Bioinformatics

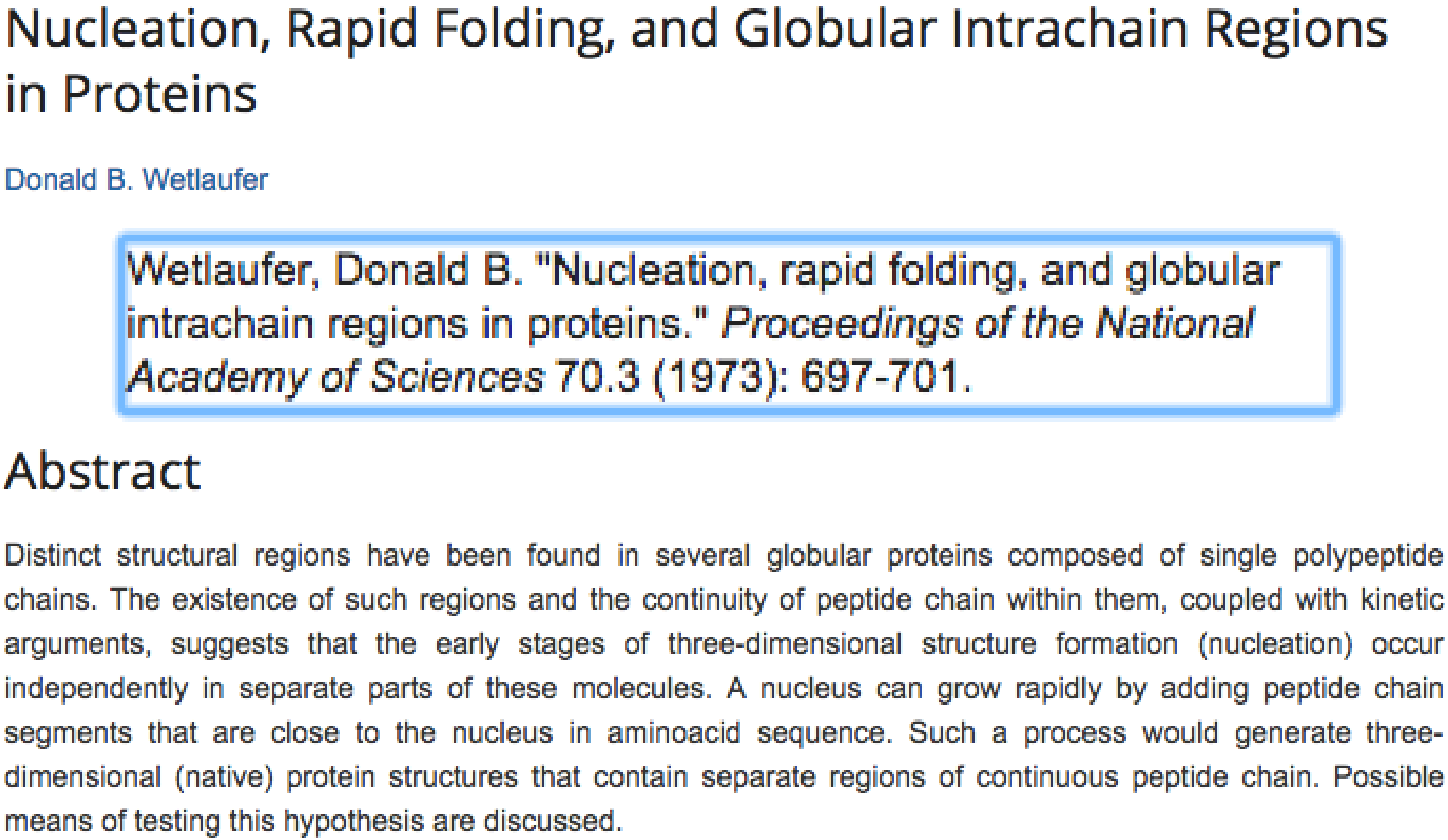
CS300

Domains according to UniProt and String

Spring 2021

Oliver BONHAM-CARTER

Proteins Fold Into Specific Structures for Functionality



Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins

Donald B. Wetlaufer

Wetlaufer, Donald B. "Nucleation, rapid folding, and globular intrachain regions in proteins." *Proceedings of the National Academy of Sciences* 70.3 (1973): 697-701.

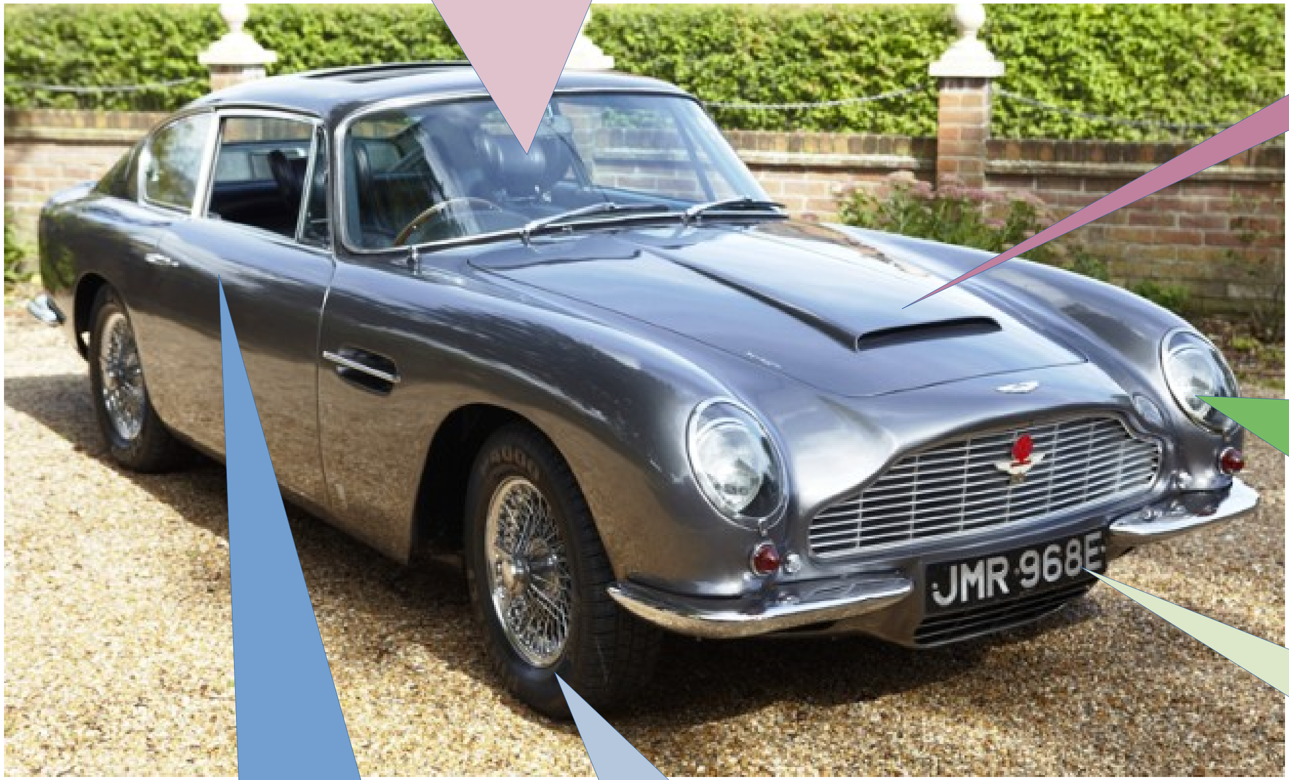
Abstract

Distinct structural regions have been found in several globular proteins composed of single polypeptide chains. The existence of such regions and the continuity of peptide chain within them, coupled with kinetic arguments, suggests that the early stages of three-dimensional structure formation (nucleation) occur independently in separate parts of these molecules. A nucleus can grow rapidly by adding peptide chain segments that are close to the nucleus in aminoacid sequence. Such a process would generate three-dimensional (native) protein structures that contain separate regions of continuous peptide chain. Possible means of testing this hypothesis are discussed.

Different regions in same protein (*domains*) performing specific tasks.

Structures For Functions

One Car, Many Functions



Windows to allow driver to see out while driving

Ventilation for cooling

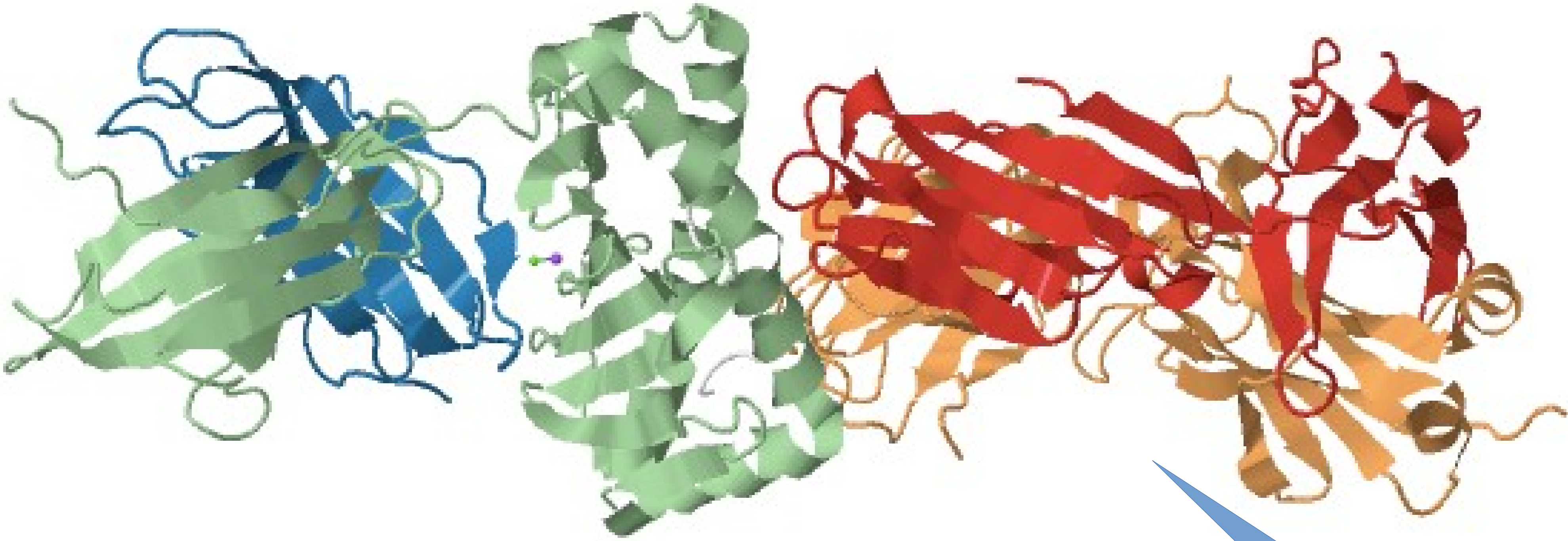
Headlights to illuminate the road when driving at night

License plate: for Identification

Wheels, necessary for mobility

Door to allow driver to enter the car

Proteins Also Have Specific Functional Regions, Too!



