TWO DIFFERENT APPROACHES TO INTERACTOME MAPPING

Demir Kurt

Introduction

Proteins are essential biomolecular players that have a direct relation to how a vast range of cellular processes and biological functions come about within living cells. However, they do not act one their own but rather rely on one another to 'drive' such processes and functions, making it crucial to explore and analyze the entirety of direct and indirect interactions between these products of gene expression – which is termed the interactome. Such an effort would hold valuable results in understanding the underlying mechanisms in phenotypic diversity and tissue-specific variability and vast implications for fields such as drug discovery, biomarker identification and functional genomics. In fact*, 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*. (Fontanillo, 2010)

Addressing this goal, the papers *'A reference map of the human reference interactome'*, published on April 8, 2020, and *'***STRING** *v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets',* published on November 22, 2018, along with the complementary paper *'The* **STRING** *database in 2021'*, published in 2020, present two different database approaches to map protein-protein interactions. Respectively, the former offer and advocate for a human-centric database that is specifically dedicated to mapping direct, physical protein-protein interactions (PPIs) comprehensively within humans and is based primarily on experimental data from high-throughput methods. The latter, on the other hand, offers a broader database that aims to collect, score and integrate publicly available information on both direct and indirect (functional) protein-protein interactions across many organisms, and to complement them with a variety of computational prediction methods, into a single body. While **HuRI** is resemblant of 'reference book' or 'dictionary' which serves as supplement in scientific research, **STRING** should be viewed as more of a tool with which scientific research can be done.

This report will delve into the generation and technical details of both databases. However, the general structures of the papers are abundantly different from one another and the scope and methodologies of the databases are dissimilar, which deems a clear and organized comparison across them inconvenient. Their details will be explained linearly, but comparisons will be made only when possible. In addition, the report will give excursuses to define and explain the terms and concept that are vital at understanding the core ideas and methodologies of the papers when seen necessary.

General Comparison: Numbers and Methodologies

The paper on **HuRI** commences by underlining a problem within molecular biology: Despite having assembled a reference human genome and having advanced in human sequencing technology, our grasp on underlying mechanisms of cellular function, organization and variability remains limited. The scrutinizing of gene expression products, namely proteins, and gathering them into a uniform reference map is seen by the authors as the only valid solution to the problem.

However, there are quite a few ways to approach this issue, namely Interactome Mapping. In the introductive part of the paper on **STRING** v11, authors shortly explain the diverse types of databases dedicated to protein association networks, list examples for each individual category and determine where **STRING** falls under among the various categories. The same process can be applied to **HuRI**, but only after each category has been explained thoroughly.

Protein-Protein Interaction (PPI) databases can be classified into three main types based on the prominent methodologies they utilize, as conveyed by the **STRING** v11 paper.

Primary Interaction Databases:

These databases collect and organize experimentally validated PPIs, providing data that is directly confirmed through laboratory techniques like **yeast two-hybrid (Y2H)** or **co-immunoprecipitation**. This ensures that the data is reliable and backed by real-world evidence. In addition, they might also maintain additional, equally important meta-data on the newly discovered interaction such as experimental conditions, methods employed and other relevant information, adding depth to the data and making comparisons and the elimination of biases easier.

As explained in the paper on **STRING** v11, some 10+ primary interaction databases are also working under the International Molecular Exchange (IMEx) Consortium, which helps to standardize the way protein interactions are documented. For example, they *develop to work a single set of curation rules when capturing data from directly deposited interaction data, preprints and publication* or make *the captured interactions available in a single search interface on a common website.* (IMeX Consortium, n.d.) This cooperation improves the overall quality and consistency of the data.

Computational Prediction Databases:

These databases predict potential PPIs using computational models and algorithms based on genetic information, protein structures, or known biological networks. They provide broader, yet sometimes less confirmed, insights into interactions. While these predictions are useful for generating hypotheses or guiding experimental research, they are less certain than direct experimental results.

A Mix of Both Worlds:

This last class of databases combines both **experimentally validated PPIs** and **computationally predicted interactions**, offering a comprehensive network of protein associations. They integrate multiple data sources, providing users with a richer, more detailed picture of the interactions, including both **physical and functional associations**.

In their paper for version 11.5, the authors of **STRING** transmit their confidence in this database class by stating that *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.* (Szklarczyk, 2020)

Where STRING and HuRI Fall in Terms of Number and Methods:

STRING is an integrative database that merges both experimentally detected and computationally predicted PPIs and, thus, falls into the third category. Its evidence sources include, but are not limited to, text-mining, lab experiments, genetic co-occurrence, co-expression, and gene fusion. In addition to the completeness of its evidence sources, **STRING** also distinguishes itself within its respective category through its usability features such as customization, enrichment detection and programmatic access, through which it paves the way for novel discoveries. It is also worth noting that **STRING** is also a member of the IMEx Consortium.

In terms of numbers, **STRING** places its focus primarily on the coverage of genome-sequenced organisms, providing coverage to more than 14000 organisms in its current version (11.5). This is a huge leap from its previous version 11.0 in 2018, when it covered only up to 5090 organisms and 24 million proteins.

On the other hand, **HuRI** is a PPI database that is based purely on experimental evidence and, thus, belongs to ‘primary interaction’ databases. It lays its focus on scrutinizing and mapping the human interactome. The PPIs within the database were captured predominantly through the yeast-two-hybrid method (explained later) which is used to identify direct physical interactions between pairs of proteins. In fact, yeast-two-hybrid

(Y2H) is highlighted within the paper on **HuRI** as the *only binary PPI assay capable of screening the human proteome with sufficient throughput.*

Previously, the authors used Y2H, followed by validation through other assays, to generate a dataset called HI-II-14 which covered approximately 14,000 PPIs involving 4000 proteins. Later, due to its relevance among other assays, the authors employed three different Y2H assay versions to screen even more PPIs to generate **HuRI**, an even larger dataset that quadrupled the identified interactions. Currently it includes 64006 verified PPIs involving 9094 Proteins.

General Comparison: Focus Data Type

The curators of **STRING** highlight a constitutive quality of the Interactome which is often perceived as a problem during Interactome mapping: the ambiguity, complexity, and diversity of the Interactome. Our current understanding of protein-protein interactions remains fragmented because of these factors. While some interactions are well-documented and comprehended, a sizable portion is still only hinted at through indirect evidence. In this ambiguous state, both **HuRI** and **STRING** encounter fluctuating variations in PPIs based on context, strength, and type of interaction and their individual focus lays on differing PPI types. Therefore, it is important to understand the varying classifications of PPIs before trying to understand the contents of the databases.

As a rule of thumb, they *(PPIs) can be classified depending on many factors, including composition, obligation or affinity, lifetime* (Creative Proteomics, 2018)*,* **and interaction type**.

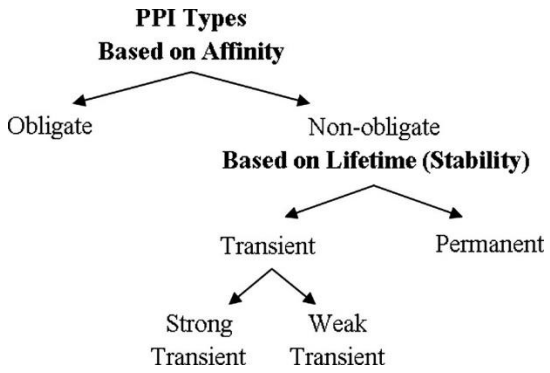


Figure 1: A table to systematically classify Protein-Protein Interactions based on Affinity and Lifetime, excluding Composition and Interaction Type. Note: Permanent PPIs are often classified under Obligate PPIs.

Composition:

- **Homo-oligomeric interactions**: These occur between **identical polypeptide chains**, forming complexes of the same protein subunits.
- **Hetero-oligomeric interactions**: These involve **non-identical polypeptide chains**, where different proteins come together to form functional complexes.

Obligation or Affinity:

- **Obligate interactions**: These occur when two proteins **cannot exist independently**. *The constituents (proteins) of the complex are unstable on their own in vivo* (Creative Proteomics, 2018) and are often found permanently bound to each other.
- **Non-obligate interactions**: In contrast, these proteins can **exist independently** and associate only when required, such as in response to certain cellular conditions.

Lifetime:

- **Transient interactions**: These are **temporary** associations, where proteins bind and dissociate depending on certain **conditions** in the cell. *These interactions are expected to control the majority of cellular processes.* (Thermo Fisher Scientific, n.d.)
- **Permanent interactions**: *These interactions are stable and irreversible, often found in obligate PPIs* (Creative Proteomics, 2018), where proteins are permanently bound and critical for maintaining cellular structures or functions.

Interaction Type:

- **Physical (direct) interactions**: These interactions involve Proteins that associate with one another physically. They bind to each other through a combination of hydrophobic bonding, Van-der-Waals-forces, and salt bridges. *The binding domains can be small clefts or large surfaces and can be a few peptides long or span hundreds of amino acids.* (Thermo Fisher Scientific, n.d.). All the previous PPI categories fall under this domain.
- **Functional (indirect) interactions**: These interactions don't involve direct physical contact but occur when proteins work together in a common pathway or process. For instance, a transcription factor regulating the expression of another protein or two enzymes sharing a common substrate.

Understanding these various categories of PPIs is essential because **HuRI** and **STRING** catalog both **direct physical and functional interactions.** providing a more complete view of cellular processes. More specifically, **STRING** integrates both physical and functional PPIs, including **predicted or indirect interactions**, whereas **HuRI** focuses more on experimentally validated **physical PPIs** in the human interactome, because the curators of **HuRI** believe that molecular mechanisms can be more readily inferred from direct than indirect PPIs. As a result, **HuRI** is better suited for detailed mechanistic studies, while **STRING** is ideal for exploratory or comparative analyses that require a wider range of interaction types. This differentiation, when thought of in the broader sense, implies that gaps across different fields of research in molecular biology can rely on a highly specialized PPI database suited for that specific problem.

