# Summary

**Github Link:** [**https://github.com/tengann/IAV\_PPI\_Graph\_Embedding\_Review**](https://github.com/tengann/IAV_PPI_Graph_Embedding_Review)

* **Review for Human-IAV host-pathogen interactions,**

**based on 4 strains of interest**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Strain** | **Subtype** | **Taxonomy ID** | **Abbreviation** | **Reason** |
| A/Puerto Rico/8/1934 | H1N1 | 211044 | PR8 | Commonly used in lab experiments |
| A/California/04/2009 | H1N1 | 641501 | CA04 | 2009 Pandemic strain |
| A/California/07/2009 | H1N1 | 641809 | CA07 | 2009 Pandemic strain  (Contains an “unusual” amino acid ‘X’ in its NP segment) |
| A/Aichi/2/1968 | H3N2 | 387139 | Aichi | H3N2 subtypes of IAV have cause seasonal epidemics since 1968 \cite{RN13} |

## Datasets

1. PPIDomainMiner (PPIDM) \cite{RN4} Construct dataset based on 4 strains of interest

Dataset of 84, 552 non-redundant DDIs

Domain interactions, interactions do not differ between strains

Domain Analysis (Superfamily VS Pfam)??

HVPPI

\cite{RN5}

* Unsupervised sequence embedding technique (doc2vec) to represent protein sequences as rich feature vectors of low dimensionality.

Training a Random Forest (RF) classifier through a training dataset that covers known PPIs between human and all viruses

* Host-pathogen PPI data from the **Host-Pathogen Interaction Database (HPIDB)** \cite{RN17, 16} V3.0
  + Contains manually curated host-pathogen interactions and also integrates corresponding molecular interactions from other public protein interaction databases
* Built based on an unbalanced human host-virus PPI dataset with positive-to-negative ratio 1:10

Performance of HVPPI (Barman et al.’s dataset)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accuracy (%) | Sensitivity (%) | Specificity (%) | F1-Score (%) | ROC-AUC |
| 79.17 | 81.85 | 76.45 | 79.79 | 0.871 |

[Show comparison of the distribution of interacting pairs across segments]

## Models

1. **Feedforward Neural Network (FNN)**

Sequential model (i.e., plain stack of layers where each layer has exactly one input tensor and one output tensor) (keras)

Same configuration as classifier used in S-VGAE \cite{RN22}

3-layers: input + 2 hidden, with 128, 64 and 32 neurons in each layer

Uses dropout technique, important trick widely used in deep learning \cite{RN12}

Train for 200 epochs using Adam algorithm with a batch size of 128

1. **Skip-GNN (pytorch) \cite{RN6}**

**FNN**

* First and simplest type of artificial neural network devised \cite{RN44}
* Connections between nodes do not form a cycles or loops in the network. Information moves in only one direction. \cite{RN45}

**Skip-GNN**

* Skip similarity (similarity in second-order proximity interactions, from second-hop neighbors) by encouraging the GNN model to embed skipped nodes close together in the embedding space
* Construct skip graph from second-hop neighbours

Element-wise matrix multiplication

A­s = sign(AAT), where sign(AAT) = 1 if AAT > 0, and sign(AAT) = 0 otherwise, where AAT= 1 indicates there exists a skipped node between nodei, j

* Iterative fusion scheme (instead of simple concatenation) to allow the skip graph and original graph to learn from each other for better integration. Model automically learn how to balance between direct similarity and skip similarity in the final embedding.

Skip-GNN and S-VGAE investigate different aspects

Skip-GNN (similarity in second-order interactions, from second-hop neighbors)

* Not quite affected by chosen embedding algorithm

S-VGAE (signed adjacency matrix) – set different signs to reinforce existing observed interactions (strengthen the negative impact of highly negative interactions)

Graph embedding methods

**[Only making use of the graph network topology]**

|  |  |
| --- | --- |
| Method Category | Method Name |
| Matrix factorization-based | GraRep \cite{RN18} |
| Random walk-based | Deepwalk \cite{RN21}  node2vec \cite{RN7}, node2vec+  struc2vec \cite{RN20}  ripple2vec |
| Neural network-based | LINE \cite{RN11}  SDNE \cite{RN10}  VAE \cite{RN46} |

Used to calculate *static* graph embeddings (i.e., computed for a graph at a fixed time period)

<https://towardsdatascience.com/lets-talk-about-graph-machine-learning-in-biomedical-networks-8a84139e970b>

Codes:

* Deepwalk, node2vec, struc2vec

<https://github.com/shenweichen/GraphEmbedding> (Repo used in Skip-GNN paper)

* LINE, SDNE, GraRep, VAE

<https://github.com/xiangyue9607/BioNEV>

* nSNE \cite{RN50}

<https://github.com/wzsong17/Signed-Network-Embedding>

* Other algorithms
* **Node2Vec+ \cite{RN41}**,implemented as part of **pecanpy** \cite{RN40}

<https://github.com/krishnanlab/PecanPy>

SparseOTF vs DenseOTF

Natural extension of node2vec and handles weighted graph more effectively using a noise threshold – Average edge weights

The node2vec+ out edges is determined by considering the edge weights

    connecting node2 (the potential next state) to the previous state. Unlinke

    node2vec, which only considers neighbors of current state that are not

    neighbors of the previous state, node2vec+ also considers neighbors of

    the previous state as out edges if the edge weight is below average.

* **ripple2vec \cite{RN37}** – Recent methodology

Node Embedding with Ripple Distance of Structures

<https://github.com/hitSongXiao/ripple2vec>

Implemented in the framework of struc2vec by adapting ripple distance to define context graphs

Helps map dis(similar) nodes to (far) near vectors

* Additional (can consider)

GCN VS Skip-GNN VS GAT

* **MixHop –** Higher order convolutional layer

Designed to capture higher order graph structure through skip connections and higher order adjacency matrix

Motivated by general network model and does not propose a solution for the specific challenge of 2-hop skip similarity in biomedical network \cite{RN6}

<https://github.com/benedekrozemberczki/MixHop-and-N-GCN>

[Skip-GNN used MixHop for comparison]

* **Edge2vec \cite{RN38}**

(To investigate the effect of edge weights)

<https://github.com/RoyZhengGao/edge2vec>

Designed for heterogeneous information networks, containing diverse biological entities (e.g., genes, proteins, drugs, phenotypes)

* **Attre2vec**

https://github.com/attre2vec/attre2vec

## Contributions: Points explored/to explore

1. **“Unique” dataset**

Curated using HVPPI web server, related to a specific problem of IAV-Human protein-protein interaction

Interaction probability used as edge weights

Also tested on an experimentally verified dataset

(i.e., real-world situation)

1. **Computation of edge embeddings (representation for pairs of nodes)**

Compare between **Concatenation, Sum, Average and Hadamard** Product

(sum and hadamard product – keep original dimensions)

[Not much difference, will just keep to using hadamard product] – Experimentally verified dataset?

*S-VGAE* – hstack (equivalent to concatenation)

*SkipGNN* – “concatenation consistently yield the best performance across different types of networks”

[Using optimized hyper-parameters]

(Briefly explored on Feedforward Neural Network,

From my results, hadamard product tends to achieve better performance instead of concatenation)

1. **With VS without edge weights**

Weights on edges are useful for graph representational learning \cite{RN33}

However, PPI networks are modelled as undirected unweighted graphs

(Use PPIDM - statistical significance (p-value) and

HVPPI - interaction probability of query protein pair as weights)

Using HVPPI prediction score to learn node features for **non-edges**

(Assigned weights to non-edges instead of simply **zero**)

*No experimentally verified data for non-interacting Human-IAV protein pairs. Therefore, may still have a slight chance of interaction?*

**Currently, HVPPI probability scores are helpful in choosing edges for constructing the dataset but not leveraged when learning graph representations**

**[Ref Github: results/Choosing\_HVPPI\_negative\_threshold.xlsx]**

* Reduce noise \cite{RN3}

1. **With VS without hyper-parameter tuning** (training of graph embeddings)

Reference to paper \cite{RN3}

Compare min and max results to show how much results can differ

Paper \cite{RN3} did not mention to optimize number and length of walks for **node2vec**

However, author of node2vec paper \cite{RN7} mentioned that both these parameters have a relatively high impact on the performance of the method.

* **HVPPI predictions dataset** – update results for node2vec (num walks & walk length)
* **Experimentally verified dataset** – Top algorithm for both FNN (SDNE) & Skip-GNN [Choose the appropriate hyper-parameters] – Can achieve better performance than protein embedding

## 4 classes

Train pos/neg, test pos

Human-human interactions (1 class) – which methods able to differentiate (?), resulting in better performance

1. **Comparison to results obtained with protein embedding**

Can only be tested on FNN, as Skip-GNN is designed to only leverage graph network topology information

Methods mentioned in DeepPPI \cite{RN24} paper

* + Protein embedding methods, iFeature \cite{RN42}
* Amino acid composition (AAC)

*Fraction of each amino acid type within a protein*

* Dipeptide Composition (DPC) – How to justify performance?

*Fraction of amino acids and their local order within a protein*

UMAP – dimensionality reduction

High-dimensional features can potentially cause over fitting and increase of redundant information. Machine learning models trained using high-dimensional initial features often perform poorly in practice \cite{RN42}

* Composition, Transition and Distribution (C/T/D)

*Amino acids are divided into three classes according to 24 attributes:*

*(13 features \* 3 classes = 39 features)*

Different classification of hydrophobicity (Amino acid index (AAindex) database) 🡪 7 different accession numbers

PRAM90010 Hydrophobicity

ARGP820101 Hydrophobicityindex

ZIMJ68010 Hydrophobicity (Isoelectric point)

PONP930101 (Hydrophobicity scales)

CASG920101 (Hydrophobicity scale from native protein structures)

ENGD860101 Hydrophobicityindex (Solvent free energy)

FASG890101 Hydrophobicityindex

6 other attributes 🡪 Normalized van der Waals volume, Polarity, Polarizability, Charge, Secondary structure, Solvent accessibility

* Quasi-Sequence-Order Descriptors (QSOrder)

*Represents amino acid distribution patterns of a specific physicochemical property along protein or peptide sequence*

*Derived from Schneider-Wrede physicochemical distance matrix and Grantham chemical distance matrix*

* Amphiphilic Pseudoamino Acid Composition (APAAC)

*Hydrophobicity and hydrophilicity*

* Conjoint Triad (CT)

*All amino acids are clustered/classified into seven categories according to their dipole and side chain volumes*

* Normalized Moreau-Broto Autocorrelation (NMBroto) \cite{RN49}

*Seven physicochemical properties*

*Hydrophobicity, hydrophilicity, side chain volume, polarity, polarizability, solvent-accessible surface area and side chain net charge index*

