Approaches to Interactome Mapping: HuRI and STRING

Demir Kurt

Bioinformatics PBL WiSe 2024/2025

Topic of the Talk

- Two databases that strive map protein-protein interactions

- Main goal for both is Interactome Mapping

- Interactome: *the entirety of interactions between biological macromolecules of a cell comprising the full spectrum from purely functional relationships to direct physical interactions between them*

- *interactome mapping has become one of the main scopes of current biological research, similar to the way “genome” projects were a driving force of molecular biology some 30 years ago.*

- *HuRI (Human Reference Interactome): serves as a ‘dictionary’, ‘reference point’ for known protein-protein interactions within homo sapiens*

- *STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins): serves as both a ‘dictionary’ and a tool for novel discovery concerning known and predicted protein-protein interactions from all organisms

Variation among Protein-Protein Interactions

Interaction Type

Physical (direct) Interactions

- Protein pairs associate physically
- *The binding domains can be small clefts or large surfaces and can be a few peptides long or span hundreds of amino acids.*

Functional (indirect) Interactions

- No direct contact involved
- Protein pairs work together in common pathway
- Exp.: A transcription factor regulating the expression of another protein

Types of Physical Interactions

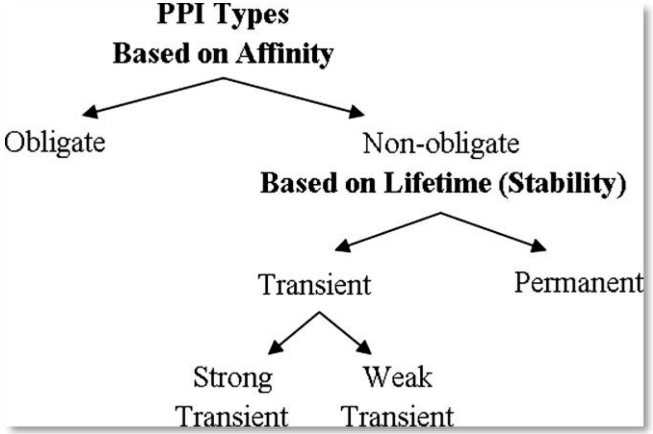


Figure 1: A table to systematically classify Protein-Protein Interactions based on Affinity and Lifetime, excluding Composition and Interaction Type. Note: Permanent PPIs are often classified under Obligate PPIs.

Composition

Homo-oligomeric:

- Identical polypeptide chains

Hetero-oligomeric:

- Non-identical polypeptide chains

Affinity

Obligate:

- Two proteins cannot exist independently
- Constituents unstable

Non-obligate:

- Association only when required

Lifetime

Transient:

- Temporary association
- Controls majority of cellular processes

Permanent:

- Stable and irreversible
- Often found in obligate PPIs

General Comparison 1: Focus Data Type

STRING

- Integrates both physical and functional PPIs
- PPIs data stem from both experiments and computational predictions
- Wide range of Interactions

HuRI

- Focuses more on physical PPIs
- Data stem only from experimental validation
- Human-centric
- Curators believe molecular mechanisms are better inferred from direct PPIs

Interactome Databases

Primary Interaction Databases

- Collection and Organization of experimentally validated PPIs
- High-throughput techniques such as Yeast-two-Hybrid or Co-Immunoprecipitation
- Reliable and backed by real-world evidence
- Might include additional, equally important meta-data such as experimental conditions, methods employed and other relevant information

IMEx Consortium

- Consultative Body for standardization of PPI data
- 10+ Primary Interaction Databases as members
- *Develop a single set of curation rules when capturing data from directly deposited interaction data, preprints and publication*
- *the captured interactions available in a single search interface on a common website.*

Computational Prediction Databases

- PPI prediction using computational models and algorithms
- Based on genetic information, protein structures or known biological networks
- Broad yet less confirmed insights
- Useful for generating hypotheses and guiding experimental research

A Mix of Both Worlds

- Combination of experimentally validated and computationally predicted PPIs
- Includes physical and functional interactions
- Integrates multiple data sources
- Offers comprehensive association network of proteins
- *Data integration across different evidence sources is known to increase the overall network quality and is also deemed necessary given the diverse modes by which proteins can be associated.*

Comparison 2: Numbers & Methodologies

STRING

- Belongs to third class
- Integrative database
- Wide variety of evidence sources
- Usability features: customization, enrichment detection and programmatic access
- Relies on data from members of IMEx Consortium

Numbers

- Primary focus on genome-sequenced organisms
- In 2018, 5090 organisms and 24 million proteins
- Currently, 14000 organisms (v. 11.5)

HuRI

- Belongs to first class
- Based purely on experimental evidence
- Predominant usage of yeast-two-hybrid method
- *only binary PPI assay capable of screening the human proteome with sufficient throughput.*

Numbers

- Primary focus is the human interactome
- Three variants of Y2H assay employed
- Currently, 64006 verified PPIs involving 9094 Proteins

STRING: Database Content

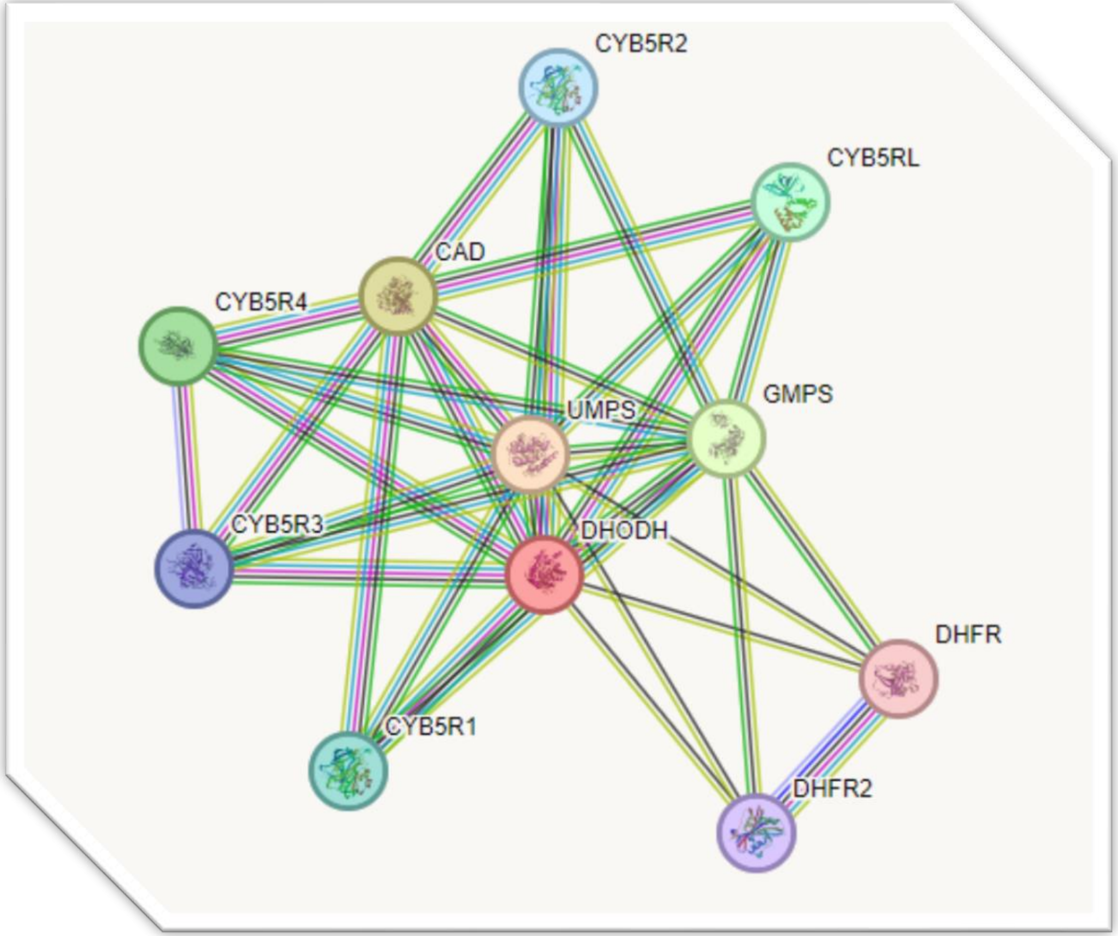


Figure:Exemplary association network of the protein ***Dihydroorotate dehydrogenase,*** *symbol (****DHODH****) . This protein catalyzes the fourth enzymatic step of de novo pyrimidine biosynthesis, taken from STRING*

The Most Basic Unit of Linkage in STRING between two proteins is functional association

Formal definition: ***Any two proteins that jointly contribute toward a specific cellular process are deemed to be functionally associated, including pairs of proteins that act antagonistically within the same process*.**

Specificity

- Overlap in function must correspond to biological pathway or function
- A general system (e.g. metabolism) is not accepted
- Physical interaction not required

Inclusiveness

- Proteins with contradicting roles are deemed related
- Inhibitors and activators in same pathways are given an association

Pathway Maps

- *a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction, and relation networks so far.*
- Maps exist for pathways in Metabolism, Genetic Information Processing, Cellular Processes, Organismal Systems etc.
- STRING refers to a specificity cutoff marking out functional associations
- The cutoff corresponds highly to those in KEGG pathway maps.
- Helps in Validation and Filtering
- Ensures Capturing of meaningful associations