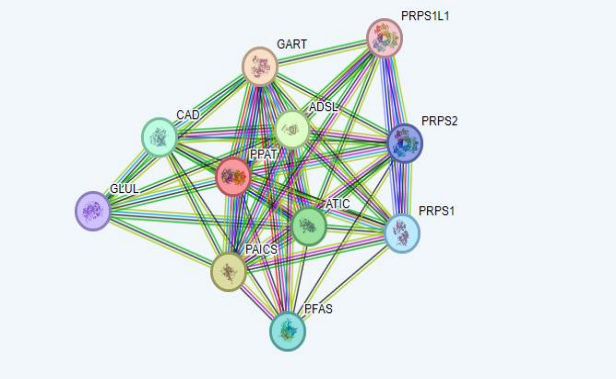


Figure 2:Exemplary association network (STRING) of the protein phosphoribosyl pyrophosphate aminotransferase, symbol PPAT. This enzyme catalyzes the first step of de novo purine nucleotide biosynthesis pathway. (National Library of Medicine, 2024)

Database Content of STRING

In **STRING**, the most basic unit of linkage between two Proteins is the functional association. As taken from the v 11.5 paper, a formal and operational definition would be: 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. Looking at this definition, there seems to be 2 fundamental criteria for a functional association to be considered. Firstly, the specificity of associations: The overlap in function must be specific enough to correspond to a distinct biological pathway or function, rather than corresponding to something very general like metabolism. Two proteins interacting physically can also be associated this way, but they do not have to. Even if some parts of their functional roles overlap, they will still be associated. Secondly, the inclusiveness of the association: Even proteins with contradicting roles within a biological process or pathway can be functionally associated. An inhibitor and an activator in the same pathway might not interact physically but still have a functional relationship. In general, such an approach to defining the association prevents overgeneralization and ensures that the network reflects biologically relevant interactions, making it more useful for understanding the complexity of cellular mechanisms.

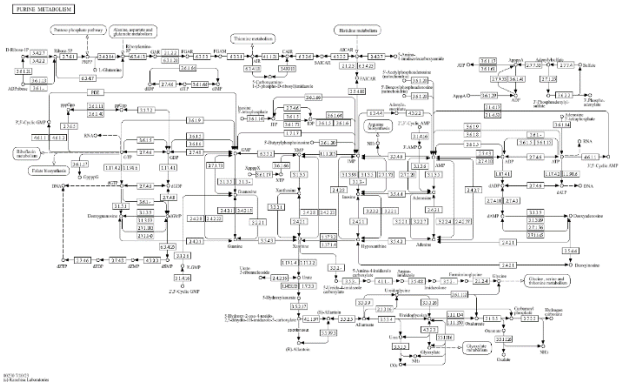


Figure 3: Purine Metabolism, KEGG PATHWAY. ***Purine metabolism*** *refers to the metabolic pathways to synthesize and break down purines that are present in many organisms.*

It is explained that **STRING** refers to a certain specificity cutoff when categorizing a certain as functional one and that this cutoff highly corresponds to those found in KEGG pathway maps. These maps *are a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction, and relation networks so far.* (Kyoto Encylopedia of Genes and Genomes, 2024) *(see Figure 2*). This alignment with KEGG's established pathway maps helps validate and filter associations, ensuring they correspond to recognized, meaningful cellular pathways, rather than overly broad or non-specific functions.

Evidence Channels

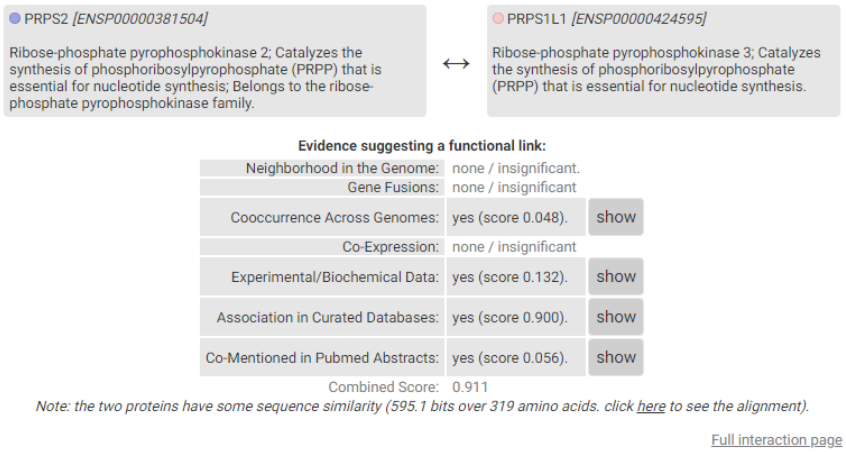


Figure 4: Exemplary Evidence breakdown of the interaction score between two Proteins (Symbols PRPS2 and PRPS1L1) within the previously mentioned PPAT protein association network, taken from STRING

The multifarious types of evidence that, when viewed individually, induce a single functional association for a protein pair can be traced back to 7 distinct types of sources, termed ‘channels’ within **STRING**: Conserved Neighborhood, Gene Fusion, Co-Occurrence, Co-Expression, Experiments, Databases and Text-Mining. For each channel, separate interaction scores and color codes are available. When multiple channels are activated by the user in combination, a combined score will be calculated and displayed. The curators of **STRING** intended these interaction scores to give a confidence approximation between 1 and 0 on whether a proposed association is biologically meaningful or not. A minimum required interaction score puts a threshold on the confidence score. Lower scores mean more interaction population within the network, but they also cause a higher percentage of false positives. This scoring system is useful because it helps ensure standardization among the different evidence types used in **STRING** and creates a uniform metric for evaluating interactions. It might also allow users to compare and integrate data from different sources more effectively.

KEGG indirectly influences the granularity of the functional associations in **STRING**, because the confidence scores are benchmarked using associations where both protein partners are functionally annotated. KEGG pathways are used as a gold standard for this.

Within each evidence channel, the origin of the evidence is further subdivided into two. The former represents evidence stemming from the organism of interest itself, while the latter represents evidence that has been transferred from other organisms. The curators of **STRING** apply the Interolog concept for the inter-organismic transfer of evidence. Here, the term ‘Interolog’ refers to *conserved interactions between a pair of proteins both of which have interacting homologs (orthologs) in another organism.* (Walhout, 2000)*(Definition from Wikipedia).*The use of this concept is beneficial because it allows **STRING** to expand the coverage of its associations by transferring evidence from well-studied to less-studied organisms with limited direct data.

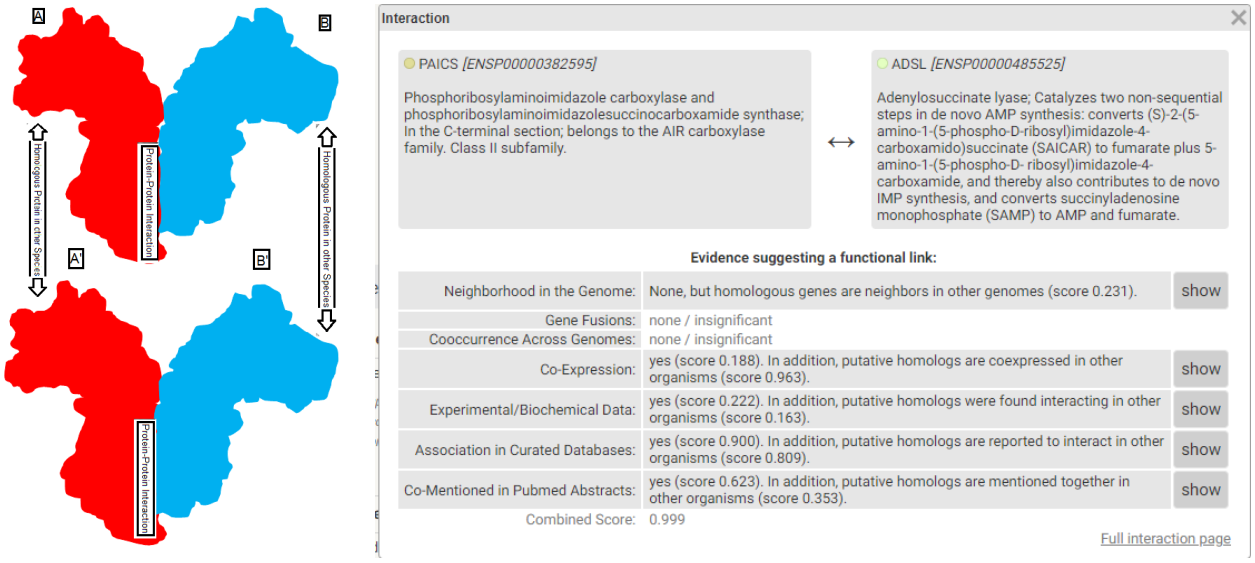


Figure 5: Visualization of the 'Interolog' Concept, taken from Wikipedia. The conserved interactions A/B are referred to as worm 'interologs' of A'/B' interactions in other species if A' and B' are orthologs of A and, respectively; Figure 6: Interaction Score breakdown of the interaction between two proteins (Symbols PAICS and ADSL). In addition to same organism evidence validating the association, evidence has also been transferred inter-genomically from other organisms for many channels.

Genomic-Based, Predictive

Conserved Neighborhood

This type of channel is the first of three genomic-based, predictive channels found in **STRING**. It is based on cluster of genes or 'runs' that occur continuously in proximity, *where the maximum allowed intergenic distance is 300 base pairs*. (STRING, n.d.). Genes located together within the boundaries of this distance are linked with a black line, as seen above. There is also a score threshold which represents the confidence level required for **STRING** to consider them within the same neighborhood. If genes fall below this score, they are canceled out from the calculation and are visually represented as small white triangles, indicating lower confidence in their association. This threshold helps filter out weaker and less conserved gene associations, allowing users to focus more on reliable and functionally relevant neighborhoods.

