

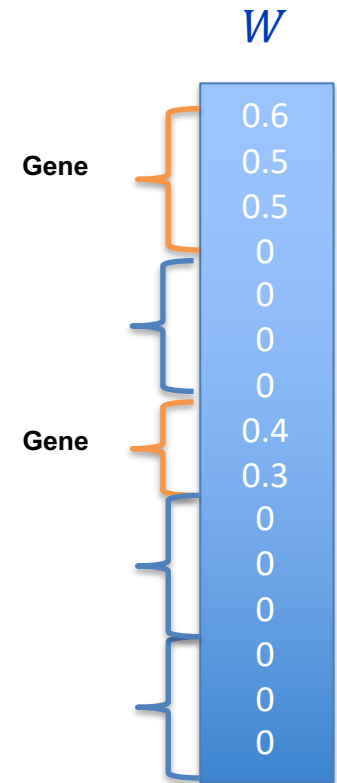
Part 1: overview

- Principal Component Analysis (PCA)
- Partial Least Squares (PLS)
- Two-way Orthogonal PLS (O2PLS)
- Post-hoc analyses using external databases
 - String-DB and gene enrichment



Background

- Data integration with OmicsPLS
 - Using the methylation and glycomics datasets, relevant genes estimates
 - These genes are important to consider
-
- **What do they tell us?**



Genes and their meaning

- Vast amount of information available on genes, DNA mutations, proteins, other biomolecules
 - Functionality, location, gene sets, interactions, etc
- Many bioinformatics databases organize this information
- Use these databases to understand significance of the selected genes



Bioinformatics database: String-DB

- We will focus on String-DB
 - A collection of gene-gene interactions based on several sources of evidence
 - <https://string-db.org/>
 - Also available as an R package
 - Connects to an online server




String-DB: entering input genes

1. Open website, go to “multiple proteins”
 - Try an example list, click on #1
 - Or: Input your list of names, and organism
2. Review the gene ID mapping and continue
3. An interaction network appears
 - Interactive, so almost everything is clickable
4. Review basic settings to customize the network
 - Most important: Interaction source and confidence
 - Typically: exclude text-mining; set confidence = 700



String-DB: input and output



plus:
My Payload
... relevant
nodes marked

X
?

SearchDownloadHelpMy Data

Protein by name>
Protein by sequence>
Multiple proteins>
Multiple sequences>
Proteins with Values/Ranks **New**>
Organisms>
Protein families ("COGs")>
Examples>
Random entry>

SEARCH

Multiple Proteins by Names / Identifiers

List Of Names: (one per line; examples: #1 #2 #3)

... or, upload a file:


Browse ...

Organism:

auto-detect ▼

Advanced Settings

SEARCH




6

String-DB: input and output

Version: 11.0

LOGIN | REGISTER



plus:
My Payload
... relevant
nodes marked

×

?

SearchDownloadHelpMy Data

The following proteins in *Escherichia coli* K12 MG1655 appear to match your input. Please review the list, then click 'Continue' to proceed.

<- BACK

↓ MAPPING

CONTINUE ->

1) 'trpA':

☒ **trpA** - Tryptophan synthase alpha chain; The alpha subunit is responsible for the aldol cleavage of indoleglycerol phosphate to indole and glyceraldehyde 3-phosphate; Belongs to the **TrpA** family

2) 'trpB':

☒ **trpB** - Tryptophan synthase beta chain; The beta subunit is responsible for the synthesis of L- tryptophan from indole and L-serine; Belongs to the **TrpB** family

3) 'TRPC_ECOLI':

☒ **trpC** - Tryptophan biosynthesis protein TrpCF; Bifunctional enzyme that catalyzes two sequential steps of tryptophan biosynthetic pathway. The first reaction is catalyzed by the isomerase, coded by the TrpF domain; the second reaction is catalyzed by the synthase, coded by the TrpC domain [a.k.a. *trpF*, *b1262*, *JW1254*, **TRPC_ECOLI**]


4) 'b1263':