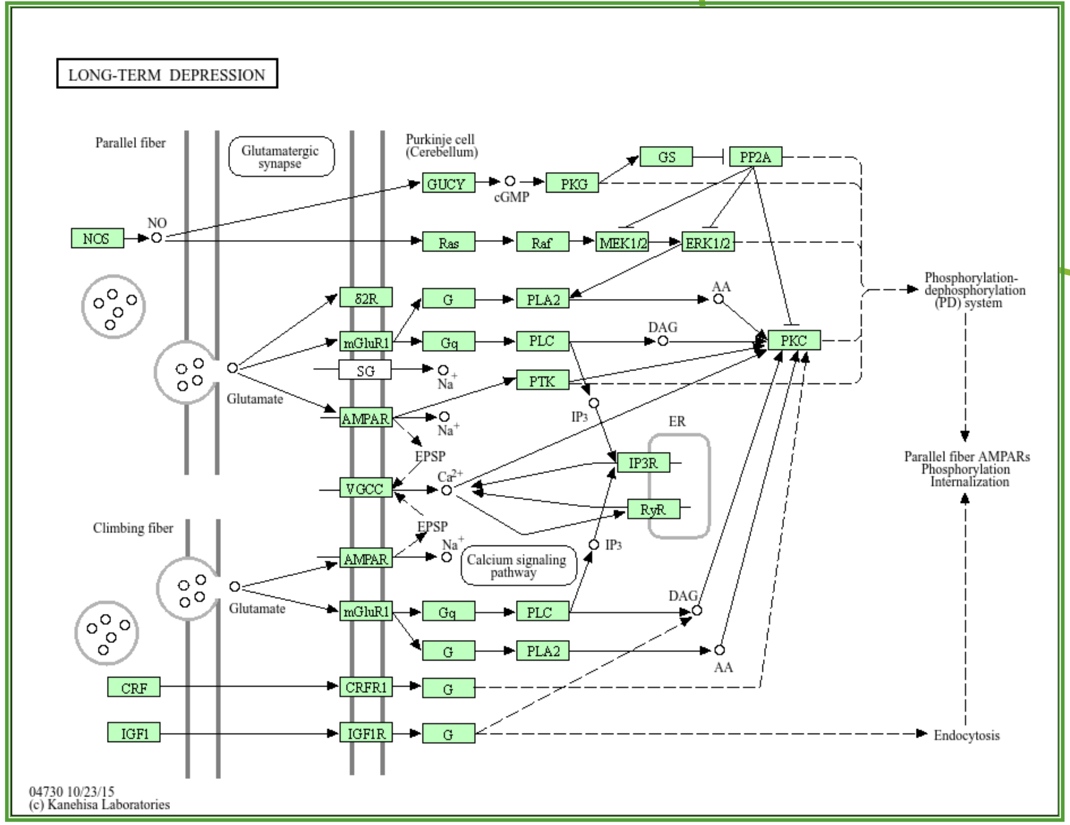
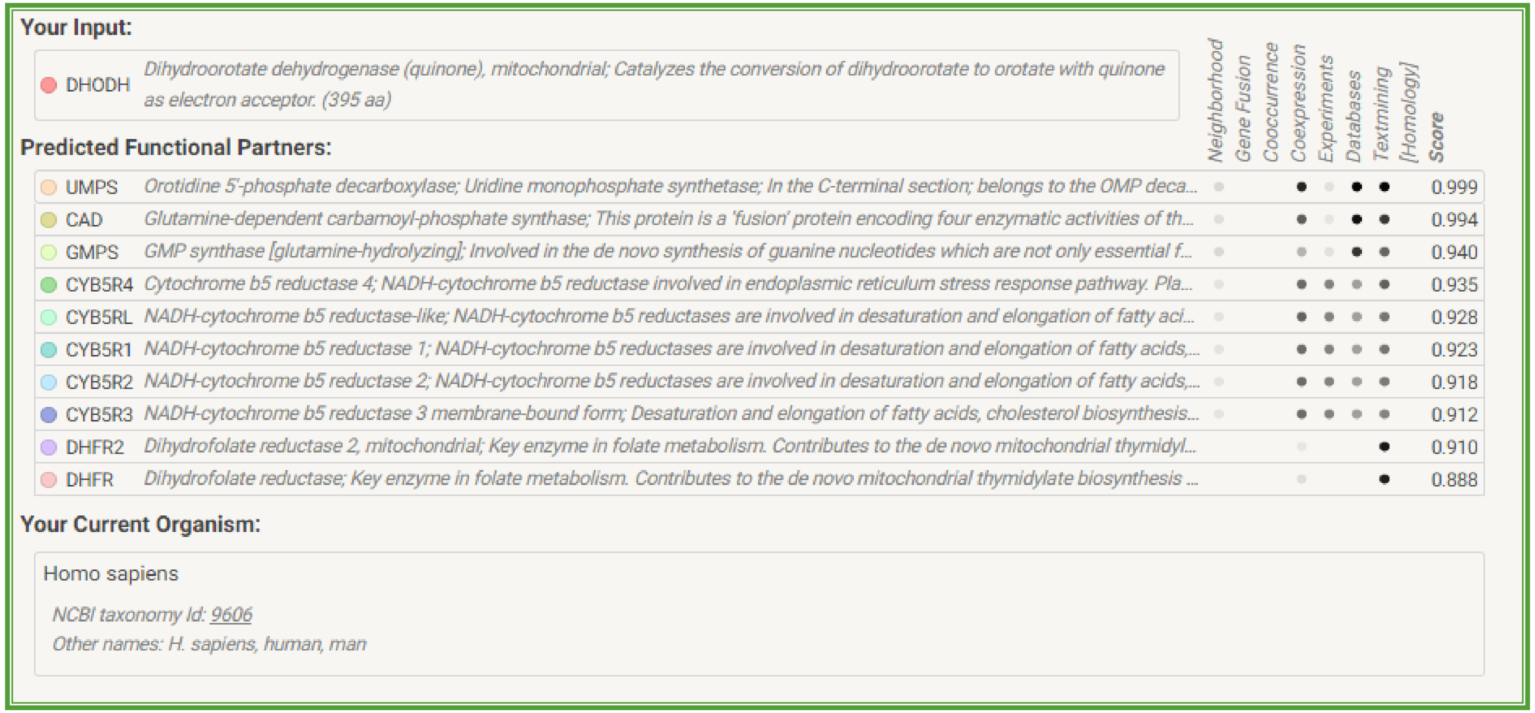


Figure: Long-term depression (LTD) pathway, Organismal Systems (KEGG). LTD refers to an activity-dependent reduction in the efficacy of neuronal synapses lasting hours or longer following a long patterned stimulus, thought to be involved in cerebellar learning.

Evidence Channels



Your Input:

● DHODH *Dihydroorotate dehydrogenase (quinone), mitochondrial; Catalyzes the conversion of dihydroorotate to orotate with quinone as electron acceptor. (395 aa)*

Predicted Functional Partners:

		Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Experiments	Databases	Textmining	[Homology]	**Score**
● UMPS	*Orotidine 5'-phosphate decarboxylase; Uridine monophosphate synthetase; In the C-terminal section; belongs to the OMP deca...*	●			●	●	●	●		0.999
● CAD	*Glutamine-dependent carbamoyl-phosphate synthase; This protein is a 'fusion' protein encoding four enzymatic activities of th...*	●			●	●	●	●		0.994
● GMPS	*GMP synthase [glutamine-hydrolyzing]; Involved in the de novo synthesis of guanine nucleotides which are not only essential f...*	●			●	●	●	●		0.940
● CYB5R4	*Cytochrome b5 reductase 4; NADH-cytochrome b5 reductase involved in endoplasmic reticulum stress response pathway. Pla...*	●			●	●	●	●		0.935
● CYB5RL	*NADH-cytochrome b5 reductase-like; NADH-cytochrome b5 reductases are involved in desaturation and elongation of fatty aci...*	●			●	●	●	●		0.928
● CYB5R1	*NADH-cytochrome b5 reductase 1; NADH-cytochrome b5 reductases are involved in desaturation and elongation of fatty acids,...*	●			●	●	●	●		0.923
● CYB5R2	*NADH-cytochrome b5 reductase 2; NADH-cytochrome b5 reductases are involved in desaturation and elongation of fatty acids,...*	●			●	●	●	●		0.918
● CYB5R3	*NADH-cytochrome b5 reductase 3 membrane-bound form; Desaturation and elongation of fatty acids, cholesterol biosynthesis...*	●			●	●	●	●		0.912
● DHFR2	*Dihydrofolate reductase 2, mitochondrial; Key enzyme in folate metabolism. Contributes to the de novo mitochondrial thymidyl...*				●			●		0.910
● DHFR	*Dihydrofolate reductase; Key enzyme in folate metabolism. Contributes to the de novo mitochondrial thymidylate biosynthesis ...*				●			●		0.888

Your Current Organism:

Homo sapiens

NCBI taxonomy Id: 9606

Other names: H. sapiens, human, man

Figure: Examplary Breakdown of the Interaction Scores for the Association network of the protein DHODH, taken from STRING

- Evidences for a single functional association can be traced back to **7 distinct sources**, termed ‘channels’
- For each channel, separate interaction scores and color codes exist
- A combined score, when multiple channels are activated
- Confidence approximation between 1 and 0: Is the association biologically meaningful?

- A minimum score puts a threshold on the confidence score
- Lower scores mean more interaction population, but it also means a higher percentage of false positives
- Scoring system ensures standardization and creates uniform metric for evaluation/integration

Role of KEGG Pathways in Evidence Channels

- KEGG indirectly influences the granularity and density of functional associations in a protein network

- Confidence scores are benchmarked using associations where both proteins are annotated.

- KEGG pathways are used as a gold standard for this

- STRING might perform better for associations involving proteins with well-established roles

- However, it may have less granularity for novel and poorly understood proteins

Organismal Origins of Evidence

Within each channel, the origin of each evidence is further subdivided into two:

- Inherent evidence
- Transferred, inter-organismic evidence

Reminder: Homology

- *Refers to biological features including genes and their products that have descended from a feature present in a common ancestor*
- **Orthologs**: Genes separated by speciation events
- **Paralogs**: Genes separated by gene duplication events

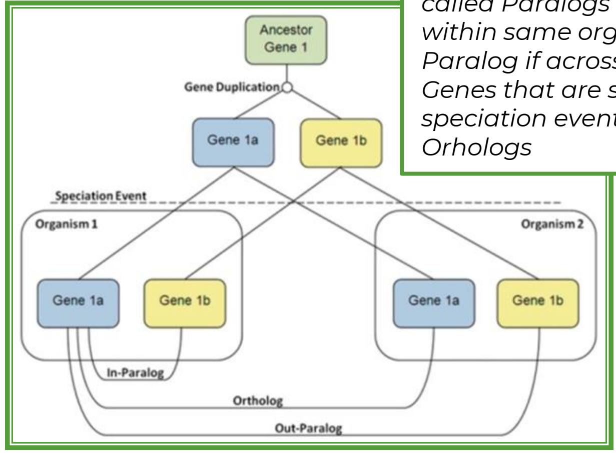


Figure: Homology Diagram; Genes that get separated by gene duplication events are called Paralogs (in-Paralog if within same organism or Out-Paralog if across organisms). Genes that are separated by speciation events are termed Orhologs

Interolog

- For the inter-organismic transfer of evidence, the Interolog concept is applied
- *conserved interactions between a pair of proteins both of which have interacting homologs (orthologs) in another organism.*
- Coverage expansion for less-studied organisms with limited direct data

