# Uniprot IDs of the Interacting Proteins (2 columns)

Names: uidA, uidB

# NCBI RefSeq Accession Number IDs (2 columns)

Names: protein\_accession\_A, protein\_accession\_B

The respective NCBI IDs of the interacting proteins (https://www.ncbi.nlm.nih.gov/books/NBK21091/table/ch18.T.refseq\_accession\_numbers\_and\_mole/?report=objectonly)

If there is no record, 'NaN' is appended

# PPI type (1 column)

Name: PPI\_type

Indicates if the PPI is positive (value is '1') or negative (value is '0').

# GO term similarity (3 columns)

Names: BP\_similarity, MF\_similarity, CC\_similarity

The similarity of the proteins in the pair based on the similarity of their respective GO terms. The GO terms are filtered, keeping only those that are indicative of a biological process ('BP\_similarity' column), of a molecular function ('MF\_similarity' column) and of a cellular component ('CC\_similarity' column).

Similarity is measured in continuous values from 0 to 1 and 'NaN'. '0' means that there is no similarity between the two proteins and ‘1’ means that the two proteins are the same. 'NaN' means that there is no information on GO terms for one of the two proteins. The tool used for measuring similarity is 'PyGOSemSim' (https://github.com/mojaie/pygosemsim) and measures semantic similarity based on Lin method (1).

# Existence of a Homologous Interacting Pair in other organisms (4 columns)

Names: Homologous in Mouse, ...Drosophila, ...Yeast, ...Ecoli

In this column it is calculated if each protein the PPI pair have a corresponding homologous protein in other organisms and if they both have, the code calculates if it the homologous pair is also an interacting pair.The values in this columns is 'NaN', '0', '1'. If a homologous pair exists, it is calculated weather this pair is also an interacting pair, and if it is, ‘1’ is appended in this column, else ‘0’ is appended. ‘NaN’ values are appended to the interacting pairs with no homology between human and the other organisms. The Datasets for the PPIs in other organisms are derived from DIP database (https://dip.doe-mbi.ucla.edu/dip/Main.cgi) and the mapping of homologous proteins was done via NCBI- HomoloGene Dataset (https://www.ncbi.nlm.nih.gov/homologene).

# Existence of the PPI in other Databases (4 columns)

Names: Exists in DIP?, ...in APID?, ... in BIOGRID?, ... in MINT?

The original Database that we used for extracting PPIs was iRefindex (for positive) and Russell's Negative PPI dataset - 16169070\_neg - (2). In these columns it is tested if the PPI is recorded in other databases as well, which is indicative of a well- documented and well-known PPI. The values in this column are '0' if the PPI isn't recorde in database 'X' and '1' if the PPI is recorded.

# Sequence Similarity (1 column)

Name: Sequence\_similarity

In this column, the E-value of the sequence similarity is measured. More specifically, for each protein in the PPI, the corresponding fasta sequence is loaded from url ("http://www.uniprot.org/uniprot/"). Then, the Similarity Score ('S- score') is calculated for each protein pair sequences. The method used is 'Pair Wise Alignement' and the Aligner assigns +1 for each match in the alignement and +0 for each mismatch. The scoring values can be fixed manually by changing 'match\_score' and 'mismatch\_score' on the Aligner function. Then, for each S-score the corresponding E-value is calculated by using the equation {E=K\*m\*n\*e^-λS}

-K,m,λ, are constants and are set as 1

-n, is the length of the query sequence. In this case there is no query seq, so length is set as the mean length of the two protein seqs of the PPI pair

-S, is the Similarity Score

The E-value is then appended as the column value.

# Domain Interactions (1 column)

Name: pfam\_interaction

Here, the presence of known domain interactions between the PPI pair is calculated. For each protein in the PPI, a list is created that contains known domains extracted from Pfam Database (www.ebi.ac.uk/interpro -Pfam) , a database of protein families and domains, each represented by multiple sequence alignments and hidden Markov models (HMMs)(3). Interactions between Pfam IDs are derived from 3did DB (4). Interaction of domains is measured in 1, 0, 'NaN'. If two proteins share at least one interacting pair of pfam IDs, this pair is marked '1'. If there are no pfam IDs in either one of the interacting proteins, the pair is marked as 'NaN'. Otherwise, the pair is marked as '0'.

# Subcellular co-localization (1 column)

Name: Subcellular Co-localization?