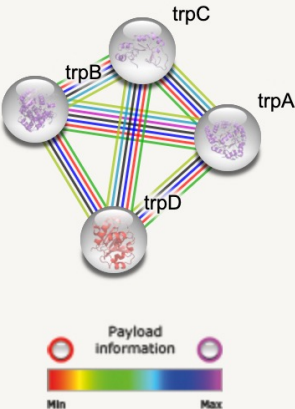
☒ **trpD** - Bifunctional protein TrpGD; Part of a heterotetrameric complex that catalyzes the two-step biosynthesis of anthranilate, an intermediate in the biosynthesis of L-tryptophan. In the first step, the glutamine- binding beta subunit (TrpG) of anthranilate synthase (AS) provides the glutamine amidotransferase activity which generates ammonia as a substrate that, along with chorismate, is used in the second step, catalyzed by the large alpha subunit of AS (TrpE) to produce anthranilate. In the absence of TrpG, TrpE can synthesize anthranilate directly from chorismate and high concentrations [...] [a.k.a. *trpGD*, **b1263**, *JW1255*]



String-DB: input and output

Version: 11.0 LOGIN REGISTER

 **STRING** plus: My Payload ... relevant nodes marked Search Download Help My Data



Viewers > Legend > Settings v Analysis > Exports > Clusters > + More - Less

Basic Settings

Network type:

☒ full network (the edges indicate both functional and physical protein associations)

☐ physical network (the edges indicate that the proteins are part of a physical complex)



meaning of network edges:

UPDATE

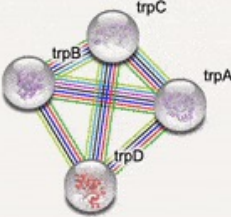



String-DB: output

Version: 11.0 LOGIN REGISTER

 plus: **My Payload**  ... relevant nodes marked ?

Search Download Help My Data


Viewers > Legend > Settings > Analysis > Exports > Clusters > More > Less >


Nodes:

Network nodes represent proteins


splice isoforms or post-translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.


Node Color

 *colored nodes: query proteins and first shell of interactors*

 *white nodes: second shell of interactors*

Node Content

 *empty nodes: proteins of unknown 3D structure*


 *filled nodes: some 3D structure is known or predicted*


Edges:

Edges represent protein-protein associations


associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding each other.


Known Interactions


 *from curated databases*

 *experimentally determined*


Predicted Interactions


 *gene neighborhood*


 *gene fusions*

 *gene co-occurrence*


Others

 *textmining*

 *co-expression*

 *protein homology*

Your Input:

 **trpA** *Tryptophan synthase alpha chain; The alpha subunit is responsible for the aldol cleavage of indoleglycerol phosphate to indole and glyceraldehyde 3- phosphate; Belongs to the TrpA family (268 aa)*



String-DB: input and output

Viewers >

Legend >

Settings

Analysis >

Exports >

Clusters >

+ More

- Less

Basic Settings

Network type:

☒ full network

(the edges indicate both functional and physical protein associations)

☐ physical network

(the edges indicate that the proteins are part of a physical complex)

meaning of network edges:

☒ evidence

(line color indicates the type of interaction evidence)

☐ confidence

(line thickness indicates the strength of data support)

active interaction sources:

☒ Textmining

☒ Experiments

☒ Databases

☒ Co-expression

☒ Neighborhood

☒ Gene Fusion

☒ Co-occurrence

minimum required interaction score:

medium confidence (0.400)

max number of interactors to show:

1st shell: - none / query proteins only -

2nd shell: - none -

UPDATE



String-DB: output

- String network
 - Network of interactions between genes
 - 7 metrics that form a combined score
 - How likely is it that this link is biologically relevant?
- Enrichment analysis
 - Are certain pre-defined gene categories overrepresented in our gene list?
 - Many types of categories (biophysical, experimental, text-mining, location)



String-DB: input and output

Viewers >

Legend >

Settings >

Analysis ▾

Exports >

Clusters >

+ More

- Less

Network Stats

number of nodes: 4

number of edges: 6

average node degree: 3

avg. local clustering coefficient: 1

expected number of edges: 1

PPI enrichment p-value: 1.75e-05

your network has significantly more interactions than expected (*what does that mean?*)

Functional enrichments in your network

explain columns

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	false discovery rate
GO:0000162	tryptophan biosynthetic process	4 of 9	2.66	2.66e-09

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	false discovery rate
GO:0004834	tryptophan synthase activity	2 of 2	3.01	0.00013
GO:0016830	carbon-carbon lyase activity	2 of 71	1.46	0.0167
GO:0016829	lyase activity	4 of 195	1.33	0.00013

