Bioinformatics

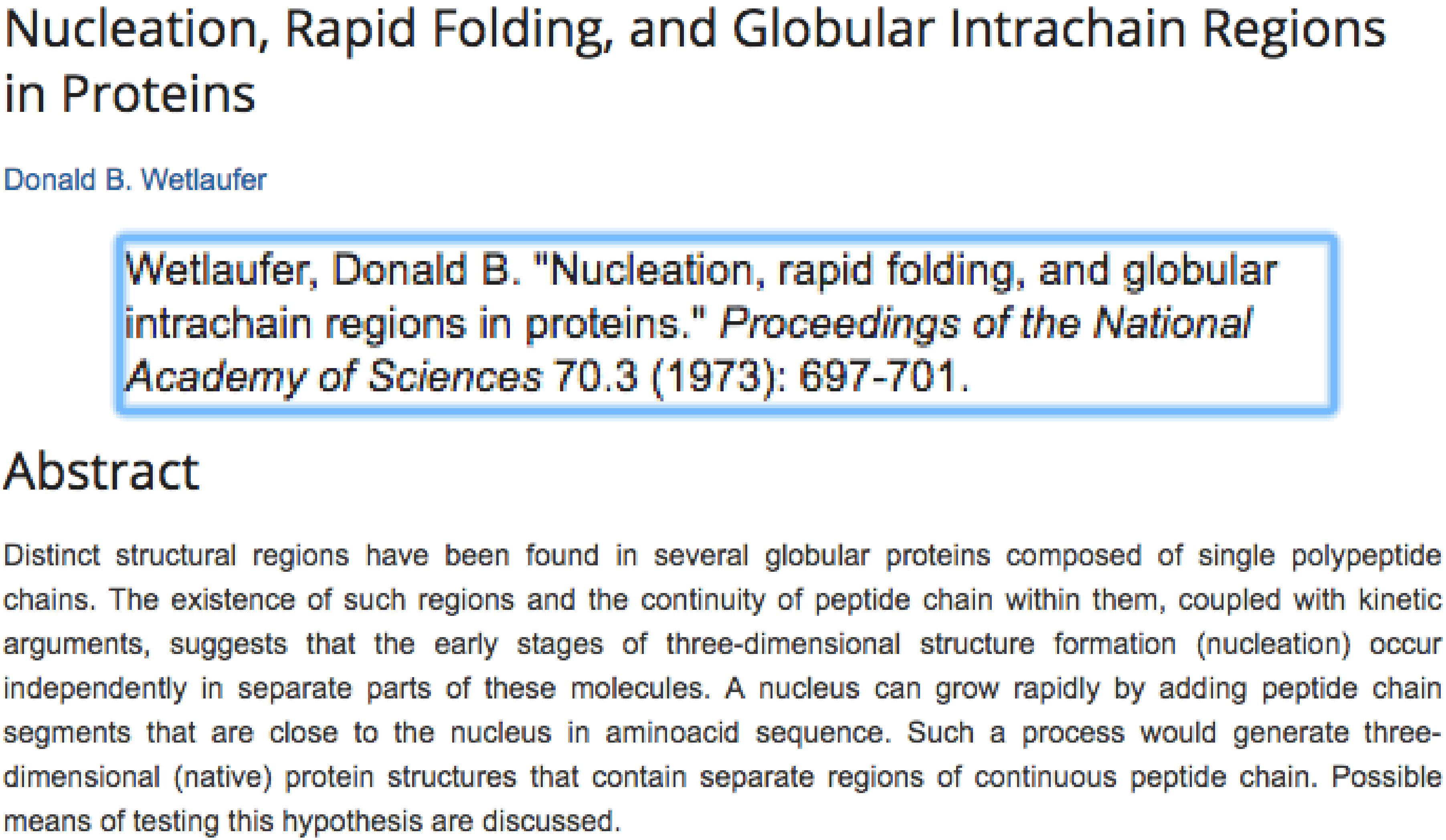
CS300

Domains according to UniProt and String

Fall 2017

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.