1. **Protein embedding – feature selection**

Rank feature importance using Chi-square test

1. **Performance on unbalanced dataset**

**Real-world application**

(i.e., different proportions of positive:negative samples) \cite{RN43} (1:3 and 1:5)

Comparison of PR-ROC scores

[Graph VS Protein embeddings trained on FNN] – show reliability of graph embeddings

[Use same hyper-parameters as in balanced dataset]

[---------------------------------------- Additional ----------------------------------------]

1. Analysis on graph constructed by HVPPI

[Network properties? (e.g. degree, clusters)]

1. **Graph embedding (Individual VS Ensemble of methods) (No longer doing)**

Finding of the best performing method

[Combination of Top 3 embedding methods & All 6 methods]

1. **Concatenating graph and protein embedding (No longer doing)**

Reviewing methodology proposed in paper \cite{RN12}

[Best performing graph + best performing protein]

(points 4 & 5 – embeddings retrieved but yet to be tested)

S-VGAE (signed adjacency matrix) vs VAE

# Section 1: Introduction

1. **Why only a specific problem of IAV-Human PPI?**

Most works tested only on benchmark datasets

1. **Significance of review**

There are still limited works that use graph embedding in bioinformatics, where most use either protein sequence or structure instead

* High research output on Graph Machine Learning, especially Graph Neural Networks (GNN) – formulating different tasks on biomedical graphs enables us to tackle some *bottlenecks* of the traditional lab experiments
* Computational link prediction: Has a high-throughput, however, often have high false positive and negative rates

**Requires a “strong” conclusion**

* **Demonstrate that graph topology embedding can achieve competitive results relative to protein sequences, using a specific problem (human-IAV PPI)**
* **Components of graph network topology that is important**

**e.g. Hops? Higher-order proximity?**

First-order proximity: Pairwise proximity (distance?) between vertices (Vi and Vj)

Second-order proximity: Similarity between neighborhood structure (NVi and NVj) of vertices Vi and Vj \cite{RN48}

Skip-GNN: Considers the similarity between neighbors of (NVi and NVj) – “Third(Higher)-order proximity”

* **Do we really need a complex classification model or just high-quality embeddings?**

1. **Pros of using graph embedding**

* Simple and low-computational complexity
* Biological features typically faces two problems: \cite{RN3}

1. May not always be available and can be hard and costly to obtain

Removing biological entities without features via pre-processing usually results in small-scale pruned datasets and thus is not pragmatic and useful in the real setting

1. Biological features & hand-crafted graph features (e.g. node degrees), may not be precise enough to represent or characterize biomedical entities

May fail to help build a robust and accurate model for many applications

\cite{RN31}

* Unusual (non-standard) amino acid residues (e.g. ‘X’) cannot be encoded. (As shown in the case of NP segment of the CA07 strain, interactions with human proteins cannot be predicted)

1. **FNN (relatively simpler model) VS Skip-GNN**

However, for Skip-GNN, results do not vary much between embedding methods.

(Possible reason is prediction is largely dependent on the skip graph instead of the original graph)

SDNE performs well on FNN when α=0 indicates that performance totally determined by the second-order proximity

1. **Past works**

Examples of past works that used protein sequence/structure information to initialize node attributes (i.e., protein sequences/structure information are projected into a homogeneous vector space)

* **HVPPI** \cite{RN5} uses doc2vec, which requires the protein sequence
* **DeepPPI** \cite{RN24} (fuse multiple protein feature extraction methods)??
* **DPPI** \cite{RN25} uses PSI-BLAST, derives a position-specific scoring matrix (PSSM)
* **Struct2Graph** \cite{RN27}– GCN with mutual attention (requires PDB structures)

Protein sequence + Graph network topology

* **S-VGAE** \cite{RN22} - Improved graph representation learning method, to incorporate graph information in PPI networks into PPI prediction. (Encoder-Decoder approach)

Abstract features are based on both sequence information and graph structure

Uses sequence representations by the conjoint triad (CT) method as input features of each node

Setting weights to edges using signed adjacency matrix, instead of only assigning 0 and 1 to the negative and positive group respectively, highly negative group assigned -1 and uncertain group assigned 0

(How was the highly negative group determined?)

* **Paper \cite{RN12} (Liu et al)** – proposed a representation method that combined amino acid sequence information and position information to generate a stronger (node) representation for the protein (direct concatenation)

Uses one-hot to encode protein sequence, GCN to capture position information

* **Topsy-Turvy \cite{RN47}**

Sequence-only model D-SCRIPT + Network-only model: GLIDE

D-SCRIPT – Protein language model + CNN (bottom-up approach)

GLIDE – scores all possible edges using a weighted combination of global and local network scores which are learned from the edges already in the training network

(top-down approach)

Examples of past works that used only graph embedding to initialize node attributes

* **HO-VGAE** \cite{RN28} - Graph embedding-based via higher-order GCN.

Combined GCN with a personalized PageRank algorithm (variant of random walk)

Connects the random walk-based propagation effect (scheme) of personalized PageRank to GCN consecutively in every convolutional layer

* **SkipGNN** \cite{RN6} – only leverage accessible network information (adjacent matrix A of the network G) to predict links. In all experiments, only **node2vec** was used to initialize the node attributes.
* Surveys on graph embedding methods conducted

\cite{RN48} and \cite{RN3}

# Section 2: Overview of graph embedding methods

**[Briefly describe algorithm/implementation of each method]**

## Key Terms

* First-order proximity

Connected nodes in a graph should have similar properties

* Second-order proximity

Nodes with similar neighborhoods should have common characteristics

1. **Random walk-based**

Random walks are computationally efficient in terms of both space and time requirements \cite{RN7}

* Deepwalk

Learns structural regularities present within short truncated random walks \cite{RN21}

* Node2vec

Flexible biased random walk procedure that can smoothly interpolate between Breadth-first Sampling (BFS) and Depth-first Sampling (DFS) to explore neighborhoods \cite{RN7}

Introduces two parameters [p (in-out) and q (return)] to control the graph structure with sampling, which import external parameters to search the structure of graph with BFS and DFS rather than the original weights in the graph \cite{RN33}.

* Struc2vec

First constructs a multi-layer weighted graph to generate context for each node.

Biased random walk is then performed on the multilayer graph to learn node sequences.

These sequences are likely to include nodes that are more structurally similar \cite{RN20}.

Applies Dynamic Time Warping (DTW) \cite{RN36} on degree sequences to measure the similarity of local structures \cite{RN37}. Nodes with high structural similarity are close to each other in the embedding space \cite{RN3}.

DTW makes distance comparisons less sensitive to signal transformations as shifting, uniform amplitude scaling or uniform time scaling \cite{RN35}.

However, DTW algorithm ignore partially the effects of connection patterns within neighborhoods. As a result, nodes with similar local structures may be mapped to far vectors \cite{RN37}. (reason for comparison to ripple2vec)

1. **Neural network-based**

* LINE

Has a carefully designed objective function, optimized by an edge-sampling algorithm, that preserves both first-order and second-order proximities \cite{RN11}

* SDNE

Semi-supervised \cite{RN10}

Siamese network consisting of two deep autoencoders along with Laplacian Eigenmaps (Matrix factorization-based) \cite{RN34}, which simultaneously optimizes the first-order and second-order proximity. The learned representations preserve the local and global network structure.

Laplacian Eigenmaps \cite{RN39} generates network representations by factorizing the Laplacian matrix of the adjacency matrix. It exploits only the first-order proximity to preserve network structure \cite{RN10}.

* VGAE

[Highlight its difference from GAE] – Purpose of variational

Utilizes a GCN encoder and an inner product decoder to learn node embeddings \cite{RN3}

Can naturally incorporate node features \cite{RN46} (e.g. protein sequence embedding)

Have “data-cleansing power”

Autoencoders distill inputs into the densest amount of data necessary to re-create a similar output

Removes data noise, transformed raw files into clean machine learning data and detect anomalies

<https://www.techtarget.com/searchenterpriseai/feature/How-to-troubleshoot-8-common-autoencoder-limitations>

Autoencoders are lossy – contain loss by aggressively pruning the problem space

1. **Matrix factorization-based**

Similarity-based learning – assumption that entities with similar interaction patterns are likely to interact

* GraRep

Extends to high-order proximity and uses the Singular Value Decomposition (SVD) to train the model \cite{RN10}.

Generalizes LINE to incorporate information from network neighborhoods beyond 2-hops, but is unable to efficiently scale to large networks \cite{RN7}.

## Effects of hyper-parameters

– Mentioned in \cite{RN3}

1. **Random walk-based**

* **Node2vec**

p (return parameter) value less than 1 encourages returning back to previous vertex, and discourage for value greater than 1

q (in-out parameter) value less than 1 encourages walks to go “outward”, and value greater than 1 encourage walking within a localized neighbourhood \cite{RN40}

Sampling strategy in Deepwalk can be seen as a special case of node2vec with p=1 and q=1 \cite{RN7}

1. **Neural network-based**

* **SDNE**
* Alpha (α)

α = 0, performance totally determined by the second-order proximity

As α increases, it indicates that model is concentrating more on the first-order proximity

* Beta (β)

1. **Matrix factorization-based**

# Section 3: Experiments

Settings (standard hyper-parameters)

*Embedding dimensions = 128 (default given by the repo) – except VAE, depends on hidden dimensions*

*All results presented are* ***mean*** *of 5 iterations*

Network properties – sparse network

(Proportion of positive (interacting) to negative (non-interacting) protein pairs in real world)

**HVPPI**

The same training set was used for all experiments, and two hold-out testsets were constructed

Number of nodes = 15, 685 (41 IAV, 15, 644 human proteins)

Total number of edges = 641, 404 (41 \* 15, 643)

However, as hvppi outputs predicted interactions, hvppi was ran twice and only edges that gave the same score based on 2 runs were considered

(change to 5 runs)

183, 615 edges (183615 / 641404 (~30%?))

with 15, 195 above the 0.212 threshold

272, 091 edges (~40%), with 48,882 interacting, (~18% of all edges)

with 18, 654 above the 0.212 threshold

Randomly sampled 4460 positive

1. Based on predicted interactions by HVPPI

**[Network reconstruction]\*\*** Existing links in the original network are known and can preserve the original network structure

|  |  |  |  |
| --- | --- | --- | --- |
|  | **No. of Samples** | **HVPPI Thresholds** | |
|  | **Training (80%) +  Validation (10%)** | **Min** | **Max** |
| Positive | 4014 | 0.212 | 0.99 |
| Negative | 4014 | 0 | 0.142 |
|  |  |  |  |
|  | **No. of Samples** | **HVPPI Thresholds** | |
|  | **Hold-out Test (10%)** | **Min** | **Max** |
| Positive | 446 | 0.212 | 0.931 |
| Negative | 446 | 0.001 | 0.141 |