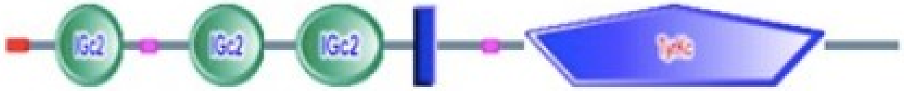
Protein Data Bank:

5WLG

Click! This is a link!

Domains

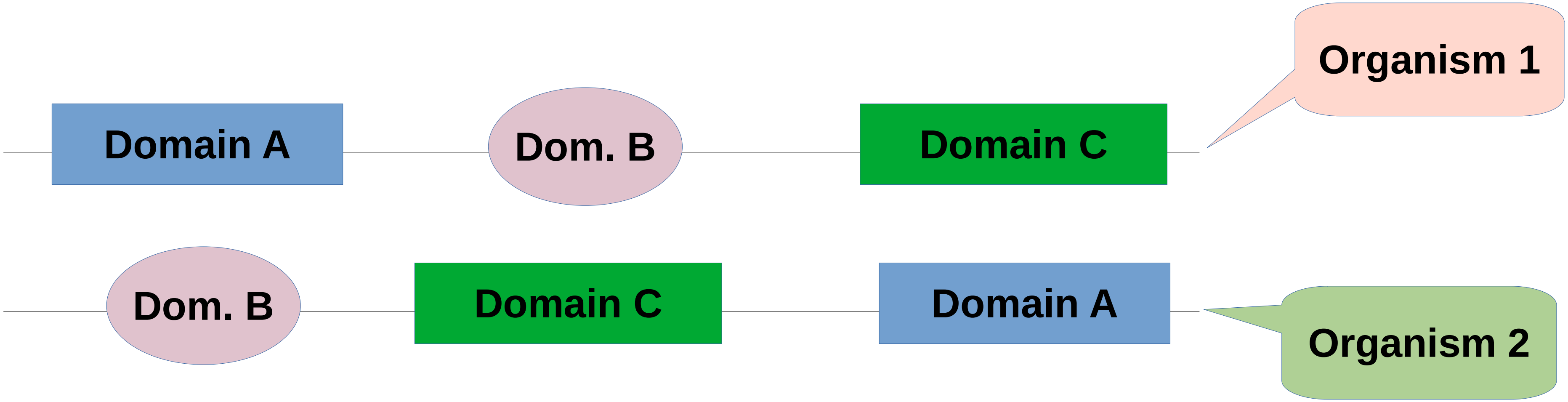
- A protein **domain** is a conserved part of a given protein sequence and (tertiary) structure.
- Can evolve, function, and exist independently of the rest of the protein chain
- Each domain forms a compact three-dimensional structure
- Often can be independently stable and folded.



SMART domain 'bubblegram' for human fibroblast growth factor (FGF) receptor 1 (type P11362 into web site: smart.embl.de)

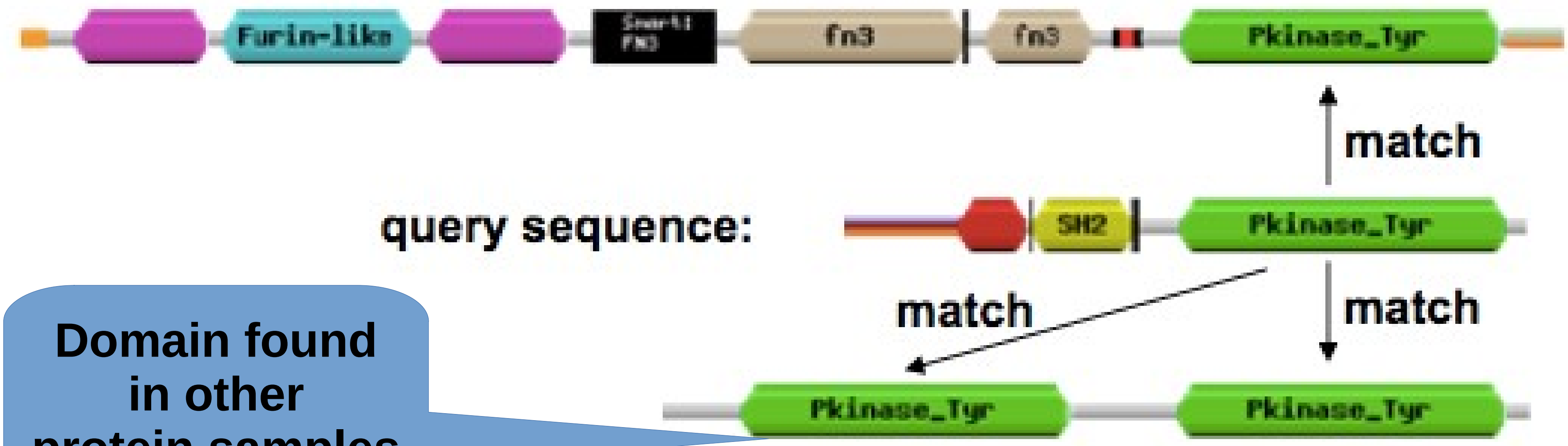
Domain Synteny

- A different order of domains in genomes
- Could provide information about relatedness across genome samples.
- Article: *Domain team: synteny of domains is a new approach in comparative genomics*
  - https://pubmed.ncbi.nlm.nih.gov/18025683/



Finding a Domain?

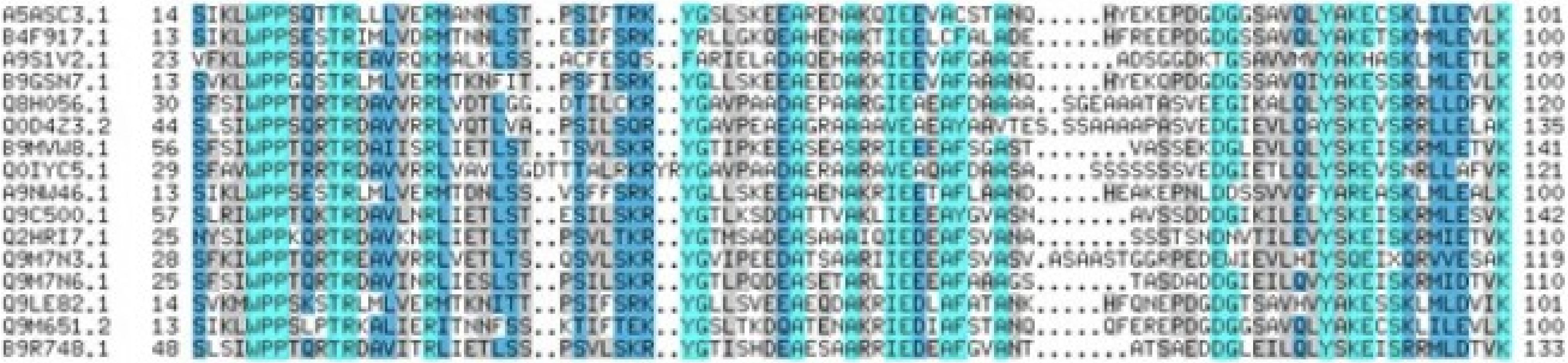
- Alignment across proteins may show domains
- Use databases to align and match protein subsections
  - Pfam, Smart, Interpro and other database tools



Domain found in other protein samples

Alignment?!