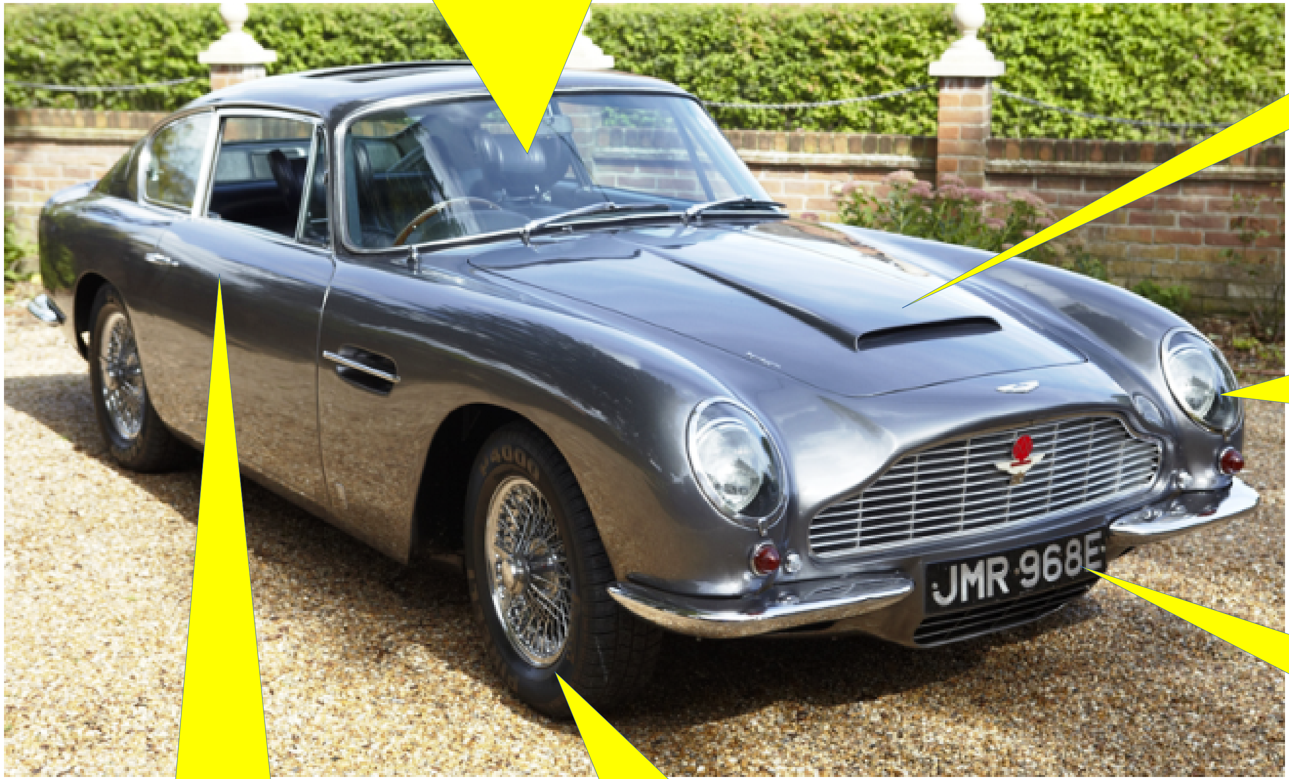
Abstract

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

Structures for Functions



Windows to allow driver to see out while driving

Ventilation for cooling

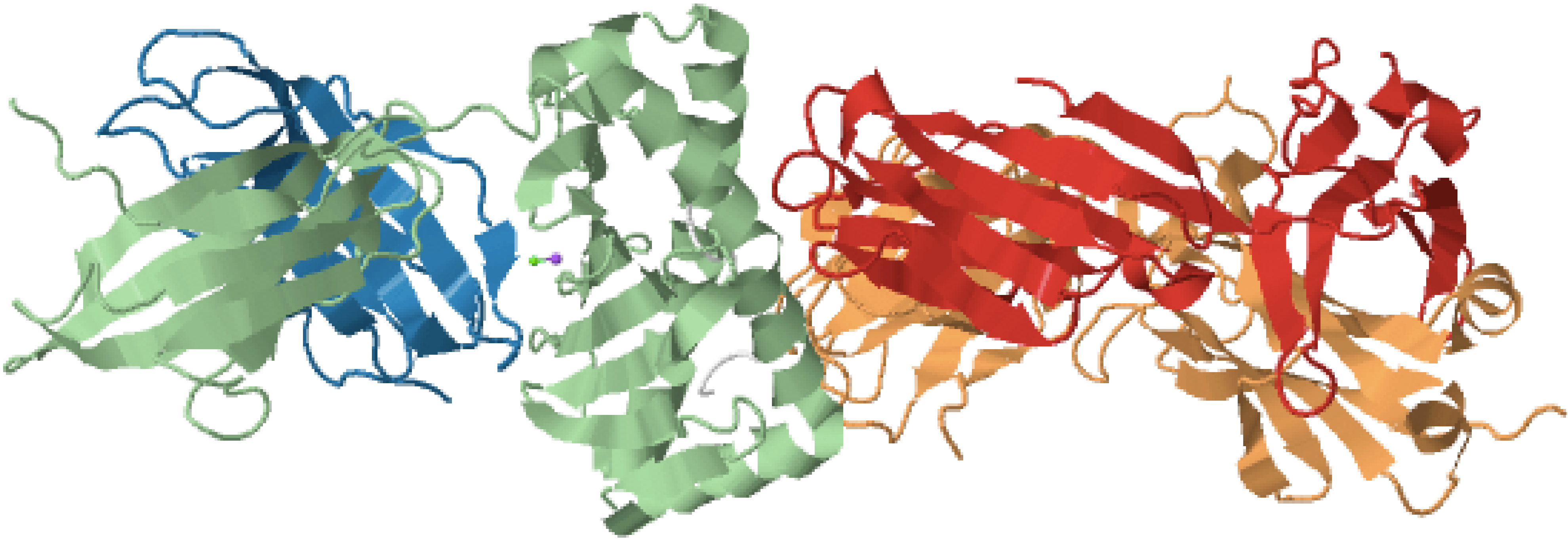
Headlights for driving at night

License plate: for Identification

Door for letting driver into the car

Wheels, necessary for mobility

Proteins Also Have Specific Functional Regions, Too!

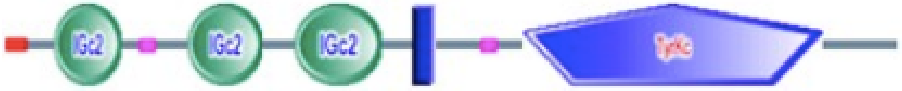


Protein Data Bank:

5WLG

Domains

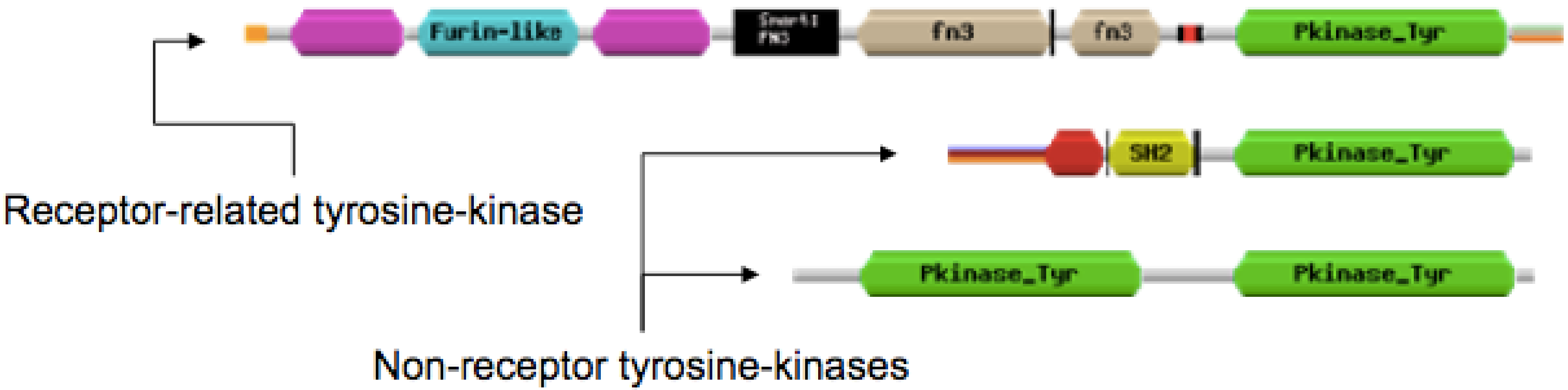
- A protein domain is a conserved part of a given protein sequence and (tertiary) structure.
- Can evolve, function, and exist independently of the rest of the protein chain
- Each domain forms a compact three-dimensional structure
- Often can be independently stable and folded.