**[Link prediction] (??)** Randomly hide a portion of the existing links and use the remaining network to train the network embedding methods

* Testset randomly sampled from the hidden portion (pairs that produced different scores on different hvppi runs)
* (did not choose threshold to allow more reliable results)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **No. of Samples** | **HVPPI Thresholds** | |
|  | **Hold-out Test (10%)** | **Min** | **Max** |
| Positive | 446 | 0.144 | 0.591 |
| Negative | 446 | 0.009 | 0.142 |

FNN vs Skip-GNN, Top 3 methods

1. Experimentally verified datasets

HPIDB 3.0 (Positive) \cite{RN17, 16} – Not in edgelist

<https://hpidb.igbb.msstate.edu/hpi.html>

Negatome 2.0 \cite{RN14} (Negative)

Negatome - negative human-human protein interactions

Performance on dataset constructed based on HVPPI predictions VS experimentally verified datasets (HPIDB and Negatome)

Explain why use only human-human PPIs for negative

1. Unbalanced dataset (Unseen edges),

Fully random, no choosing of thresholds

[Complete set] 🡪 When number of negative samples is more than 5 times that of positive samples

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | | # Samples | HVPPI Thresholds | |
| MIN | MAX |
| Positive | Train + Validation (90%) | 49, 688 | 0.144 | 0.665 |
| Holdout Test (10%) | 5521 | 0.144 | 0.59 |
| Negative | Train + Validation (90%) | 281, 032 | 0.003 | 0.142 |
| Holdout Test (10%) | 31, 226 | 0.009 | 0.142 |

**PPIDM**

10,420 protein pairs, formed by 354 domain pairs (240 bronze, 114 silver)

4900 protein pairs, formed by 114 **silver** domain pairs

Negative (Randomly sampled from complement of positive dataset)

Positive train: 4279 protein pairs

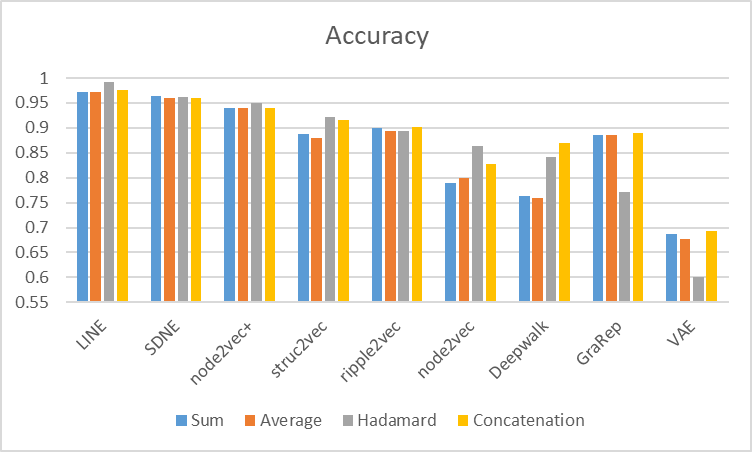
1. Compute edge embeddings

Link prediction requires computation of representation for pairs of nodes

Edge embeddings – compare between Concatenation, Sum, Hadamard Product

(sum and hadamard product – keep original dimensions)

[Plot embeddings? + Results]



1. Visualization using UMAP

**Clustering performance evaluation**

(Higher score means better defined clusters, Scores are higher when clusters are dense and well separated)

Silhouette score (sklearn.metrics.silhouette\_score)

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html>

Score bounded between -1 for incorrect clustering and +1 for highly dense clustering

Scores around zero indicate overlapping clusters

Peter J. Rousseeuw (1987). [“Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis”](https://doi.org/10.1016/0377-0427(87)90125-7) . Computational and Applied Mathematics 20: 53–65.

Calinski-Harabasz Index (sklearn.metrics.calinski\_harabasz\_score)

Variance Ratio Criterion

Ratio of the sum between-clusters dispersion and within-cluster dispersion for all clusters (where dispersion is defined as the sum of distances squared)

Caliński, T., & Harabasz, J. (1974). [“A Dendrite Method for Cluster Analysis”](https://www.researchgate.net/publication/233096619_A_Dendrite_Method_for_Cluster_Analysis). [Communications in Statistics-theory and Methods 3: 1-27](https://doi.org/10.1080/03610927408827101).

* Both metrics gives generally higher scores for convex clusters (Assigns each observation to a point called the “cluster centroid”. Two observations are said to belong to the same cluster if they share the same cluster centroid

<https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation>

[Ground truth labels need to be known]

Homogeneity – each cluster contains only members of a single class

Completeness – all members of a given class are assigned to the same cluster

1. with VS without edge weights

node2vec VS node2vec+

1. with VS without hyper-param tuning (effects of hyper-param tuning)

[training of graph embeddings]

With reference to paper \cite{RN3}

1. Graph embedding

First-order proximity VS First + Second order proximity

(Using LINE algorithm to demonstrate)

1. Comparison to protein embedding
2. Performance on unbalanced dataset
3. Analysis on graph network properties?? (e.g. degree, clustering)

[Aid in discussion of best-performing method?]

# Section 4: Discussion

Clustering PPI networks

* Proteins interacting with each other often participates in the same biological processes or can be associated with specific biological functions being strongly related {Tornow, 2003 #52} (i.e., more similar, shorter distance between proteins)

**[Highlight Pros & Cons of each method]**

**Deepwalk**

* Does not have an explicit objective function to capture the network structure
* Uses random walk to enrich the vertexes, which introduces a lot of noises due to the randomness, especially for vertexes which have high degrees \cite{RN10}

**Node2vec**

**Struc2vec**

* Determines a structural similarity between two nodes without any node or edge attributes \cite{RN20}

**LINE**

* Adopts shallow structure, which is difficult to capture the highly non-linear structure in the underlying network \cite{RN10}
* After optimizing the loss functions (defined separately to preserve the first-order and second-order proximity), LINE directly concatenates the representations, which is sub-optimal than jointly optimizing them in SDNE \cite{RN10}

**SDNE \cite{RN10}**

* Capture highly non-linear network structure

Semi-supervised deep model with multiple layers of non-linear functions

Address structure-preserving and sparsity problem

* Jointly optimize local and global network structure, learnt representations are local-global structure preserved and robust to sparse networks

**GraRep**

* Considers the high-order proximity of the network \cite{RN10, 3}
* Directly concatenates the representations of first-order and high-order \cite{RN10}

**SDNE best performing graph embedding algorithm on our dataset**

**Why VAE performance is bad compared to others (without features?)**

**[Discussion points]**

1. Skip-GNN VS FNN

Why choice of embedding algorithm does not affect Skip-GNN much (original graph)?

1. Edge weights

**Deepwalk** – only concerns that whether there is a connection between two nodes

**Node2vec**

**LINE** – mentioned that proposed model is able to apply to weighted graph, however, does not explicitly leverage the weights on graph

1. Hyper-parameter tuning

Graph embeddings are sensitive to hyperparameters

(Dependent on individual graphs)

Whereas, hyper-param tuning is not required for protein embedding

1. Ensemble of graph embeddings

Too large number of dimensions may introduce noises

e.g. performance drops when other methods added to SDNE

1. Comparison with protein embeddings

### Limitations

* HVPPI was trained solely on human-hosts and it is a challenge to find sufficient training data for other hosts
* For experimentally verified dataset, human-human only

Different from IAV-human? – High specificity and precision (i.e., low number of false positives) 0 identified as 1

# Section 5: Future work

Expansion to other strains

# Dataset

* Overview of constructed datasets

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset |  | | | | Link Prediction  (Unbalanced dataset)5 | |
| Train1 | Test 1  (Network Reconstruction)1 | Test 2  (Experimentally Verified) | Test 3  (Experimentally Verified) | Train | Test |
| # nodes involved in interaction | 1098  (19 IAV, 1079 Human) | 376  (16 IAV, 360 Human) | 361  (14 IAV, 347 Human) | 230  (14 IAV, 216 Human) | 6522  (30 IAV,  6492 Human) | 2681  (28 IAV,  2653 Human) |
| # Positive samples  (i.e., interacting pairs) | 4014 | 446 | 4462 | 2704 | 91, 217 - 48, 882 = 42, 335  42, 335 \* 0.9 = 38, 101 | 42, 335 \* 0.1 = 4234 |
| # Negative samples  (i.e., non-interacting pairs) | 4014 | 446 | 4463 | 176 (IAV-Human)4 + 446 (Human-human) = 622 | 38, 101 \* 4 = 152, 404 | 4234 \* 9 =  38, 106 |
| HVPPI Score (MIN) | 0.000 | 0.001 | 0.023  (excluding negative samples) | 0.143 | 0.004 | 0.004 |
| HVPPI Score (MAX) | 0.99 | 0.931 | 0.981 | 0.981 | 0.845 | 0.743 |
| Average Degree | 7.3115 | 2.3723 | 2.4709 |  | 11.6838 | 3.1585 |

**1 Edges are found in edgelist.**

2 Positive samples from HPIDB 3.0 \cite{}

3 Negative samples from negatome 2.0 \cite{} database. Human-Human protein interactions only. Involving 482 Human Protein Nodes

4 Positive samples from HPIDB 3.0 checked againt HVPPI (i.e., also indicated by HVPPI to be interacting). Labels of non-interacting indicated by HVPPI (Score < 0.143) changed to 0. (176 samples)

5 All remaining **interacting** protein pairs not in edgelist made up the positive samples in this dataset.

Ratio (Positive: Negative)

Training + Validation 1 (20%): 4 (80%)

Test 1 (10%): 9 (90%)

To (further) demonstrate robustness of different (graph embedding/NRL) methods on incomplete interaction network. [Does not need all edges to achieve “good quality” graph embeddings].