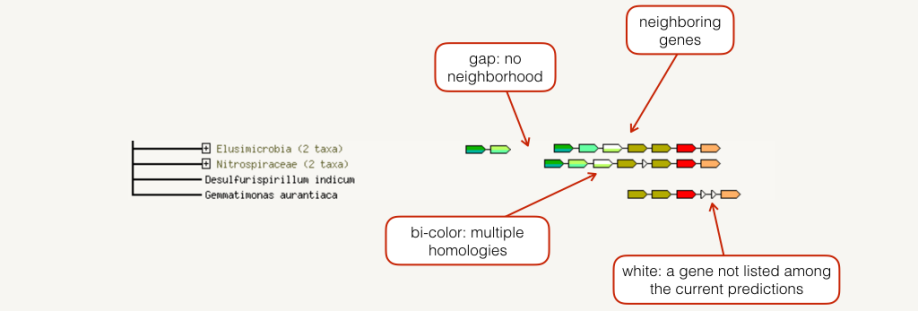


Figure 7: Visualization of the Gene Neighborhood Evidence Channel

It is important to note that this type of evidence can only be derived from prokaryotic genomes because a *genetic regulatory system in which genes coding for functionally related proteins are clustered along the DNA, is found in bacteria and viruses* (Brittanica, n.d.).This system, called the Operon, *allows protein synthesis to be controlled coordinately and allows the cell to conserve energy* (Brittanica, n.d.), by having the genes in the same pathway spatially close to one another.

Gene Fusion

Another predictive channel of genomic nature is gene fusion. This refers to the formation of *a gene made by joining parts of two different genes*. This may occur artificially *in the laboratory or naturally in the body when parts of DNA from one chromosome moves to another chromosome. If the rearrangement places two, originally independent genes in close proximity, the cell may read them as one,* resulting in a *hybrid protein containing the features of both genes.* (CARIS Life Sciences, n.d.) (National Cancer Institute, n.d.)

Figure 8: Exemplary Visualization of the gene fusion channel. List of Taxa on the left, list of proteins that are infused on the right. Functional association exists for proteins, which exist as fused, hybrid proteins in other organisms.

When two genes are found to be fused into a single gene in one organism, it implies that the functions of those genes are related or complementary. Since the fused gene is creating a single hybrid protein, this indicates that these two proteins participated in the same biological pathway. Based on gene fusion in one organism, **STRING** gives an association score to the non-fused, constituent proteins in the genomes of other organisms. If the fused genes have a strong orthology to their counterparts in other organisms, they will end up getting a higher association score.

Gene Co-Occurrence

Gene Occurrence is the third and last channel that is based on genomic predictions. This 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. Thus, when two proteins significantly co-occur in a large number of genomes or can be observed as fusion proteins in a subset of genomes, they are likely to be binding partners or enzymes needed for a specific metabolic pathway* (Müller, 2008). Based on this observation, it may be concluded that a functional relationship exists between these two proteins.

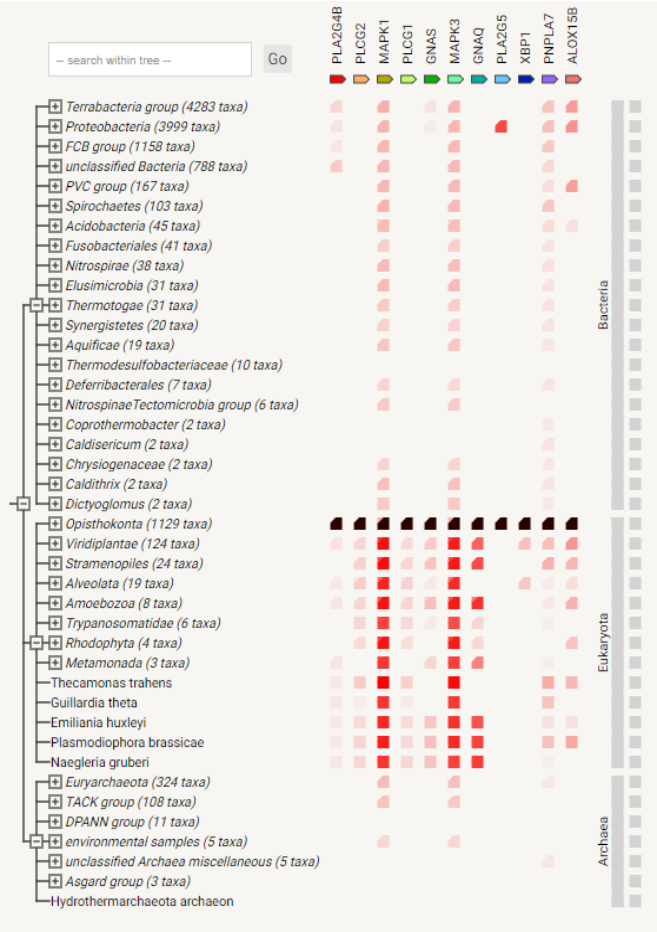


Figure 9: Gene Co-Occurrence grid for the association network of protein PLA2G4B in homo sapiens. Taxa list on the left, Protein Occurrence grid on the right. Proteins MAPK1 and MAPK3 always co-occurs, resulting in a functional association.

This channel comes with a grid diagram and a phylogeny tree of the organisms in which the proteins in question occur, as visible in the Figure above. White space signifies that a specific protein in the list above (not visible in the figure) is absent in that species, whereas a red square signifies the presence of that protein in that species. The intensity of red is correlated with how evolutionarily preserved the homologous protein is.

For this diagram, 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.

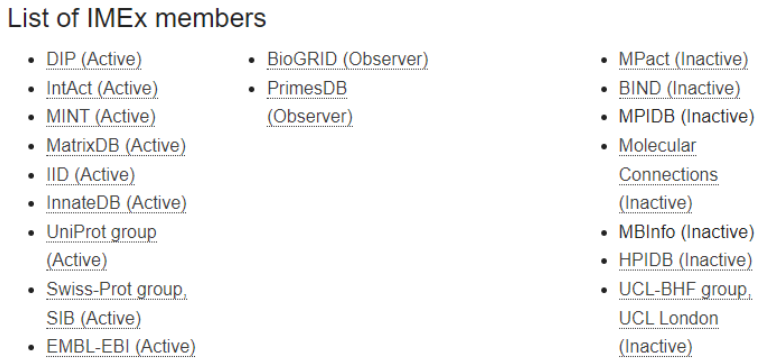
This method is advantageous because it provides insights into evolutionarily conserved relationships, enhancing the understanding of protein functions across species.

It's considered reliable when proteins co-occur consistently across many species, reflecting their potential involvement in essential, conserved processes.

Known

Experiments

The experiments channel is the first channel to fall into the ‘known’ channels category. This channel collects data on physical (direct) protein-protein interactions from high-quality and, more importantly, primary interaction databases included in the IMEx Consortium (plus BioGRID), as these databases contain *a non-redundant set of physical molecular interaction data from a broad taxonomic range of organisms.* (IMeX Consortium, n.d.)


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 10: List Showing members of the IMEx Consortium, taken from its official website.

Once **STRING** pulls the interaction data from these databases, it goes through a remapping and reprocessing phase. First **STRING** merges or removes duplicate records to avoid counting the same interactions multiple times, since some interactions might be reported in more than one database or publication. Then, the information on naming and annotation gets standardized to avoid confusion or redundancy. Once the interaction data is cleaned, **STRING** compares and evaluates all records on interactions against known elements in the functional pathways from KEGG pathway maps to ensure the quality of the experimental data.

Databases

In this second ‘known’ channel, the data is gathered from manually curated interaction records, specifically from those at KEGG, Reactome, Gene Ontology and a few other, all of which are reviewed, covered and compiled by expert curators from alternating fields.

Since the interaction information within the records is carefully selected and confirmed by experts, this channel in **STRING** is considered highly dependable.

One thing to note is that, for this channel, **STRING** only maintains associations between proteins within the same biological pathway or protein complex. By excluding broader or less direct associations, **STRING** ensures that the associations included are highly specific and biologically meaningful.

Since the high-quality manual curation means that data is considered inherently reliable, data pertained in this channel is assigned a high-confidence score of 0.9 without further re-calibration.

Purely Predictive

Text-Mining

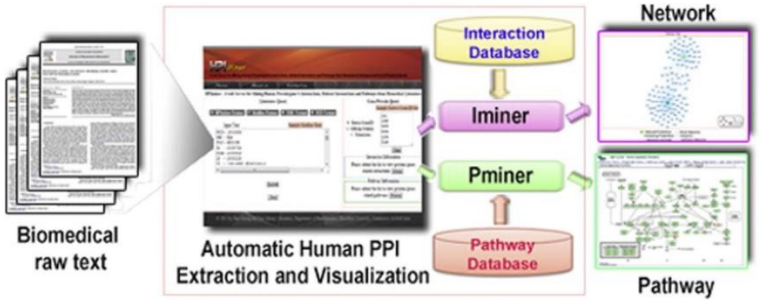


Figure 11: Depiction of the Text-Mining pipeline for Protein-Protein Interaction Network and Pathway Generation.

The text mining channel in **STRING** is a purely computational prediction channel. Text mining methods, in general, *automate the extraction of interconnected proteins through their coexistence in sentences, abstracts or paragraphs within text corpuses.* **STRING** does this by conducting co-citation analysis, in which it *searches for statistically significant co-occurrences between gene names in public repositories and online resources.* Additionally, **STRING** utilizes an advanced Text Mining techniques called Natural Language Processing (NLP) of text, *considering gene names as nodes and verbs as edges giving them a semantic notion on the graphs.* (Nikolas Papanikolaou, 2015)

In versions prior to 10.5, **STRING** used to parse and conduct co-citation analysis only on the abstracts of PubMed articles. Version 10.5 saw the addition of a subset of full-text-articles to the corpus. Version 11.0 expanded upon this basis by including open access and author-submitted full-text-articles from PubMed Central, available in BioC XML format-a simple for designed specifically for text-mining and information retrieval

search. To ensure the relevance and quality of the extracted interactions, **STRING** applied 2 filtering criteria to the text corpus:

- Language Filtering: Articles that were not in English were excluded using a model sensitive to over 176 languages.
- General Content Filtering: Articles mentioning over 200 relevant biomedical entities were removed to prevent general results and increase precision.