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

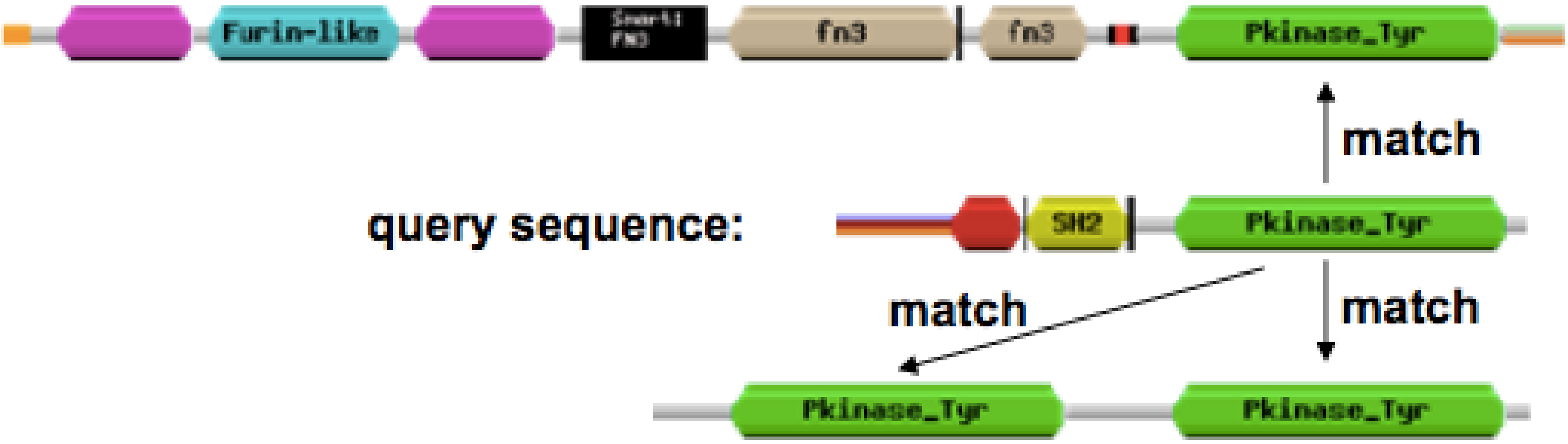
Protein Modularity

- Discrete functional units
- Found domains do not occur in the same order across proteins.
- Domains are considered separately in protein function predictions



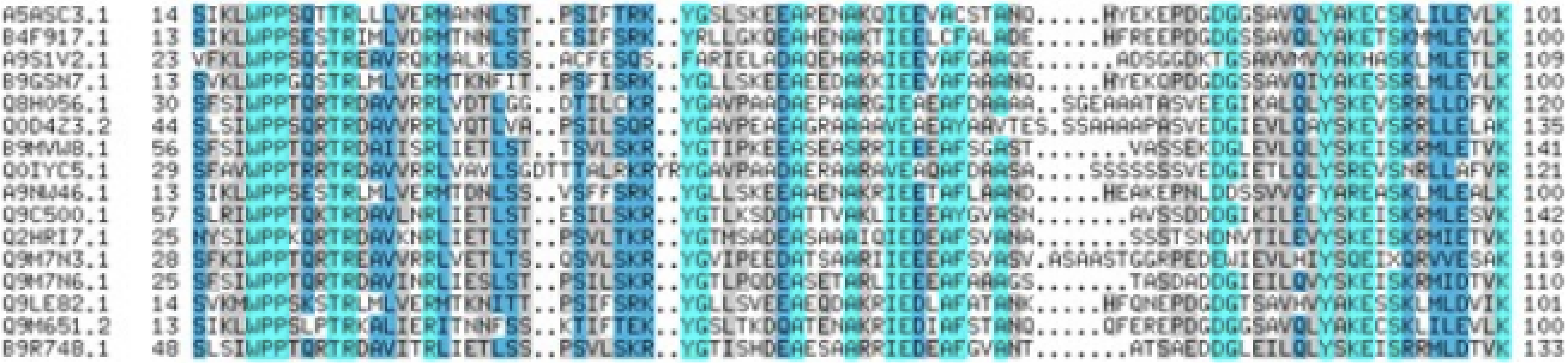
Finding a Domain?

- Alignment across proteins may show domains
- Use databases to match similar parts of proteins
  - Pfam, Smart, Interpro and others



Alignment?!