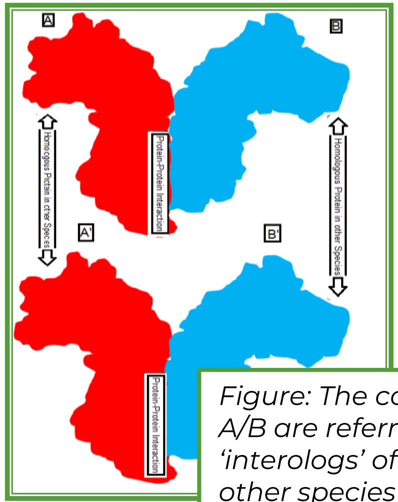
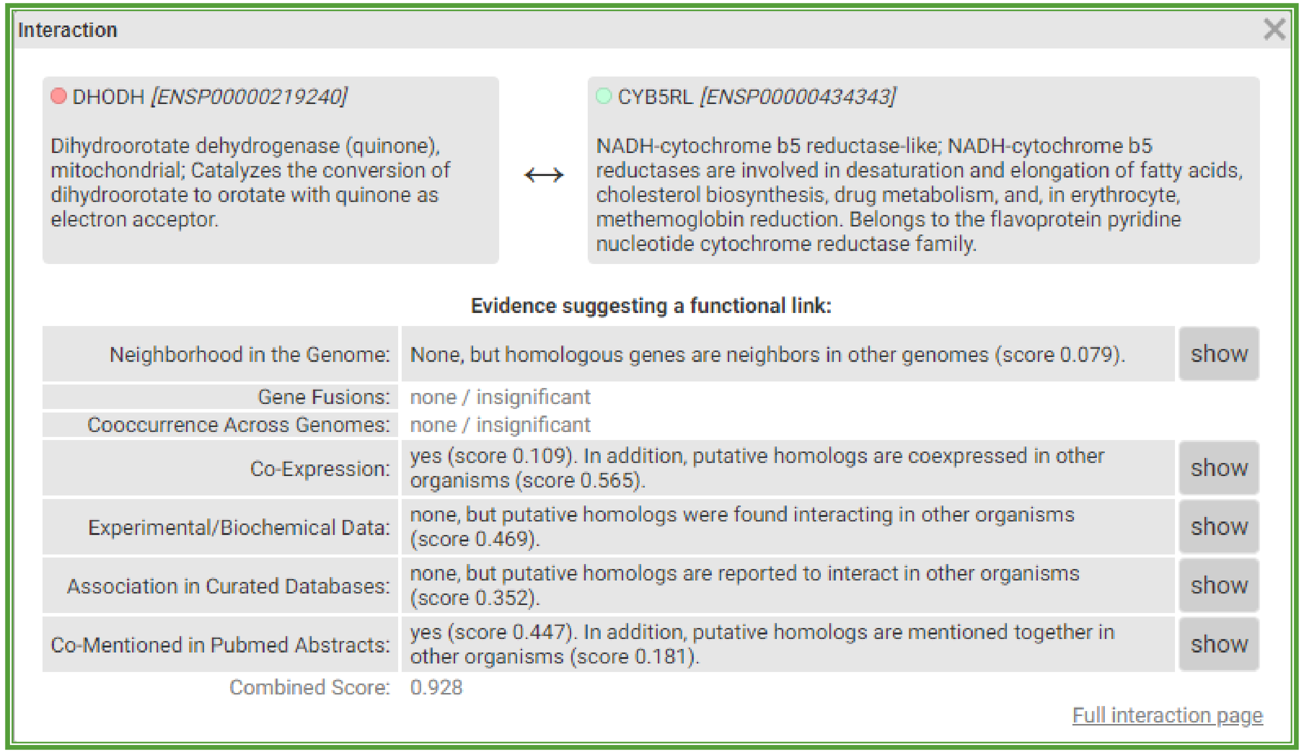

A
B
Homologous Protein in other Species
Protein-Protein Interaction
A'
B'
Homologous Protein in other Species
Protein-Protein Interaction

Figure: The conserved interactions A/B are referred to as worm 'interologs' of A'/B' interactions in other species if A' and B' are orthologs of A and, respectively B.

Organismal Origin of Evidence: Example



● DHODH *[ENSP00000219240]*

Dihydroorotate dehydrogenase (quinone), mitochondrial; Catalyzes the conversion of dihydroorotate to orotate with quinone as electron acceptor.

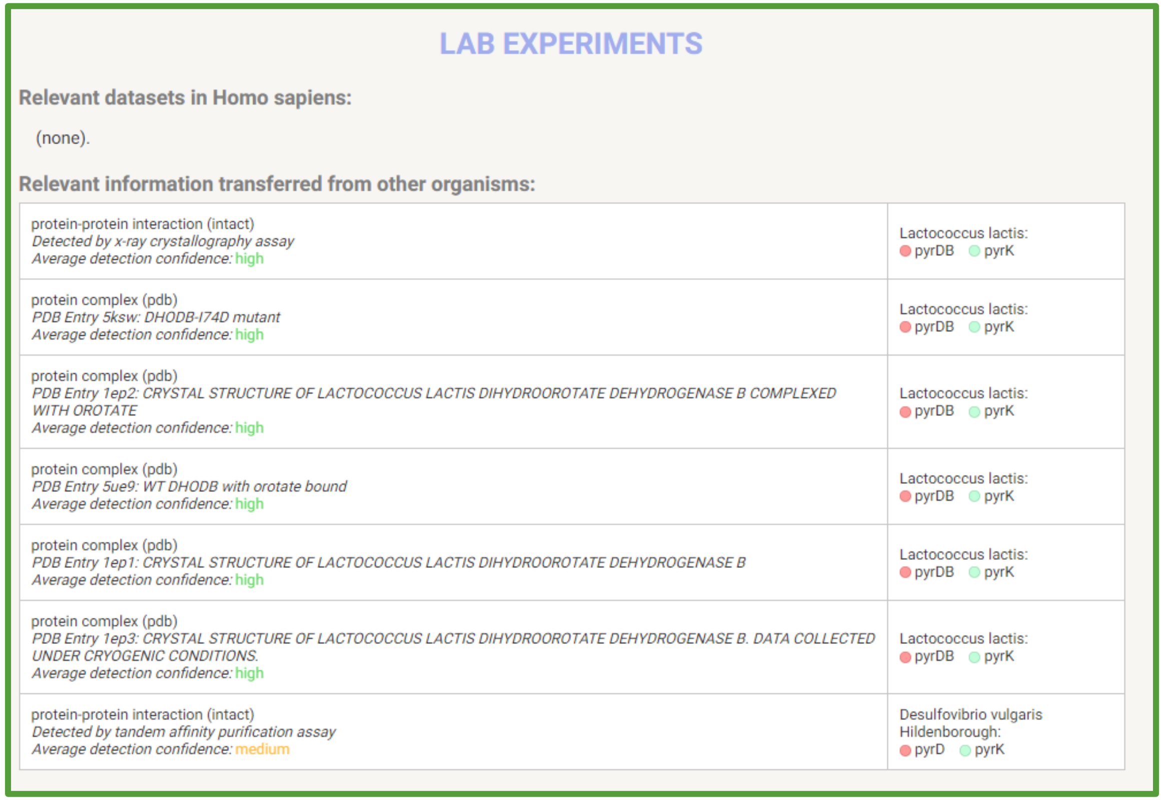
↔

○ CYB5RL *[ENSP00000434343]*

NADH-cytochrome b5 reductase-like; NADH-cytochrome b5 reductases are involved in desaturation and elongation of fatty acids, cholesterol biosynthesis, drug metabolism, and, in erythrocyte, methemoglobin reduction. Belongs to the flavoprotein pyridine nucleotide cytochrome reductase family.

Evidence suggesting a functional link:

Neighborhood in the Genome:	None, but homologous genes are neighbors in other genomes (score 0.079).	show
Gene Fusions:	none / insignificant	
Cooccurrence Across Genomes:	none / insignificant	
Co-Expression:	yes (score 0.109). In addition, putative homologs are coexpressed in other organisms (score 0.565).	show
Experimental/Biochemical Data:	none, but putative homologs were found interacting in other organisms (score 0.469).	show
Association in Curated Databases:	none, but putative homologs are reported to interact in other organisms (score 0.352).	show
Co-Mentioned in Pubmed Abstracts:	yes (score 0.447). In addition, putative homologs are mentioned together in other organisms (score 0.181).	show
Combined Score:	0.928	



LAB EXPERIMENTS

Relevant datasets in Homo sapiens:

(none).

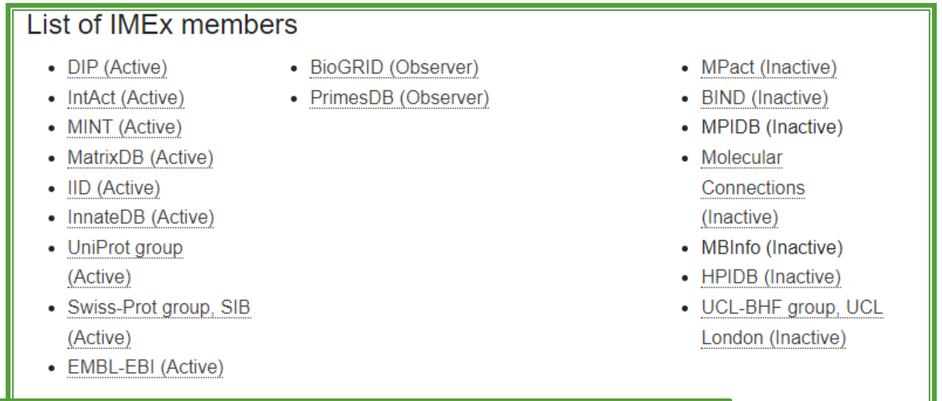
Relevant information transferred from other organisms:

protein-protein interaction (intact) *Detected by x-ray crystallography assay* *Average detection confidence:* high	Lactococcus lactis: ● pyrDB ○ pyrK
protein complex (pdb) *PDB Entry 5ksw: DHODB-I74D mutant* *Average detection confidence:* high	Lactococcus lactis: ● pyrDB ○ pyrK
protein complex (pdb) *PDB Entry 1ep2: CRYSTAL STRUCTURE OF LACTOCOCCUS LACTIS DIHYDROOROTATE DEHYDROGENASE B COMPLEXED WITH OROTATE* *Average detection confidence:* high	Lactococcus lactis: ● pyrDB ○ pyrK
protein complex (pdb) *PDB Entry 5ue9: WT DHODB with orotate bound* *Average detection confidence:* high	Lactococcus lactis: ● pyrDB ○ pyrK
protein complex (pdb) *PDB Entry 1ep1: CRYSTAL STRUCTURE OF LACTOCOCCUS LACTIS DIHYDROOROTATE DEHYDROGENASE B* *Average detection confidence:* high	Lactococcus lactis: ● pyrDB ○ pyrK
protein complex (pdb) *PDB Entry 1ep3: CRYSTAL STRUCTURE OF LACTOCOCCUS LACTIS DIHYDROOROTATE DEHYDROGENASE B. DATA COLLECTED UNDER CRYOGENIC CONDITIONS.* *Average detection confidence:* high	Lactococcus lactis: ● pyrDB ○ pyrK
protein-protein interaction (intact) *Detected by tandem affinity purification assay* *Average detection confidence:* medium	Desulfovibrio vulgaris Hildenborough: ● pyrD ○ pyrK

- Putative Genes: An alignment segment of the DNA that is believed to be a gene but the function of which remains unknown
- PDB: Protein Data Bank

Evidence Channels in Detail

‘Known’ Evidences: Experiments Channel



List of IMEx members

- DIP (Active)
- IntAct (Active)
- MINT (Active)
- MatrixDB (Active)
- IID (Active)
- InnateDB (Active)
- UniProt group (Active)
- Swiss-Prot group, SIB (Active)
- EMBL-EBI (Active)
- BioGRID (Observer)
- PrimesDB (Observer)
- MPact (Inactive)
- BIND (Inactive)
- MPIDB (Inactive)
- Molecular Connections (Inactive)
- MBInfo (Inactive)
- HPIDB (Inactive)
- UCL-BHF group, UCL London (Inactive)

Figure: Members of IMEx Consortium

- This channel collects data physical (direct) PPIs from primary interaction databases in the IMEx Consortium
- They contain *a non-redundant set of physical molecular interaction data from a broad taxonomic range of organisms.*
- High Quality and prior, consolidated knowledge

- Data pulled from the databases goes to remapping and reprocessing
- Duplicate records get merged or removed to avoid redundancy
- The information on naming and annotation gets standardized
- After cleaning, STRING benchmarks and evaluates all records against known elements in the functional pathways from KEGG maps

‘Known’ Evidences: Databases Channel

- Based on manually curated interaction records in well-established databases reviewed, covered and assembled by expert curators from alternating fields
- These databases include KEGG, Reactome, Gene Ontology and a few others
- This channel is considered highly dependable
- For this channel, STRING induces a functional association for proteins within the same biological pathway or protein complex.
- Associations are highly specific and biologically meaningful
- All data pertained in this channel is assigned the standard high-confidence score of 0.9, no further score calibration is applied.