The final text corpus after **STRING**'s text mining efforts currently consists of 28,579,637 scientific publications with:

- 2,106,542 of them being full-text articles

- The rest being abstracts from PubMed and other sources.

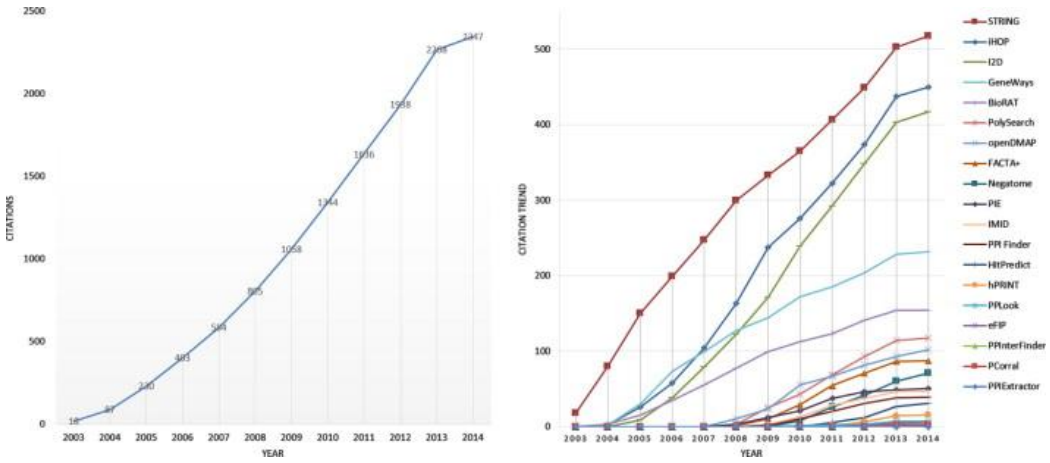


Figure 12: Left: Google-Scholar Citations for text-mining Tools. Right: Google-Scholar citation trends for each text-mining tool (Nikolas Papanikolaou, 2015). The left graph shows a steady increase in the number of citations from 2003 up until 2014. The rights graph depicts STRING always in the first place from 2003 up to 2014, with works citing its text mining channel continuously increasing.

Since the early 2000 there has been a steady increase in general interest for text mining, reaching nearly 2500 citations in the year 2014 (Figure 12). Through the years, **STRING** retains its reputation as having the highest and fastest growing citation rate compared to other text-mining tools. These citation trends emphasize the reliability and popularity of **STRING**'s text-mining channel, showing that it is a well-recognized and widely used tool in the scientific community for extracting and analyzing protein-protein interactions from scientific literature. The high citation count reflects **STRING**'s utility in the research community, particularly its integration of text-mining with other methods to provide comprehensive coverage for protein-protein association data.

Co-Expression:

Co-Expression channel is the seventh and last channel covered by **STRING**, also belonging to the 'purely predictive' channels alongside text-mining. Co-expression is *based on the concept that* expression *profiles of time series, or result of specific perturbations, may be indicative of* similarities *and differences between transcripts, implying their regulation* (Danila Vella, 2017) . **STRING** applies this concept by, first, gathering expression data from both transcriptome and proteome measurements and then, conducting gene-by-gene correlation analysis on the data.

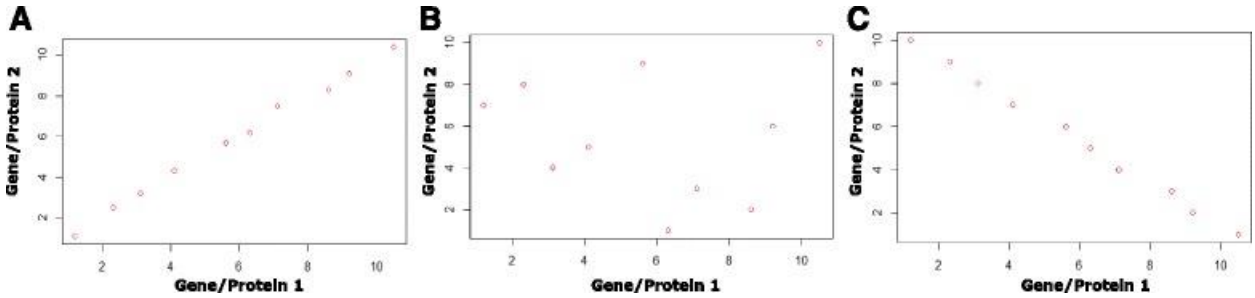


Figure 13: Graphical depiction of different values for Pearson Correlation Coefficient. Graph A depicts a correlation near 1, indicating positive correlation, graph C depicts a correlation near -1, implying a negative correlation, and last, graph B depicts a correlation near 0, implying no correlation.

For the transcriptome-wide co-expression analysis, data is gathered from experiments archived in the NCBI Gene Expression Omnibus, *which 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.* (NCBI, n.d.) **STRING** normalizes, prunes and compares the data on expression profiles over a large variety of conditions. Pairs of transcripts that show consistent correlation are given a functional association score. For determining whether a correlation exists, **STRING** measures the Pearson Correlation Coefficient, which is a statistical measure that *helps us understand the relationship between two quantitative variables when the relationship between them is assumed to take a linear pattern* (McClenaghan, 2024) . If two genes show positive or negative correlation in their expression patterns across many conditions, as seen in the Figure (13), they are likely to be functionally related.

On the other hand, Proteome-wide co-expression analysis is a newly introduced feature in version 11. Currently, this analysis is limited to ProteomeHD, a dataset that stores data on *protein abundance changes in response to biological perturbations* (ProteomeHD, n.d.) from 294 different biological conditions in human cells. Expression Data relating to each

condition is this database is measured with *Stable Isotope labeling by amino acids in cell culture* which is *an excellent approach for high-accuracy detection of protein quantities,* involving *culturing cells in a medium supplemented with amino acids containing either normal or heavy stable isotopes. Since the amino acids are metabolically incorporated into the proteins of the cells through protein synthesis, the relative abundance of the proteins is measured based on the intensities of the light and heavy peptides* (Guoan Zhang, 2009). This technique is highly dependable and advantageous, because introducing labeled amino acids delivers excellent prediction results of mass-labeled-proteins.

One thing to note is that, unlike co-expression analysis of transcriptome data, ProteomeHD utilizes the treeClust Algorithm from R's treeClust software package to determine the correlation between two variables. 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* (Buttrey, 2018) across many trees. In the case protein data, it measures how frequently two proteins end up in the same leaf in the decision trees. If they consistently group together, they are considered co-expressed and correlated. This algorithm is more robust than Pearson Correlation for handling complex, non-linear relationships and noisy data, making it a reliable method for proteomics data, where non-linear relationships are common.

Overall, the co-expression channel has high trustworthiness, especially for large-scale exploratory analyses. It uses well-established methods for transcript data and more advanced, robust approaches for proteomics, ensuring that the results are trustworthy, though some inherent limitations exist.

Enrichment and Bias Analysis Tool for STRING

As mentioned previously, **STRING** is not solely a protein-protein association database, but also a handy tool that can mediate novel discovery. For users that query on **STRING** with a set/list of proteins rather than a single protein of interest, **STRING** can compute functional enrichment analyses in the background. Starting from version 11, **STRING** offers users two types of tests to analyze their input-proteins for functional enrichment, both of which are performed for a total of eleven functional pathway frameworks such as Gene Ontology, KEGG PATHWAY, etc.

The first one is a straightforward over-representation (ORA) analysis using hypergeometric distribution that gives an enrichment p-value. *ORA is 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* (Huang da W, 2008) (Sagendorf, 2022). If any of the gene sets are significantly more 'present' in the protein list, it is concluded that this gene set is over-represented.

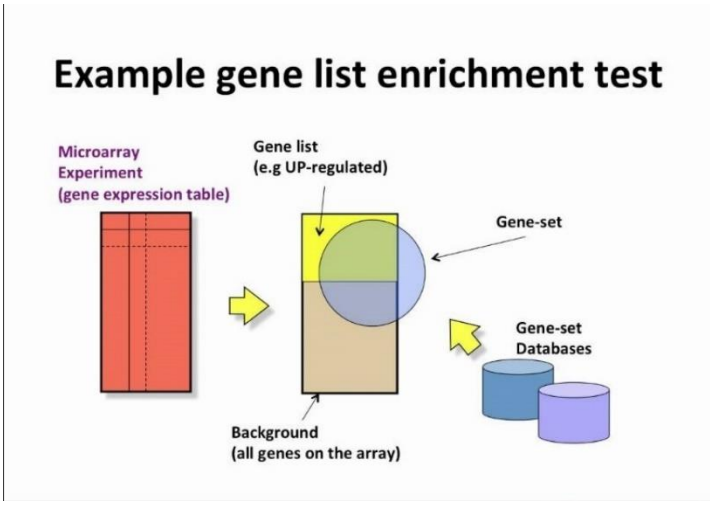


Figure 14: Overrepresentation analysis for tested gene set, where a certain threshold differs between up regulated (significant) and downregulated (insignificant) genes. The analysis mentioned here for STRING is done differently, where there is no differentiation between the input proteins.

To calculate the enrichment p-value for a particular gene set using Hypergeometric Distribution, **STRING** applies the formula, where:

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

N is the number of background genes, n is the number of "interesting" genes, M is the number of genes that are annotated to a particular gene set S, and x is the number of "interesting" genes that are annotated to S. (Sagendorf, 2022)

However, this type of functional enrichment analysis may fall short of reliably presenting an accurate result as it discards a large portion of the information the user and the list might have had, such as:

- 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

All of these alterations and deficiencies in the processed data might provide highly inaccurate and misleading results, especially for large lists.

