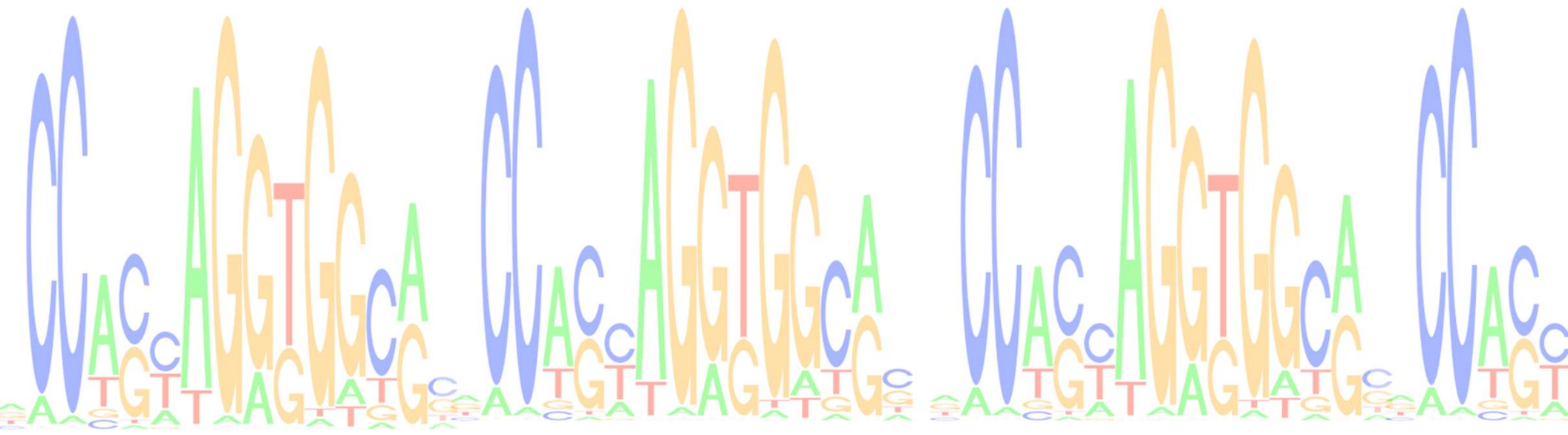


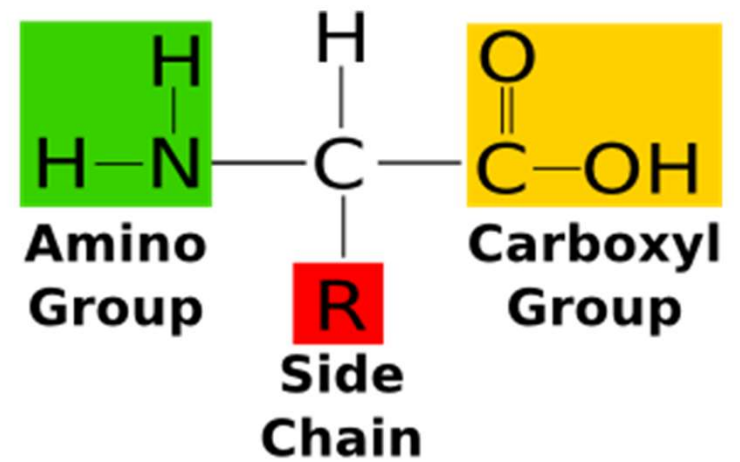
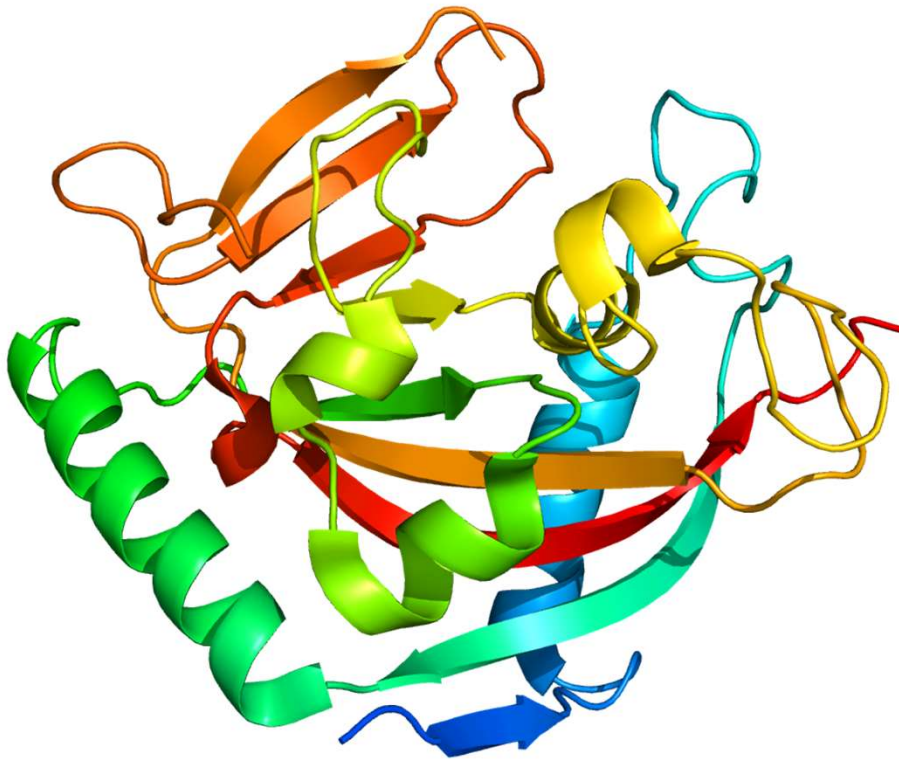
# Motif & Domain Discovery