Genomic-based, Predictive Evidences: Gene Fusion Channel

- Gene Fusion: *Formation of a gene made by joining parts of two different genes*
- May occur artificially in the laboratory or naturally in the body
- Resulting hybrid protein contains features of both, originally independent genes
- Since the fused gene expresses a single hybrid protein, it indicates that the constituent genes participated in the same pathway
- STRING gives an association score to the non-fused constituents in the genomes of other organisms
- If there are many gene fusions and the fused genes have strong orthology to their counterparts, they will get a higher association score

Exemplary Gene Fusion Table

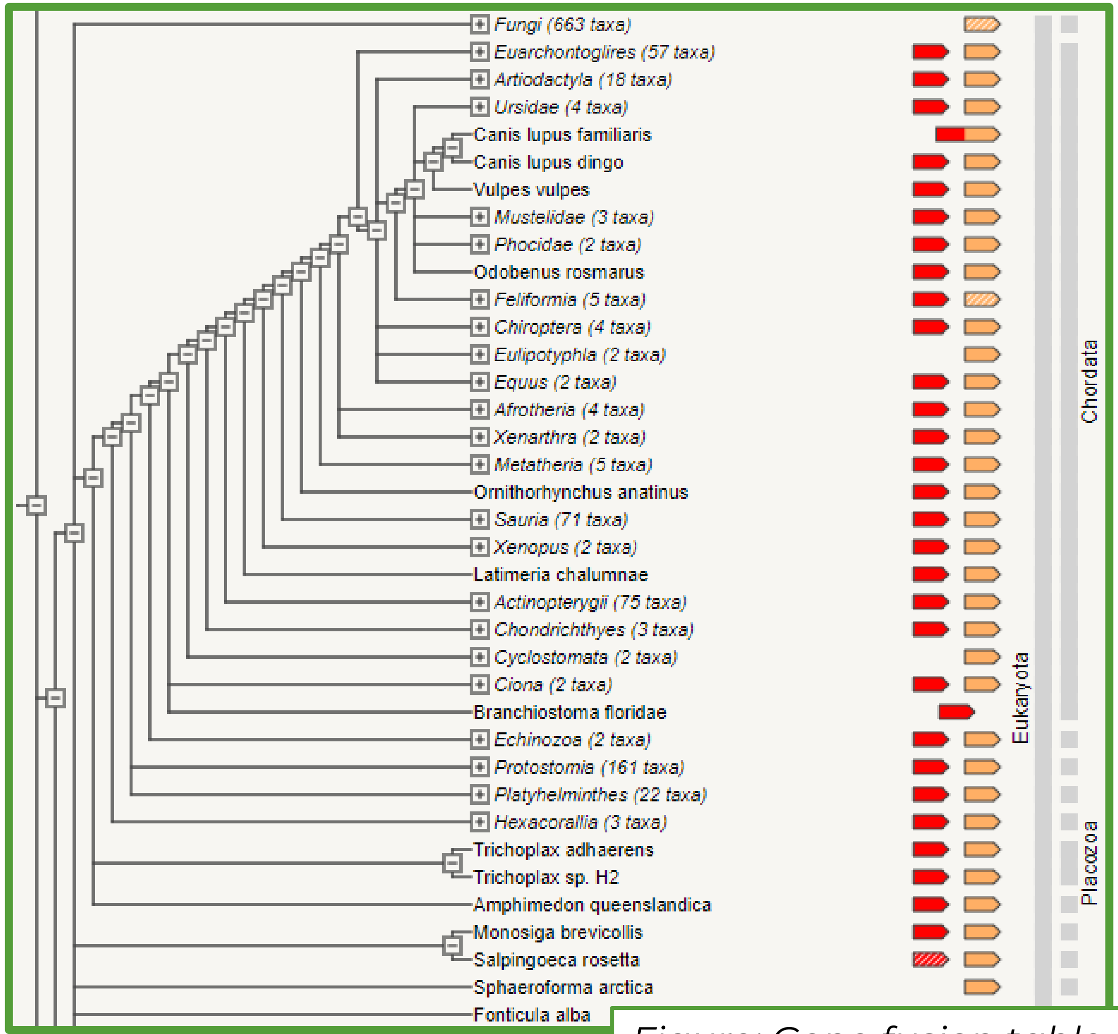
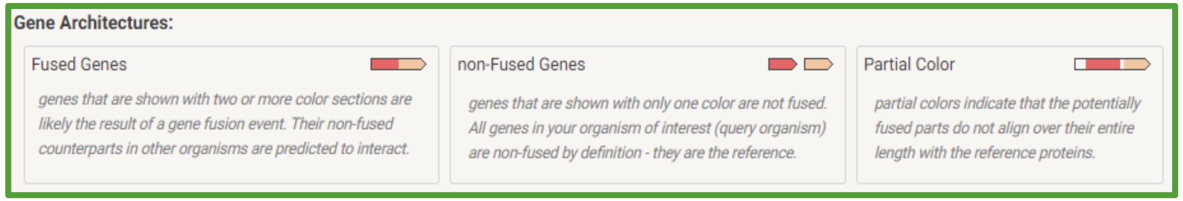


Figure: *Gene fusion table between proteins CCDC183 and RABL6*

- ▪Canis Lupus Familiaris: The dog (domesticated descendant of the wolf)

- ▪Chordata: Third largest phylum of the animal kingdom. Humans also belong to this phylum.



Gene Architectures:

Fused Genes

genes that are shown with two or more color sections are likely the result of a gene fusion event. Their non-fused counterparts in other organisms are predicted to interact.

non-Fused Genes

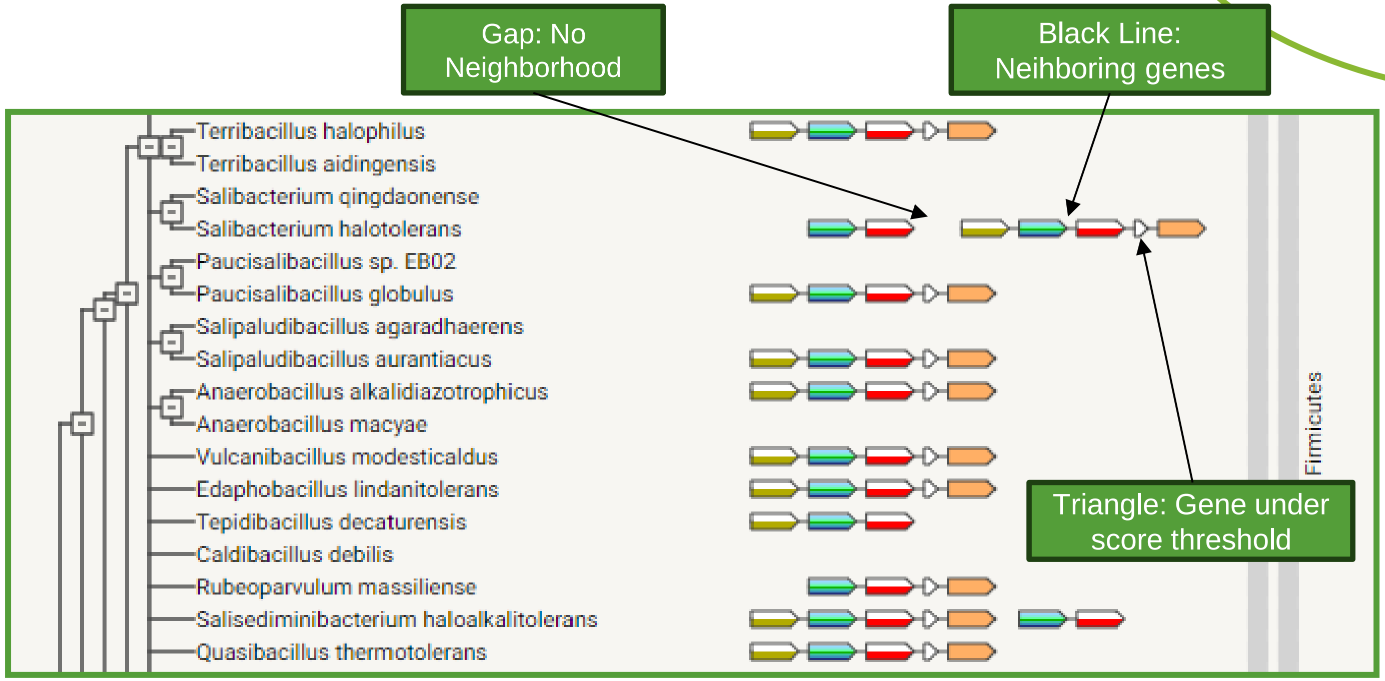
genes that are shown with only one color are not fused. All genes in your organism of interest (query organism) are non-fused by definition - they are the reference.

Partial Color

partial colors indicate that the potentially fused parts do not align over their entire length with the reference proteins.

Genomic-based, Predictive Evidences: Gene Neighborhood Channel

- Based on ‘runs’ (clusters) of genes that occur continuously in proximity
- Maximum intergenic distance is 300 base pairs
- A score threshold represents the confidence level and acts as a filter for weaker and less conserved gene associations



Gap: No Neighborhood

Black Line: Neihboring genes

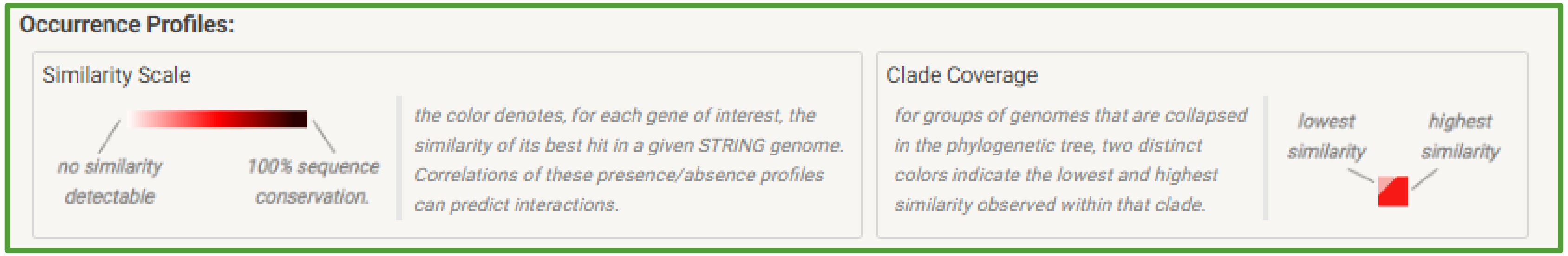
Triangle: Gene under score threshold

Terribacillus halophilus
Terribacillus aidingensis
Salibacterium qingdaonense
Salibacterium halotolerans
Paucisalibacillus sp. EB02
Paucisalibacillus globulus
Salipaludibacillus agaradhaerens
Salipaludibacillus aurantiacus
Anaerobacillus alkalidiazotrophicus
Anaerobacillus macyae
Vulcanibacillus modesticaldus
Edaphobacillus lindanitolerans
Tepidibacillus decaturensis
Caldibacillus debilis
Rubeoparvulum massiliense
Salisediminibacterium haloalkalitolerans
Quasibacillus thermotolerans
Firmicutes

- This evidence can only be derived from prokaryotic genomes
- Operon: *a genetic regulatory system in which genes coding for functionally related proteins are clustered along the DNA, only found in bacteria and viruses*
- This system helps the cell to conserve energy and controls protein synthesis
- Genes in same pathway spatially closer

Genomic-based, Predictive Evidences: Gene Co-Occurrence Channel

- relies on the fact that proteins do not function in isolation and are dependent on other proteins, either as direct binding partners, or as catalysts of substrates.
- If two proteins significantly occur together in many genomes, they are likely to be binding partners or enzymes needed for a specific metabolic pathway
- Based on this observation, a functional association may be concluded
- This channel generates a grid diagram together with a phylogeny tree that marks the presence or absence of each protein in a species



Occurrence Profiles:

Similarity Scale

the color denotes, for each gene of interest, the similarity of its best hit in a given STRING genome. Correlations of these presence/absence profiles can predict interactions.