# References

@article{RN4,

author = {Alborzi, S. Z. and Ahmed Nacer, A. and Najjar, H. and Ritchie, D. W. and Devignes, M. D.},

title = {PPIDomainMiner: Inferring domain-domain interactions from multiple sources of protein-protein interactions},

journal = {PLoS Comput Biol},

volume = {17},

number = {8},

pages = {e1008844},

note = {Alborzi, Seyed Ziaeddin

Ahmed Nacer, Amina

Najjar, Hiba

Ritchie, David W

Devignes, Marie-Dominique

eng

Research Support, Non-U.S. Gov't

2021/08/10

PLoS Comput Biol. 2021 Aug 9;17(8):e1008844. doi: 10.1371/journal.pcbi.1008844. eCollection 2021 Aug.},

abstract = {Many biological processes are mediated by protein-protein interactions (PPIs). Because protein domains are the building blocks of proteins, PPIs likely rely on domain-domain interactions (DDIs). Several attempts exist to infer DDIs from PPI networks but the produced datasets are heterogeneous and sometimes not accessible, while the PPI interactome data keeps growing. We describe a new computational approach called "PPIDM" (Protein-Protein Interactions Domain Miner) for inferring DDIs using multiple sources of PPIs. The approach is an extension of our previously described "CODAC" (Computational Discovery of Direct Associations using Common neighbors) method for inferring new edges in a tripartite graph. The PPIDM method has been applied to seven widely used PPI resources, using as "Gold-Standard" a set of DDIs extracted from 3D structural databases. Overall, PPIDM has produced a dataset of 84,552 non-redundant DDIs. Statistical significance (p-value) is calculated for each source of PPI and used to classify the PPIDM DDIs in Gold (9,175 DDIs), Silver (24,934 DDIs) and Bronze (50,443 DDIs) categories. Dataset comparison reveals that PPIDM has inferred from the 2017 releases of PPI sources about 46% of the DDIs present in the 2020 release of the 3did database, not counting the DDIs present in the Gold-Standard. The PPIDM dataset contains 10,229 DDIs that are consistent with more than 13,300 PPIs extracted from the IMEx database, and nearly 23,300 DDIs (27.5%) that are consistent with more than 214,000 human PPIs extracted from the STRING database. Examples of newly inferred DDIs covering more than 10 PPIs in the IMEx database are provided. Further exploitation of the PPIDM DDI reservoir includes the inventory of possible partners of a protein of interest and characterization of protein interactions at the domain level in combination with other methods. The result is publicly available at <http://ppidm.loria.fr/>.},

keywords = {Algorithms

Computational Biology

Data Mining/statistics & numerical data

Databases, Protein/statistics & numerical data

Humans

\*Protein Interaction Domains and Motifs

Protein Interaction Mapping/\*statistics & numerical data

\*Protein Interaction Maps

Software},

ISSN = {1553-7358 (Electronic)

1553-734X (Linking)},

DOI = {10.1371/journal.pcbi.1008844},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/34370723>},

year = {2021},

type = {Journal Article}

}

@article{RN17,

author = {Ammari, Mais G. and Gresham, Cathy R. and McCarthy, Fiona M. and Nanduri, Bindu},

title = {HPIDB 2.0: a curated database for host–pathogen interactions},

journal = {Database},

volume = {2016},

pages = {baw103},

abstract = {Identification and analysis of host–pathogen interactions (HPI) is essential to study infectious diseases. However, HPI data are sparse in existing molecular interaction databases, especially for agricultural host–pathogen systems. Therefore, resources that annotate, predict and display the HPI that underpin infectious diseases are critical for developing novel intervention strategies. HPIDB 2.0 ( <http://www.agbase.msstate.edu/hpi/main.html> ) is a resource for HPI data, and contains 45, 238 manually curated entries in the current release. Since the first description of the database in 2010, multiple enhancements to HPIDB data and interface services were made that are described here. Notably, HPIDB 2.0 now provides targeted biocuration of molecular interaction data. As a member of the International Molecular Exchange consortium, annotations provided by HPIDB 2.0 curators meet community standards to provide detailed contextual experimental information and facilitate data sharing. Moreover, HPIDB 2.0 provides access to rapidly available community annotations that capture minimum molecular interaction information to address immediate researcher needs for HPI network analysis. In addition to curation, HPIDB 2.0 integrates HPI from existing external sources and contains tools to infer additional HPI where annotated data are scarce. Compared to other interaction databases, our data collection approach ensures HPIDB 2.0 users access the most comprehensive HPI data from a wide range of pathogens and their hosts (594 pathogen and 70 host species, as of February 2016). Improvements also include enhanced search capacity, addition of Gene Ontology functional information, and implementation of network visualization. The changes made to HPIDB 2.0 content and interface ensure that users, especially agricultural researchers, are able to easily access and analyse high quality, comprehensive HPI data. All HPIDB 2.0 data are updated regularly, are publically available for direct download, and are disseminated to other molecular interaction resources. Database URL:<http://www.agbase.msstate.edu/hpi/main.html>},

ISSN = {1758-0463},

DOI = {10.1093/database/baw103},

url = {<https://doi.org/10.1093/database/baw103>},

year = {2016},

type = {Journal Article}

}

@article{RN27,

author = {Baranwal, Mayank and Magner, Abram and Saldinger, Jacob and Turali-Emre, Emine S. and Elvati, Paolo and Kozarekar, Shivani and VanEpps, J. Scott and Kotov, Nicholas A. and Violi, Angela and Hero, Alfred O.},

title = {Struct2Graph: a graph attention network for structure based predictions of protein–protein interactions},

journal = {BMC Bioinformatics},

volume = {23},

number = {1},

pages = {370},

abstract = {Development of new methods for analysis of protein–protein interactions (PPIs) at molecular and nanometer scales gives insights into intracellular signaling pathways and will improve understanding of protein functions, as well as other nanoscale structures of biological and abiological origins. Recent advances in computational tools, particularly the ones involving modern deep learning algorithms, have been shown to complement experimental approaches for describing and rationalizing PPIs. However, most of the existing works on PPI predictions use protein-sequence information, and thus have difficulties in accounting for the three-dimensional organization of the protein chains.},

ISSN = {1471-2105},

DOI = {10.1186/s12859-022-04910-9},

url = {<https://doi.org/10.1186/s12859-022-04910-9>},

year = {2022},

type = {Journal Article}

}

@article{RN39,

author = {Belkin, M. and Niyogi, P.},

title = {Laplacian Eigenmaps for Dimensionality Reduction and Data Representation},

journal = {Neural Computation},

volume = {15},

number = {6},

pages = {1373-1396},

ISSN = {0899-7667},

DOI = {10.1162/089976603321780317},

year = {2003},

type = {Journal Article}

}

@article{RN14,

author = {Blohm, P. and Frishman, G. and Smialowski, P. and Goebels, F. and Wachinger, B. and Ruepp, A. and Frishman, D.},

title = {Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis},

journal = {Nucleic Acids Res},

volume = {42},

number = {Database issue},

pages = {D396-400},

note = {Blohm, Philipp

Frishman, Goar

Smialowski, Pawel

Goebels, Florian

Wachinger, Benedikt

Ruepp, Andreas

Frishman, Dmitrij

eng

Research Support, Non-U.S. Gov't

England

2013/11/12

Nucleic Acids Res. 2014 Jan;42(Database issue):D396-400. doi: 10.1093/nar/gkt1079. Epub 2013 Nov 8.},

abstract = {Knowledge about non-interacting proteins (NIPs) is important for training the algorithms to predict protein-protein interactions (PPIs) and for assessing the false positive rates of PPI detection efforts. We present the second version of Negatome, a database of proteins and protein domains that are unlikely to engage in physical interactions (available online at <http://mips.helmholtz-muenchen.de/proj/ppi/negatome>). Negatome is derived by manual curation of literature and by analyzing three-dimensional structures of protein complexes. The main methodological innovation in Negatome 2.0 is the utilization of an advanced text mining procedure to guide the manual annotation process. Potential non-interactions were identified by a modified version of Excerbt, a text mining tool based on semantic sentence analysis. Manual verification shows that nearly a half of the text mining results with the highest confidence values correspond to NIP pairs. Compared to the first version the contents of the database have grown by over 300%.},

keywords = {Data Mining

\*Databases, Protein

Internet

Molecular Sequence Annotation

Protein Conformation

\*Protein Interaction Domains and Motifs

\*Protein Interaction Mapping},

ISSN = {1362-4962 (Electronic)

0305-1048 (Linking)},

DOI = {10.1093/nar/gkt1079},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/24214996>},

year = {2014},

type = {Journal Article}

}

@misc{RN18,

author = {Cao, Shaosheng and Lu, Wei and Xu, Qiongkai},

title = {GraRep: Learning Graph Representations with Global Structural Information},

pages = {891-900},

DOI = {10.1145/2806416.2806512},

year = {2015},

type = {Conference Paper}

}

@inbook{RN35,

author = {Cassisi, Carmelo and Montalto, Placido and Aliotta, Marco and Cannata, Andrea and Pulvirenti, Alfredo},

title = {Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining},

ISBN = {978-953-51-0748-4},

DOI = {10.5772/49941},

year = {2012},

type = {Book Section}

}

@article{RN42,

author = {Chen, Z. and Zhao, P. and Li, F. and Leier, A. and Marquez-Lago, T. T. and Wang, Y. and Webb, G. I. and Smith, A. I. and Daly, R. J. and Chou, K. C. and Song, J.},

title = {iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences},

journal = {Bioinformatics},

volume = {34},

number = {14},

pages = {2499-2502},

note = {Chen, Zhen

Zhao, Pei

Li, Fuyi

Leier, Andre

Marquez-Lago, Tatiana T

Wang, Yanan

Webb, Geoffrey I

Smith, A Ian

Daly, Roger J

Chou, Kuo-Chen

Song, Jiangning

eng

R01 AI111965/AI/NIAID NIH HHS/

Research Support, N.I.H., Extramural

Research Support, Non-U.S. Gov't

England

2018/03/13

Bioinformatics. 2018 Jul 15;34(14):2499-2502. doi: 10.1093/bioinformatics/bty140.},

abstract = {Summary: Structural and physiochemical descriptors extracted from sequence data have been widely used to represent sequences and predict structural, functional, expression and interaction profiles of proteins and peptides as well as DNAs/RNAs. Here, we present iFeature, a versatile Python-based toolkit for generating various numerical feature representation schemes for both protein and peptide sequences. iFeature is capable of calculating and extracting a comprehensive spectrum of 18 major sequence encoding schemes that encompass 53 different types of feature descriptors. It also allows users to extract specific amino acid properties from the AAindex database. Furthermore, iFeature integrates 12 different types of commonly used feature clustering, selection and dimensionality reduction algorithms, greatly facilitating training, analysis and benchmarking of machine-learning models. The functionality of iFeature is made freely available via an online web server and a stand-alone toolkit. Availability and implementation: <http://iFeature.erc.monash.edu/>; <https://github.com/Superzchen/iFeature/>. Supplementary information: Supplementary data are available at Bioinformatics online.},

keywords = {Machine Learning

\*Molecular Sequence Annotation

Peptides/chemistry/\*metabolism/physiology

Protein Conformation

Proteins/chemistry/\*metabolism/physiology

Sequence Analysis, Protein/\*methods

\*Software},

ISSN = {1367-4811 (Electronic)