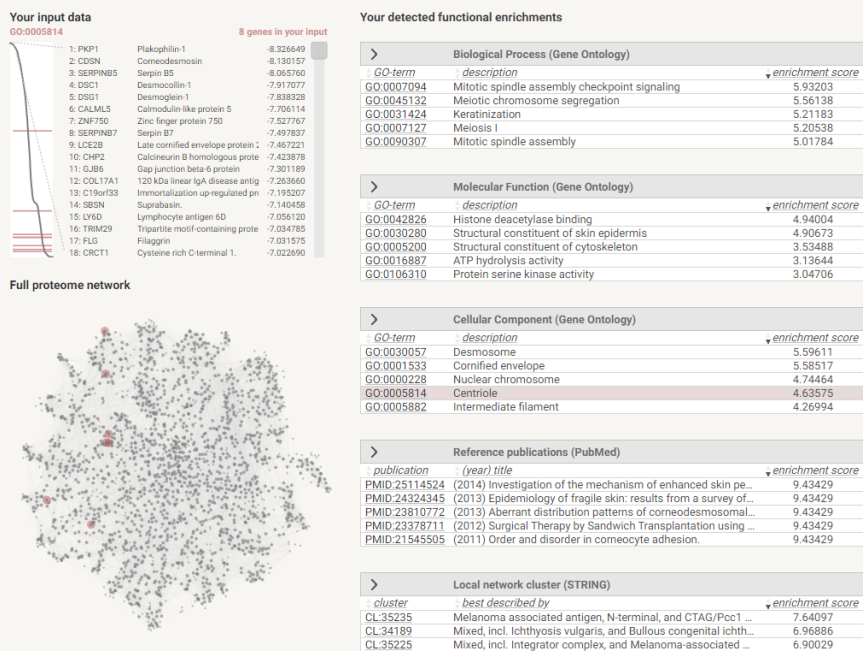


Figure 15: Three-Panel-View showing input list of proteins and numerical values on the top-left, functional enrichment results on the right and the place of each protein in the input highlighted within the full proteome network on the left below. Functional enrichment results only show the top remaining discoveries after the Benjamini-Hochberg Procedure.

To battle these shortcomings, version 11 of **STRING** offers users a second option for analyzing their protein lists for functional enrichment. In this alternative, **STRING** expects a genome-scale input, wherein each gene or protein should be assigned a numerical value. This value could be any measurement or statistical score, like gene expression levels or p-values from an experiment. For the enrichment analysis itself, **STRING** uses a permutation-based method called Aggregate Fold Change. In each gene set being tested, **STRING** calculates the averages of the numerical values assigned to the genes. This average is then compared to averages of randomized gene sets of the same size within each functional pathway framework (Gene Ontology, KEGG, InterPro).

During functional enrichment, **STRING** tests multiple gene sets to determine if they are significantly enriched in the user-provided gene set. Each test yields a p-value, but testing many gene sets increases the likelihood of false positives, which are incorrectly identified enrichment due to chance. Thus, **STRING** ranks the p-values within each framework from lower to higher and applies Bejamini-Hochberg procedure to adjust each p-value and to control the false discovery rate among the p-values (respective formulas below). The remaining p-values, along with the gene sets they represent, are considered correct. **STRING** applies the correction only within individual frameworks 21 but not across frameworks because they often overlap.

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

- 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.)

$$False\ Discovery\ Rate = \frac{False\ Positives}{(True\ Positives + False\ Positives)}$$

Unlike ORA, which can sometimes miss subtle yet important differences between gene sets, as it treats all genes equally, AFC leverages the magnitude of changes, identifying biological processes or pathways that are not just statistically enriched, but also biologically relevant in terms of gene activity.

In version 11.5, a new function has been added to support these functional enrichment analyses. During enrichment analysis, **STRING** performs an automated bias analysis on the input protein list to detect biases of both technical and biological origin. The analysis is based on inherent properties of the genes/proteins and data gathered from external databases. The results of the analysis are then displayed in a graphical report that includes 7 graphics, as seen in the Figure below. The first graph in the report Input Data Histogram, shows the distribution of your input values across the whole value range. The second graph charts the numerical values passed by the user against their rank in the input dataset. The rest of the graphs show the results of the bias analyses conducted for 5 variables:

- Average Protein Abundance, which is *defined as the number of copies of a protein molecule in a cell, which is the result of the dynamic balance among* (Mehdi AM, 2014) subprocesses within Protein Synthesis. *Protein abundance here is present in ppm which is short for parts per million.* (PaxDB, n.d.)
- Protein Length/Size
- Number of Publications mentioning the protein in PubMed literature
- Protein Disorder, which refers to the amount of intrinsically disordered proteins within the list. These *do not have a highly populated stable*

secondary and tertiary structure under physiological conditions, but full essential biological functions and are ubiquitous. (Lieutaud P, 2015)

- Average GC Content of the encoding transcript

All in all, This Automated bias analysis offers a quality check for the dataset, providing users with graphical reports that visualize these biases. This transparency allows users to assess the robustness of their data and enrichments before proceeding to further validation experiments. If biases are detected, users can adjust their experimental design or data interpretation strategies, accordingly, enhancing confidence in the findings.

Additional Gene Sets

In addition to the usual functional classification frameworks above (Gene Ontology, KEGG, InterPro) used in the functional enrichment analyses explained above, **STRING** provides users with two further gene set collections, both of which are generated based on data that has already been mapped to **STRING**.

For the generation of the first gene set collection, **STRING** once again makes use of its text mining channel, which includes an already mapped text corpus filled with PubMed Article abstracts and full-text-articles. With individual publications assuming the role of a pathway, **STRING** forms a gene set with all proteins discussed within, which is later employed to test the user's input for functional enrichment. The sheer quantity of scientific publications available for gene set generation makes it obligatory to correct the discoveries for multiple testing, yet the curators of **STRING** refrain from doing so, because they uphold the idea that the generated gene sets may shed light on uncharted and controversial pathways that have not been curated in any pathway database.

The second gene set collection is generated by hierarchically clustering all the tightly connected proteins within the entire association network in **STRING** to functional modules (units) which are then admitted as gene sets. Hierarchical clustering *is a connectivity-based clustering model that groups data points together that are close to each other based on the measure of similarity or distance. The assumption is that data points that are close to each other are more similar or related than data points that are farther apart* (Geeks For Geeks, 2024 (Last Updated)). Among its variations, **STRING** employes the agglomerative approach in this case, where, at the beginning, *each data point is defined as a cluster and existing clusters get combined at each iterative step* (Penn State University, n.d.). Roughly speaking, the agglomerative clustering steps are as follows:

1. Compute the dissimilarity matrix by using a particular distance metric.

2. Assign each data point to a cluster.

3. Merge the clusters based on a linkage criterion for the similarity between clusters.

4. Update the distance matrix.

5. Repeat steps 3 and 4 until a single cluster remains or any stop criterion is met. (Noble, 2024)

STRING only allows clusters of sizes between 5 and 200 to be utilized as gene sets.

The distance metric in the dissimilarity matrix mentioned herein is based on a Diffusion State Distance (DSD). *Given some fixed* $k > 0$*, we define* $He^{\{k\}}(A, B)$ *to be the expected number of times that a random walk starting at A and proceeding for k steps, will visit B.* This definition accounts for the existence of low-degree and high-degree nodes, or hubs (in our context, hub proteins), between two nodes A and B, which attract random walkers due to their high connectivity, which reduces the likelihood of reaching specific nodes, such as B from A. *Node pairs connected by many short paths of low degree nodes will tend to have high* $He^{\{k\}}()$ *values.* Thus, in this metric, pairs of nodes that have a high h $He^{\{k\}}()$ value will be considered 'similar'. In the case of the variable k, **STRING** assigns a large value to it because, *as the k in* $He^{\{k\}}()$ *goes to infinity, DSD converges, allowing us to define the distance independently from the value of k.* (Cao M, 2013)

Hierarchical clustering analysis has high computational costs and is less scalable for bigger datasets, because this type of clustering is ''greedy'', meaning that the algorithm *decides which clusters to merge or split by making the locally optimal choice at each stage of the process* (Noble, 2024)*,* (Kuchciak, 2024). To counterbalance the computational complexity, **STRING** utilizes *HPC-CLUST, a highly optimized software pipeline that can cluster large numbers of pre-aligned DNA sequences by running on distributed computing hardware. It allocates both memory and computing resources efficiently and can process more than a million sequences in a few hours on a small cluster.* (João F Matias Rodrigues, 2013)

In contrast to other gene set collections, this second gene set generation approach also includes proteins that are less studied and understood within its clusters and may partition functional subsystems differently. As this approach takes the 'bigger picture', that being

the entire protein association network, into consideration, it provides a broader scope above the already known pathways. This may, in return, reveal meaningful, novel pathways and make way for new discoveries.

Yeast-two-Hybrid Screening System

Before advancing to the Human Reference Interactome (HuRI), a general explanation of the Yeast-Two-Hybrid System (abbreviated as Y2H) is required, since this method is essential for capturing and registering unprecedented physical protein-protein interactions, especially in the case of HuRI.

Normally, in any given DNA sequence, a Promoter region is located upstream of the gene sequence, to which transcription factors bind to initialize the transcription of DNA into messenger-RNA. Many transcription factors consist of two domains: a DNA-binding domain (DBD) and an activation domain (AD). The DBD helps the transcription factor bind to the promoter region, but it cannot activate transcription on its own. Here, the AD (Activation Domain) recruits RNA Polymerase II, after which this Enzyme reads the sequence and synthesizes a messenger-RNA. Both domains rely on each other for a successful initiation of the transcription. (Yang, n.d.)