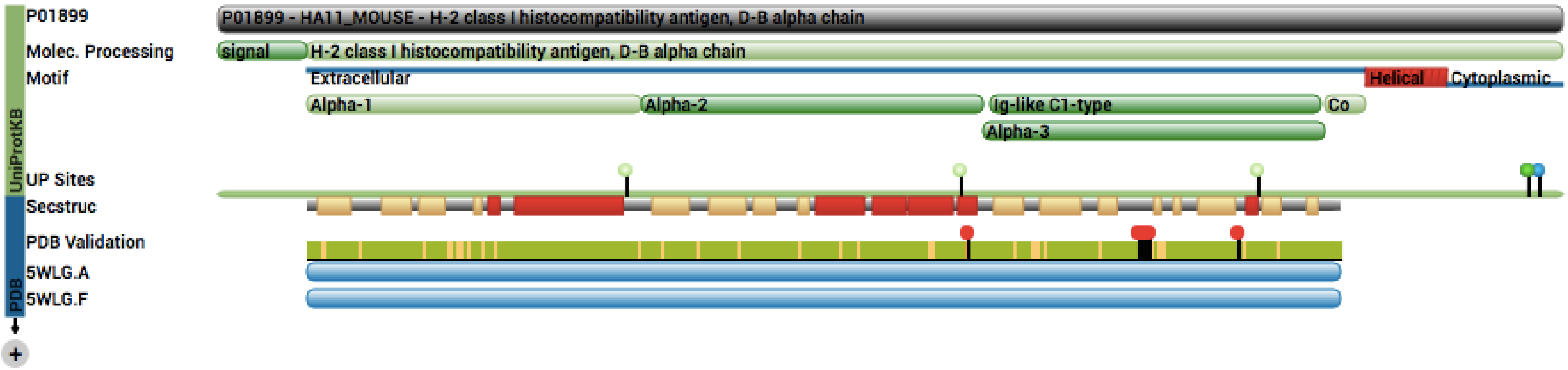
- Provide more info about a protein’s family, relatedness and other details.
- Domain landmarks include: low-complexity or disorder to suggest that these regions may have a specific syntax or pronounced grammar.



```
A5ASC3.1   14 SIKLWPPSQTTRLLLVERMANNLST..PSIFTRK..YGSLSKEEARENAKQIEEVACSTANQ.....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 SFAVWPPTRRTRDAVVRRLVAVLSGDTTTALRKRYRYGAVPAADAERAARAVEAQAFDARSA....SSSSSSSVEDGIETLQLYSREVSNRLLAFVR 121
A9NW46.1   13 SIKLWPPSESTRLMLVERMTDNLSS..VSFFSRK..YGLLSKEEAAENAKRIEETAFLAAND.....HEAKEPNLDDSSVVQFYAREASKLMLEALK 100