1367-4803 (Linking)},

DOI = {10.1093/bioinformatics/bty140},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/29528364>},

year = {2018},

type = {Journal Article}

}

@article{RN24,

author = {Du, X. and Sun, S. and Hu, C. and Yao, Y. and Yan, Y. and Zhang, Y.},

title = {DeepPPI: Boosting Prediction of Protein-Protein Interactions with Deep Neural Networks},

journal = {J Chem Inf Model},

volume = {57},

number = {6},

pages = {1499-1510},

note = {Du, Xiuquan

Sun, Shiwei

Hu, Changlin

Yao, Yu

Yan, Yuanting

Zhang, Yanping

eng

2017/05/18

J Chem Inf Model. 2017 Jun 26;57(6):1499-1510. doi: 10.1021/acs.jcim.7b00028. Epub 2017 May 26.},

abstract = {The complex language of eukaryotic gene expression remains incompletely understood. Despite the importance suggested by many proteins variants statistically associated with human disease, nearly all such variants have unknown mechanisms, for example, protein-protein interactions (PPIs). In this study, we address this challenge using a recent machine learning advance-deep neural networks (DNNs). We aim at improving the performance of PPIs prediction and propose a method called DeepPPI (Deep neural networks for Protein-Protein Interactions prediction), which employs deep neural networks to learn effectively the representations of proteins from common protein descriptors. The experimental results indicate that DeepPPI achieves superior performance on the test data set with an Accuracy of 92.50%, Precision of 94.38%, Recall of 90.56%, Specificity of 94.49%, Matthews Correlation Coefficient of 85.08% and Area Under the Curve of 97.43%, respectively. Extensive experiments show that DeepPPI can learn useful features of proteins pairs by a layer-wise abstraction, and thus achieves better prediction performance than existing methods. The source code of our approach can be available via <http://ailab.ahu.edu.cn:8087/DeepPPI/index.html> .},

keywords = {\*Neural Networks, Computer

Protein Interaction Mapping/\*methods

Saccharomyces cerevisiae/metabolism},

ISSN = {1549-960X (Electronic)

1549-9596 (Linking)},

DOI = {10.1021/acs.jcim.7b00028},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/28514151>},

year = {2017},

type = {Journal Article}

}

@article{RN43,

author = {Dunham, B. and Ganapathiraju, M. K.},

title = {Benchmark Evaluation of Protein-Protein Interaction Prediction Algorithms},

journal = {Molecules},

volume = {27},

number = {1},

note = {Dunham, Brandan

Ganapathiraju, Madhavi K

eng

T15 LM007059/LM/NLM NIH HHS/

T15LM007059/ National Library of Medicine

Switzerland

2022/01/12

Molecules. 2021 Dec 22;27(1). pii: molecules27010041. doi: 10.3390/molecules27010041.},

abstract = {Protein-protein interactions (PPIs) perform various functions and regulate processes throughout cells. Knowledge of the full network of PPIs is vital to biomedical research, but most of the PPIs are still unknown. As it is infeasible to discover all of them experimentally due to technical and resource limitations, computational prediction of PPIs is essential and accurately assessing the performance of algorithms is required before further application or translation. However, many published methods compose their evaluation datasets incorrectly, using a higher proportion of positive class data than occuring naturally, leading to exaggerated performance. We re-implemented various published algorithms and evaluated them on datasets with realistic data compositions and found that their performance is overstated in original publications; with several methods outperformed by our control models built on 'illogical' and random number features. We conclude that these methods are influenced by an over-characterization of some proteins in the literature and due to scale-free nature of PPI network and that they fail when tested on all possible protein pairs. Additionally, we found that sequence-only-based algorithms performed worse than those that employ functional and expression features. We present a benchmark evaluation of many published algorithms for PPI prediction. The source code of our implementations and the benchmark datasets created here are made available in open source.},

keywords = {\*Algorithms

Computational Biology/\*methods

Databases, Genetic

Humans

Protein Interaction Mapping/\*methods

ROC Curve

Reproducibility of Results

\*Software

computational prediction

evaluation

interactome

protein-protein interactions},

ISSN = {1420-3049 (Electronic)

1420-3049 (Linking)},

DOI = {10.3390/molecules27010041},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/35011283>},

year = {2021},

type = {Journal Article}

}

@article{RN38,

author = {Gao, Z. and Fu, G. and Ouyang, C. and Tsutsui, S. and Liu, X. and Yang, J. and Gessner, C. and Foote, B. and Wild, D. and Ding, Y. and Yu, Q.},

title = {edge2vec: Representation learning using edge semantics for biomedical knowledge discovery},

journal = {BMC Bioinformatics},

volume = {20},

number = {1},

pages = {306},

note = {Gao, Zheng

Fu, Gang

Ouyang, Chunping

Tsutsui, Satoshi

Liu, Xiaozhong

Yang, Jeremy

Gessner, Christopher

Foote, Brian

Wild, David

Ding, Ying

Yu, Qi

eng

71573162/National Natural Science Foundation of China

England

2019/06/27

BMC Bioinformatics. 2019 Jun 10;20(1):306. doi: 10.1186/s12859-019-2914-2.},

abstract = {BACKGROUND: Representation learning provides new and powerful graph analytical approaches and tools for the highly valued data science challenge of mining knowledge graphs. Since previous graph analytical methods have mostly focused on homogeneous graphs, an important current challenge is extending this methodology for richly heterogeneous graphs and knowledge domains. The biomedical sciences are such a domain, reflecting the complexity of biology, with entities such as genes, proteins, drugs, diseases, and phenotypes, and relationships such as gene co-expression, biochemical regulation, and biomolecular inhibition or activation. Therefore, the semantics of edges and nodes are critical for representation learning and knowledge discovery in real world biomedical problems. RESULTS: In this paper, we propose the edge2vec model, which represents graphs considering edge semantics. An edge-type transition matrix is trained by an Expectation-Maximization approach, and a stochastic gradient descent model is employed to learn node embedding on a heterogeneous graph via the trained transition matrix. edge2vec is validated on three biomedical domain tasks: biomedical entity classification, compound-gene bioactivity prediction, and biomedical information retrieval. Results show that by considering edge-types into node embedding learning in heterogeneous graphs, edge2vec significantly outperforms state-of-the-art models on all three tasks. CONCLUSIONS: We propose this method for its added value relative to existing graph analytical methodology, and in the real world context of biomedical knowledge discovery applicability.},

keywords = {Algorithms

Biomedical Research

Humans

Informatics/\*methods

\*Knowledge

\*Learning

Neural Networks, Computer

Semantics

Applied machine learning

Biomedical knowledge discovery

Data science

Edge semantics

Graph embedding

Heterogeneous network

Knowledge graph

Linked data

Network science

Node embedding

Representation learning

Semantic web

Systems biology},

ISSN = {1471-2105 (Electronic)

1471-2105 (Linking)},

DOI = {10.1186/s12859-019-2914-2},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/31238875>},

year = {2019},

type = {Journal Article}

}

@misc{RN34,

author = {Goyal, Palash and Hosseinmardi, Homa and Ferrara, Emilio and Galstyan, Aram},

title = {Embedding Networks with Edge Attributes},

pages = {38-42},

DOI = {10.1145/3209542.3209571},

year = {2018},

type = {Conference Paper}

}

@article{RN7,

author = {Grover, A. and Leskovec, J.},

title = {node2vec: Scalable Feature Learning for Networks},

journal = {KDD},

volume = {2016},

pages = {855-864},

note = {Grover, Aditya

Leskovec, Jure

eng

U54 EB020405/EB/NIBIB NIH HHS/

2016/11/18

KDD. 2016 Aug;2016:855-864. doi: 10.1145/2939672.2939754.},

abstract = {Prediction tasks over nodes and edges in networks require careful effort in engineering features used by learning algorithms. Recent research in the broader field of representation learning has led to significant progress in automating prediction by learning the features themselves. However, present feature learning approaches are not expressive enough to capture the diversity of connectivity patterns observed in networks. Here we propose node2vec, an algorithmic framework for learning continuous feature representations for nodes in networks. In node2vec, we learn a mapping of nodes to a low-dimensional space of features that maximizes the likelihood of preserving network neighborhoods of nodes. We define a flexible notion of a node's network neighborhood and design a biased random walk procedure, which efficiently explores diverse neighborhoods. Our algorithm generalizes prior work which is based on rigid notions of network neighborhoods, and we argue that the added flexibility in exploring neighborhoods is the key to learning richer representations. We demonstrate the efficacy of node2vec over existing state-of-the-art techniques on multi-label classification and link prediction in several real-world networks from diverse domains. Taken together, our work represents a new way for efficiently learning state-of-the-art task-independent representations in complex networks.},

keywords = {Algorithms

Experimentation

Feature learning

Graph representations

Information networks

Node embeddings},

ISSN = {2154-817X (Print)

2154-817X (Linking)},

DOI = {10.1145/2939672.2939754},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/27853626>},

year = {2016},

type = {Journal Article}

}

@article{RN31,

author = {Hamilton, William and Ying, Rex and Leskovec, Jure},

title = {Representation Learning on Graphs: Methods and Applications},

year = {2017},

type = {Journal Article}

}

@article{RN25,

author = {Hashemifar, S. and Neyshabur, B. and Khan, A. A. and Xu, J.},

title = {Predicting protein-protein interactions through sequence-based deep learning},

journal = {Bioinformatics},

volume = {34},

number = {17},

pages = {i802-i810},

note = {Hashemifar, Somaye

Neyshabur, Behnam

Khan, Aly A

Xu, Jinbo

eng

R01 GM089753/GM/NIGMS NIH HHS/

Research Support, N.I.H., Extramural

Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, Non-P.H.S.