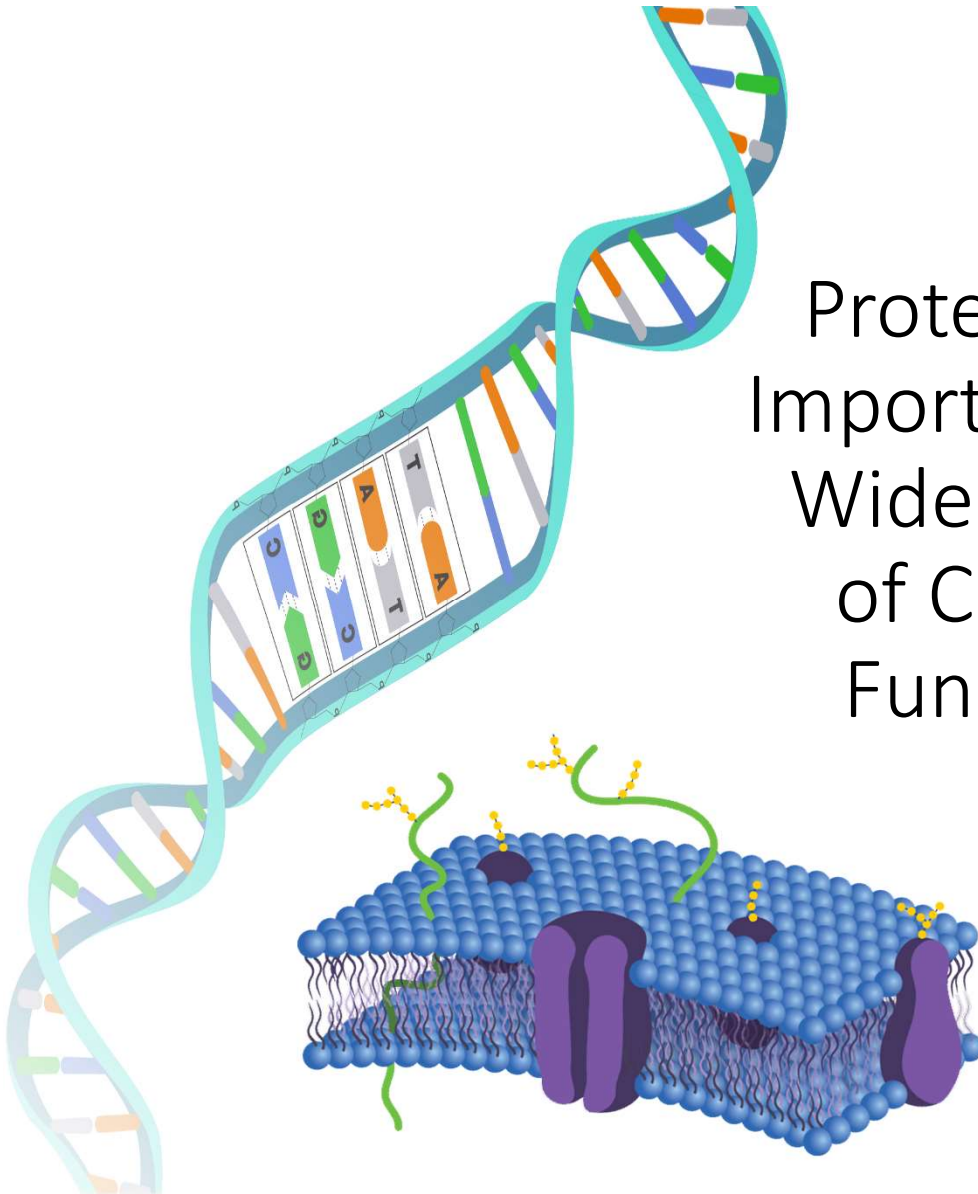
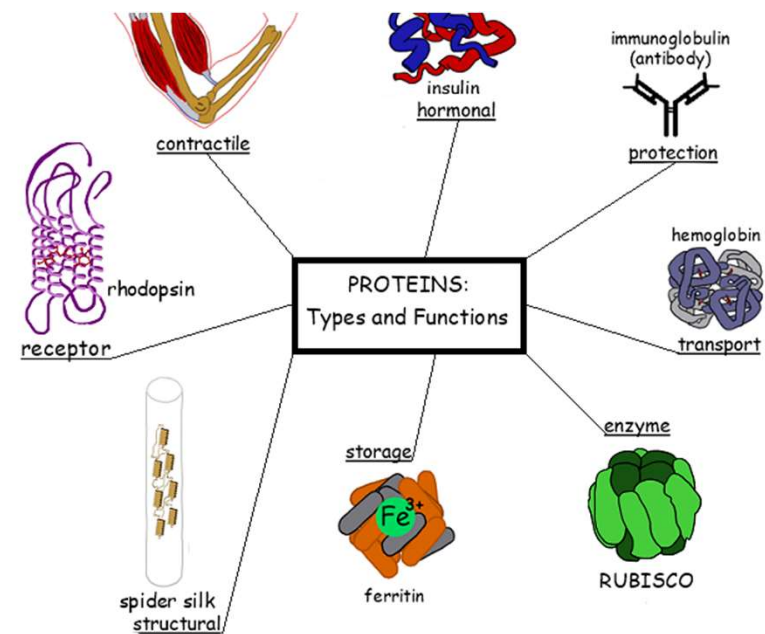
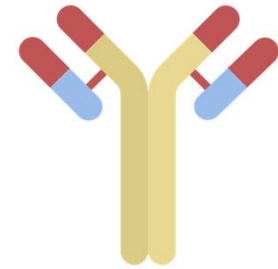
Shailaja Singh & Jeff Pietroske