In the Y2H system the transcription factor is manipulated and split into the two previously explained domains, namely the DNA-Binding and the Activation Domain. The proteins of interest (often referred to as bait and prey) are then genetically fused to DBD and AD. In the ordinary way the DBD, with now the bait protein attached to it, binds to the promoter region yet again. Here, two outcomes can be expected:

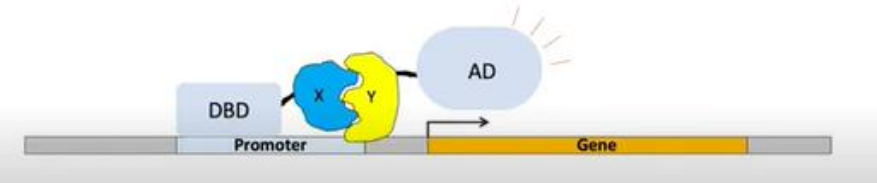


Figure 16: Interacting Bait X and Prey Y. (Henrik, 2019)

If the bait and prey proteins interact with one another, the AD attached to the prey protein will come in proximity of the gene region. In this way, RNA Polymerase II can be recruited which leads to the transcription of the gene

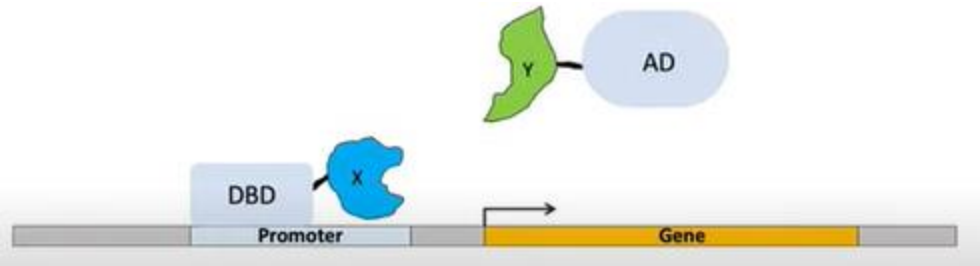


Figure 17: Not-interacting Bait X and Prey Y. (Henrik, 2019)

If the prey protein chooses not to interact with the bait protein, the AD will not approach the gene, thus making the recruitment of RNA Polymerase II and, ultimately, transcription impossible.

Generation and Characterization of HuRI

Moving on to the Human Reference Interactome (**HuRI**), it becomes clear that its generation shows little to no resemblance to that of **STRING**. As previously mentioned, the curators of **HuRI** dominantly employ and rely on the yeast-two-hybrid assay (Y2H), three variants of it, to be more specific. Its high sensitivity, along with its clear inclination towards low false-positive rates, has been validated against two reference sets: the Positive and Random Reference Set (PRSv1 and RRSv1, respectively). The first one is a curated *collection of well-documented protein-protein interactions* that are known to be true positives, while the latter is a *set of protein pairs that are randomly chosen.* (Braun P, 2009)

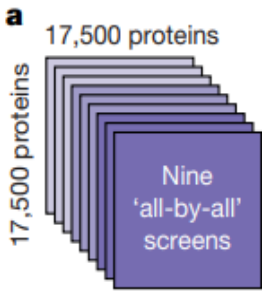


Figure 18: Visualization of the ''all-by-all" screenings concept.

The three Y2H variants were utilized to screen the proteins encoded by the open reading frames, *DNA sequence that lack a stop codon and potentially codes for a protein* (Synapse, 2023), in the newly established ORFeome v9.1 which covers 17,408 protein-

coding genes. Figure 1.a visualizes the scale and coverage of this process, showing that all these Proteins were analyzed using nine ‘’all-by-all’ screenings. It means that the approach involved testing each protein against every other protein to identify PPIs. For a better understanding the total number of screenings can be calculated as follows:

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

In total, approximately a multitude of 1,3 billion screenings were done.

During data generation the authors switched the Y2H variant being used after every three screenings. The decision turned out to be beneficial, because, just as Figure d depicts it, the cumulative number of unique and new PPIs detected continued to rise, as more screens are conducted. The Y2H variants appear to be complementary in their results.

wo additional binary PPI assays, the Mammalian Protein-Protein Interaction Trap (MAPPIT) and the Gaussia Princeps Complementation Assay (GPCA), were utilized to authenticate the newly detected PPIs and to further increase the confidence percentage of the results. These assays leverage different biological mechanisms to validate interactions, thereby adding a new layer of experimental proof to the PPIs. MAPPIT utilizes cytkone receptor signaling, while GPCA relies on splitting a certain enzyme called Gaussia princeps luciferase, just like the yeast-two-hybrid method.

The binary protein interactome network mapped from these efforts, denoted here as HI-III-20 but known as **HuRI**, currently includes 64006 verified PPIs involving 9094 proteins. To determine whether these numbers are significant and promising, the magnitude of the entire human interactome must be looked at. It is predicted that the entire human *interactome contains 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.* (Uros Kuzmanov, 2013). Based on these numbers, **HuRI** is estimated to represent 2-11% of the entire binary protein interactome.

In the face of the results, a comparison of **HuRI** with previous efforts of interactome mapping is required to get a glimpse of their significance. As seen in Figures (e,f), it is abundantly clear that **HuRI** outperforms them, both in terms of binary (physical) protein protein interactions captured and the total number of PPIs detected.

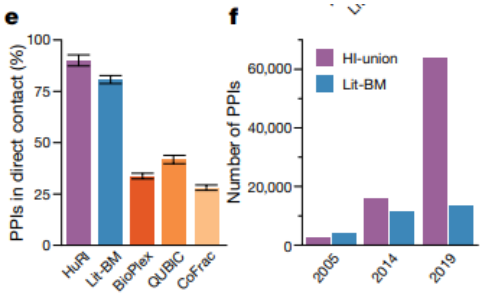


Figure 19: Left: Percentage of direct (physical) PPIs to the whole across multiple Protein-Protein Interaction Databases. Right: Number of PPIs detected in HuRI and Lit-BM in three different time periods.

As for the rest of the remaining, undetected binary interactions, it can be said that *no high-throughput technique can detect all interactions, and false negatives are unavoidable. Consequently, a variety of methods must be considered when working with interactome mapping, and new strategies should be employed* (Laure Sambourg, 2010). Additionally, the curators of **HuRI** firmly hold the persuasion that the interactome is dominated by weaker and more transient, difficult-to-detect PPIs, which is another reason why a huge portion of them remain undetected.

Functional Relationships in HuRI

The curators of **HuRI** also highlight the existence of functional relationships between proteins with **(not to be confused with functional associations found in HuRI, this refers to two proteins functioning in a similar pattern)** with respect to their interaction interfaces. An interaction interface is the *specific residual regions or surface areas of a protein that contact with residues from the other interacting protein* (Yan C, 2008). In **HuRI**, Proteins with similar interfaces have been observed sharing the same interaction partners. A profile similarity network (PSN) was then composed which indicated that this observed relationship between the proteins was, by no means, a coincidence but rather a legitimate relationship. As can be concluded from Figure (18), the relationship network appears to be as complex as the protein pairwise combination dataset itself.

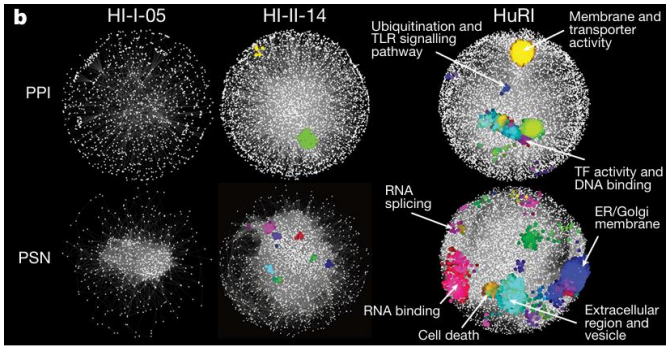


Figure 20: Comparisons between the entire PPI network in HuRI and the Profile Similarity Network of the Protein within.

Additionally, it was observed that proteins sharing interaction partners only exhibited similar interfaces rather than complementary ones, meaning that they did not interact with one another, except in a few cases where the proteins stem from a common ancestor.

The curators also emphasize that proteins with similar interaction interfaces are not detectable solely by their sequence identity. Even if their interfaces have high similarity, they do not appear to share a sequence similarity exceeding 20%. Thus, global sequence identity only remains to be an indicator and not a solid evidence of interface similarity.

Applications of HuRI

Application 1: Uncharted Disease-Related Interactions

HuRI's contribution to the biomedical sciences is also unmissable. It provided high-quality screenings and PPI data on biomedically significant genes, for which previous data was sparse, by exploring uncharted territory of the protein interactome.

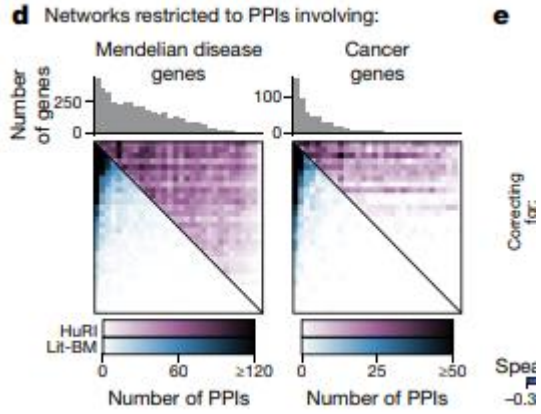


Figure 21: Intensity of Charted Proteins in different biomedical territories. Explained Below.

Figure d (3) serves to prove this contribution. The heatmaps in this figure show the density of interactions captured in two different interactomes for two biomedically interesting gene types, with the color gradient showing the number of PPIs. Data from **HuRI** is compared to that from Lit-BM. In both the panels, it is abundantly clear that the distribution of PPIs captured in **HuRI** is both denser and more extensive, making it a more dominant identifier.

Furthermore, a faulty correlation in other protein complex and interaction networks was disproven with the help of **HuRI**. The authors conducted and analysis on these networks and saw that protein popularity among scientific publications and expression levels correlated with each other, the number of interaction partners and some other variables (explained below), which was not the case in **HuRI**. After correcting the first two factors, this correlation was deemed illegitimate.