Q9C5D0.1   57 SLRIWPPTQKTRDAVLNRLIETLST..ESILSKR..YGTLKSDDATTVAKLIEEEAYGVASN.......AVSSDDDGIKILELYSKEISKRMLESVK 142
Q2HRI7.1   25 NYSIWPPKQRTRDAVKNRLIETLST..PSVLTKR..YGTMSADEASAAAIQIEDEAFSVANA.......SSSTSNDNVTILEVYSKEISKRMIETVK 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*

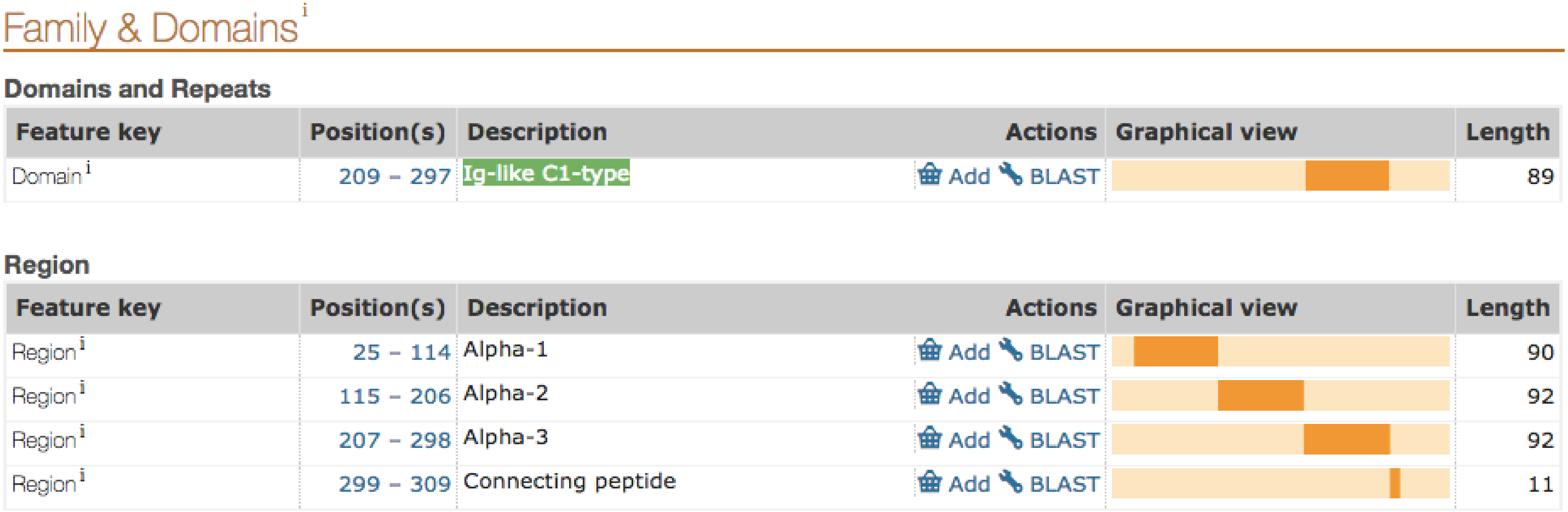


Protein Data Bank ID: 5WLG

https://www.rcsb.org/pdb/explore/explore.do?structureId=5WLG

Domains By Uniprot

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



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

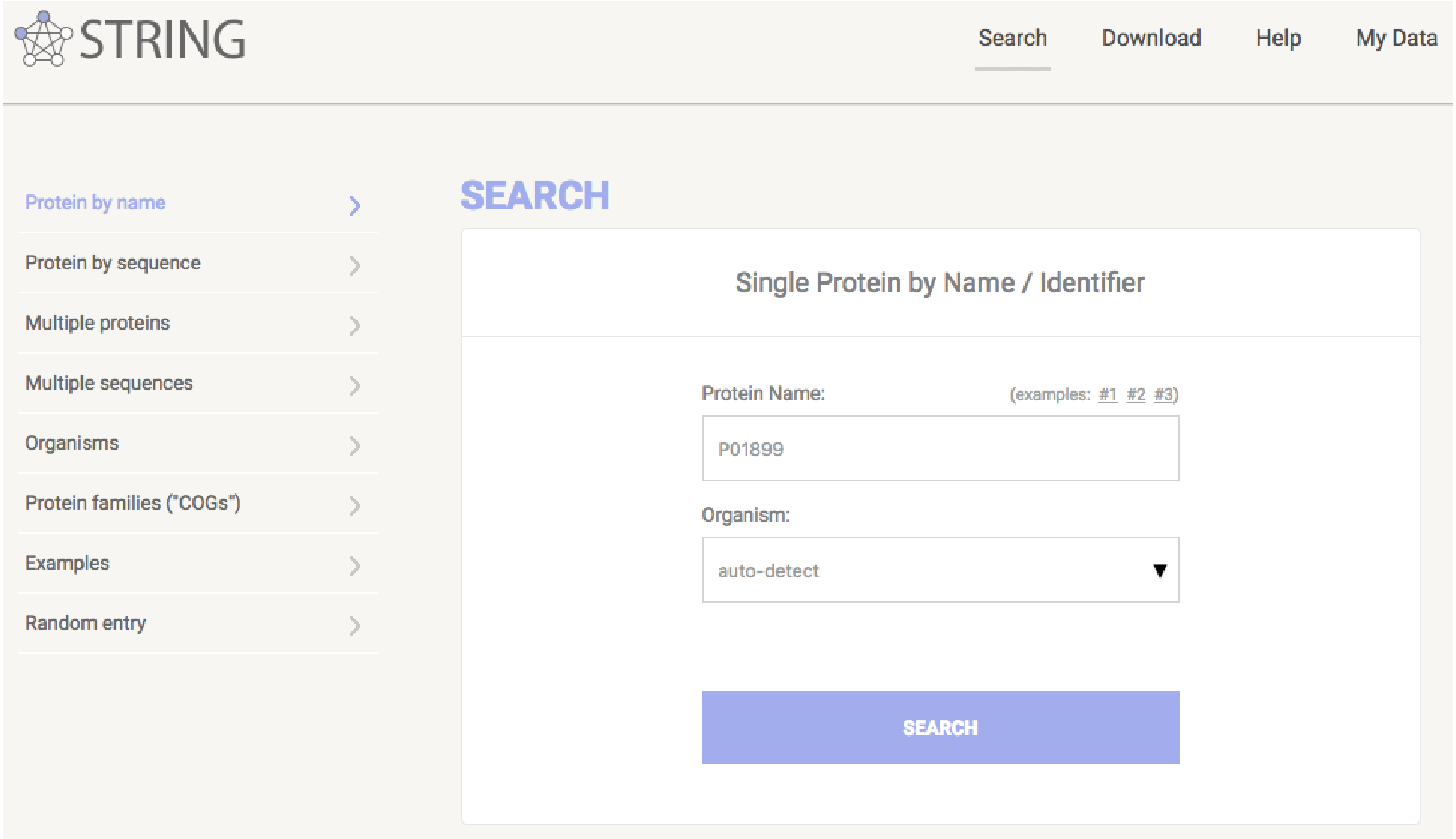
Feature key	Position(s)	Description	Actions	Graphical view	Length