Clade Coverage

for groups of genomes that are collapsed in the phylogenetic tree, two distinct colors indicate the lowest and highest similarity observed within that clade.

Exemplary Gene Occurrence Tables

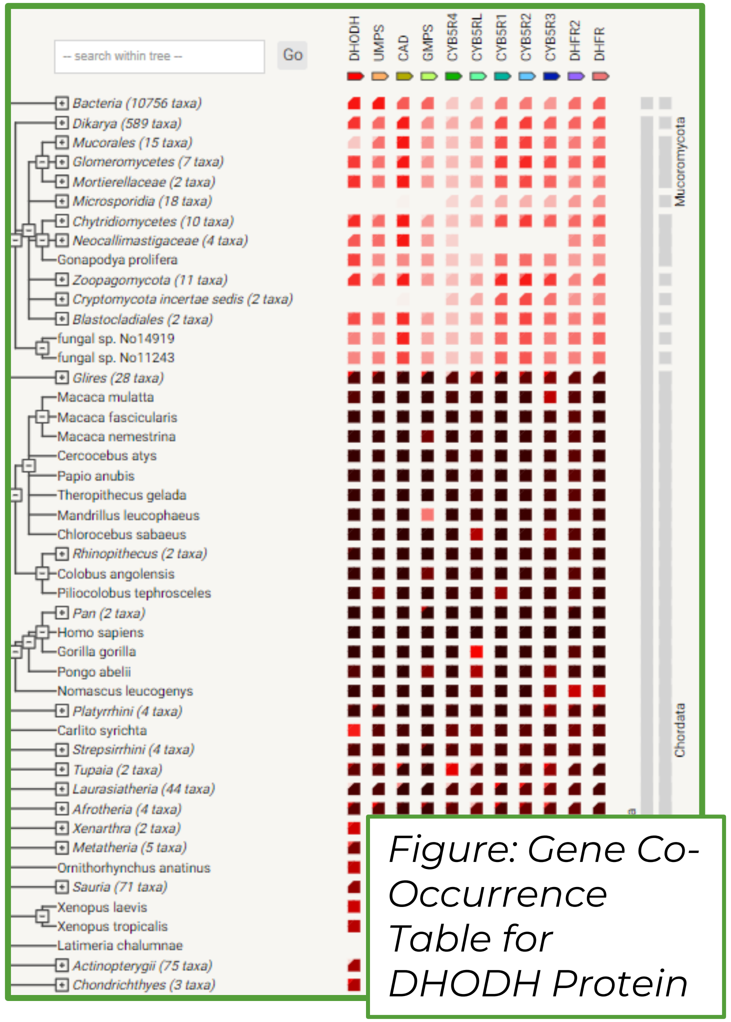


Figure: Gene Co-Occurrence Table for DHODH Protein

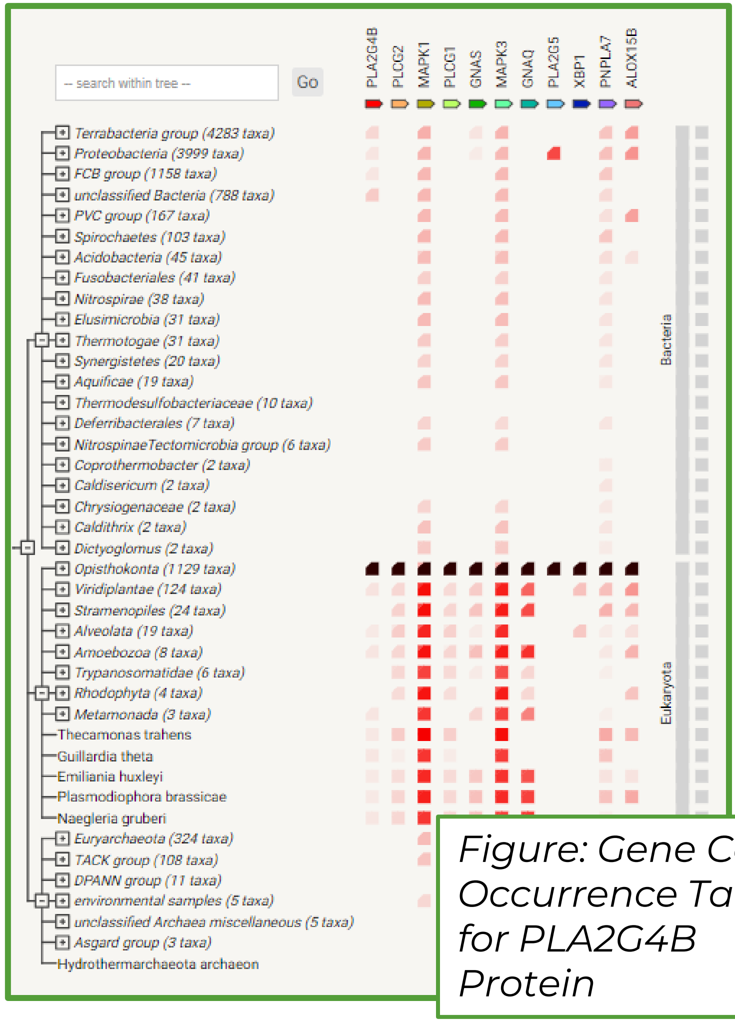


Figure: Gene Co-Occurrence Table for PLA2G4B Protein

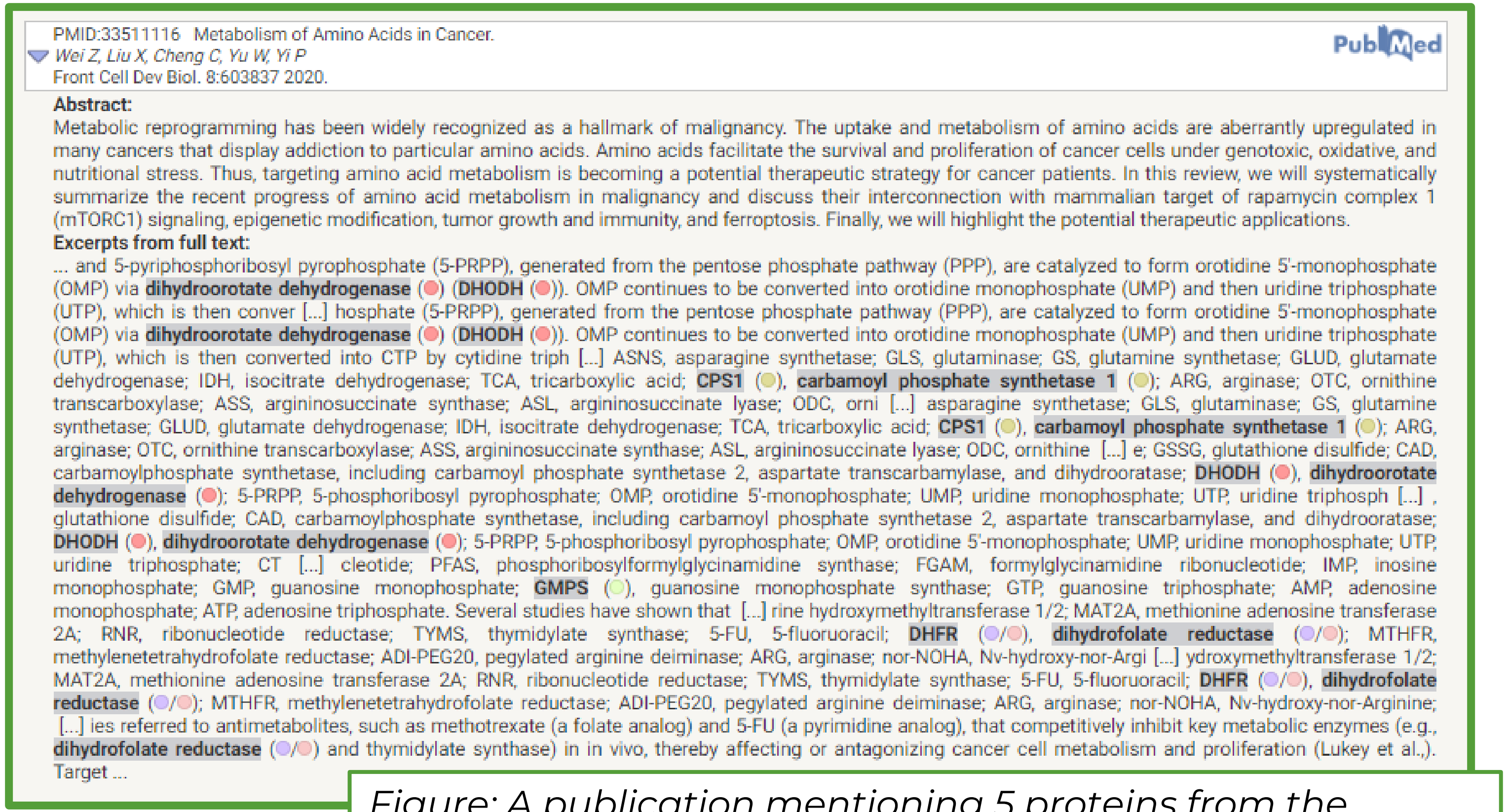
For these diagrams, the following rule of thumb applies:

- If two proteins co-occur or are “co absent” in many species, it may imply that they work together in a common biological process.
- Conversely, if one protein is present without the latter in several species, it may rule out the possibility of a functional linkage.

Purely Predictive: Text Mining Channel

- In general, text mining methods *automate the extraction of interconnected proteins through their coexistence in sentences, abstracts or paragraphs within text corpuses.*
- STRING conducts a co-citation analysis: It searches for statistically significant cases where gene names occur together in public repositories and online resources
- An advanced Text Mining technique, Natural Language Processing (NLP) of text, is also utilized: Gene names are considered as nodes and verbs as edges
- This gives proteins semantic notion on the graphs
- STRING's current text corpus for the text mining channel consists of:
    - 2,106,542 full-text articles
    - 26,473,095 PubMed abstracts and other sources

Exemplary Text Mining Query



PMID:33511116 Metabolism of Amino Acids in Cancer.
Wei Z, Liu X, Cheng C, Yu W, Yi P
Front Cell Dev Biol. 8:603837 2020.