2/16/23

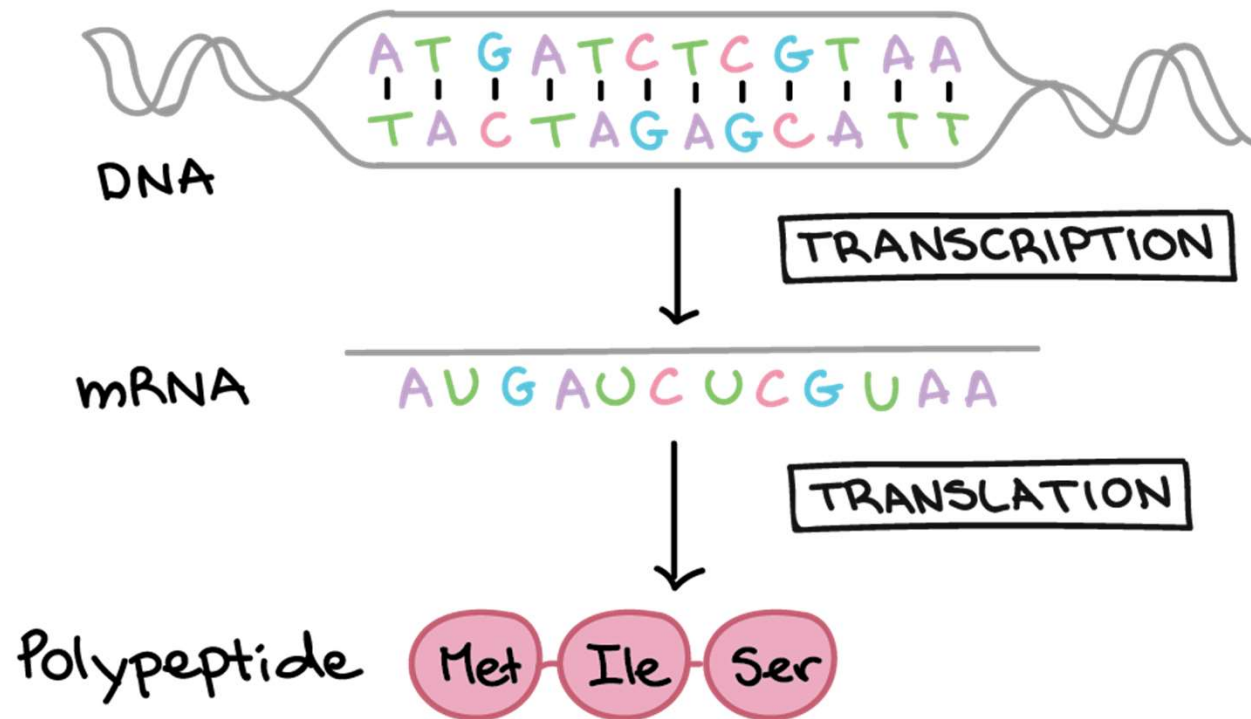
# What is a protein?