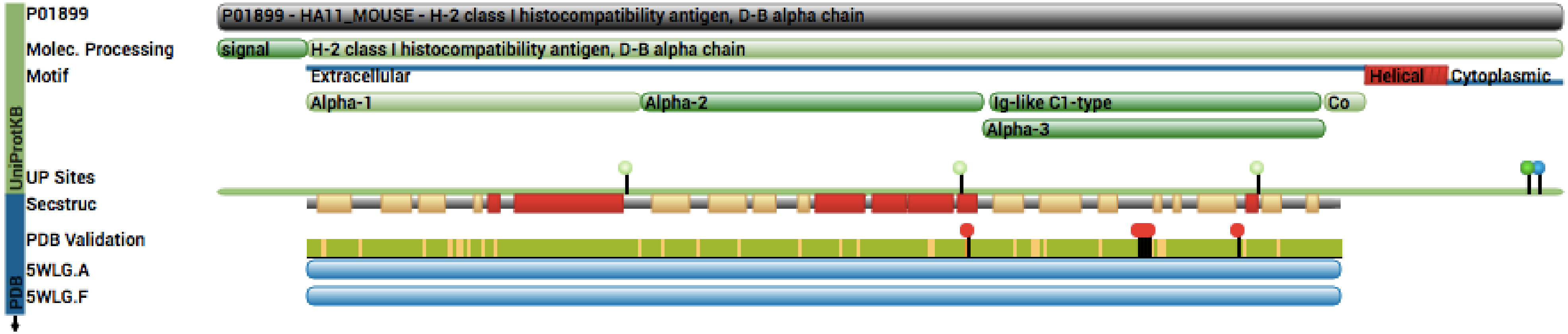
- Provide more info about a protein’s family, relatedness and other details.
- Domain landmarks include: low-complexity or disorder to suggest that these regions may have a specific syntax or pronounced grammar.



```
A5ASC3.1   14 SIKLWPPSQTTRLLLVERMANNLST..PSIFTRK..YGSLSKEEAPENAKQIEEVACSTANQ.....HYEKEPDGDGGSAVQLYAKECSKLILEVLK 101
B4F917.1   13 SIKLWPPSESTRIMLVDRMTNNLST..ESIFSRK..YRLLGKQEAHENAKTIEELCFALADE.....HFREEPDGDGSSAVQLYAKETSKMMLEVLK 100
A9S1V2.1   23 VFKLWPPSQGTREAVRQKMALKLSS..ACFESQS..FARIELADAQEHARAIEEVAFGAAQE......ADSGGDKTGSAVVMVYAKHASKLMLETLR 109
B9GSN7.1   13 SVKLWPPGQSTRLMLVERMTKNFIT..PSFISRK..YGLLSKEEAEEDAKKIEEVAFAAANQ.....HYEKQPDGDGSSAVQIYAKESSRLMLEVLK 100
Q8H056.1   30 SFSIWPPTQRTRDAVVRRLVDTLGG..DTILCKR..YGAVPAADAEPAARGIEAEAFDAAAA..SGEAAATASVEEGIKALQLYSKEVSRRLLDFVK 120
Q0D4Z3.2   44 SLSIWPPSQRTRDAVVRRLVQTLVA..PSILSQR..YGAVPEAEAGRAAAAVEAEAYAAVTES.SSAAAAPASVEDGIEVLQAYSKEVSRRLLELAK 135
B9MVW8.1   56 SFSIWPPTQRTRDAIISRLIETLST..TSVLSKR..YGTIPKEEASEASRRIEEEAFSGAST.......VASSEKDGLEVLQLYSKEISKRMLETVK 141
Q0IYC5.1   29 SFAVWPPTRRTRDAVVRRLVAVLSGDTTTALPKRYRYGAVPAADAERAARAVEAQAFDAASA....SSSSSSSVEDGIETLQLYSREVSNRLLAFVR 121
A9NW46.1   13 SIKLWPPSESTRLMLVERMTDNLSS..VSFFSRK..YGLLSKEEAAENAKRIEETAFLAAND.....HEAKEPNLDDSSVVQFYAREASKLMLEALK 100
Q9C500.1   57 SLRIWPPTQKTRDAVLNRLIETLST..ESILSKR..YGTLKSDDATTVAKLIEEEAYGVASN.......AVSSDDDGIKILELYSKEISKRMLESVK 142
Q2HRI7.1   25 NYSIWPPKQRTRDAVKNRLIETLST..PSVLTKR..YGTHSADEASAAAIQIEDEAFSVANA.......SSSTSNDHVTILEVYSKEISKRMIETVK 110
Q9M7N3.1   28 SFKIWPPTQRTREAVVRRLVETLTS..QSVLSKR..YGVIPEEDATSAARIIEEEAFSVASV.ASAASTGGRPEDEWIEVLHIYSQEIXQRVVESAK 119
Q9M7N6.1   25 SFSIWPPTQRTRDAVINRLIESLST..PSILSKR..YGTLPQDEASETARLIEEEAFAAAGS.......TASDADDGIEILQVYSKEISKRMIDTVK 110
Q9LE82.1   14 SVKMWPPSKSTRLMLVERMTKNITT..PSIFSRK..YGLLSVEEAEQDAKRIEDLAFATANK.....HFQNEPDGDGTSAVHVYAKESSKLMLDVIK 101
Q9M651.2   13 SIKLWPPSLPTRKALIERITNNFSS..KTIFTEK..YGSLTKDQATENAKRIEDIAFSTANQ.....QFEREPDGDGGSAVQLYAKECSKLILEVLK 100
B9R748.1   48 SLSIWPPTQRTRDAVITRLIETLSS..PSVLSKR..YGTISHDEAESAARRIEDEAFGVANT.......ATSAEDDGLEILQLYSKEISRRMLDTVK 133
```

Domains By PDB

- Domains give the protein special qualities:
  - Domain Names: *Alpha1, Alpha2, Alpha3, Ig-like C1-type*

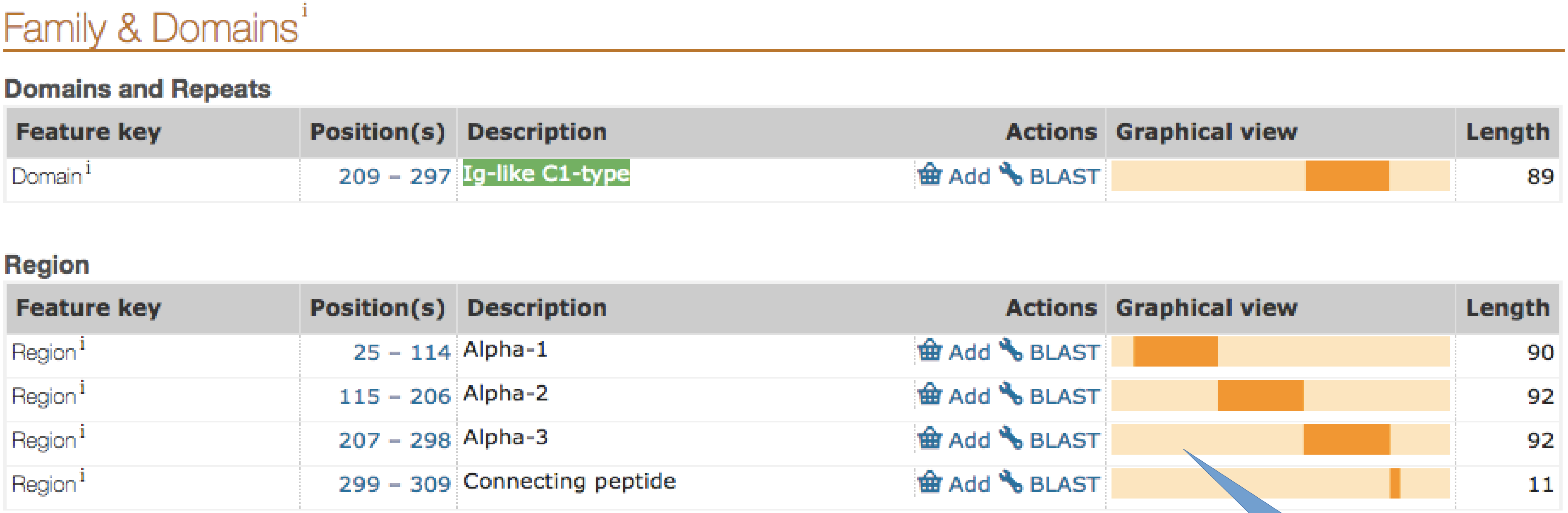


Protein Data Bank ID: 5WLG

- This protein:
  - https://www.rcsb.org/pdb/explore/explore.do?structureId=5WLG
- Domains:
  - https://www.rcsb.org/Annotations/5WLG
- Help with features
  - https://www.rcsb.org/pages/help/featureView

Domains By Uniprot

- Domains give the protein special qualities:
  - Domain Names *Alpha1 (and etc.) can be Blasted to find copies in other proteins*



Family & Domains

Domains and Repeats

Feature key	Position(s)	Description	Actions	Graphical view	Length
Domain	209 – 297	Ig-like C1-type	Add BLAST		89

Region

Feature key	Position(s)	Description	Actions	Graphical view	Length
Region	25 – 114	Alpha-1	Add BLAST		90
Region	115 – 206	Alpha-2	Add BLAST		92
Region	207 – 298	Alpha-3	Add BLAST		92
Region	299 – 309	Connecting peptide	Add BLAST		11

UniProt ID: P01899

A Protein Knowledge Base

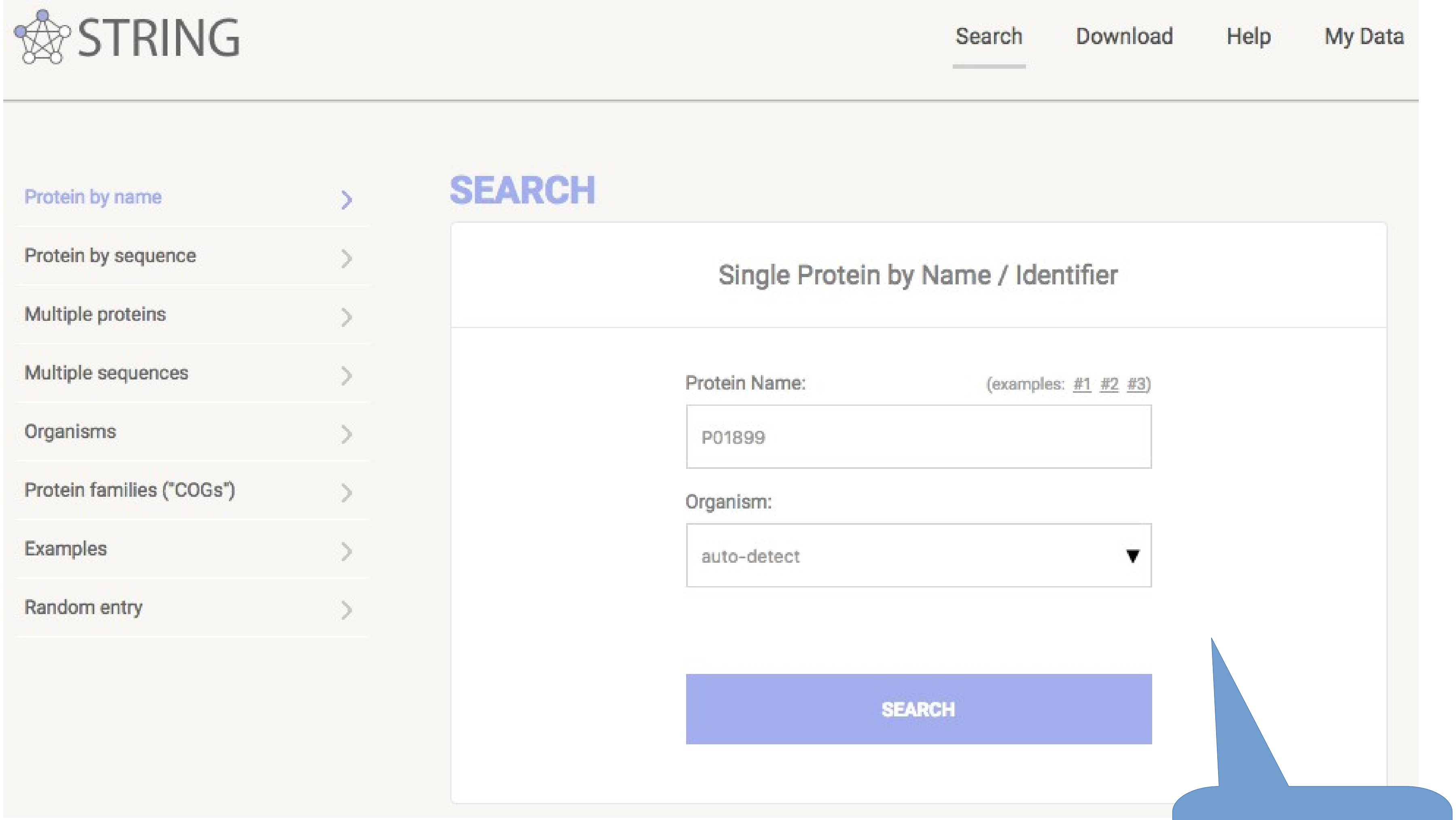
http://www.uniprot.org/uniprot/P01899#family_and_domains

Click! This is a link!