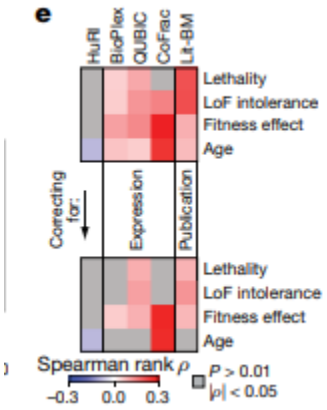


Figure 22: Bias Correction for different PPI databases (before and after). Explained Below.

This figure e (20) displays the correlation between popularity (here publication), expression levels and the previously mentioned 'other variables' (lethality, loss-of-function tolerance, fitness effect, and age) across different datasets. The top section shows captured variable correlations before, whereas the bottom section shows variable correlations after the correction for popularity and expression levels. The color gradients within the panels indicate how high or low the correlation between the variables is. Overall, **HuRI** maintains consistently weak correlations throughout, whereas the other datasets see a drastic shift from some or high correlation to minimal or no correlation after the correction. This highlights the relative lack of bias in **HuRI**.

Application 2: PPI Roles Within Cell Compartments

Proteins are localized to specific subregions within the cell known as compartments. These intercellular membrane systems create enclosed compartments *that are separate from the cytosol. Providing the cell with functionally specialized aqueous spaces* (Alberts B, 2002), the functions of these compartments are dependent on both the subcellular environment and the local PPI network, where **HuRI** might be of help. Even though the experimental detection of these networks is seen as challenging, the authors believe that many of these localized PPI networks can be generated by integrating **HuRI** in available localized protein data.

An exemplary application of **HuRI** in compartment-specific PPI discovery is the extracellular vesicles. These are known to *facilitate intercellular communication and traffic of proteins, lipids and DNA* (abcam, 2023). Specific to these particles, **HuRI** revealed a highly connected subnetwork of proteins, more so than randomized networks. Many of the PPIs are already known to play vital roles in vesicle biogenesis and cargo recruitment.

Syntenin-1, a protein with 48 different instances of interactions with other vesicle proteins, is particularly mentioned by the authors, as it is thought to have a role in protein recruitment. To test this dependency to Syntenin-1, it was knocked within its cell line, which resulted in reduced protein levels and lingering extracellular vesicle production. To the authors, the result suggests that it is indeed involved in the production line. This underscores **HuRI**'s potential in studying protein function in subcellular environments.

However, Proteins also tend to shift their localization, called co-localization, in which they localize multiple subcellular compartments, causing some compartments to share more proteins (crosstalk). This complicates the capturing process, but the authors are confident that **HuRI** can still show high applicability in such 'inter-compartmental' contexts. In fact, **HuRI** has revealed a positive correlation between non-co-localized interacting proteins and compartments with significant crosstalk.

Application 3: Principles of Tissue-Specific Function

Analysis and conclusion-drawing on tissue-specific functions is yet another field of study where **HuRI** has shown high return. As denoted by the authors, the lack of comprehension on how tissue-preferentially expressed (TiP) genes co-operate and facilitate tissue-specific function withing this field is still an unresovled problem. The authors preach that experimental methods will fall short of providing this uniform coverage and affirm, on the other hand, that TiP proteins are well-represented in **HuRI**. Figure a (21), which compares the percentage of TiP proteins relative to all proteins in

HuRI and various other protein interaction networks, attest to this remark. Even as the tissue-preferential expression cutoff, meaning the stringency in defining TiP proteins based on the expression levels in specific tissue, increases, **HuRI** still manages to represent more TiP proteins than any other datasets, maintaining a percentage closer to the expected value.

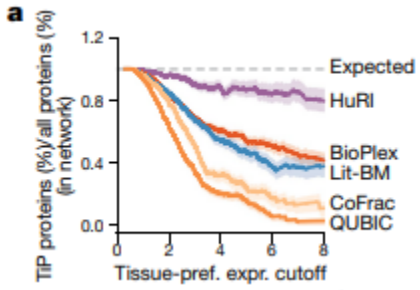


Figure 23: Change in Proportion of TiP proteins to all the proteins in individual PPI databases with rising expression cutoff. Expression levels are normalized to proteins per million (ppm) between 0 and 8. TiP proteins are expected to represent %100 percent (1.0) of the entire databases.

Upon being limited to proteins expressed within the same tissue, **HuRI** revealed that:

- TiP proteins engage in as many PPIs as proteins which are expressed uniformly

-TiP PPIs in **HuRI** are higher in numbers when compared to TiP genes.

From these insights the authors concluded that TiP PPIs will be much more informative about tissue-specific functions. Thus, as a rich and extensive interactome, **HuRI** can provide a detailed picture of the subject.

To further understand the significance of TiP proteins in tissue-specific functions and to generate data on local TiP PPIs **HuRI** was utilized to derive separate interactomes for 35 tissues, where the mean of all the tissue-specific subnetworks corresponded to 25,000 PPIs per tissue.

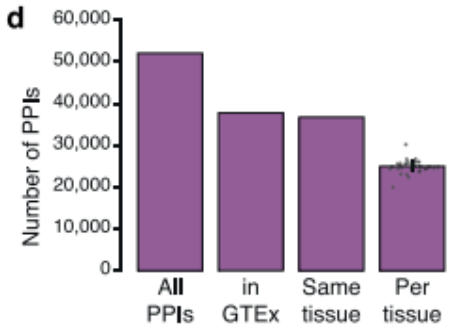


Figure 24: Number of PPIs in HuRI, involving proteins in GTEx, in which both proteins are expressed in the same tissue, and the mean of the tissue-specific subnetworks.

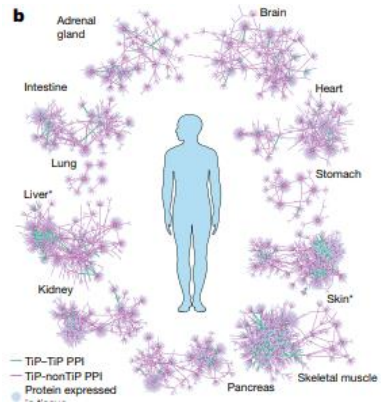


Figure 25: Tissue-preferential subnetworks

The authors filtered their findings to PPIs involving at least one TiP within each tissue, revealing a vast network of interaction partners as seen in figure b (23). Strikingly, it was also identified that TiP proteins tend to interact more with non-TiP proteins, whereas very few TiP-TiP protein interactions are recorded. This led the authors to the conclusion that tissue-specific functions are caused by TiP to Non-TiP interactions. Even in such specific biological contexts **HuRI** still manages to provide efficient data which emphasizes underlying patterns and characteristics. The role of the protein OTUD6A in cell death was mentioned as an exemplary point of interest, where **HuRI** reinforced the analysis and helped verify this protein as an indirect inducer of cell death.

Application 4: Mechanisms Of Tissue-specific Disease

Another facet of tissue-specific medium where **HuRI** can be utilized is Mendelian tissue-specific diseases that exhibit phenotypes. When the uniformly expressed disease-causal protein products of the mutated genes interact with their usual TiP protein partners, the PPI gets disrupted or ‘perturbated’, as the authors choose to word it. The resulting Perturbations are thought to be the direct reason behind the tissue-specific phenotypic aftermath of these diseases. **HuRI** already provides insights to such PPIs within its tissue-specific PPI networks, but it showed no enrichment for PPIs occurring between causal proteins and TIP proteins in this case.

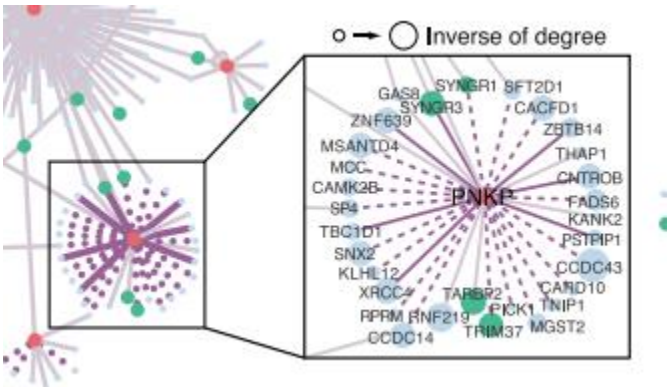


Figure 26:PPI network of the PNKP protein, which has been modified to account for 'perturbated' PPIs, when the disease-causing Glu325Lys mutation occurs. Dotted lines indicate perturbated PPIs, whereas undotted ones indicate PPIs that remain undisturbed after the mutation.

The authors chose to tackle this hypothesis, and they tested it experimentally with 10 disease-causal proteins. 7 proteins out of this set displayed disrupted the PPIs in the corresponding disease tissue. Again, out of this set, a variant of the protein PNKP with the mutation Glu326Lys is chosen as an example. Normally, PNKP partakes in PPIs within the brain, but its mutational variant *can lead to microcephaly or neurodegeneration* (Dumitrache LC, 2017). As can be seen in this concentric graph taken from Extensive Figure 9c, many of PPIs involving this PNKP get perturbated. Specifically, to the authors, the perturbated PPI between the variant PNKP and the protein TRIM37 is a striking discovery because this protein is known to facilitate DNA repair. This implies that perturbation of the PPI might affect functions specific to the brain. It is an important implication because it emphasizes how **HuRI** can still provide coverage and a point of reference in mutational and disease-related context where it involves the PPIs having mutated interaction partners.

HuRI's Limitations and Future

In this very last part, the weaknesses and limitations of HuRI are accentuated by the authors to the addresses, along with what future possibilities HuRI holds. Although HuRI expands upon previous efforts to generate a consistent protein-protein interactome, there is still room for improvement. In this sense, the authors identify two abundant limitations:

1.The cellular functions of most of the detected PPIs mediate are unknown. The authors seem to be positive towards this limitation. They claim that, in follow-up studies, HuRI can be integrated into and reinforced by transcriptome and subject-related proteome data to increase precision in function detection. Additionally, they advise against opting for the removal of PPIs with

unknown function, because, in their opinion, they can still offer useful insights into phenotypic aftermaths of disease-related gene and protein expressions.