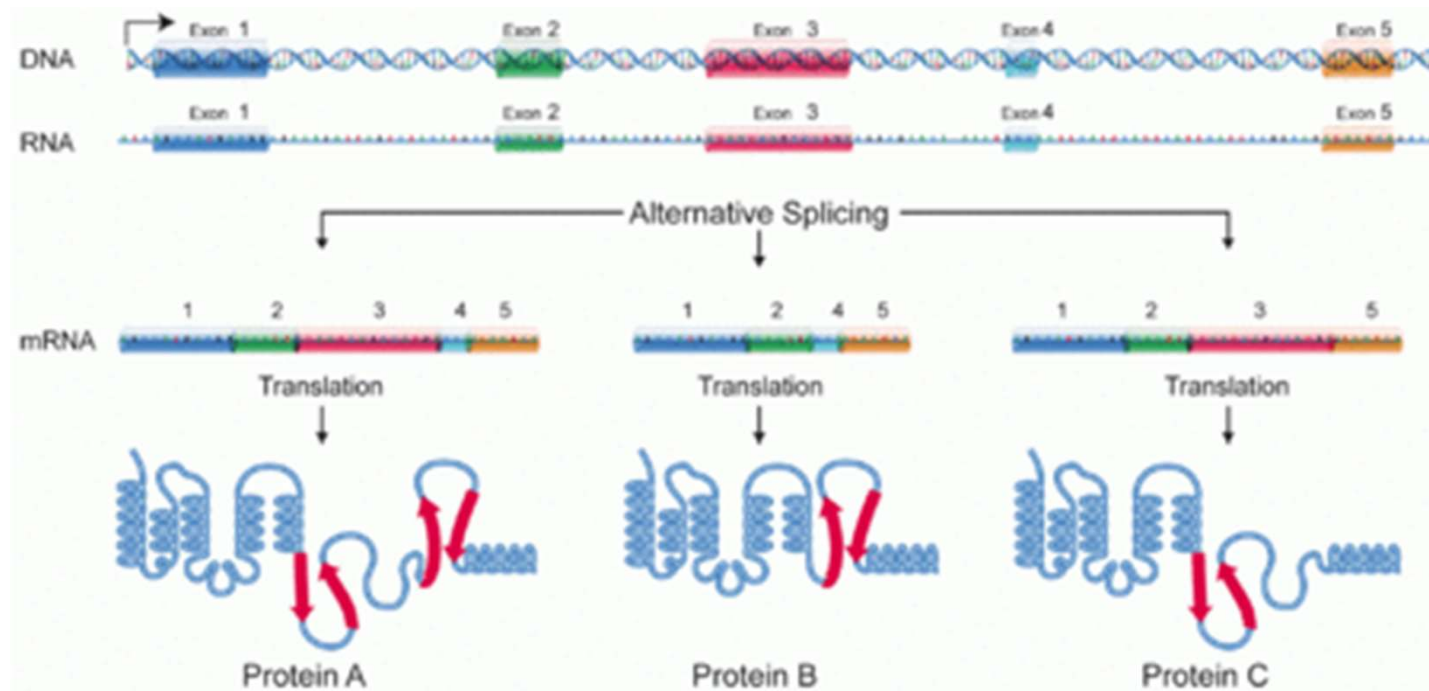
Proteins are  
Important for a  
Wide Variety  
of Cellular  
Functions