Abstract:
Metabolic reprogramming has been widely recognized as a hallmark of malignancy. The uptake and metabolism of amino acids are aberrantly upregulated in many cancers that display addiction to particular amino acids. Amino acids facilitate the survival and proliferation of cancer cells under genotoxic, oxidative, and nutritional stress. Thus, targeting amino acid metabolism is becoming a potential therapeutic strategy for cancer patients. In this review, we will systematically summarize the recent progress of amino acid metabolism in malignancy and discuss their interconnection with mammalian target of rapamycin complex 1 (mTORC1) signaling, epigenetic modification, tumor growth and immunity, and ferroptosis. Finally, we will highlight the potential therapeutic applications.

Excerpts from full text:
... and 5-pyriphosphoribosyl pyrophosphate (5-PRPP), generated from the pentose phosphate pathway (PPP), are catalyzed to form orotidine 5'-monophosphate (OMP) via **dihydroorotate dehydrogenase** (●) (**DHODH** (●)). OMP continues to be converted into orotidine monophosphate (UMP) and then uridine triphosphate (UTP), which is then conver [...] hosphate (5-PRPP), generated from the pentose phosphate pathway (PPP), are catalyzed to form orotidine 5'-monophosphate (OMP) via **dihydroorotate dehydrogenase** (●) (**DHODH** (●)). OMP continues to be converted into orotidine monophosphate (UMP) and then uridine triphosphate (UTP), which is then converted into CTP by cytidine triph [...] ASNS, asparagine synthetase; GLS, glutaminase; GS, glutamine synthetase; GLUD, glutamate dehydrogenase; IDH, isocitrate dehydrogenase; TCA, tricarboxylic acid; **CPS1** (●), **carbamoyl phosphate synthetase 1** (●); ARG, arginase; OTC, ornithine transcarboxylase; ASS, argininosuccinate synthase; ASL, argininosuccinate lyase; ODC, orni [...] asparagine synthetase; GLS, glutaminase; GS, glutamine synthetase; GLUD, glutamate dehydrogenase; IDH, isocitrate dehydrogenase; TCA, tricarboxylic acid; **CPS1** (●), **carbamoyl phosphate synthetase 1** (●); ARG, arginase; OTC, ornithine transcarboxylase; ASS, argininosuccinate synthase; ASL, argininosuccinate lyase; ODC, ornithine [...] e; GSSG, glutathione disulfide; CAD, carbamoylphosphate synthetase, including carbamoyl phosphate synthetase 2, aspartate transcarbamylase, and dihydrooratase; **DHODH** (●), **dihydroorotate dehydrogenase** (●); 5-PRPP, 5-phosphoribosyl pyrophosphate; OMP, orotidine 5'-monophosphate; UMP, uridine monophosphate; UTP, uridine triphosph [...] , glutathione disulfide; CAD, carbamoylphosphate synthetase, including carbamoyl phosphate synthetase 2, aspartate transcarbamylase, and dihydrooratase; **DHODH** (●), **dihydroorotate dehydrogenase** (●); 5-PRPP, 5-phosphoribosyl pyrophosphate; OMP, orotidine 5'-monophosphate; UMP, uridine monophosphate; UTP, uridine triphosphate; CT [...] cleotide; PFAS, phosphoribosylformylglycinamidine synthase; FGAM, formylglycinamidine ribonucleotide; IMP, inosine monophosphate; GMP, guanosine monophosphate; **GMPS** (●), guanosine monophosphate synthase; GTP, guanosine triphosphate; AMP, adenosine monophosphate; ATP, adenosine triphosphate. Several studies have shown that [...] rine hydroxymethyltransferase 1/2; MAT2A, methionine adenosine transferase 2A; RNR, ribonucleotide reductase; TYMS, thymidylate synthase; 5-FU, 5-fluoruoracil; **DHFR** (●/●), **dihydrofolate reductase** (●/●); MTHFR, methylenetetrahydrofolate reductase; ADI-PEG20, pegylated arginine deiminase; ARG, arginase; nor-NOHA, Nv-hydroxy-nor-Argi [...] ydroxymethyltransferase 1/2; MAT2A, methionine adenosine transferase 2A; RNR, ribonucleotide reductase; TYMS, thymidylate synthase; 5-FU, 5-fluoruoracil; **DHFR** (●/●), **dihydrofolate reductase** (●/●); MTHFR, methylenetetrahydrofolate reductase; ADI-PEG20, pegylated arginine deiminase; ARG, arginase; nor-NOHA, Nv-hydroxy-nor-Arginine; [...] ies referred to antimetabolites, such as methotrexate (a folate analog) and 5-FU (a pyrimidine analog), that competitively inhibit key metabolic enzymes (e.g., **dihydrofolate reductase** (●/●) and thymidylate synthase) in in vivo, thereby affecting or antagonizing cancer cell metabolism and proliferation (Lukey et al.). Target ...

Figure: A publication mentioning 5 proteins from the association network of DHODH, serving as evidence for the text mining channel.

- Each publication in this channel mentions at least two proteins in the queried protein network
- Associations are given higher confidence scores if they appear in proximity (same paragraph or sentence)
- In ambiguities regarding homology, a gene name may be assigned more than one nodes (proteins)

Rising Interest for Text Mining Tools

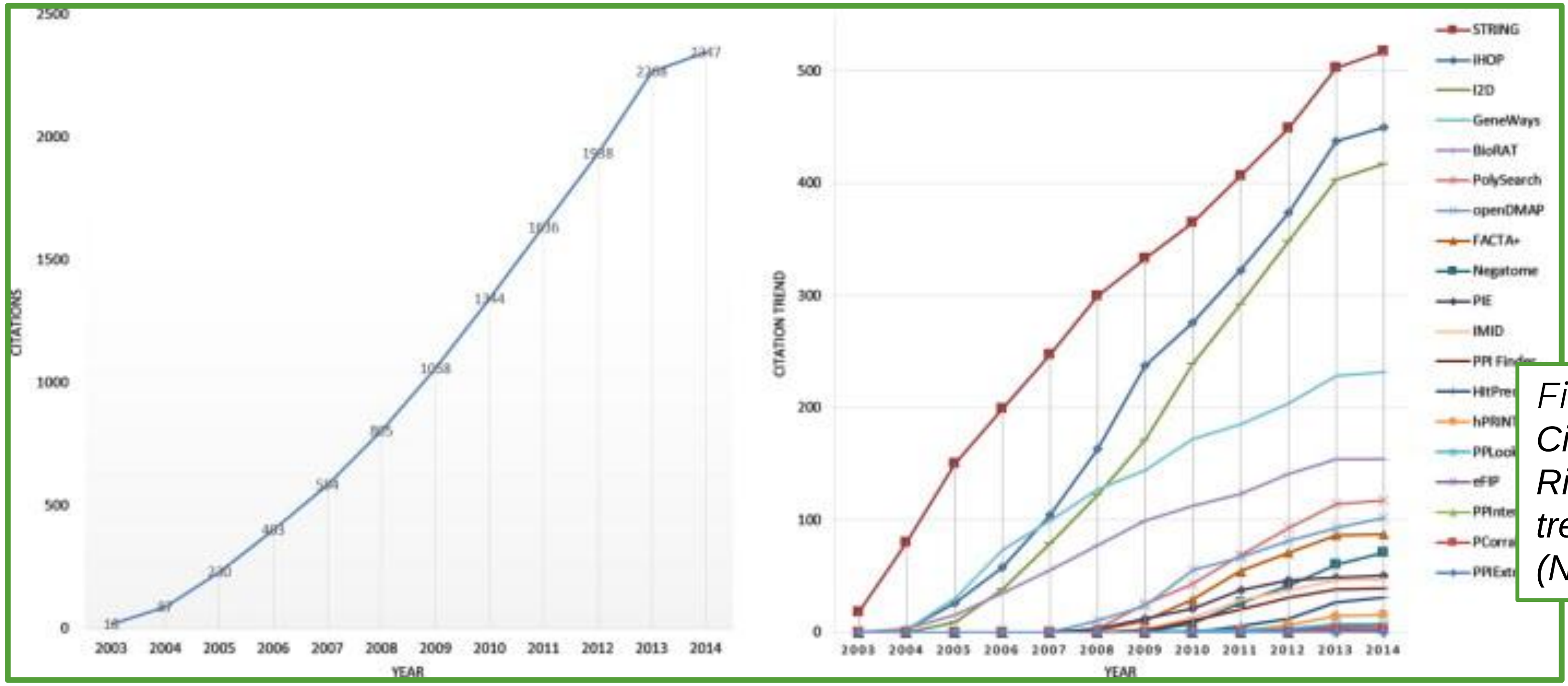
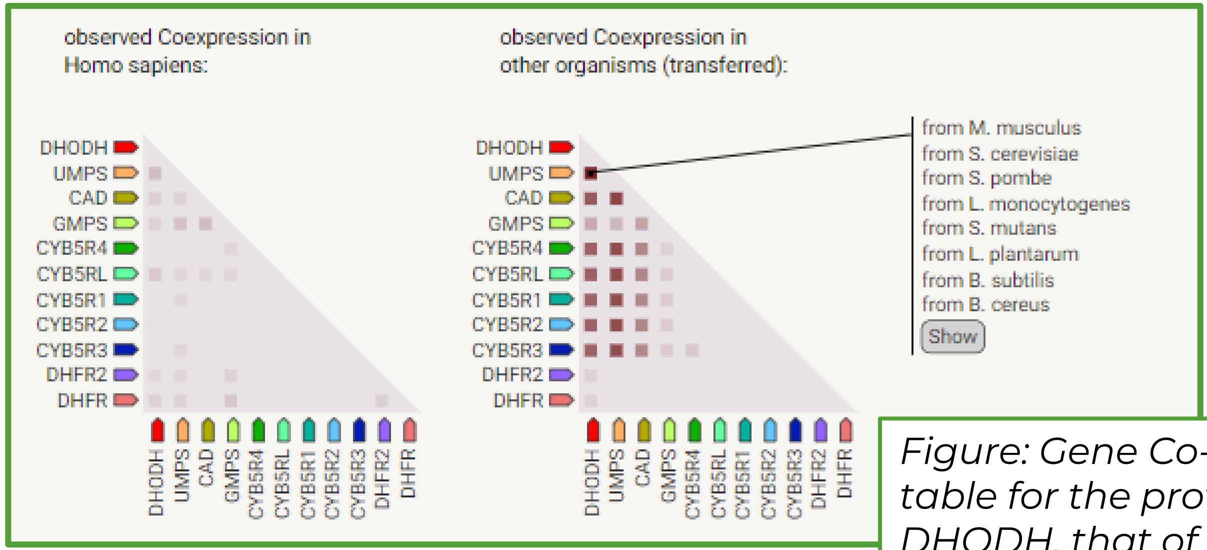


Figure: Left: Google-Scholar Citations for text-mining Tools. Right: Google-Scholar citation trends for each text-mining tool (Nikolas Papanikolaou, 2015)

- STRING retains its reputation as the most cited text mining tool through the years
- STRING is a well-recognized, widely-used and utile tool in the scientific community for extracting and analyzing PPIs from literature
- Provides comprehensive coverage for protein-protein association data

Purely Predictive: Co-Expression Channel

- Co-Expression analysis involves *identifying group of genes that show similar expression patterns under* different biological conditions *across tens or hundreds of experiments regardless of their differential expression*
- STRING gathers expression data from proteome and transcriptome measurements and conducts gene-by-gene correlation analysis
- This analysis results in a co-expression table that displays a combined score for each protein pair



observed Coexpression in Homo sapiens:

observed Coexpression in other organisms (transferred):