England

2018/11/14

Bioinformatics. 2018 Sep 1;34(17):i802-i810. doi: 10.1093/bioinformatics/bty573.},

abstract = {Motivation: High-throughput experimental techniques have produced a large amount of protein-protein interaction (PPI) data, but their coverage is still low and the PPI data is also very noisy. Computational prediction of PPIs can be used to discover new PPIs and identify errors in the experimental PPI data. Results: We present a novel deep learning framework, DPPI, to model and predict PPIs from sequence information alone. Our model efficiently applies a deep, Siamese-like convolutional neural network combined with random projection and data augmentation to predict PPIs, leveraging existing high-quality experimental PPI data and evolutionary information of a protein pair under prediction. Our experimental results show that DPPI outperforms the state-of-the-art methods on several benchmarks in terms of area under precision-recall curve (auPR), and computationally is more efficient. We also show that DPPI is able to predict homodimeric interactions where other methods fail to work accurately, and the effectiveness of DPPI in specific applications such as predicting cytokine-receptor binding affinities. Availability and implementation: Predicting protein-protein interactions through sequence-based deep learning): <https://github.com/hashemifar/DPPI/>. Supplementary information: Supplementary data are available at Bioinformatics online.},

keywords = {Animals

Area Under Curve

\*Deep Learning

Humans

Mice

Protein Binding

Protein Interaction Mapping/\*methods

Proteins/chemistry/\*metabolism

Software},

ISSN = {1367-4811 (Electronic)

1367-4803 (Linking)},

DOI = {10.1093/bioinformatics/bty573},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/30423091>},

year = {2018},

type = {Journal Article}

}

@article{RN6,

author = {Huang, K. and Xiao, C. and Glass, L. M. and Zitnik, M. and Sun, J.},

title = {SkipGNN: predicting molecular interactions with skip-graph networks},

journal = {Sci Rep},

volume = {10},

number = {1},

pages = {21092},

note = {Huang, Kexin

Xiao, Cao

Glass, Lucas M

Zitnik, Marinka

Sun, Jimeng

eng

England

2020/12/05

Sci Rep. 2020 Dec 3;10(1):21092. doi: 10.1038/s41598-020-77766-9.},

abstract = {Molecular interaction networks are powerful resources for molecular discovery. They are increasingly used with machine learning methods to predict biologically meaningful interactions. While deep learning on graphs has dramatically advanced the prediction prowess, current graph neural network (GNN) methods are mainly optimized for prediction on the basis of direct similarity between interacting nodes. In biological networks, however, similarity between nodes that do not directly interact has proved incredibly useful in the last decade across a variety of interaction networks. Here, we present SkipGNN, a graph neural network approach for the prediction of molecular interactions. SkipGNN predicts molecular interactions by not only aggregating information from direct interactions but also from second-order interactions, which we call skip similarity. In contrast to existing GNNs, SkipGNN receives neural messages from two-hop neighbors as well as immediate neighbors in the interaction network and non-linearly transforms the messages to obtain useful information for prediction. To inject skip similarity into a GNN, we construct a modified version of the original network, called the skip graph. We then develop an iterative fusion scheme that optimizes a GNN using both the skip graph and the original graph. Experiments on four interaction networks, including drug-drug, drug-target, protein-protein, and gene-disease interactions, show that SkipGNN achieves superior and robust performance. Furthermore, we show that unlike popular GNNs, SkipGNN learns biologically meaningful embeddings and performs especially well on noisy, incomplete interaction networks.},

ISSN = {2045-2322 (Electronic)

2045-2322 (Linking)},

DOI = {10.1038/s41598-020-77766-9},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/33273494>},

year = {2020},

type = {Journal Article}

}

@inbook{RN13,

author = {Kakkola, Laura and Ikonen, Niina and Julkunen, Ilkka},

title = {Influenza A Viruses (Orthomyxoviridae)},

booktitle = {Encyclopedia of Virology (Fourth Edition)},

editor = {Bamford, Dennis H. and Zuckerman, Mark},

publisher = {Academic Press},

address = {Oxford},

pages = {551-560},

abstract = {Influenza A viruses (IAV) cause annual epidemics and occasional pandemics in the human population. IAVs are subtyped according to the viral envelope glycoproteins hemagglutinin (HA) and neuraminidase (NA) proteins. IAVs, seasonal H1N1 and H3N2 as well as some other subtypes cause respiratory tract infections that can occasionally be severe leading to pneumonia or acute respiratory distress syndrome (ARDS) in humans. IAV infections can be prevented with vaccines or treated with antiviral drugs. However, the vaccine composition needs to be updated every year due to antigenic variation of IAVs.},

keywords = {Antigenic variation

Antivirals

ARDS

Influenza A virus

Innate immune response

Pandemic

Pneumonia

Vaccine},

ISBN = {978-0-12-814516-6},

DOI = {<https://doi.org/10.1016/B978-0-12-814515-9.00046-1>},

url = {<https://www.sciencedirect.com/science/article/pii/B9780128145159000461>},

year = {2021},

type = {Book Section}

}

@article{RN36,

author = {Keogh, Eamonn and Ratanamahatana, Chotirat Ann},

title = {Exact indexing of dynamic time warping},

journal = {Knowledge and Information Systems},

volume = {7},

number = {3},

pages = {358-386},

abstract = {The problem of indexing time series has attracted much interest. Most algorithms used to index time series utilize the Euclidean distance or some variation thereof. However, it has been forcefully shown that the Euclidean distance is a very brittle distance measure. Dynamic time warping (DTW) is a much more robust distance measure for time series, allowing similar shapes to match even if they are out of phase in the time axis. Because of this flexibility, DTW is widely used in science, medicine, industry and finance. Unfortunately, however, DTW does not obey the triangular inequality and thus has resisted attempts at exact indexing. Instead, many researchers have introduced approximate indexing techniques or abandoned the idea of indexing and concentrated on speeding up sequential searches. In this work, we introduce a novel technique for the exact indexing of DTW. We prove that our method guarantees no false dismissals and we demonstrate its vast superiority over all competing approaches in the largest and most comprehensive set of time series indexing experiments ever undertaken.},

ISSN = {0219-3116},

DOI = {10.1007/s10115-004-0154-9},

url = {<https://doi.org/10.1007/s10115-004-0154-9>},

year = {2005},

type = {Journal Article}

}

@article{RN46,

author = {Kipf, Thomas N. and Welling, Max},

title = {Variational Graph Auto-Encoders},

DOI = {10.48550/ARXIV.1611.07308},

url = {<https://arxiv.org/abs/1611.07308>},

year = {2016},

type = {Journal Article}

}

@article{RN16,

author = {Kumar, Ranjit and Nanduri, Bindu},

title = {HPIDB - a unified resource for host-pathogen interactions},

journal = {BMC Bioinformatics},

volume = {11},

number = {6},

pages = {S16},

abstract = {Protein-protein interactions (PPIs) play a crucial role in initiating infection in a host-pathogen system. Identification of these PPIs is important for understanding the underlying biological mechanism of infection and identifying putative drug targets. Database resources for studying host-pathogen systems are scarce and are either host specific or dedicated to specific pathogens.},

ISSN = {1471-2105},

DOI = {10.1186/1471-2105-11-S6-S16},

url = {<https://doi.org/10.1186/1471-2105-11-S6-S16>},

year = {2010},

type = {Journal Article}

}

@article{RN12,

author = {Liu, L. and Zhu, X. and Ma, Y. and Piao, H. and Yang, Y. and Hao, X. and Fu, Y. and Wang, L. and Peng, J.},

title = {Combining sequence and network information to enhance protein-protein interaction prediction},

journal = {BMC Bioinformatics},

volume = {21},

number = {Suppl 16},

pages = {537},

note = {Liu, Leilei

Zhu, Xianglei

Ma, Yi

Piao, Haiyin

Yang, Yaodong

Hao, Xiaotian

Fu, Yue

Wang, Li

Peng, Jiajie

eng

Grant Nos. U1836214/National Natural Science Foundation of China (CN)

19JCYBJC16300/Tianjin Nature fund, Research on Data Platform Technology Based on Automotive Electronic Identification System

England

2020/12/17

BMC Bioinformatics. 2020 Dec 16;21(Suppl 16):537. doi: 10.1186/s12859-020-03896-6.},

abstract = {BACKGROUND: Protein-protein interactions (PPIs) are of great importance in cellular systems of organisms, since they are the basis of cellular structure and function and many essential cellular processes are related to that. Most proteins perform their functions by interacting with other proteins, so predicting PPIs accurately is crucial for understanding cell physiology. RESULTS: Recently, graph convolutional networks (GCNs) have been proposed to capture the graph structure information and generate representations for nodes in the graph. In our paper, we use GCNs to learn the position information of proteins in the PPIs networks graph, which can reflect the properties of proteins to some extent. Combining amino acid sequence information and position information makes a stronger representation for protein, which improves the accuracy of PPIs prediction. CONCLUSION: In previous research methods, most of them only used protein amino acid sequence as input information to make predictions, without considering the structural information of PPIs networks graph. We first time combine amino acid sequence information and position information to make representations for proteins. The experimental results indicate that our method has strong competitiveness compared with several sequence-based methods.},

keywords = {Amino Acid Sequence

Databases, Protein

Humans

Protein Interaction Mapping/\*methods

Proteins/\*chemistry/\*metabolism

Saccharomyces cerevisiae/metabolism

Saccharomyces cerevisiae Proteins/metabolism

Graph convolutional networks

Protein-protein interactions},

ISSN = {1471-2105 (Electronic)

1471-2105 (Linking)},

DOI = {10.1186/s12859-020-03896-6},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/33323120>},

year = {2020},

type = {Journal Article}

}

@article{RN41,

author = {Liu, Renming and Hirn, Matthew J. and Krishnan, Arjun},

title = {Accurately Modeling Biased Random Walks on Weighted Graphs Using Node2vec+},

journal = {CoRR},

volume = {abs/2109.08031},

url = {<https://arxiv.org/abs/2109.08031>},

year = {2021},

type = {Journal Article}

}

@article{RN40,

author = {Liu, Renming and Krishnan, Arjun},

title = {PecanPy: a fast, efficient and parallelized Python implementation of node2vec},

journal = {Bioinformatics},

volume = {37},

number = {19},

pages = {3377-3379},

abstract = {Learning low-dimensional representations (embeddings) of nodes in large graphs is key to applying machine learning on massive biological networks. Node2vec is the most widely used method for node embedding. However, its original Python and C++ implementations scale poorly with network density, failing for dense biological networks with hundreds of millions of edges. We have developed PecanPy, a new Python implementation of node2vec that uses cache-optimized compact graph data structures and precomputing/parallelization to result in fast, high-quality node embeddings for biological networks of all sizes and densities.PecanPy software is freely available at <https://github.com/krishnanlab/PecanPy.Supplementary> data are available at Bioinformatics online.},