# How are proteins made?

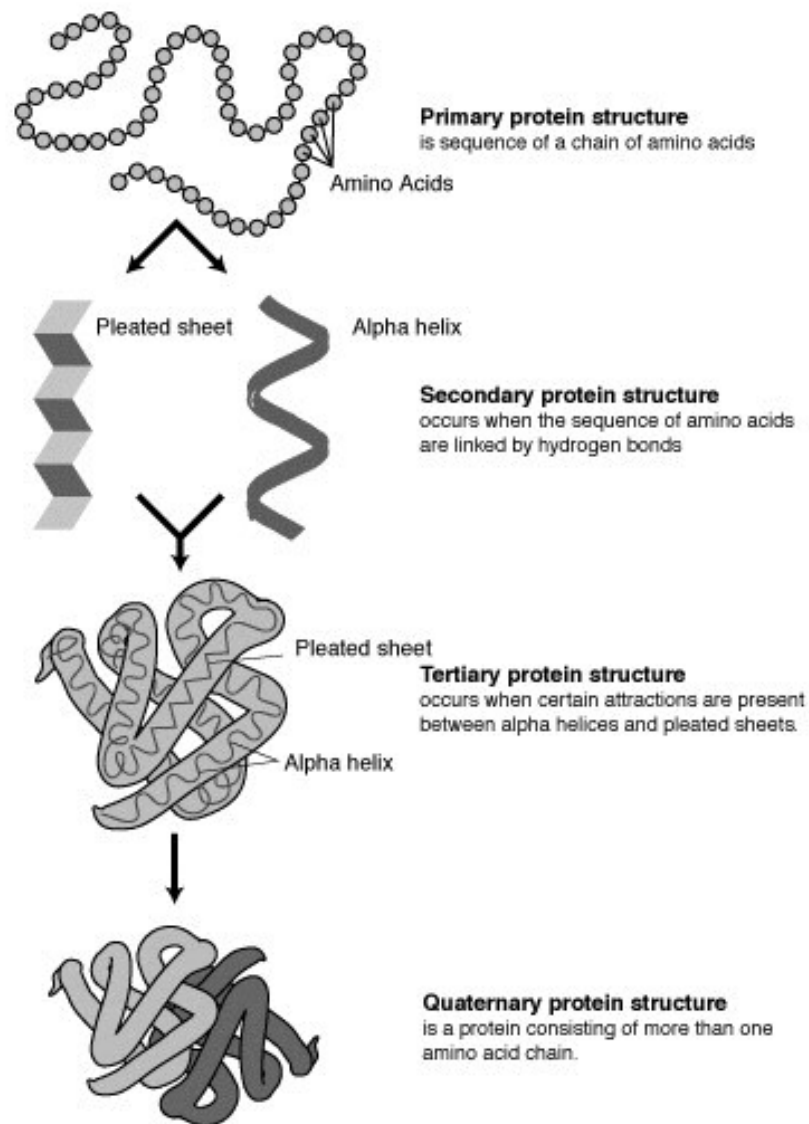


# How do genes code for multiple proteins?



**Alternative Splicing**