Bring the Tool!

Up Next!

STRING: Functional Protein Association Networks



Protein Name: (examples: #1 #2 #3)

P01899

Organism:

auto-detect ▼

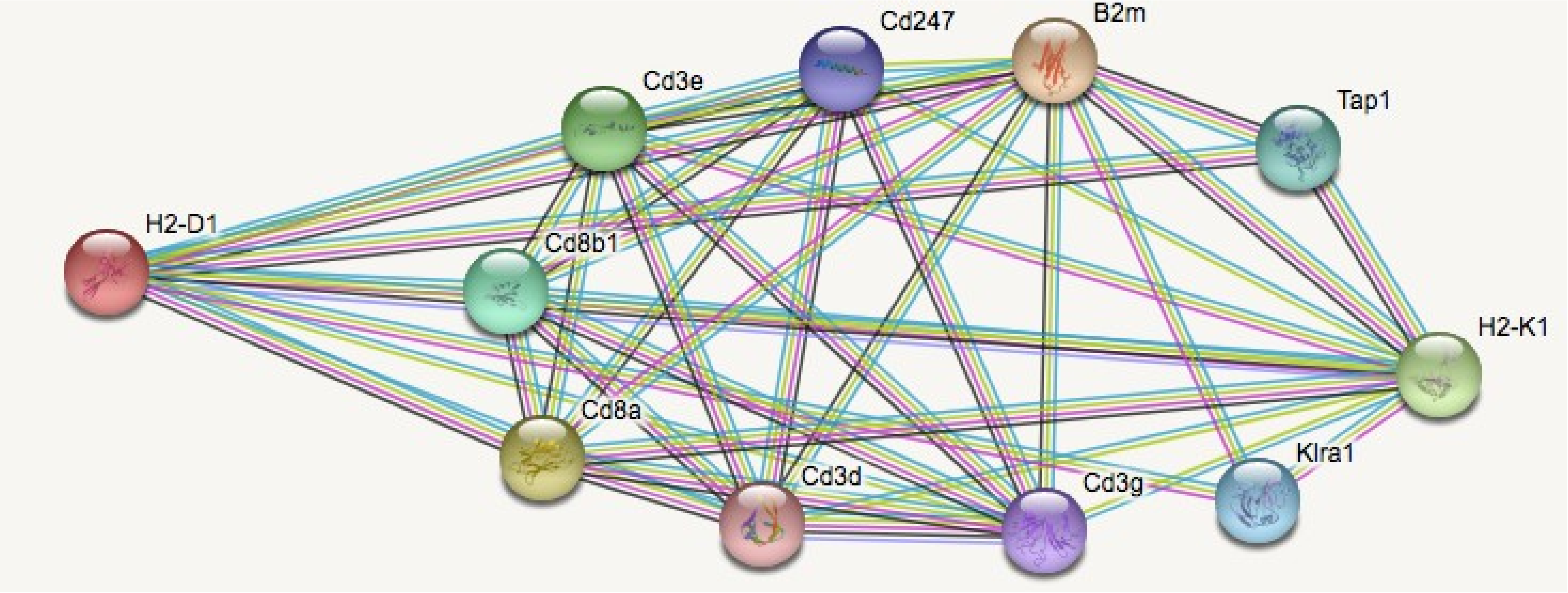
String DB ID P01899

http://string-db.org/

Click! This is a link!

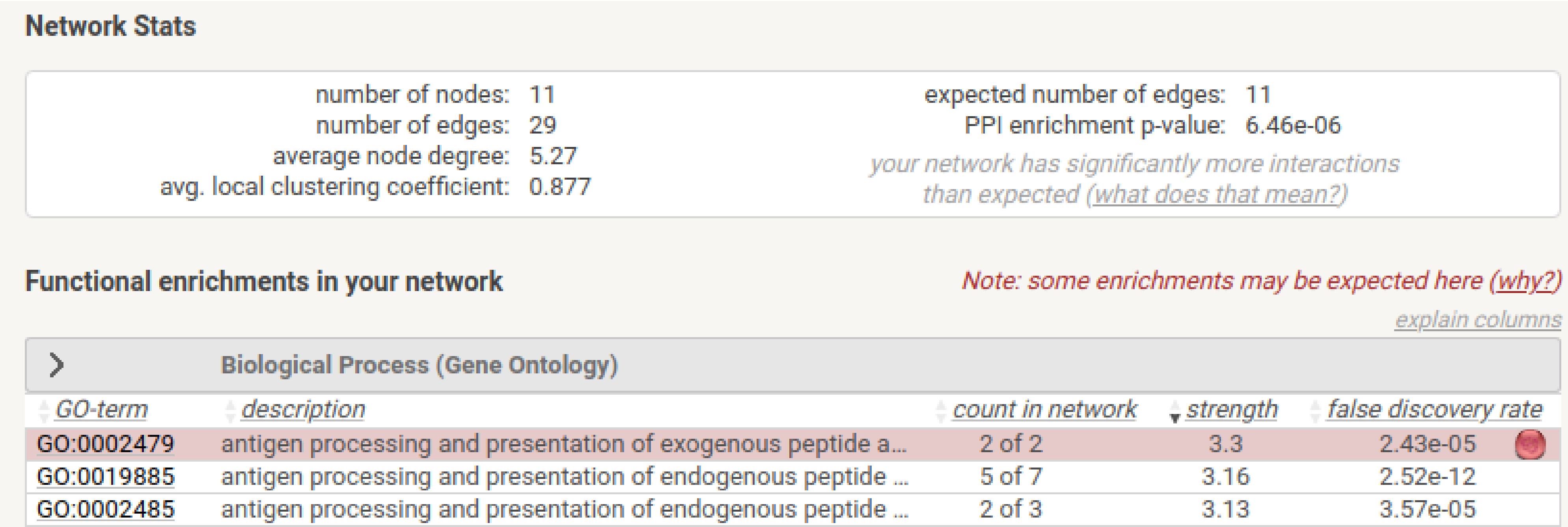
STRING: Functional Protein Association Networks

- Known and predicted protein-protein interactions
- How does a protein interact with others?
- What types of interactions are these (across all known genomes, of any organism)?



STRING: Functional Protein Association Networks

- What types of interactions are happening and where?



Network Stats

number of nodes: 11
number of edges: 29
average node degree: 5.27
avg. local clustering coefficient: 0.877

expected number of edges: 11
PPI enrichment p-value: 6.46e-06
your network has significantly more interactions than expected (what does that mean?)

Functional enrichments in your network

Note: some enrichments may be expected here (why?)

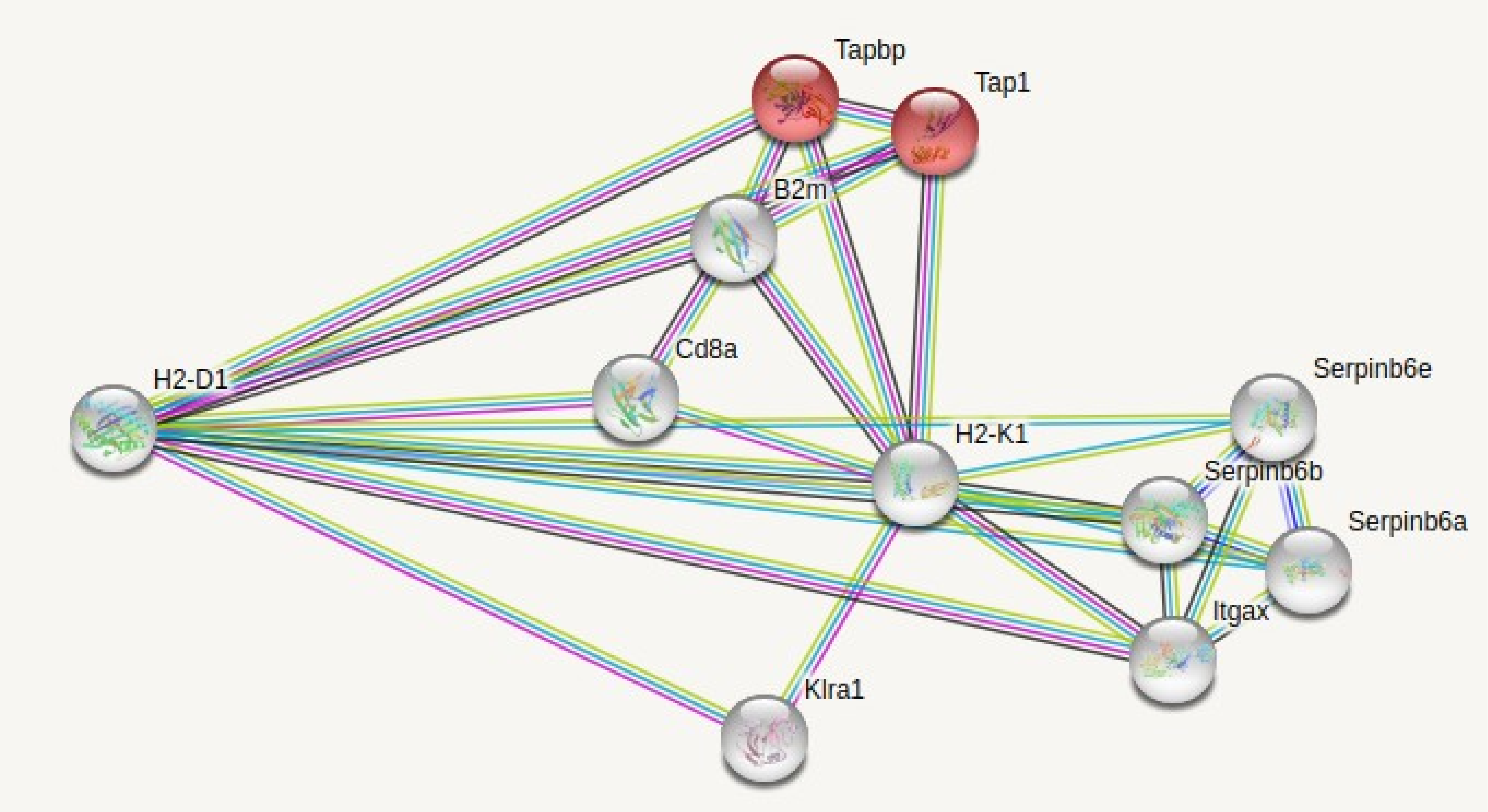
explain columns

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	false discovery rate
GO:0002479	antigen processing and presentation of exogenous peptide a...	2 of 2	3.3	2.43e-05 ●
GO:0019885	antigen processing and presentation of endogenous peptide ...	5 of 7	3.16	2.52e-12
GO:0002485	antigen processing and presentation of endogenous peptide ...	2 of 3	3.13	3.57e-05