Figure: Gene Co-expression table for the protein DHODH, that of both the target organism and other organisms

- The intensity of color red indicates the level of confidence that two proteins are functionally associated, given the overall expression data

Transcriptome-Wide Co-Expression Analysis

- Refers to the comprehensive analysis of complete set of RNA transcripts (mainly mRNA) produced by the genome in a cell
- Measurement of the mRNA levels show the potential for protein production, not the actual quantity of proteins
- Data for this type of analysis gathered from experiments archived in the NCBI Gene Expression Omnibus
- This *is an international public repository that archives and freely distributes microarray, next-generation-sequencing and other forms of high-throughput functional genomics data submitted by the research community.*

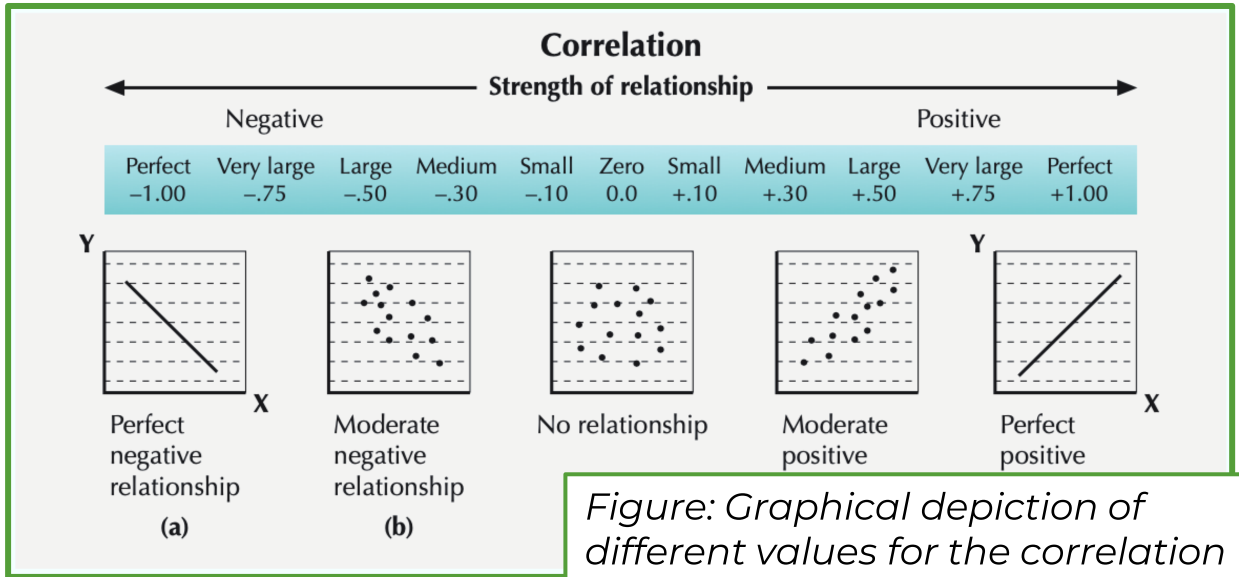


Figure: Graphical depiction of different values for the correlation coefficient between -1.0 and 1.0

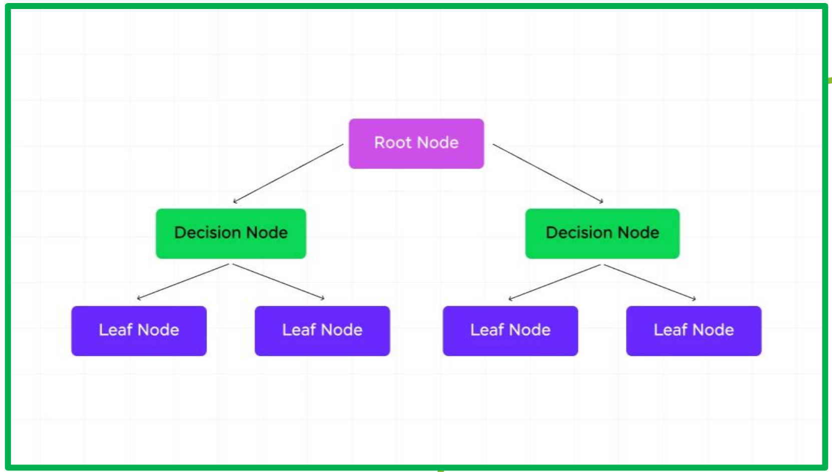
- STRING normalizes, prunes and compares the data on expression profiles over a large variety of conditions.

- For determining whether a correlation exists, STRING measures the Pearson Correlation Coefficient between different transcript levels

Proteome-Wide Co-Expression Analysis

- Refers to the quantitative analysis of the complete set of proteins expressed in cell, tissue or organism
- Measurement of protein levels show actual biological activity within the cell
- Currently, co-expression data can only be retrieved from Proteome HD
- This is a dataset that *stores data on protein abundance changes in response to biological perturbations from 294 different biological conditions in human cells*

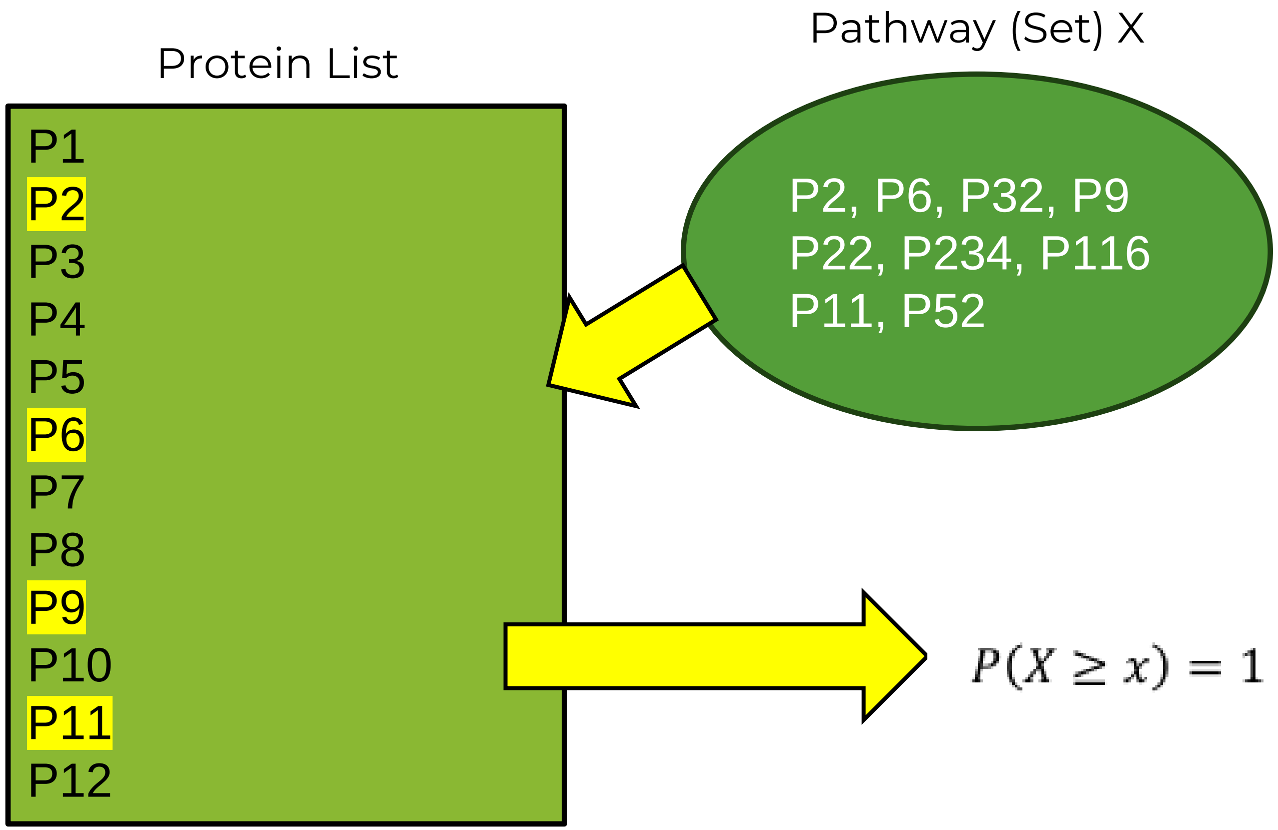
- *To determine correlation between variables, ProteomeHD utilizes the treeClust Algorithm from R's treeClust package*
- *This algorithm uses a set of classification or regression trees to build an inter-point dissimilarity in which two points are similar when they tend to fall in the same leaves of trees*
- *Statistical question: How frequently does proteins end up in the same leaf?*
- If they consistently group together, they are considered co expressed and correlated.



STRING: Enrichment Mode

Overrepresentation Analysis (ORA)

- *used to determine which a priori defined gene sets are more present (over-represented) in a subset of "interesting" genes than what would be expected by chance*
- Only allows a protein list as input



N: #Background Proteins
(e.g. whole Proteome)

M: #Proteins in Reference Set X

n: # ''interesting'' Proteins from the List in Pathway X

P(X>x): Probability of observing x or more "interesting" genes (evidence for enrichment)

Hypergeometric distribution

$$P(X \geq x) = 1 - P(X \leq x - 1) = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{N-M}{i}}{\binom{N}{n}}$$

Limitations of ORA

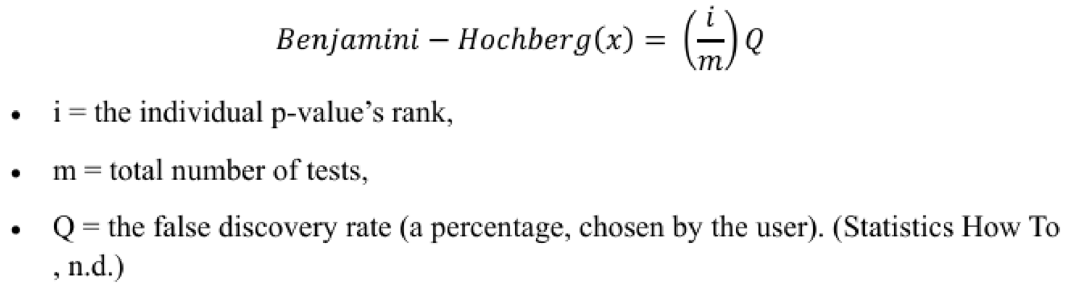
- ORA may fall short of reliably presenting accurate results, especially for large lists
- It discards 3 information regarding the data the list might have had:
  - The original list might have been much longer, and the user would have had to trim it.
  - The items in the list might have been ranked meaningfully, or
  - Each protein might have been assigned some numerical information or meta-data from the underlying experiment

Aggregate Fold Change

- Permutation-based method
- Allows genome-scale, large input, where each gene/protein should be assigned a numerical value
- Example numerical values: gene expression, protein abundance, fold change, p-values etc.
- STRING calculates averages of the numerical values assigned to proteins representing the overall behavior, for each gene set to be tested
- This average is compared to averages of randomized gene sets of the same size in each functional pathway framework (KEGG, Gene Ontology, InterPro, etc.)