ISSN = {1367-4803},

DOI = {10.1093/bioinformatics/btab202},

url = {<https://doi.org/10.1093/bioinformatics/btab202>},

year = {2021},

type = {Journal Article}

}

@article{RN37,

author = {Luo, Jizhou and Xiao, Song and Jiang, Shouxu and Gao, Hong and Xiao, Yinuo},

title = {ripple2vec: Node Embedding with Ripple Distance of Structures},

journal = {Data Science and Engineering},

volume = {7},

number = {2},

pages = {156-174},

ISSN = {2364-1185

2364-1541},

DOI = {10.1007/s41019-022-00184-6},

year = {2022},

type = {Journal Article}

}

@article{RN48,

author = {Makarov, I. and Kiselev, D. and Nikitinsky, N. and Subelj, L.},

title = {Survey on graph embeddings and their applications to machine learning problems on graphs},

journal = {PeerJ Comput Sci},

volume = {7},

pages = {e357},

note = {Makarov, Ilya

Kiselev, Dmitrii

Nikitinsky, Nikita

Subelj, Lovro

eng

2021/04/06

PeerJ Comput Sci. 2021 Feb 4;7:e357. doi: 10.7717/peerj-cs.357. eCollection 2021.},

abstract = {Dealing with relational data always required significant computational resources, domain expertise and task-dependent feature engineering to incorporate structural information into a predictive model. Nowadays, a family of automated graph feature engineering techniques has been proposed in different streams of literature. So-called graph embeddings provide a powerful tool to construct vectorized feature spaces for graphs and their components, such as nodes, edges and subgraphs under preserving inner graph properties. Using the constructed feature spaces, many machine learning problems on graphs can be solved via standard frameworks suitable for vectorized feature representation. Our survey aims to describe the core concepts of graph embeddings and provide several taxonomies for their description. First, we start with the methodological approach and extract three types of graph embedding models based on matrix factorization, random-walks and deep learning approaches. Next, we describe how different types of networks impact the ability of models to incorporate structural and attributed data into a unified embedding. Going further, we perform a thorough evaluation of graph embedding applications to machine learning problems on graphs, among which are node classification, link prediction, clustering, visualization, compression, and a family of the whole graph embedding algorithms suitable for graph classification, similarity and alignment problems. Finally, we overview the existing applications of graph embeddings to computer science domains, formulate open problems and provide experiment results, explaining how different networks properties result in graph embeddings quality in the four classic machine learning problems on graphs, such as node classification, link prediction, clustering and graph visualization. As a result, our survey covers a new rapidly growing field of network feature engineering, presents an in-depth analysis of models based on network types, and overviews a wide range of applications to machine learning problems on graphs.},

keywords = {Geometric deep learning

Graph embedding

Graph neural networks

Graph visualization

Knowledge representation

Link prediction

Machine learning

Network science

Node classification

Node clustering},

ISSN = {2376-5992 (Electronic)

2376-5992 (Linking)},

DOI = {10.7717/peerj-cs.357},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/33817007>},

year = {2021},

type = {Journal Article}

}

@article{RN21,

author = {Perozzi, Bryan and Al-Rfou, Rami and Skiena, Steven},

title = {DeepWalk: Online Learning of Social Representations},

journal = {Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining},

DOI = {10.1145/2623330.2623732},

year = {2014},

type = {Journal Article}

}

@misc{RN20,

author = {Ribeiro, Leonardo F. R. and Saverese, Pedro H. P. and Figueiredo, Daniel R.},

title = {struc2vec: Learning Node Representations from Structural Identity},

pages = {385-394},

DOI = {10.1145/3097983.3098061},

year = {2017},

type = {Conference Paper}

}

@article{RN44,

author = {Schmidhuber, Jürgen},

title = {Deep learning in neural networks: An overview},

journal = {Neural Networks},

volume = {61},

pages = {85-117},

abstract = {In recent years, deep artificial neural networks (including recurrent ones) have won numerous contests in pattern recognition and machine learning. This historical survey compactly summarizes relevant work, much of it from the previous millennium. Shallow and Deep Learners are distinguished by the depth of their credit assignment paths, which are chains of possibly learnable, causal links between actions and effects. I review deep supervised learning (also recapitulating the history of backpropagation), unsupervised learning, reinforcement learning & evolutionary computation, and indirect search for short programs encoding deep and large networks.},

keywords = {Deep learning

Supervised learning

Unsupervised learning

Reinforcement learning

Evolutionary computation},

ISSN = {0893-6080},

DOI = {<https://doi.org/10.1016/j.neunet.2014.09.003>},

url = {<https://www.sciencedirect.com/science/article/pii/S0893608014002135>},

year = {2015},

type = {Journal Article}

}

@article{RN47,

author = {Singh, Rohit and Devkota, Kapil and Sledzieski, Samuel and Berger, Bonnie and Cowen, Lenore},

title = {Topsy-Turvy: integrating a global view into sequence-based PPI prediction},

journal = {Bioinformatics},

volume = {38},

number = {Supplement\_1},

pages = {i264-i272},

abstract = {Computational methods to predict protein–protein interaction (PPI) typically segregate into sequence-based ‘bottom-up’ methods that infer properties from the characteristics of the individual protein sequences, or global ‘top-down’ methods that infer properties from the pattern of already known PPIs in the species of interest. However, a way to incorporate top-down insights into sequence-based bottom-up PPI prediction methods has been elusive. We thus introduce Topsy-Turvy, a method that newly synthesizes both views in a sequence-based, multi-scale, deep-learning model for PPI prediction. While Topsy-Turvy makes predictions using only sequence data, during the training phase it takes a transfer-learning approach by incorporating patterns from both global and molecular-level views of protein interaction. In a cross-species context, we show it achieves state-of-the-art performance, offering the ability to perform genome-scale, interpretable PPI prediction for non-model organisms with no existing experimental PPI data. In species with available experimental PPI data, we further present a Topsy-Turvy hybrid (TT-Hybrid) model which integrates Topsy-Turvy with a purely network-based model for link prediction that provides information about species-specific network rewiring. TT-Hybrid makes accurate predictions for both well- and sparsely-characterized proteins, outperforming both its constituent components as well as other state-of-the-art PPI prediction methods. Furthermore, running Topsy-Turvy and TT-Hybrid screens is feasible for whole genomes, and thus these methods scale to settings where other methods (e.g. AlphaFold-Multimer) might be infeasible. The generalizability, accuracy and genome-level scalability of Topsy-Turvy and TT-Hybrid unlocks a more comprehensive map of protein interaction and organization in both model and non-model organisms.<https://topsyturvy.csail.mit.edu.Supplementary> data are available at Bioinformatics online.},

ISSN = {1367-4803},

DOI = {10.1093/bioinformatics/btac258},

url = {<https://doi.org/10.1093/bioinformatics/btac258>},

year = {2022},

type = {Journal Article}

}

@article{RN50,

author = {Song, Wenzhuo and Wang, Shengsheng and Yang, Bo and Lu, You and Zhao, Xuehua and Liu, Xueyan},

title = {Learning node and edge embeddings for signed networks},

journal = {Neurocomputing},

volume = {319},

pages = {42-54},

ISSN = {09252312},

DOI = {10.1016/j.neucom.2018.08.072},

year = {2018},

type = {Journal Article}

}

@misc{RN11,

author = {Tang, Jian and Qu, Meng and Wang, Mingzhe and Zhang, Ming and Yan, Jun and Mei, Qiaozhu},

title = {LINE: Large-scale Information Network Embedding},

pages = {1067-1077},

DOI = {10.1145/2736277.2741093},

year = {2015},

type = {Conference Paper}

}

@misc{RN10,

author = {Wang, Daixin and Cui, Peng and Zhu, Wenwu},

title = {Structural Deep Network Embedding},

pages = {1225-1234},

DOI = {10.1145/2939672.2939753},

year = {2016},

type = {Conference Paper}

}

@article{RN33,

author = {Wu, Xiaohua and Pang, Hong and Fan, Youping and Linghu, Yang and Luo, Yu},

title = {ProbWalk: A random walk approach in weighted graph embedding},

journal = {Procedia Computer Science},

volume = {183},

pages = {683-689},

abstract = {There are many weighted graphs in the real-world networks, such as social networks, communication networks, citation networks, etc. Along with successful application of deep learning in graph embedding, we study how to embed weight graph, because weights on the edges also play an important role in the graph. We propose a novel algorithm called ProbWalk, which take advantage of edge weights and convert the weights into transition probabilities. Our proposed method specifies the strategy of sampling the surrounding vertices by weights and generate the random walk for graph embedding according to transition probability. We evaluate our methods on tasks including multi-label classification and link prediction. Experimental results show that our method performs better than competed method on several weighted graph datasets.},

keywords = {graph embedding

machine learning

network representation learning

weighted graph embedding},

ISSN = {1877-0509},

DOI = {<https://doi.org/10.1016/j.procs.2021.02.115>},

url = {<https://www.sciencedirect.com/science/article/pii/S1877050921005913>},

year = {2021},

type = {Journal Article}

}

@article{RN28,

author = {Xiao, Z. and Deng, Y.},

title = {Graph embedding-based novel protein interaction prediction via higher-order graph convolutional network},

journal = {PLoS One},

volume = {15},

number = {9},

pages = {e0238915},

note = {Xiao, Ze

Deng, Yue

eng

Research Support, Non-U.S. Gov't

2020/09/25

PLoS One. 2020 Sep 24;15(9):e0238915. doi: 10.1371/journal.pone.0238915. eCollection 2020.},

abstract = {Protein-protein interactions (PPIs) are essential for most biological processes. However, current PPI networks present high levels of noise, sparseness and incompleteness, which limits our ability to understand the cell at the system level from the PPI network. Predicting novel (missing) links in noisy PPI networks is an essential computational method for automatically expanding the human interactome and for identifying biologically legitimate but undetected interactions for experimental determination of PPIs, which is both expensive and time-consuming. Recently, graph convolutional networks (GCN) have shown their effectiveness in modeling graph-structured data, which employ a 1-hop neighborhood aggregation procedure and have emerged as a powerful architecture for node or graph representations. In this paper, we propose a novel node (protein) embedding method by combining GCN and PageRank as the latter can significantly improve the GCN's aggregation scheme, which has difficulty in extending and exploring topological information of networks across higher-order neighborhoods of each node. Building on this novel node embedding model, we develop a higher-order GCN variational auto-encoder (HO-VGAE) architecture, which can learn a joint node representation of higher-order local and global PPI network topology for novel protein interaction prediction. It is worth noting that our method is based exclusively on network topology, with no protein attributes or extra biological features used. Extensive computational validations on PPI prediction task demonstrate our method without leveraging any additional biological information shows competitive performance-outperforms all existing graph embedding-based link prediction methods in both accuracy and robustness.},

keywords = {Algorithms

Computational Biology/\*methods

Humans

Neural Networks, Computer

Protein Interaction Mapping/\*methods},

ISSN = {1932-6203 (Electronic)