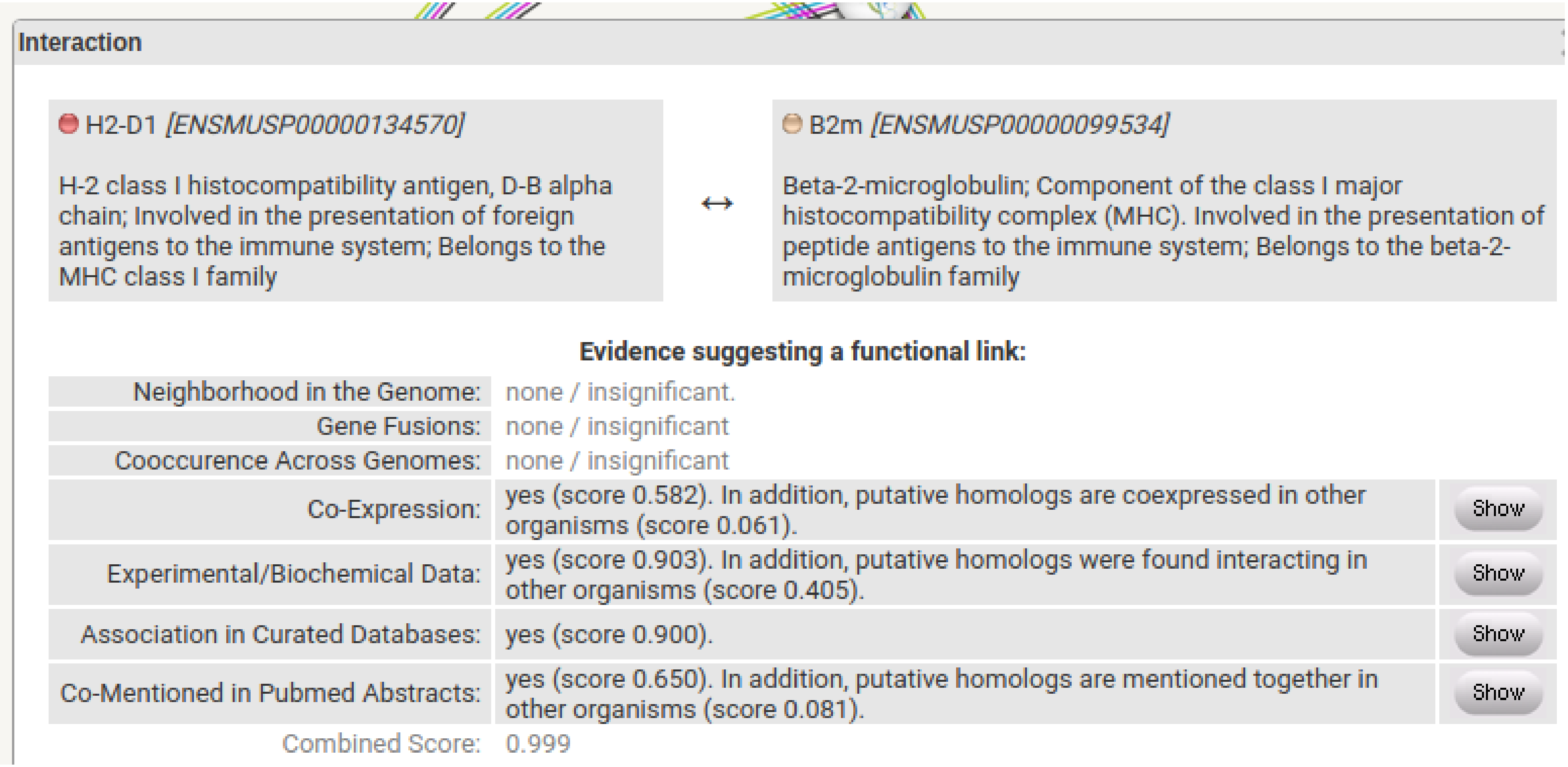
STRING: Functional Protein Association Networks

- These nodes play roles in the interaction.



STRING: Functional Protein Association Networks

- Click on an edge to see the type of interaction



Interaction

● H2-D1 *[ENSMUSP00000134570]*

H-2 class I histocompatibility antigen, D-B alpha chain; Involved in the presentation of foreign antigens to the immune system; Belongs to the MHC class I family

↔

● B2m *[ENSMUSP00000099534]*

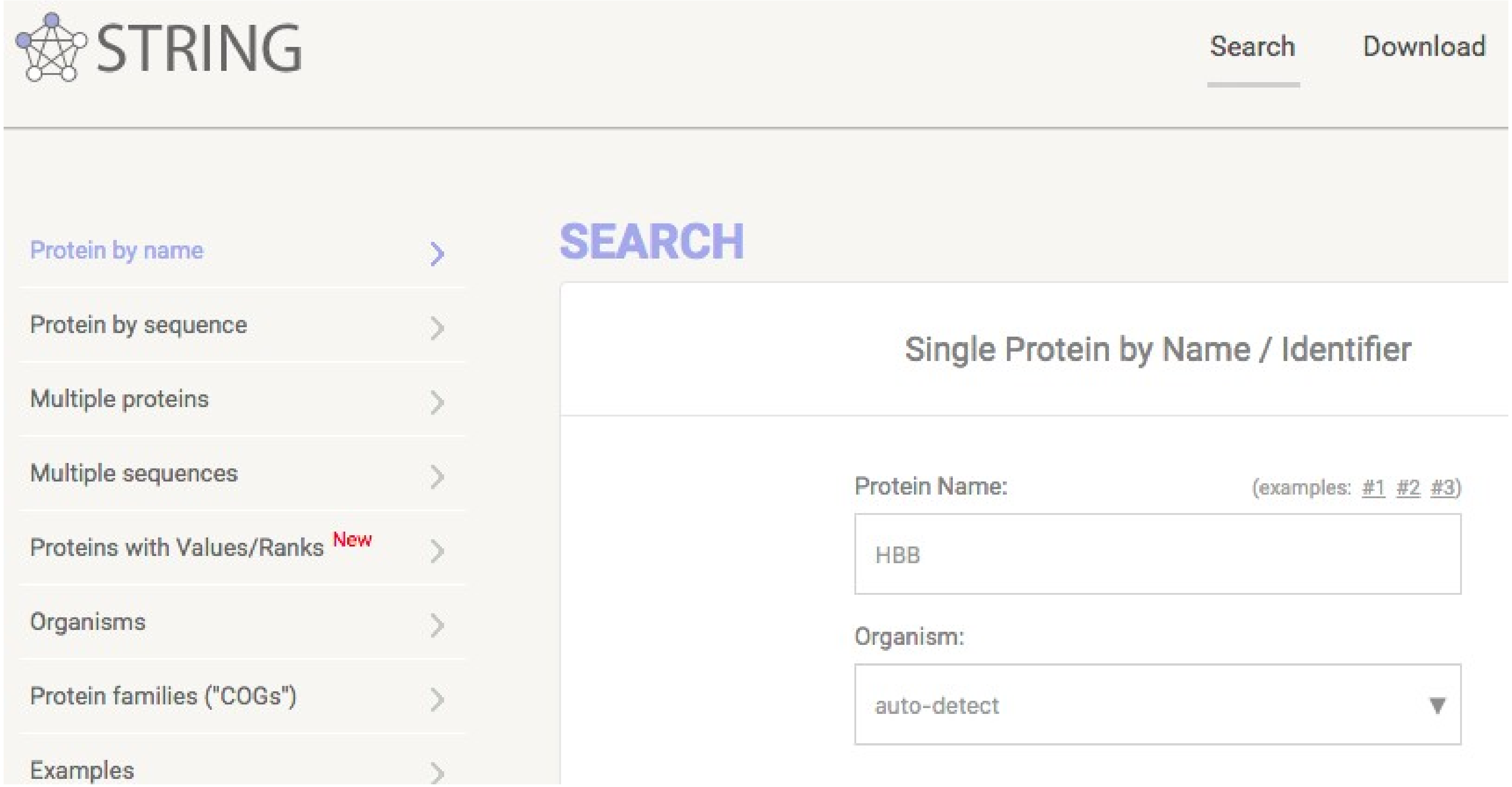
Beta-2-microglobulin; Component of the class I major histocompatibility complex (MHC). Involved in the presentation of peptide antigens to the immune system; Belongs to the beta-2-microglobulin family

Evidence suggesting a functional link:

Neighborhood in the Genome:	none / insignificant.	
Gene Fusions:	none / insignificant	
Cooccurence Across Genomes:	none / insignificant	
Co-Expression:	yes (score 0.582). In addition, putative homologs are coexpressed in other organisms (score 0.061).	Show
Experimental/Biochemical Data:	yes (score 0.903). In addition, putative homologs were found interacting in other organisms (score 0.405).	Show
Association in Curated Databases:	yes (score 0.900).	Show
Co-Mentioned in Pubmed Abstracts:	yes (score 0.650). In addition, putative homologs are mentioned together in other organisms (score 0.081).	Show
Combined Score:	0.999	

STRING: Functional Protein Association Networks

- Question: What proteins (from genes) interact with **HBB** protein (from the gene)?