Region	25 – 114	Alpha-1	Add BLAST		90
Region	115 – 206	Alpha-2	Add BLAST		92
Region	207 – 298	Alpha-3	Add BLAST		92
Region	299 – 309	Connecting peptide	Add BLAST		11

UniProt ID: P01899

A Protein Knowledge Base

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

The String Database For Analysis



Protein Name: P01899

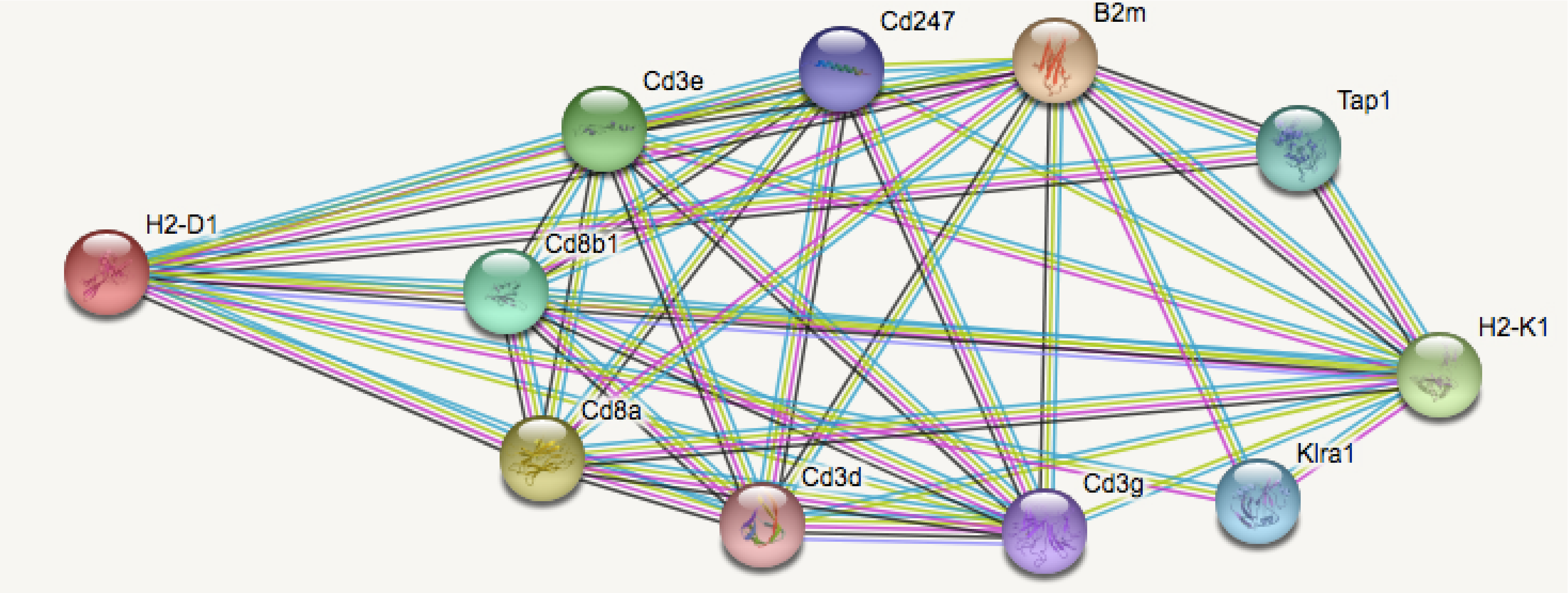
Organism: auto-detect

String DB ID P01899

http://string-db.org/

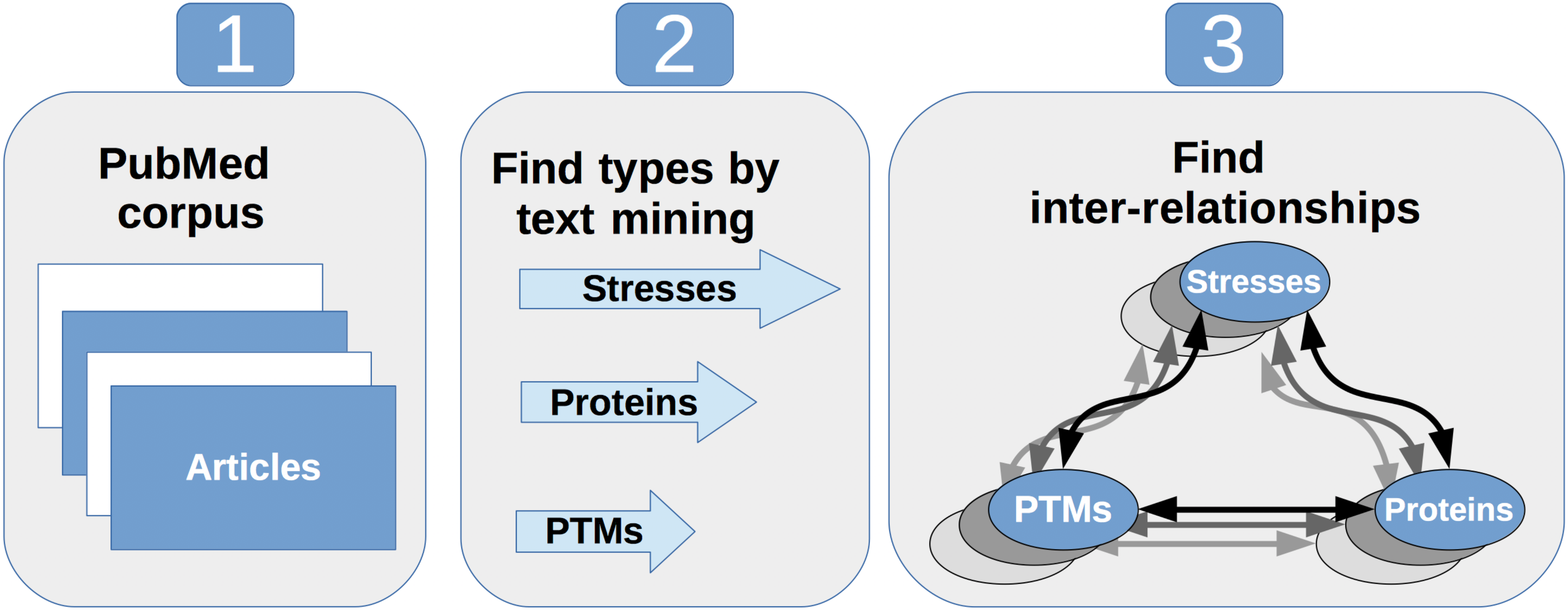
The String Database For Analysis

- How is a particular protein *related* to others?
- Can we determine this by criteria?