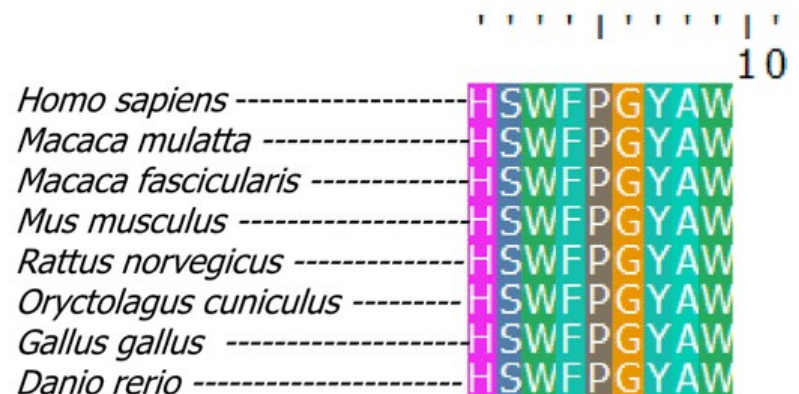
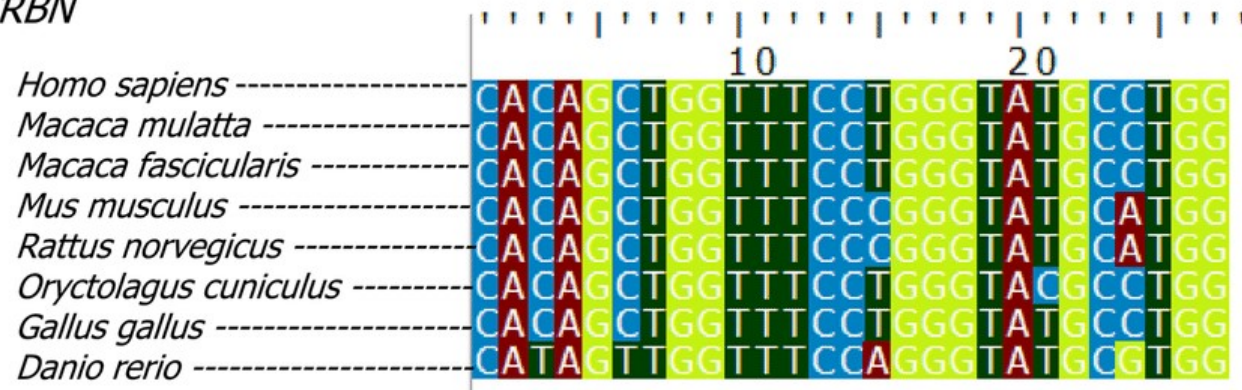
What types  
of protein  
structure  
are there?