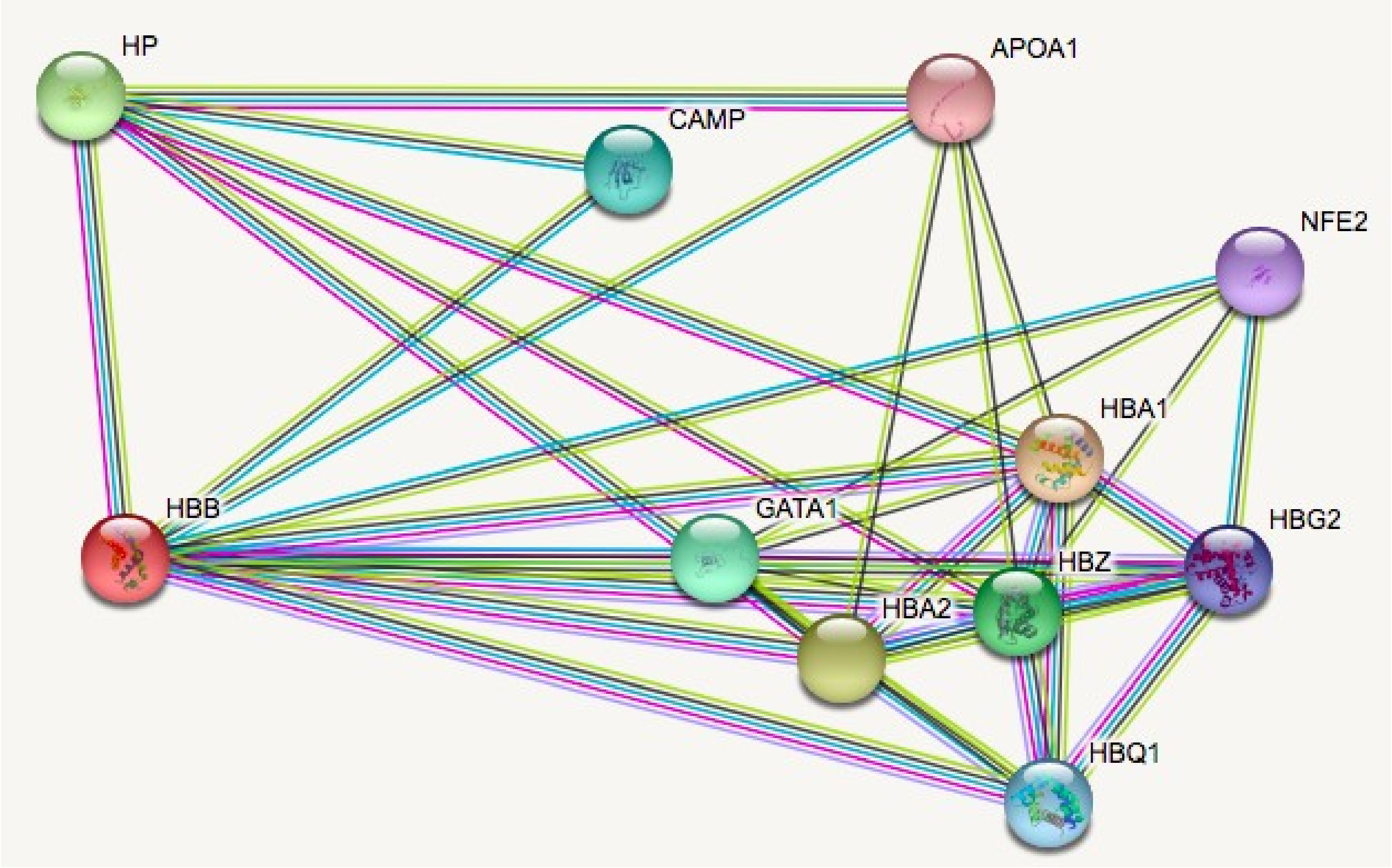
Protein Name: HBB

Organism: auto-detect ▼

https://string-db.org/

STRING: Functional Protein Association Networks

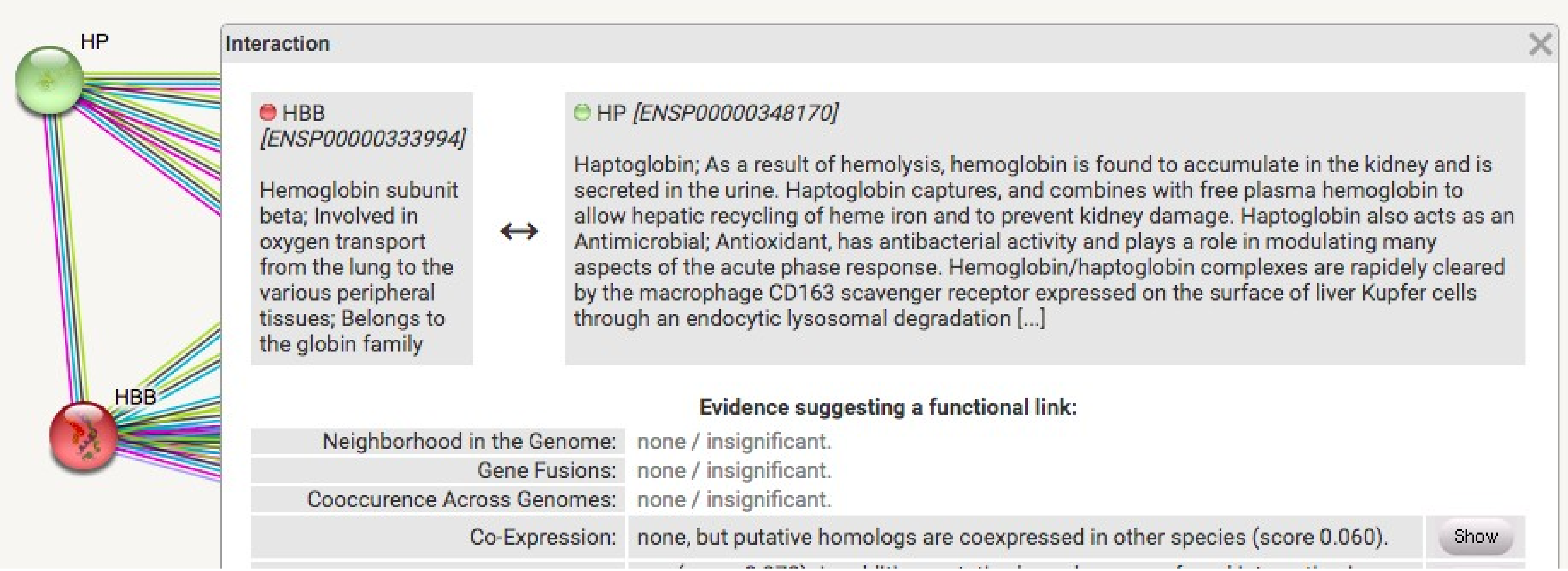
- Answer: Lots!



https://string-db.org/

STRING: Functional Protein Association Networks

- What kinds of interactions?



🟥 HBB *[ENSP00000333994]*

Hemoglobin subunit beta; Involved in oxygen transport from the lung to the various peripheral tissues; Belongs to the globin family

↔

🟩 HP *[ENSP00000348170]*

Haptoglobin; As a result of hemolysis, hemoglobin is found to accumulate in the kidney and is secreted in the urine. Haptoglobin captures, and combines with free plasma hemoglobin to allow hepatic recycling of heme iron and to prevent kidney damage. Haptoglobin also acts as an Antimicrobial; Antioxidant, has antibacterial activity and plays a role in modulating many aspects of the acute phase response. Hemoglobin/haptoglobin complexes are rapidly cleared by the macrophage CD163 scavenger receptor expressed on the surface of liver Kupfer cells through an endocytic lysosomal degradation [...]

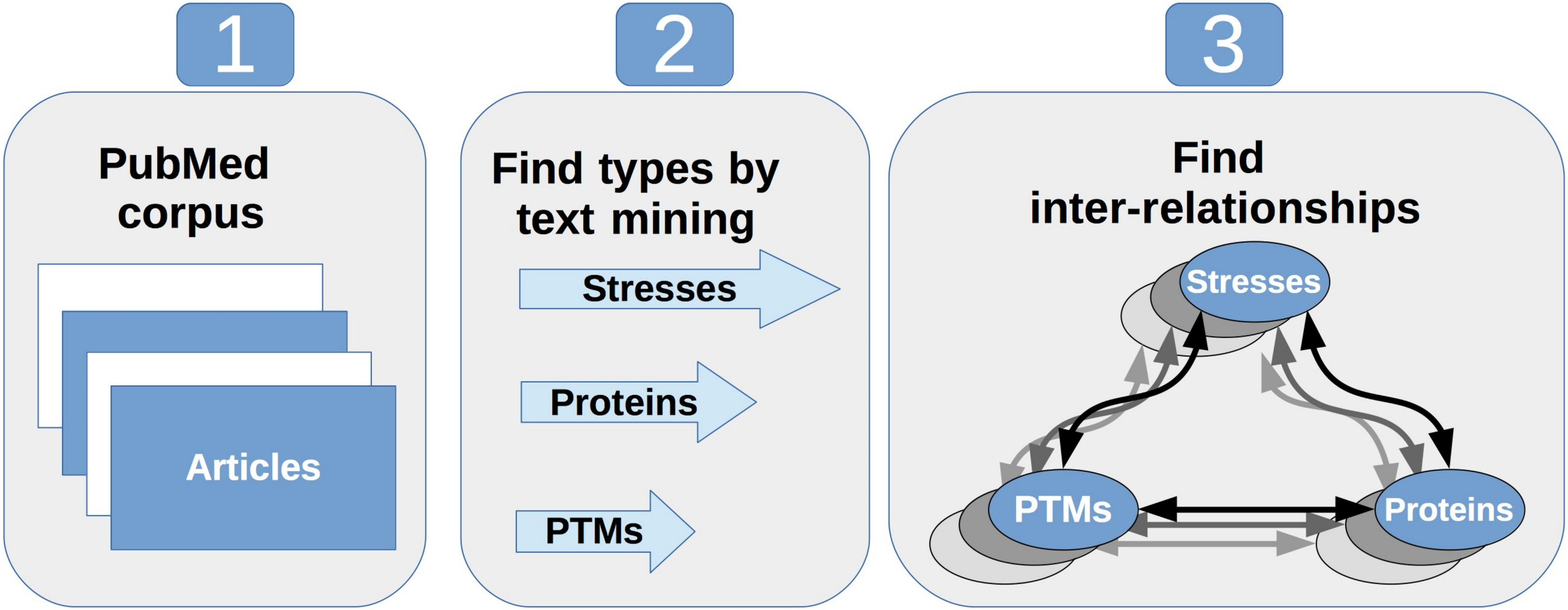
Evidence suggesting a functional link:

Neighborhood in the Genome:	none / insignificant.
Gene Fusions:	none / insignificant.
Cooccurence Across Genomes:	none / insignificant.
Co-Expression:	none, but putative homologs are coexpressed in other species (score 0.060).

https://string-db.org/

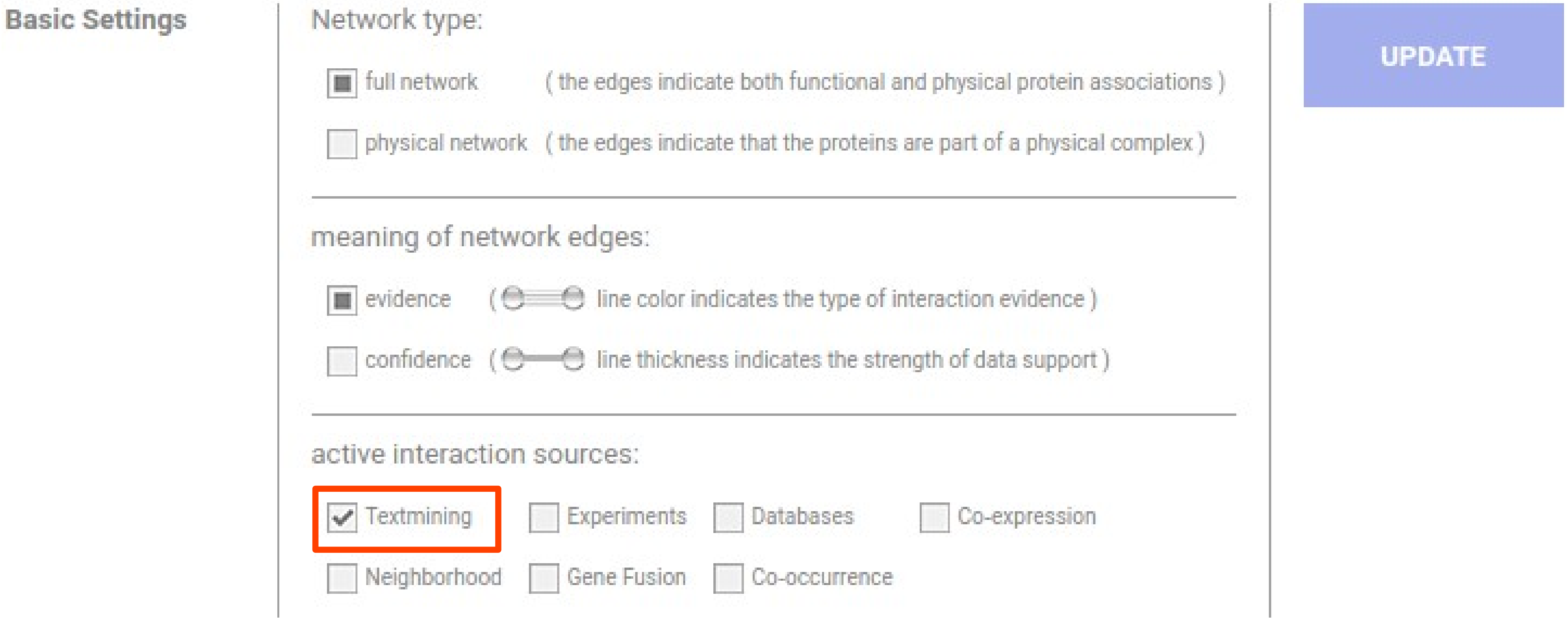
Criteria to Determine Relations

- There are many ways to measure the distance between two different proteins
  - Text Mining



String: by Text Mining

- HBB’s interactions according to the literature
- Go to SETTINGs and select only “Textmining”



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:

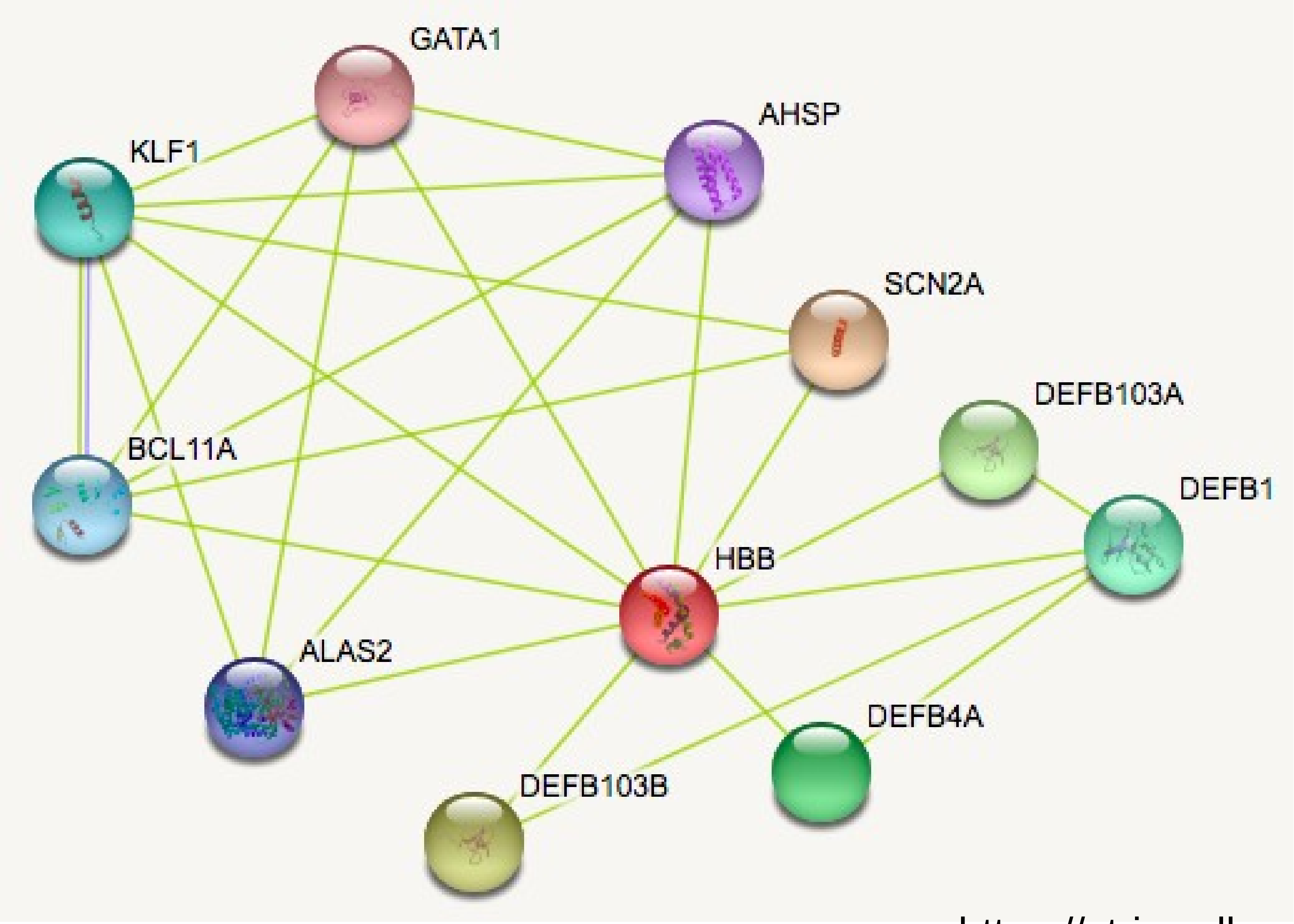
☑ Textmining □ Experiments □ Databases □ Co-expression

□ Neighborhood □ Gene Fusion □ Co-occurrence

https://string-db.org/

String: by Text Mining

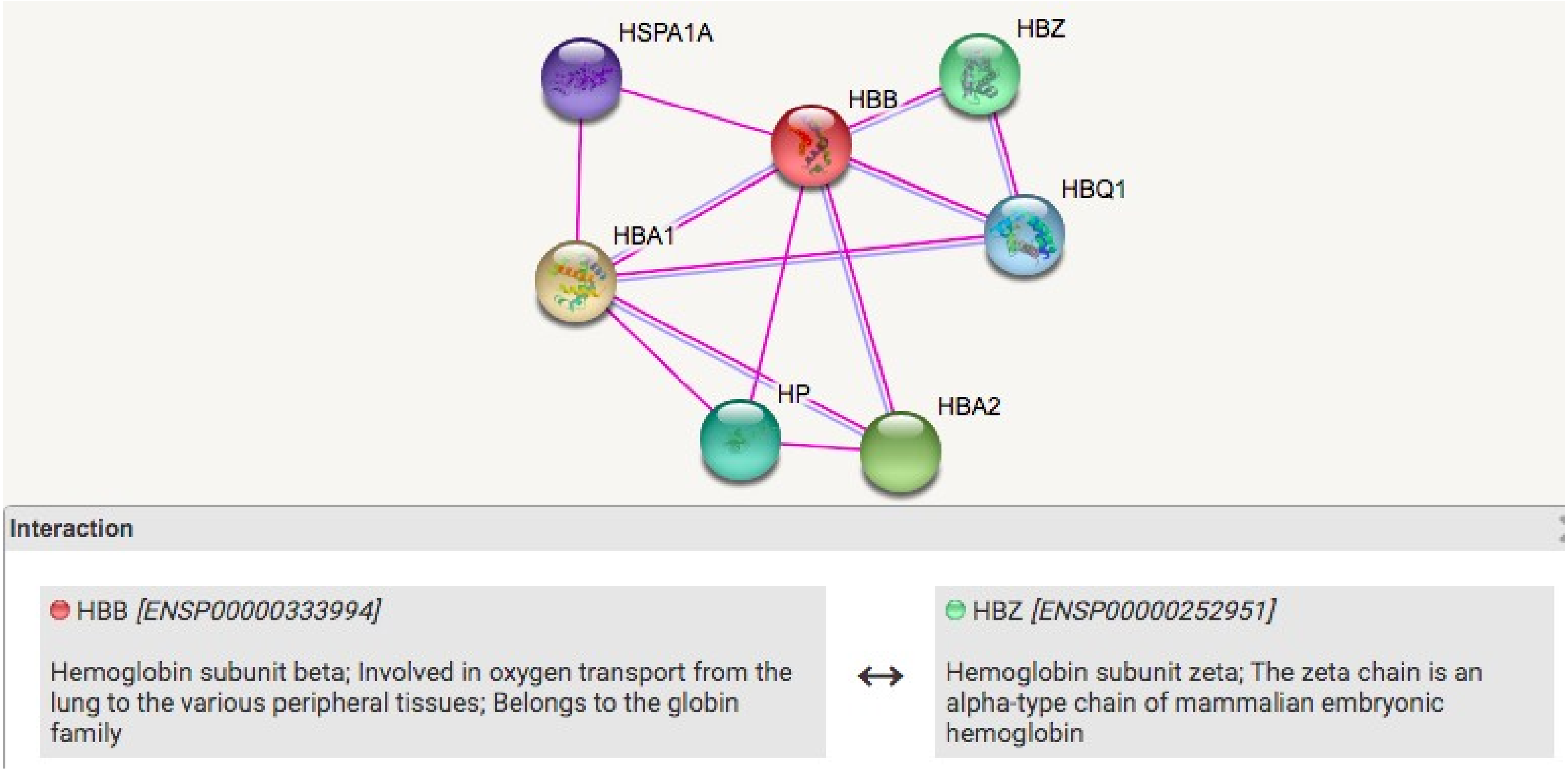
- HBB’s interactions according to the literature