Reference publications (PubMed)

publication	(year) title	count in network	strength	false discovery rate
PMID:2838460	(1988) Cosmid cloning of five Zymomonas trp genes by com...	4 of 4	3.01	5.33e-09
PMID:2113923	(1990) Nucleotide sequences and genomic constitution of fiv...	4 of 4	3.01	5.33e-09
PMID:6276387	(1982) Yeast gene TRP5: structure, function, regulation.	3 of 3	3.01	3.14e-07
PMID:403178	(1977) Immunochemical comparison of phosphoribosylanthr...	3 of 3	3.01	3.14e-07
PMID:3146017	(1988) Cloning of the trp genes from the archaeobacterium M...	3 of 3	3.01	3.14e-07

(more ...)

local network cluster (STRING)

cluster	description	count in network	strength	false discovery rate
CL:1345	Tryptophan biosynthesis	4 of 5	2.92	7.28e-11

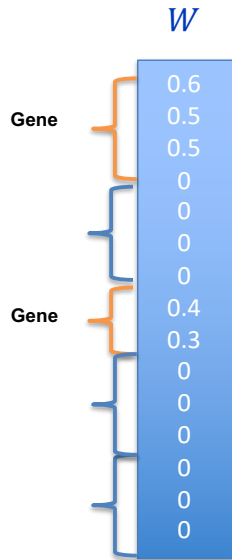
KEGG Pathways

pathway	description	count in network	strength	false discovery rate
eco00400	Phenylalanine, tyrosine and tryptophan biosynthesis	4 of 21	2.29	6.26e-09
eco00260	Glycine, serine and threonine metabolism	2 of 37	1.75	0.00062
eco01230	Biosynthesis of amino acids	4 of 116	1.55	2.03e-06
eco01130	Biosynthesis of antibiotics	4 of 209	1.3	1.38e-05
eco01110	Biosynthesis of secondary metabolites	4 of 301	1.14	4.38e-05

(more ...)



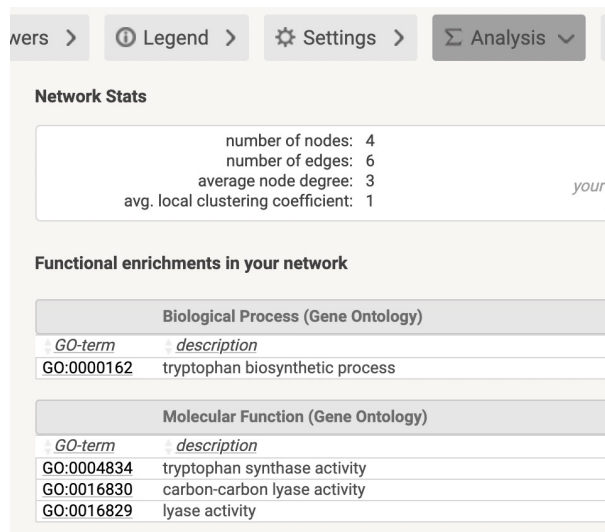
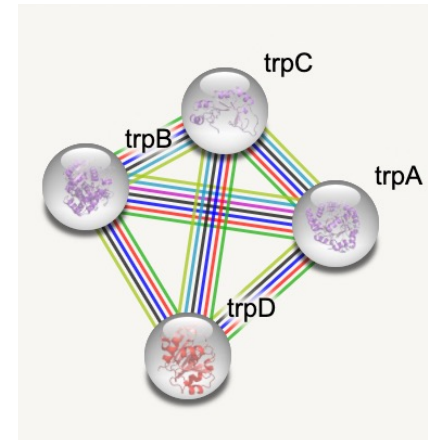
Omic analysis workflow



Estimate
important
genes/proteins

Input in String-
DB

Interpret
clusters and
annotations



Other gene enrichment databases available

We will also consider DisGeNet

- A database of gene-disease associations
- Contains information about potential links between gene sets and diseases
- <https://www.disgenet.org/>
- Also available as an R package
 - Nowadays need to register an account



Exercise: Down syndrome case study

- Apply some of the techniques to investigate the relation between Down syndrome and multi-omics data
- Use statistical measures and bioinformatics databases to interpret the results



Part 1: overview

- Principal Component Analysis (PCA)
- Partial Least Squares (PLS)
- Two-way Orthogonal PLS (O2PLS)
- Post-hoc analyses using external databases
 - String-DB and gene enrichment