Criteria to Determine Relations

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



1

PubMed corpus

Articles

2

Find types by text mining

Stresses

Proteins

PTMs

3

Find inter-relationships

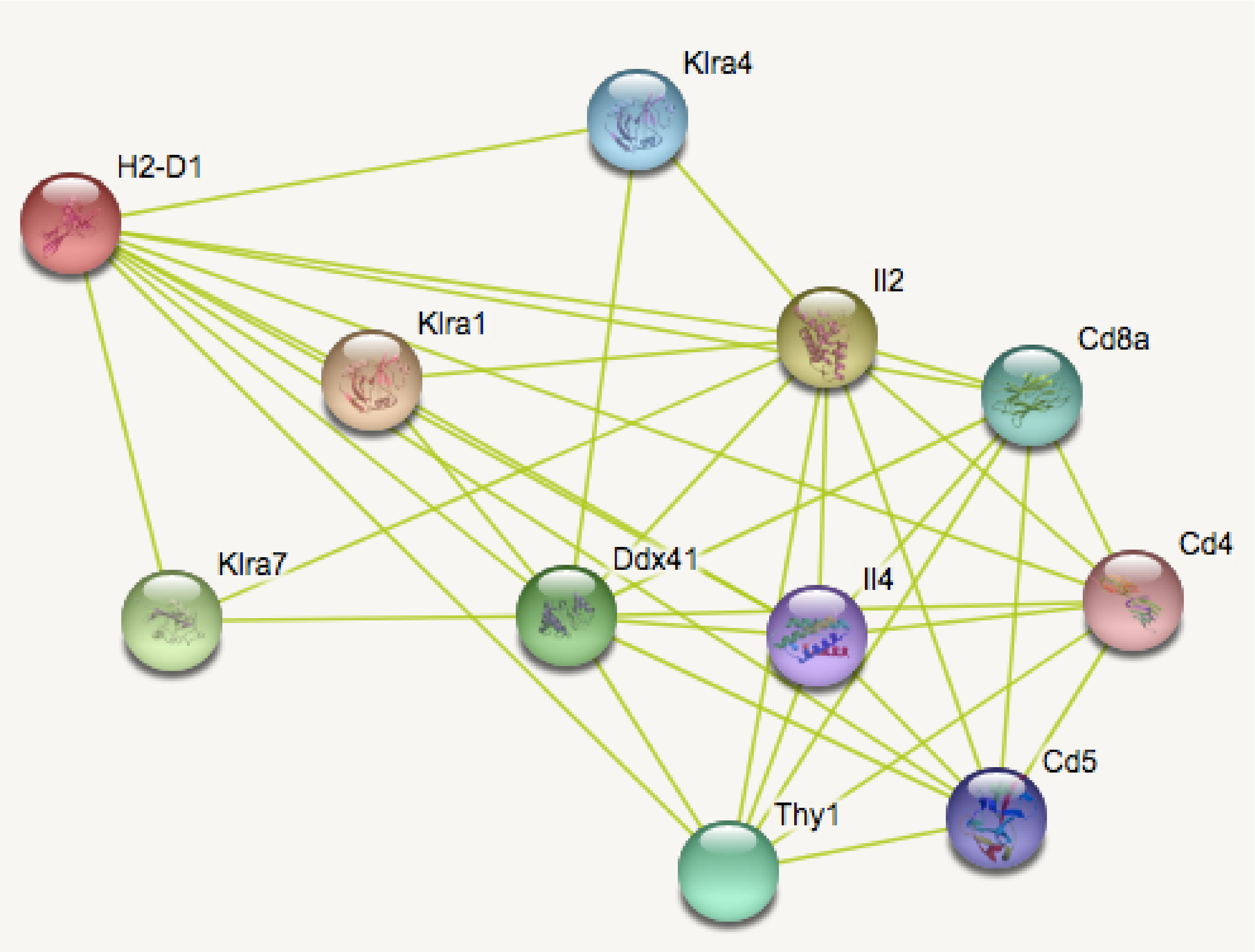
Stresses

PTMs

Proteins

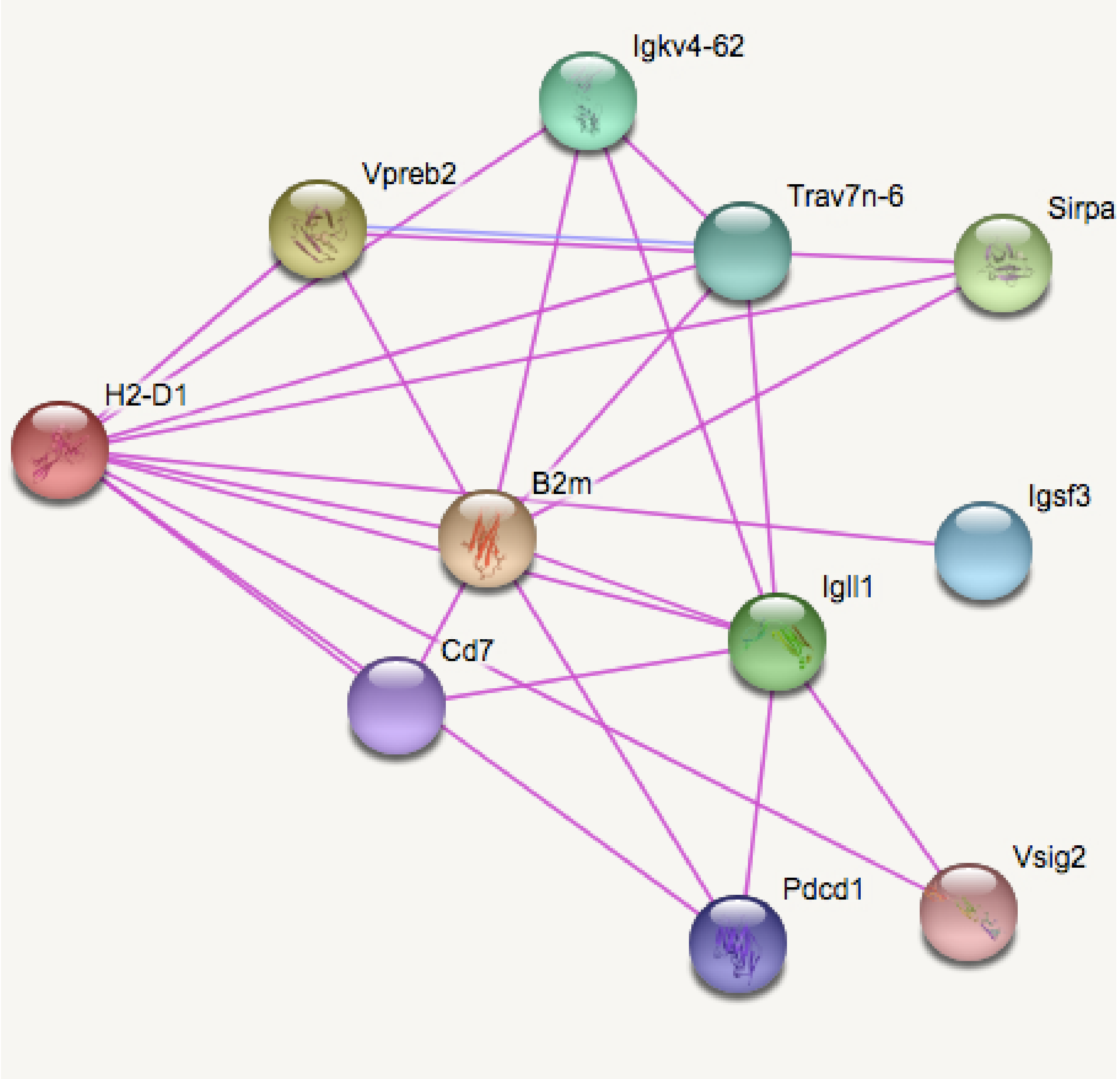
String: by Text Mining

- Mentioned by name in articles



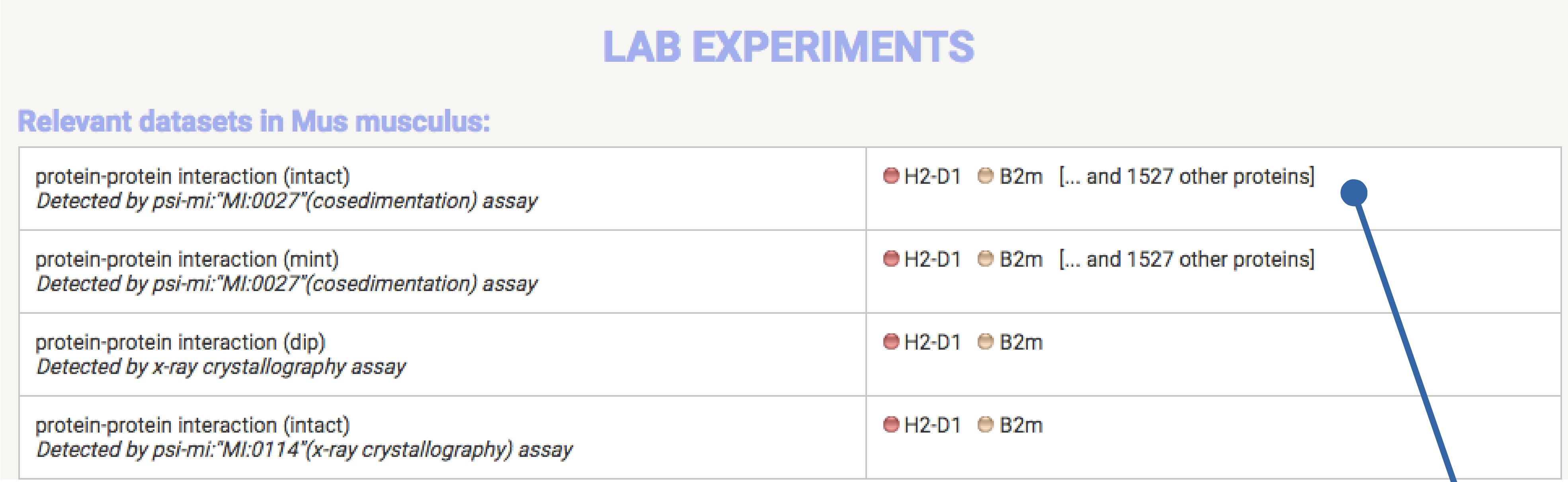
String: Linked Experimentally

- Experiments performed to show one protein is related to another



String: Linked Experimentally

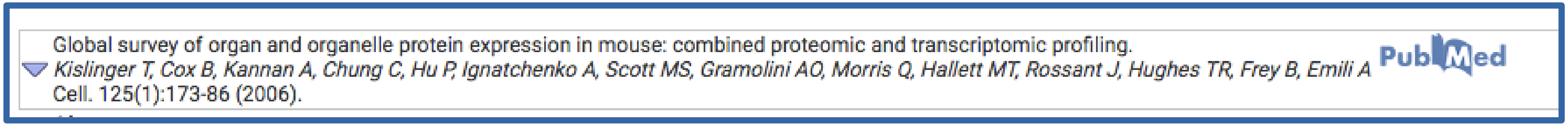
- Learn about the experiments