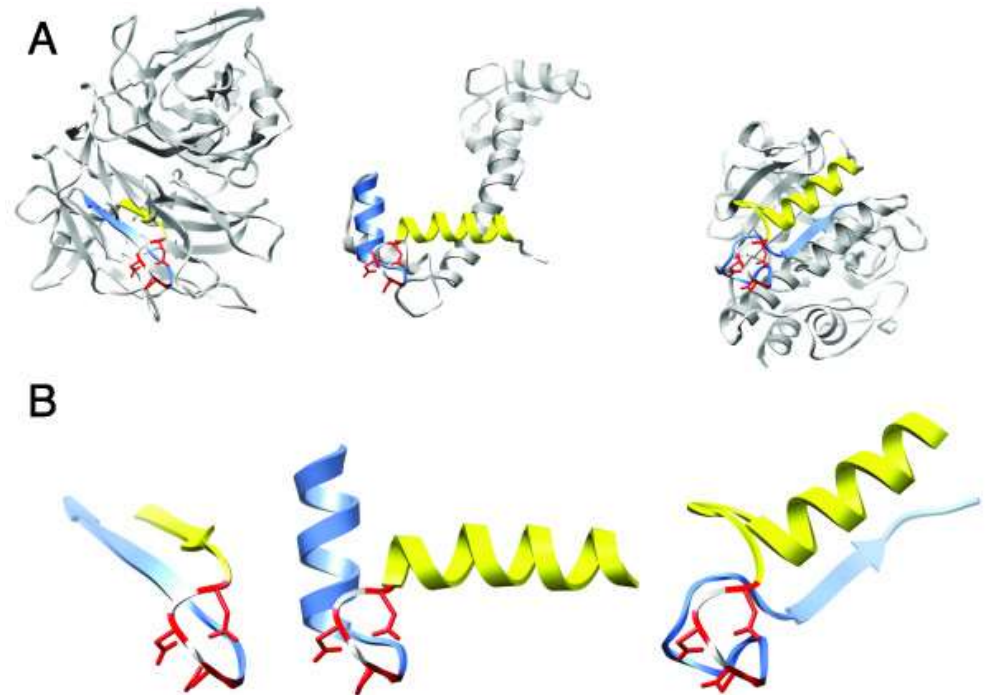
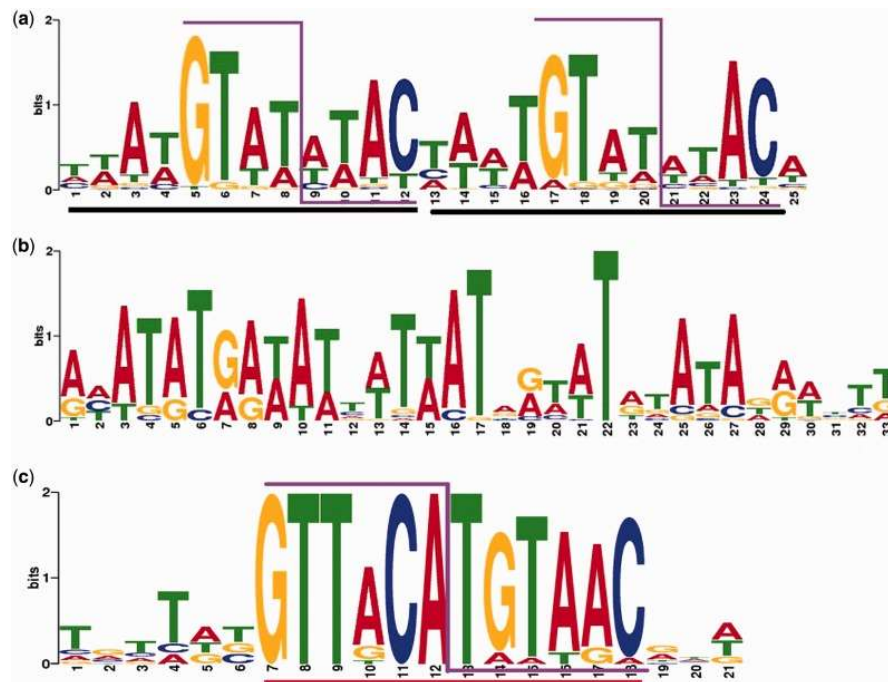