2. A very large portion of the PPIs failed to be detected. This second deficiency is caused by the limitations of Y2H, the assay the authors relied on the most to generate this interactome. Contradictory to being a widely preferred tool, *many natural PPIs cannot be detected* using this method. In fact, a study found that the Y2H *assay can only detect about 23% of all the PPIs in certain bacteria (Treponema pallidum)* (Seesandra V Rajagopala, 2007). The Y2H takes place in the yeast cell nucleus. Proteins that don't localize in the nucleus, such as *secretory or membrane proteins, don't interact within the nucleus. Proteins affected by post-translational motifs are also unavailable to yeast. Some proteins are even toxic to yeast (Manfred Koegl, 2008).* For this reason, the rate of interactions not detectable by this assay is substantial.

Regarding the entirety of the interactome, the authors have concluded that HuRI represents only a small fraction of it and that the rest unstable and functionally not conserved. But they are confident that HuRI provides a solid and contemporary reference point for follow-up studies on cellular function.

Conclusion

In conclusion, both HuRI and STRING databases offer invaluable resources for understanding protein-protein interactions, but they take distinct approaches to mapping the interactome. HuRI, with its focus on experimentally validated physical interactions in the human proteome, provides a robust foundation for studying molecular mechanisms at a high resolution. It excels in its ability to map tissue-specific and disease-related interactions, particularly through its use of the yeast-two-hybrid (Y2H) system, which allows for direct detection of PPIs. STRING, on the other hand, integrates both experimentally validated and computationally predicted interactions across multiple organisms, offering a broader and more exploratory tool for functional genomics. This combination of physical and functional data in STRING enables users to make novel discoveries, particularly in large-scale or comparative analyses.

Both databases play complementary roles in interactome research. HuRI is better suited for detailed, mechanistic studies of human-specific interactions, while STRING's expansive dataset is ideal for identifying functional relationships across species and pathways. Despite its limitations, such as incomplete coverage due to reliance on Y2H, HuRI remains a key reference point for follow-up studies in cellular function and disease. Meanwhile, STRING's integrative approach and extensive coverage across organisms

ensure that it remains a powerful tool for hypothesis generation and pathway discovery. Together, these databases significantly advance our understanding of protein interactions, with important implications for fields such as drug discovery, biomarker identification, and functional genomics.

Bibliography

(2024). Retrieved from Kyoto Encylopedia of Genes and Genomes: https://www.kegg.jp/kegg/

abcam. (2023, May 15). *Extracellular vesicles*. Retrieved from abcam: https://www.abcam.com/en-de/technical-resources/pathways/extracellular-vesicles

Alberts B, J. A. (2002). The Compartmentalization of Cells. *Molecular Biology of the Cell. 4th edition.*

Braun P, T. M.-R. (2009, January 6). An experimentally derived confidence score for binary protein-protein interactions. *Nat Methods*. doi:10.1038/nmeth.1281.

Brittanica. (n.d.). *Operon, Genetics*. Retrieved from Brittanica: https://www.britannica.com/science/operon

Buttrey, S. (2018, 05 7). *Package 'treeClust'.* Retrieved from R-Project: https://cran.r-project.org/web/packages/treeClust/treeClust.pdf

Cao M, Z. H. (2013, October 23). Going the Distance for Protein Function Prediction: A New Distance Metric for Protein Interaction Networks. *PLoS One.* doi:10.1371/journal.pone.0076339.

CARIS Life Sciences. (n.d.). *CARIS Life Sciences, What is gene fusion?* Retrieved from CARIS Life Sciences: https://www.carislifesciences.com/what-is-gene-fusion/

Creative Proteomics. (2018, September 18). Retrieved from Creative Proteomics Blog: https://www.creative-proteomics.com/blog/index.php/brief-introduction-of-protein-protein-interaction-ppi/

Danila Vella, I. Z. (2017). From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *EURASIP J Bioinform Syst Biol.*

Dumitrache LC, M. P. (2017, February 18). Polynucleotide kinase-phosphatase (PNKP) mutations and neurologic disease. *Mechanisms of Ageing and Development*. doi:10.1016/j.mad.2016.04.009

Fontanillo, J. D. (2010). Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2891586/

Geeks For Geeks. (2024 (Last Updated), March 11). *Hierarchical Clustering in Machine Learning*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/hierarchical-clustering/

Guoan Zhang, T. A. (2009). Use of Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) for Phosphotyrosine Protein Identification and Quantitation. *Methods Mol Biol. 2009;527:79-92, xi.* doi:doi: 10.1007/978-1-60327-834-8_7.

Henrik, E. (Director). (2019). *Henrik's Lab: Yeast-two-hybrid screen (Y2H)* [Motion Picture]. Retrieved from https://www.youtube.com/watch?v=w5Pvri4-cUA

Huang da W, S. B. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res. 2009* . Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2615629/

IMeX Consortium. (n.d.). Retrieved from IMeX Consortium: https://www.imexconsortium.org/about/

João F Matias Rodrigues, C. v. (2013, November 9). HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics. 2014 Jan*. doi:10.1093/bioinformatics/btt657

Kuchciak, M. (2024, January). *Hierarchical clustering – pros, cons, interpretation, application*. Retrieved from R Publications by RStudio: https://rpubs.com/TusVasMit/HierarchicalClusteringOverwiev#:~:text=Cons%20of%20hierarchical%20clustering,-Scalability%20and%20computational&text=It%20is%20less%20scalable%20for,of%20the%20dendrogram%2C%20misleading%20interpretations.

Laure Sambourg, N. T.-M. (2010). New insights into protein-protein interaction data lead to increased estimates of the S. cerevisiae interactome size. *BMC Bioinformatics*. Retrieved from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-605

Lieutaud P, F. F. (2015, December 21). How disordered is my protein and what is its disorder for? A guide through the “dark side” of the protein universe. *Intrinsically Disord Proteins.* doi:10.1080/21690707.2016.1259708.

Manfred Koegl, P. U. (2008, January 24). Improving yeast two-hybrid screening systems . *Briefings in Functional Genomics, Volume 6* . doi:10.1093/bfgp/elm035

McClenaghan, E. (2024, April 29). *Pearson Correlation*. Retrieved from Technology Networks: https://www.technologynetworks.com/tn/articles/pearson-correlation-385871

Mehdi AM, P. R. (2014, Feb 16). Predicting the dynamics of protein abundance. *Mol Cell Proteomics. 2014 May*. doi:10.1074/mcp.M113.033076

Müller, H. (2008, Sep 10). Identification and Analysis of Co-Occurrence Networks with NetCutter. doi:10.1371/journal.pone.0003178

National Cancer Institute. (n.d.). Retrieved from National Institute of Health: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/fusion-gene

National Library of Medicine. (2024, September 19). Retrieved from National Libary of Medicine: https://www.ncbi.nlm.nih.gov/gene/5471

NCBI. (n.d.). *GEO Overview*. Retrieved from NCBI Gene Expression Omnibus: https://www.ncbi.nlm.nih.gov/geo/info/overview.html

Nikolas Papanikolaou, G. A. (2015, March 1). Protein–protein interaction predictions using text mining methods. doi:10.1016/j.ymeth.2014.10.026

Noble, J. (2024, August 5). *What is hierarchical clustering?* Retrieved from IBM: https://www.ibm.com/think/topics/hierarchical-clustering

PaxDB. (n.d.). *PaxDb User Documentation, Getting Started*. Retrieved from PaxDB Organization: https://pax-db.org/help

Penn State University. (n.d.). *Applied Multivariate Statistical Analysis*. Retrieved from Penn State Eberly College of Science: https://online.stat.psu.edu/stat505/lesson/14/14.4#:~:text=Average%20Linkage%3A%20In%20average%20linkage,points%20in%20the%20second%20cluster.

ProteomeHD. (n.d.). Retrieved from ProteomeHD: https://www.proteomehd.net/documentation

Sagendorf, T. (2022, May 27). *Overrepresentation Analysis*. Retrieved from Proteomics Data Analysis in R : https://pnnl-comp-mass-spec.github.io/proteomics-data-analysis-tutorial/ora.html#ref-huang_bioinformatics_2009

Seesandra V Rajagopala, B. T. (2007, July 31). The protein network of bacterial motility. *Molecular Systems Biology* . doi:10.1038/msb4100166

Statistics How To . (n.d.). *What is the Benjamini-Hochberg Procedure?* Retrieved from Statistics How To: https://www.statisticshowto.com/benjamini-hochberg-procedure/

Stephanie Susnjara, I. S. (2024, July 9). *What is high-performance computing(HPC)?* Retrieved from IBM: https://www.ibm.com/topics/hpc

STRING. (n.d.). *STRING Database User Documentation*. Retrieved from STRING Database: https://string-db.org/help//getting_started

Synapse. (2023, November 23). *Biological Glossary | What is Open Reading Frame (ORF)?* Retrieved from Synapse by patsnap: https://synapse.patsnap.com/blog/biological-glossary-what-is-open-reading-frame-orf

Szklarczyk, D. (2020). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets.

Thermo Fisher Scientific. (n.d.). Retrieved from Thermp Fisher Scientific: https://www.thermofisher.com/tr/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-protein-protein-interaction-analysis.html

Uros Kuzmanov, A. E. (2013). Protein-protein interaction networks: probing disease mechanisms using model systems. *Genome Medicine*. Retrieved from https://genomemedicine.biomedcentral.com/articles/10.1186/gm441#:~:text=Protein%2Dprotein%20interactions%20(PPIs)%20are%20central%20to%20the%20proper,600%2C000%20%5B1%2C%202%5D.

Walhout, A. J. (2000, Jan 7). Protein Interaction Mapping in C. elegans Using Proteins Involved in Vulval Development. doi:DOI: 10.1126/science.287.5450.116

Yan C, W. F. (2008, Oct 10). Interfaces, Characterization of Protein–Protein. *Protein J. 2008 Jan*. doi: 10.1007/s10930-007-9108-x

Yang, J. (n.d.). *Yeast-Two-Hybrid*. Retrieved from Singer Instruments: https://www.singerinstruments.com/resource/yeast-2-hybrid/