Subcellular co-localization of the two proteins in the pair is measured in this field. The database used for extracting information about protein localizations in the eukaryotic cell, is eSLDB (5). For each protein, there is a list of cellular localizations and then, if the two proteins are share at least 1 common subcellular localization, '1' is appended in this this field (ex. [P1->['Cytoplasm', ' Nucleus'], P2-> ['Nucleus']]). If there is no information for at least one of the proteins, 'NaN' is appended. Else, '0' is appended.

# Gene expression profile similarity (15 columns)

Names: 0,1,....15

This feature represents the similarity of the two proteins in terms of their Gene Expression among fifteen gene expression datasets. The exact gene expression datasets are: GDS531, GDS534, GDS596, GDS651, GDS806, GDS807, GDS843, GDS987, GDS1085, GDS2855, GDS1402, GDS181, GDS1088, GDS841, GDS3257, from NCBI Gene Expression Omnibus. In each dataset, each protein is matched with the corresponding expression profile and then spearman correlation indexes are calculated between the two proteins of each PPI pair. Then the index is used as the column value.

# Aboslute difference in numerical features (30 columns)

In the following columns the absolute difference is calculated between the following numerical features for each PPI pair:

- The difference in the percentage of every amino acid (columns A%.....G %)(20 columns)

- The difference in molecular weigth (MW dif column)

- The difference in aromaticity index according to Lobry, 1994 (6) (Aromaticity dif column)

- The difference in instability index according to Guruprasad et al, 1990 (7) (Instability dif column)

- The difference in fraction of total amino acids that are contained in 3 areas: (3 features)

\*The fraction of aa in helix (helix\_fraction\_dif column)

\*The fraction of aa in turn (turn\_fraction\_dif column)

\*The fraction of aa in sheet (sheet\_fraction\_dif column)

- The difference in molar extinction coefficient (2 columns), when:

(a) The molar extinction coefficient is calculated assuming cysteines(reduced)(cys\_reduced\_dif column), and

(b) The molar extinction coefficient is calculated assuming cystines residues (Cys-Cys-bond) (cys\_residues\_dif column)

- The difference in GRAVY (Grand Average of Hydropathy) according to Kyte and Doolitle, 1982 (8) (gravy\_dif column)

- The difference in protein charge at pH=7 (ph7\_charge\_dif column)

# RNA expression profile similarity (2 columns)

Names: GSE227375\_spearman, GSE228702\_spearman

Similarly to the Gene Expression profile similarity, this feature calculates the similarity between each protein in the pair, in terms of their RNA expression profiles. The RNA expression datasets used were GSE227375 and GSE228702 from NCBI GeneExpression Omnibus. The expression similarity between each protein in the PPI pair is measured using spearman correlation index and then appended to the corresponding column, according to the dataset used (GSE227375 -> GSE227375\_spearman, GSE228702 -> GSE228702\_spearman).

(1) D. Lin. Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconson, USA, July 24-27, 1998 , page 296-304. Morgan Kaufmann, (1998)

(2) Trabuco, L. G., Betts, M. J., & Russell, R. B. (2012). Negative protein–protein interaction datasets derived from large-scale two-hybrid experiments. Methods, 58(4), 343–348. doi:10.1016/j.ymeth.2012.07.028

(3) Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A (2008). "The Pfam protein families database". Nucleic Acids Res. 36 (Database issue): D281–8. doi:10.1093/nar/gkm960. PMC 2238907. PMID 18039703.

(4) Roberto Mosca, Arnaud Ceol, Amelie Stein, Roger Olivella & Patrick Aloy, 3did: a catalogue of domain-based interactions of known three-dimensional structure, Nucleic Acids Research 2014, 42(D1):D374-D379, doi:10.1093/nar/gkt887

(5) Pierleoni A, Martelli PL, Fariselli P, Casadio R. eSLDB: eukaryotic subcellular localization database. Nucleic Acids Res. 2007 Jan;35(Database issue):D208-12. doi: 10.1093/nar/gkl775. Epub 2006 Nov 15. PMID: 17108361; PMCID: PMC1669738.

(6)Lobry JR, Gautier C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Res. 1994 Aug 11;22(15):3174-80. doi: 10.1093/nar/22.15.3174. PMID: 8065933; PMCID: PMC310293.

(7) Guruprasad K, Reddy BV, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng. 1990 Dec;4(2):155-61. doi: 10.1093/protein/4.2.155. PMID: 2075190.

(8) Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982 May 5;157(1):105-32. doi: 10.1016/0022-2836(82)90515-0. PMID: 7108955.