https://string-db.org/

String: Linked Experimentally

- Experiments performed to show that protein are related



Interaction

● HBB *[ENSP00000333994]*

Hemoglobin subunit beta; Involved in oxygen transport from the lung to the various peripheral tissues; Belongs to the globin family

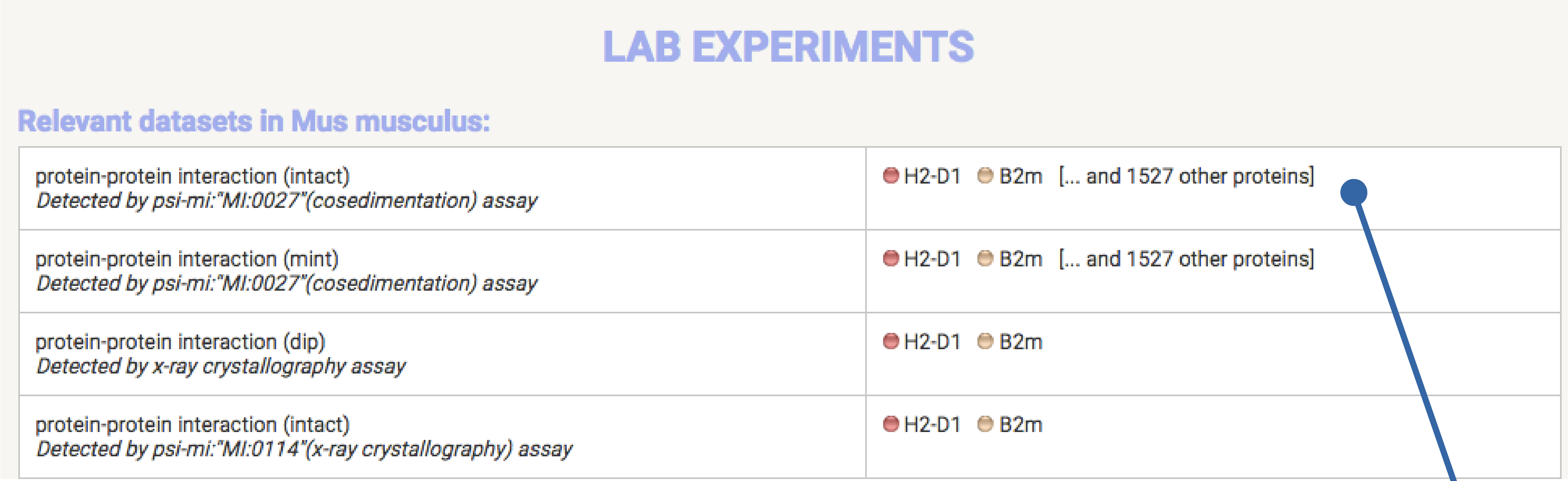
↔

● HBZ *[ENSP00000252951]*

Hemoglobin subunit zeta; The zeta chain is an alpha-type chain of mammalian embryonic hemoglobin

String: Linked Experimentally

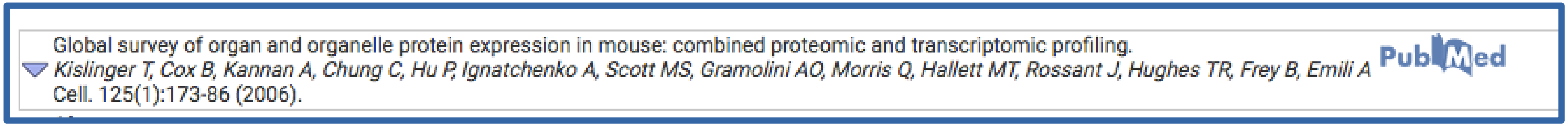
- Learn about the experiments



LAB EXPERIMENTS

Relevant datasets in Mus musculus:

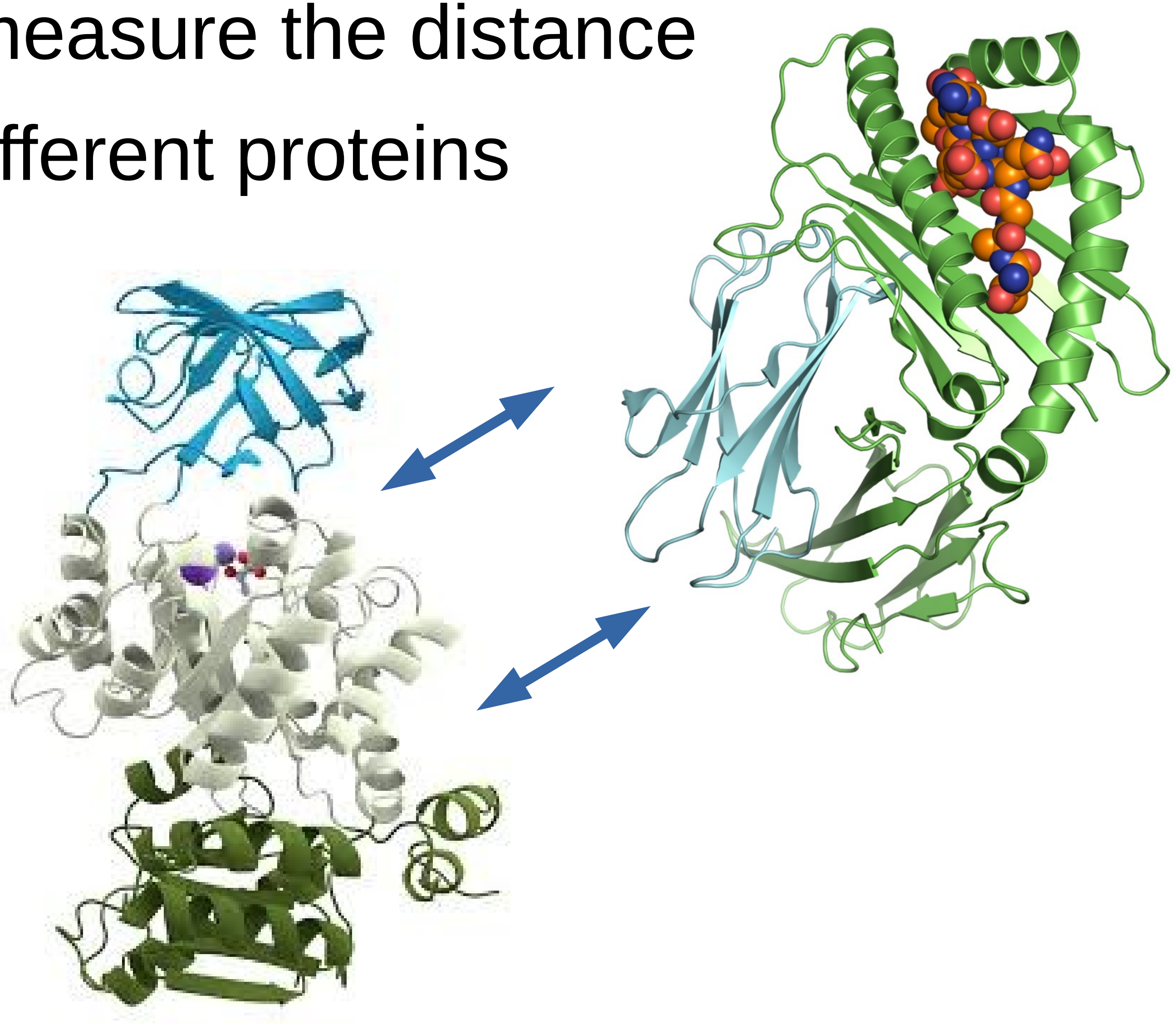
protein-protein interaction (intact) *Detected by psi-mi:"MI:0027"(cosedimentation) assay*	● H2-D1 ● B2m [... and 1527 other proteins]
protein-protein interaction (mint) *Detected by psi-mi:"MI:0027"(cosedimentation) assay*	● H2-D1 ● B2m [... and 1527 other proteins]
protein-protein interaction (dip) *Detected by x-ray crystallography assay*	● H2-D1 ● B2m
protein-protein interaction (intact) *Detected by psi-mi:"MI:0114"(x-ray crystallography) assay*	● H2-D1 ● B2m



Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling.
Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, Scott MS, Gramolini AO, Morris Q, Hallett MT, Rossant J, Hughes TR, Frey B, Emili A
Cell. 125(1):173-86 (2006).

Criteria to Determine Relations

- Other ways to measure the distance between two different proteins
  - Neighborhood
  - Experiments
  - Databases
  - Co-Expression
  - And others...



More Information?

- Unify the representation of gene and gene product attributes across all species information
  - **AmiGO 2: Gene ontology**
    - http://amigo.geneontology.org/amigo/landing

- Information of effects of genetic variation on human health
  - **Genetics Home Reference**
    - https://ghr.nlm.nih.gov/

Go Play!

- Pick your favorite protein and get gene name
  - http://www.uniprot.org/

    example: P01899, gene name: H2-D1
- Then check out its networks at:
- https://string-db.org/