LAB EXPERIMENTS

Relevant datasets in Mus musculus:

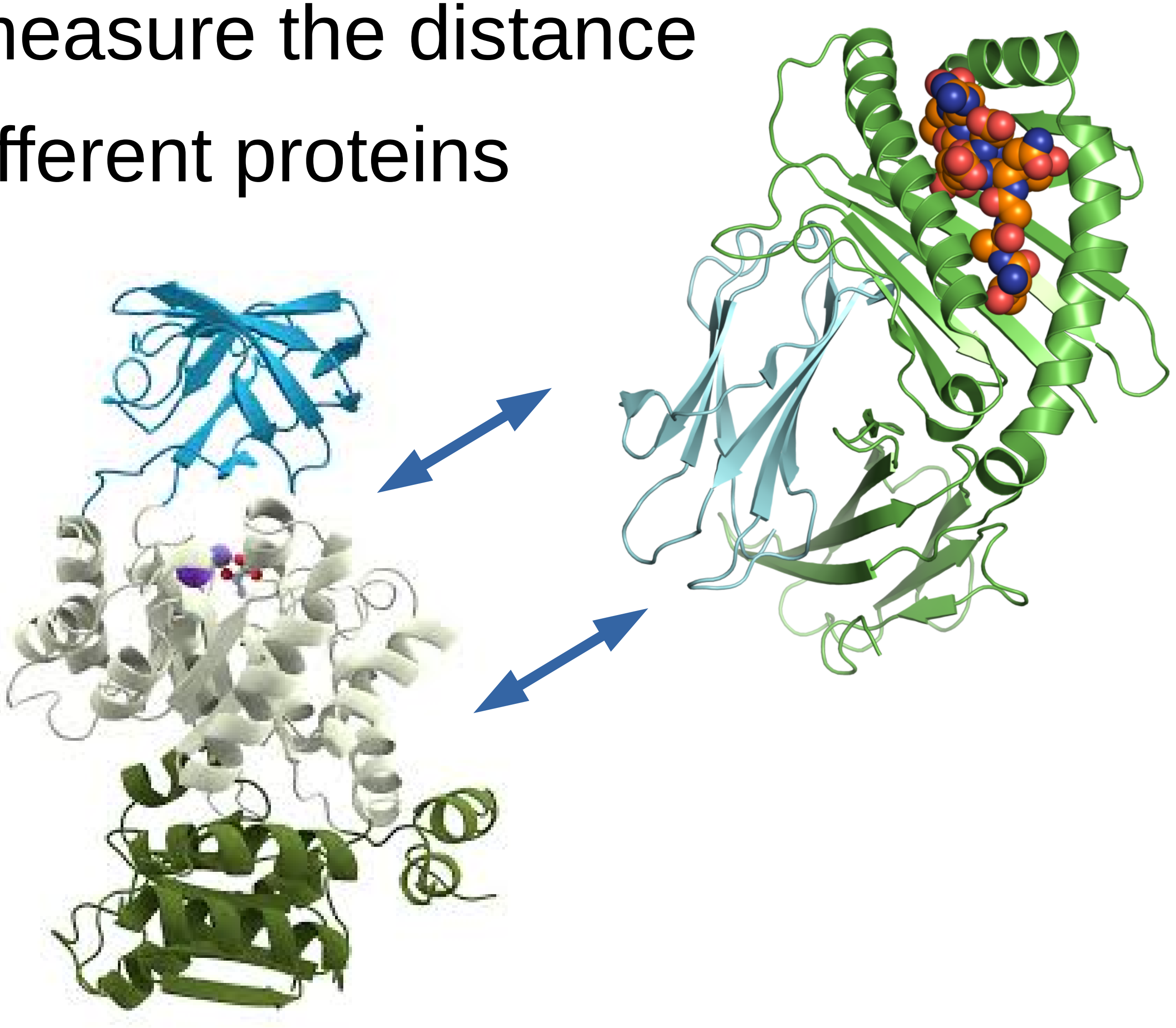
protein-protein interaction (intact) *Detected by psi-mi:"MI:0027"(cosedimentation) assay*	● H2-D1 ● B2m [... and 1527 other proteins]
protein-protein interaction (mint) *Detected by psi-mi:"MI:0027"(cosedimentation) assay*	● H2-D1 ● B2m [... and 1527 other proteins]
protein-protein interaction (dip)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



Header

- Pick your favorite protein and head over to:
  - http://www.uniprot.org/

    example: P01899
- Then check out the networks at:
- https://string-db.org/