1932-6203 (Linking)},

DOI = {10.1371/journal.pone.0238915},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/32970681>},

year = {2020},

type = {Journal Article}

}

@article{RN22,

author = {Yang, F. and Fan, K. and Song, D. and Lin, H.},

title = {Graph-based prediction of Protein-protein interactions with attributed signed graph embedding},

journal = {BMC Bioinformatics},

volume = {21},

number = {1},

pages = {323},

note = {Yang, Fang

Fan, Kunjie

Song, Dandan

Lin, Huakang

eng

England

2020/07/23

BMC Bioinformatics. 2020 Jul 21;21(1):323. doi: 10.1186/s12859-020-03646-8.},

abstract = {BACKGROUND: Protein-protein interactions (PPIs) are central to many biological processes. Considering that the experimental methods for identifying PPIs are time-consuming and expensive, it is important to develop automated computational methods to better predict PPIs. Various machine learning methods have been proposed, including a deep learning technique which is sequence-based that has achieved promising results. However, it only focuses on sequence information while ignoring the structural information of PPI networks. Structural information of PPI networks such as their degree, position, and neighboring nodes in a graph has been proved to be informative in PPI prediction. RESULTS: Facing the challenge of representing graph information, we introduce an improved graph representation learning method. Our model can study PPI prediction based on both sequence information and graph structure. Moreover, our study takes advantage of a representation learning model and employs a graph-based deep learning method for PPI prediction, which shows superiority over existing sequence-based methods. Statistically, Our method achieves state-of-the-art accuracy of 99.15% on Human protein reference database (HPRD) dataset and also obtains best results on Database of Interacting Protein (DIP) Human, Drosophila, Escherichia coli (E. coli), and Caenorhabditis elegans (C. elegan) datasets. CONCLUSION: Here, we introduce signed variational graph auto-encoder (S-VGAE), an improved graph representation learning method, to automatically learn to encode graph structure into low-dimensional embeddings. Experimental results demonstrate that our method outperforms other existing sequence-based methods on several datasets. We also prove the robustness of our model for very sparse networks and the generalization for a new dataset that consists of four datasets: HPRD, E.coli, C.elegan, and Drosophila.},

keywords = {Animals

Caenorhabditis elegans/metabolism

Computer Simulation

Databases, Protein

Drosophila/metabolism

Escherichia coli/metabolism

Humans

Machine Learning

Neural Networks, Computer

Protein Interaction Mapping/\*methods

Network embedding

Protein-protein interaction

Representation learning

Variational graph auto-encoder},

ISSN = {1471-2105 (Electronic)

1471-2105 (Linking)},

DOI = {10.1186/s12859-020-03646-8},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/32693790>},

year = {2020},

type = {Journal Article}

}

@article{RN5,

author = {Yang, X. and Yang, S. and Li, Q. and Wuchty, S. and Zhang, Z.},

title = {Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method},

journal = {Comput Struct Biotechnol J},

volume = {18},

pages = {153-161},

note = {Yang, Xiaodi

Yang, Shiping

Li, Qinmengge

Wuchty, Stefan

Zhang, Ziding

eng

Netherlands

2020/01/24

Comput Struct Biotechnol J. 2019 Dec 26;18:153-161. doi: 10.1016/j.csbj.2019.12.005. eCollection 2020.},

abstract = {The identification of human-virus protein-protein interactions (PPIs) is an essential and challenging research topic, potentially providing a mechanistic understanding of viral infection. Given that the experimental determination of human-virus PPIs is time-consuming and labor-intensive, computational methods are playing an important role in providing testable hypotheses, complementing the determination of large-scale interactome between species. In this work, we applied an unsupervised sequence embedding technique (doc2vec) to represent protein sequences as rich feature vectors of low dimensionality. Training a Random Forest (RF) classifier through a training dataset that covers known PPIs between human and all viruses, we obtained excellent predictive accuracy outperforming various combinations of machine learning algorithms and commonly-used sequence encoding schemes. Rigorous comparison with three existing human-virus PPI prediction methods, our proposed computational framework further provided very competitive and promising performance, suggesting that the doc2vec encoding scheme effectively captures context information of protein sequences, pertaining to corresponding protein-protein interactions. Our approach is freely accessible through our web server as part of our host-pathogen PPI prediction platform (<http://zzdlab.com/InterSPPI/>). Taken together, we hope the current work not only contributes a useful predictor to accelerate the exploration of human-virus PPIs, but also provides some meaningful insights into human-virus relationships.},

keywords = {AC, Auto Covariance

ACC, Accuracy

AUC, area under the ROC curve

AUPRC, area under the PR curve

Adaboost, Adaptive Boosting

CT, Conjoint Triad

Doc2vec

Embedding

Human-virus interaction

LD, Local Descriptor

MCC, Matthews correlation coefficient

ML, machine learning

MLP, Multiple Layer Perceptron

MS, mass spectroscopy

Machine learning

PPIs, protein-protein interactions

PR, Precision-Recall

Prediction

Protein-protein interaction

RBF, radial basis function

RF, Random Forest

ROC, Receiver Operating Characteristic

SGD, stochastic gradient descent

SVM, Support Vector Machine

Y2H, yeast two-hybrid},

ISSN = {2001-0370 (Print)

2001-0370 (Linking)},

DOI = {10.1016/j.csbj.2019.12.005},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/31969974>},

year = {2020},

type = {Journal Article}

}

@article{RN49,

author = {Yu, Bin and Chen, Cheng and Wang, Xiaolin and Yu, Zhaomin and Ma, Anjun and Liu, Bingqiang},

title = {Prediction of protein–protein interactions based on elastic net and deep forest},

journal = {Expert Systems with Applications},

volume = {176},

pages = {114876},

abstract = {Prediction of protein–protein interactions (PPIs) helps to grasp molecular roots of disease. However, web-lab experiments to predict PPIs are limited and costly. Using machine-learning-based frameworks can not only automatically identify PPIs, but also provide new ideas for drug research and development from a promising alternative. We present a novel deep-forest-based method for PPIs prediction. Firstly, pseudo amino acid composition (PAAC), autocorrelation descriptor (Auto), multivariate mutual information (MMI), composition-transition-distribution (CTD), amino acid composition position-specific scoring matrix (AAC-PSSM), and dipeptide composition PSSM (DPC-PSSM) are adopted to extract and construct the pattern of PPIs. Secondly, elastic net is utilized to optimize the initial feature vectors and boost the predictive performance. Finally, we ensemble XGBoost, random forest, and extremely randomized trees to construct deep forest model via cascade architecture for PPIs prediction (GcForest-PPI). Benchmark experiments reveal that the proposed approach outperforms other state-of-the-art predictors on Saccharomyces cerevisiae and Helicobacter pylori. We also apply GcForest-PPI on independent test sets, CD9-core network, crossover network, and cancer-specific network. The evaluation shows that GcForest-PPI can boost the prediction accuracy, complement experiments and improve drug discovery.},

keywords = {Protein-protein interactions

Multi-information fusion

Elastic net

Deep forest},

ISSN = {0957-4174},

DOI = {<https://doi.org/10.1016/j.eswa.2021.114876>},

url = {<https://www.sciencedirect.com/science/article/pii/S0957417421003171>},

year = {2021},

type = {Journal Article}

}

@article{RN3,

author = {Yue, X. and Wang, Z. and Huang, J. and Parthasarathy, S. and Moosavinasab, S. and Huang, Y. and Lin, S. M. and Zhang, W. and Zhang, P. and Sun, H.},

title = {Graph embedding on biomedical networks: methods, applications and evaluations},

journal = {Bioinformatics},

volume = {36},

number = {4},

pages = {1241-1251},

note = {Yue, Xiang

Wang, Zhen

Huang, Jingong

Parthasarathy, Srinivasan

Moosavinasab, Soheil

Huang, Yungui

Lin, Simon M

Zhang, Wen

Zhang, Ping

Sun, Huan

eng

Research Support, Non-U.S. Gov't

England

2019/10/05

Bioinformatics. 2020 Feb 15;36(4):1241-1251. doi: 10.1093/bioinformatics/btz718.},

abstract = {MOTIVATION: Graph embedding learning that aims to automatically learn low-dimensional node representations, has drawn increasing attention in recent years. To date, most recent graph embedding methods are evaluated on social and information networks and are not comprehensively studied on biomedical networks under systematic experiments and analyses. On the other hand, for a variety of biomedical network analysis tasks, traditional techniques such as matrix factorization (which can be seen as a type of graph embedding methods) have shown promising results, and hence there is a need to systematically evaluate the more recent graph embedding methods (e.g. random walk-based and neural network-based) in terms of their usability and potential to further the state-of-the-art. RESULTS: We select 11 representative graph embedding methods and conduct a systematic comparison on 3 important biomedical link prediction tasks: drug-disease association (DDA) prediction, drug-drug interaction (DDI) prediction, protein-protein interaction (PPI) prediction; and 2 node classification tasks: medical term semantic type classification, protein function prediction. Our experimental results demonstrate that the recent graph embedding methods achieve promising results and deserve more attention in the future biomedical graph analysis. Compared with three state-of-the-art methods for DDAs, DDIs and protein function predictions, the recent graph embedding methods achieve competitive performance without using any biological features and the learned embeddings can be treated as complementary representations for the biological features. By summarizing the experimental results, we provide general guidelines for properly selecting graph embedding methods and setting their hyper-parameters for different biomedical tasks. AVAILABILITY AND IMPLEMENTATION: As part of our contributions in the paper, we develop an easy-to-use Python package with detailed instructions, BioNEV, available at: <https://github.com/xiangyue9607/BioNEV>, including all source code and datasets, to facilitate studying various graph embedding methods on biomedical tasks. SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics online.},

keywords = {Drug Interactions

\*Neural Networks, Computer

Proteins

Semantics

\*Software},

ISSN = {1367-4811 (Electronic)

1367-4803 (Linking)},

DOI = {10.1093/bioinformatics/btz718},

url = {<https://www.ncbi.nlm.nih.gov/pubmed/31584634>},

year = {2020},

type = {Journal Article}

}

@inproceedings{RN45,

author = {Zell, Andreas},

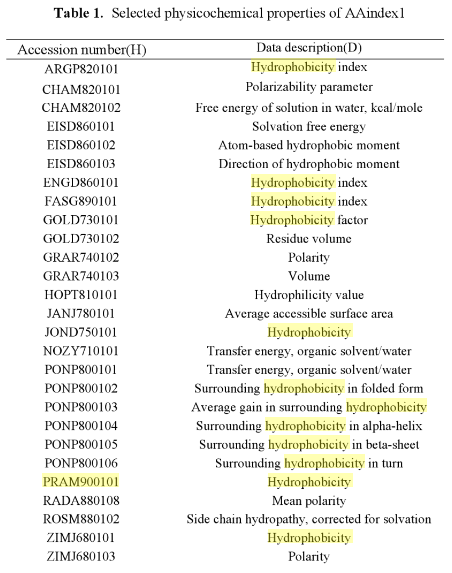
title = {Simulation Neuronaler Netze [Simulation of Neural Networks] (in German)},

pages = {73},

ISBN = {ISBN 3-89319-554-8},

type = {Conference Proceedings}

}



https://books.google.com.sg/books?id=pKYarYSGCTUC&pg=PA523&lpg=PA523&dq=what+is+hydrophobicity+PRAM900101&source=bl&ots=qAxfP0AX-r&sig=ACfU3U2luP9jMEXMGCZmiBgkxVKxyPZxpA&hl=en&sa=X&ved=2ahUKEwi21ZLewdX7AhX763MBHX-nAKoQ6AF6BAgdEAM#v=onepage&q=what%20is%20hydrophobicity%20PRAM900101&f=false