Multiple Testing Correction

- Multiple gene sets are being tested for significant enrichment based on the user provided set
- Each test yields a p-value, but testing many gene sets increases the likelihood of false discoveries (The observation is actually a result of chance)
- STRING ranks p-values within each pathway framework from higher to lower and applies Benjamini-Hochberg procedure
- This is done to adjust p-values and control false discovery rate among all p-values
- Remaining p-values after the correction are considered correct (The largest p-value that is smaller than the critical value (FDR) is the basepoint for trimming)



$$Benjamini - Hochberg(x) = \left(\frac{i}{m}\right)Q$$

- i = the individual p-value’s rank,
- m = total number of tests,
- Q = the false discovery rate (a percentage, chosen by the user). (Statistics How To , n.d.)

Yeast-2-Hybrid Screening System

Yeast-two-Hybrid Screening

- Captures unprecedented physical protein-protein interactions

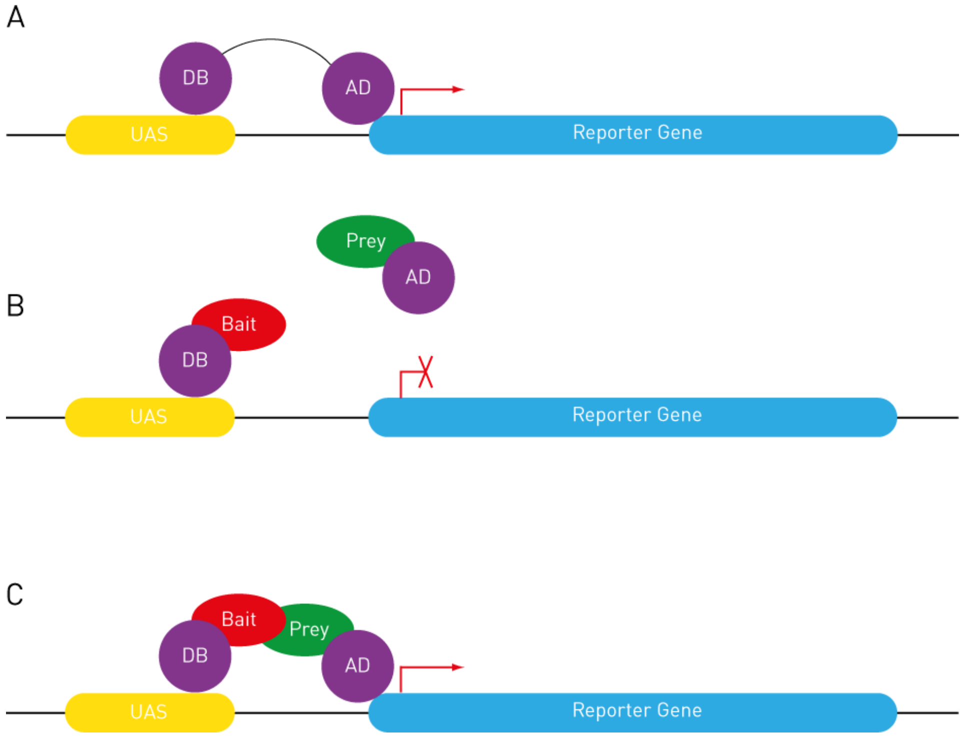
- Makes use of transcription factors

- Upstream Activating sequence (Transcription factors bind to this region)

- Transcription factors consist of: DNA-Binding Domain (DBD) and Activation Domain (AD)

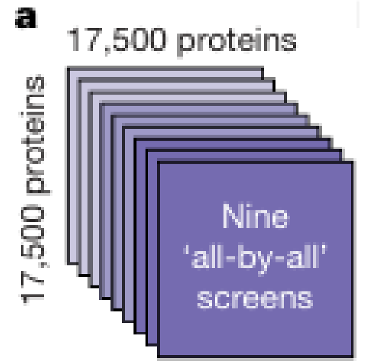
- If Bait and Prey interact, the gene gets transcribed -> this implies that a PPI exists

- If Bait and Prey don't interact, the gene is not transcribed -> this implies that a PPI does not exist



HuRI: Generation and Characterization

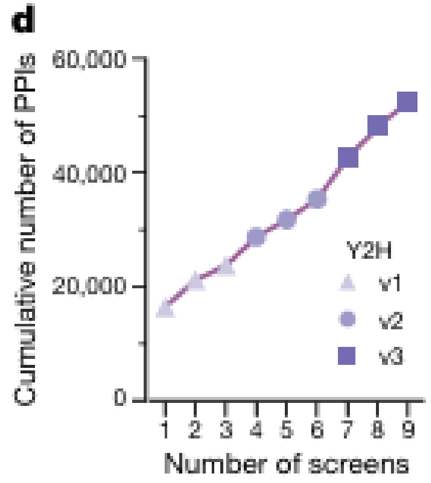
Generation of HuRI



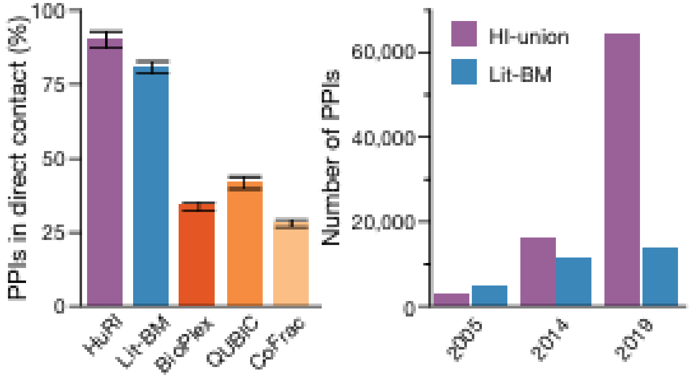
- Three variants of the Yeast-two-Hybrid were utilized
- 17,408 Open reading frames were scanned (ORFeome v9.1)
- Open Reading Frame: spans of DNA sequence between start and stop codons that potentially code for a protein
- 9 ''all-by-all" screenings of the search space (1,3 billion screenings)

$$\boxed{\frac{17.500\ \times 17500}{2} \times 9 = \ 1378125000}$$

- After every three screening, Y2H variants were switched
- Y2H variants are complementary and PPIs detected rise with every screening done cumulatively



Coverage of HuRI



- 64,056 PPIs captured involving 9,094 Proteins
- Most complete collection of high-quality, direct PPI data up-to-date

However …

- The Human Interactome is estimated to contain between 130,000 and 600,000 PPIs
- *These include interactions of structural proteins inside the cell, and multi-protein complexes that are involved in core processes such as transcription and translation, cell-cell adhesion and communication, protein synthesis and degradation, cell cycle control and signaling cascades.*
- *Based on these numbers, HuRI represents 2-11% of the entire binary protein Interactome*

Functional Relationships in HuRI

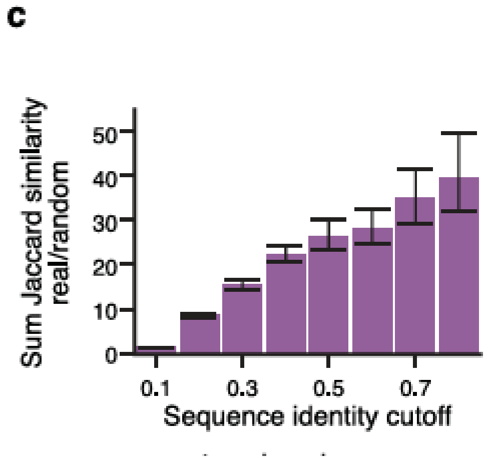
- Proteins with similar interaction interfaces tend to share interaction partners

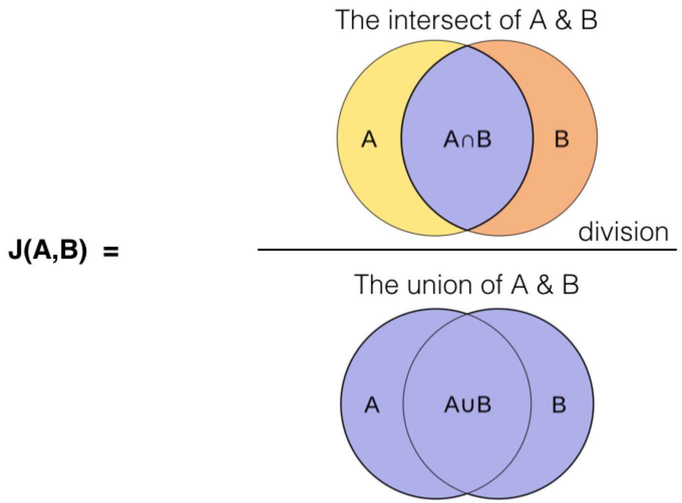
- Interaction Interface: *specific residual regions or surface areas of a protein that contact with residues from the other interacting protein*

- Sequence identity does not directly imply functional similarity (Jaccard Index)

- In fact, profile similarity ≥ 0.5 only exhibit ≤ 0.2 sequence identity

- Example: Protein TMEM258 and C19ORF18 share 80% of their partners but only have 10% sequence identity





Where is HuRI applicable?

Mechanims of Tissue-Specific Disease (Mendelian)

- Mendelian Diseases are caused by mutations in one gene

- When the abnormal, uniformly expressed protein interacts with its TiP interaction partner, the PPI gets perturbated

- These perturbations are thought to be the reason behind tissue-specific phenotypes of Mendelian diseases

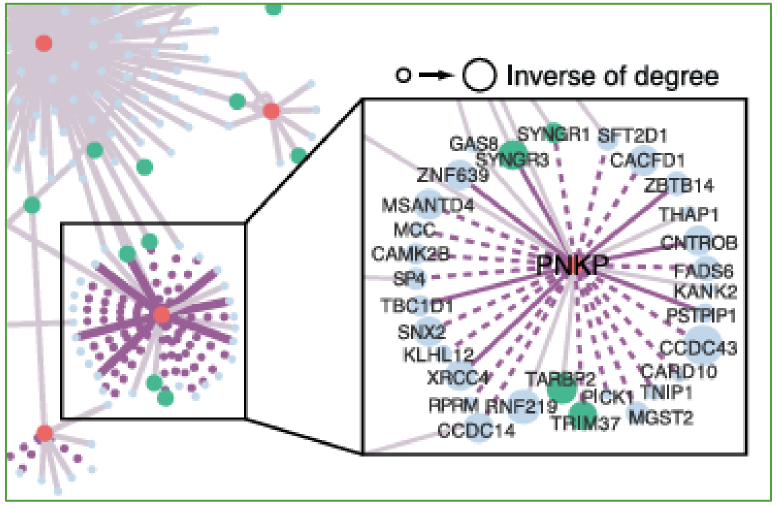
- PNKP normally partakes in PPIs in the brain

- PNKP with mutation Glu326LYs causes microcephaly

- In the concentric Graph many PPIs are perturbated

- TRIM37 is known to facilitate DNA Repair

- HuRI can provide coverage and a point of reference in disease related contexts


Inverse of degree
PNKP
GAS8
SYNGR1
SFT2D1
SYNGR3
CACFD1
ZNF639
ZBTB14
MSANTD4
THAP1
MCC
CNTROB
CAMK2B
FADS6
SP4
KANK2
TBC1D1
PSTPIP1
SNX2
CCDC43
KLHL12
CARD10
XRCC4
TARBP2
TNIP1
PICK1
RPRM
RNF219
MGST2
TRIM37
CCDC14