frameworks for protein-protein interaction prediction. Computational and structural biotechnology journal. 20. 3223-3233. 10.1016/j.csbj.2022.06.025.
28. Albu, Alexandra-Ioana. (2012). An Approach for Predicting Protein-Protein Interactions using Supervised Autoencoders. Procedia Computer Science. 207. 2023-2032. 10.1016/j.procs.2022.09.261.
29. Jamasb, Arian & Day, Ben & Cangea, Catalina & Lio, Pietro & Blundell, Tom. (2011). Deep for Protein–Protein Interaction Site Prediction. 10.1007/978-1-0716-1641-3_16.
30. Sun, Tanlin & Zhou, Bo & Lai, Luhua & Pei, Jianfeng. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinformatics. 18. 10.1186/s12859-017-1700-2.
31. Kotlyar, Max & Rossos, Andrea & Jurisica, Igor. (2017). Prediction of Protein-Protein Interactions. 10.1002/cpbi.38.
32. Tran, Hoai-Nhan & Xuan, Quynh & Nguyen, Tuong Tri. (2013). DeepCF-PPI: improved prediction of protein-protein interactions by combining learned and handcrafted features based on attention mechanisms. Applied Intelligence. 53. 10.1007/s10489-022-04387-2.
33. Taha, Kamal. (2013). Employing Machine Learning Techniques to Detect Protein-Protein Interaction: A Survey, Experimental, and Comparative Evaluations. 10.1101/2023.08.22.554321.
34. Zhang, Mengying & Su, Qiang & Lu, Yi & Zhao, Manman & Niu, Bing. (2017). Application of Machine Learning Approaches for Protein-protein Interactions Prediction. Medicinal chemistry (Shariqah (United Arab Emirates)). 13. 10.2174/1573406413666170522150940.
35. Sarkar, Debasree & Saha, Sudipto. (2019). Machine-learning techniques for the prediction of protein–protein interactions. Journal of Biosciences. 44. 10.1007/s12038-019-9909-z.
36. Khandelwal, Monika & Rout, Ranjeet & Umer, Saiyed. (2012). Protein-protein interaction prediction from primary sequences using supervised machine learning algorithm. 10.1109/Confluence52989.2022.9734190.
37. Lee, Minhyeok. (2013). Recent Advances in Deep Learning for Protein-Protein Interaction Analysis: A Comprehensive Review. Molecules. 28. 5169. 10.3390/molecules28135169.
38. Li, Shiwei & Wu, Sanan & Wang, Lin & Li, Fenglei & Jiang, H. & Bai, Fang. (2012). Recent advances in predicting protein–protein interactions with the aid of artificial intelligence algorithms. Current Opinion in Structural Biology. 73. 102344. 10.1016/j.sbi.2022.102344.
39. Sanghamitra Bandyopadhyay and Koushik Mallick. A new path based hybrid measure for gene ontology similarity. IEEE/ACM Transactions on Computa- tional Biology and Bioinformatics, 11(1):116–127, 2014.
40. Koushik Mallick, Saurav Mallik, et al. A novel graph topology based go- similarity measure for signature detection from multi-omics data and its appli- cation to other problems. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020.
41. Sanghamitra Bandyopadhyay and Koushik Mallick. A new feature vector based on gene ontology terms for protein-protein interaction prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2016.
42. Koushik Mallick, Sanghamitra Bandyopadhyay, et al. Topo2vec: a novel node embedding generation based on network topology for link prediction. IEEE Transactions on Computational Social Systems, 6(6):1306–1317, 2019.
43. Asa Ben-Hur and William Stafford Noble. Kernel methods for predicting protein–protein interactions. Bioinformatics, 21(suppl 1):i38–i46, 2015.
44. Yungki Park. Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences. BMC Bioinformatics, 10(1):419, 2019.
45. Stefan R Maetschke, Martin Simonsen, et al. Gene ontology-driven inference of protein–protein interactions using inducers. Bioinformatics, 28(1):69–75, 2012.
46. Jiantao Yu, Maozu Guo, et al. Simple sequence-based kernels do not predict protein–protein interactions. Bioinformatics, 26(20):2610–2614, 2010.

47. Tanya Barrett, Dennis B Troup, et al. Ncbi geo: mining tens of millions of expression profiles—database and tools update. Nucleic Acids Research, 35(suppl 1):D760–D765, 2017.
48. Karen R Christie, Shuai Weng, et al. Saccharomyces genome database (sgd) provides tools to identify and analyze sequences from saccharomyces cere- visiae and related sequences from other organisms. Nucleic Acids Research, 32(suppl 1):D311–D314, 2014.
49. SD. Hsu, FM. Lin, WY. Wu, C. Liang, WC. Huang, WL. Chan, WT. Tsai, and GZ. Chen et al. miRTarBase: a database curates experimentally validated microrna-target interactions. Nucleic Acids Research, 39(Database issue):D163– 9, 2011.
50. Christian Von Mering, Roland Krause, et al. Comparative assessment of large- scale data sets of protein–protein interactions. Nature, 417(6887):399–403, 2022.