# What does it mean when proteins are conserved?

*CRBN*



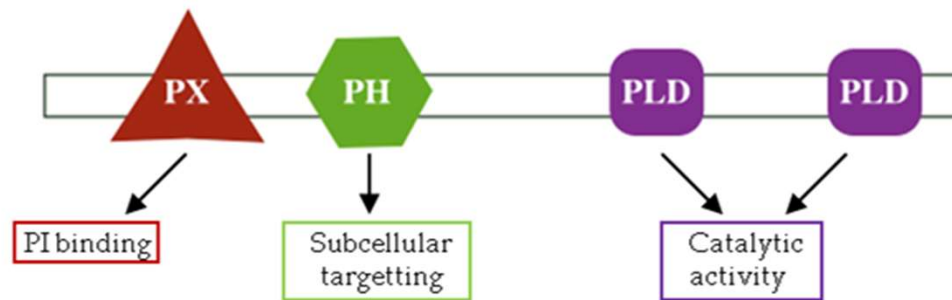
# What is a Protein Motif?



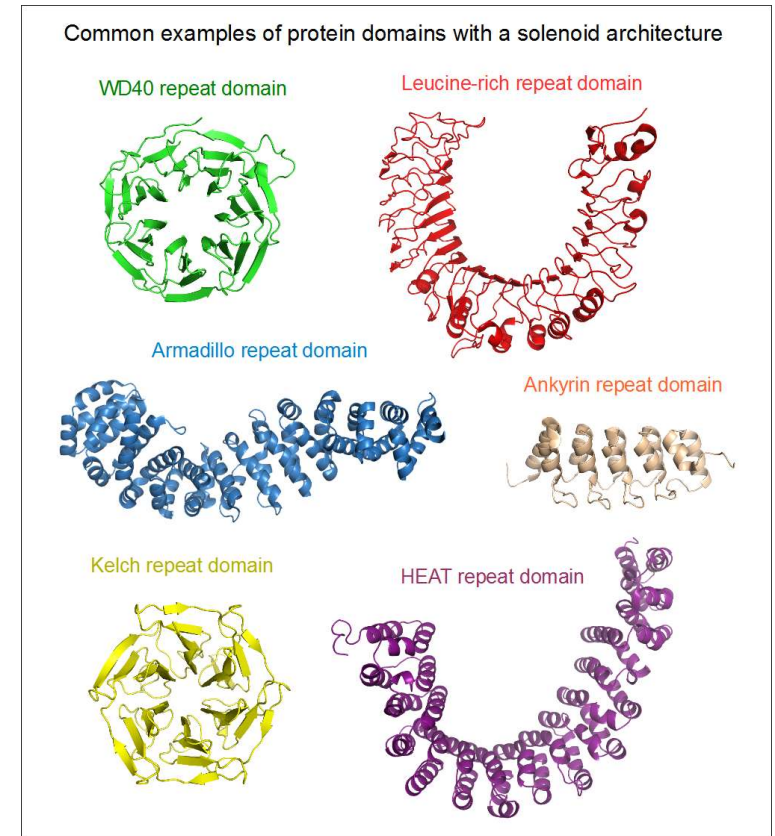
Small conserved regions of protein structure



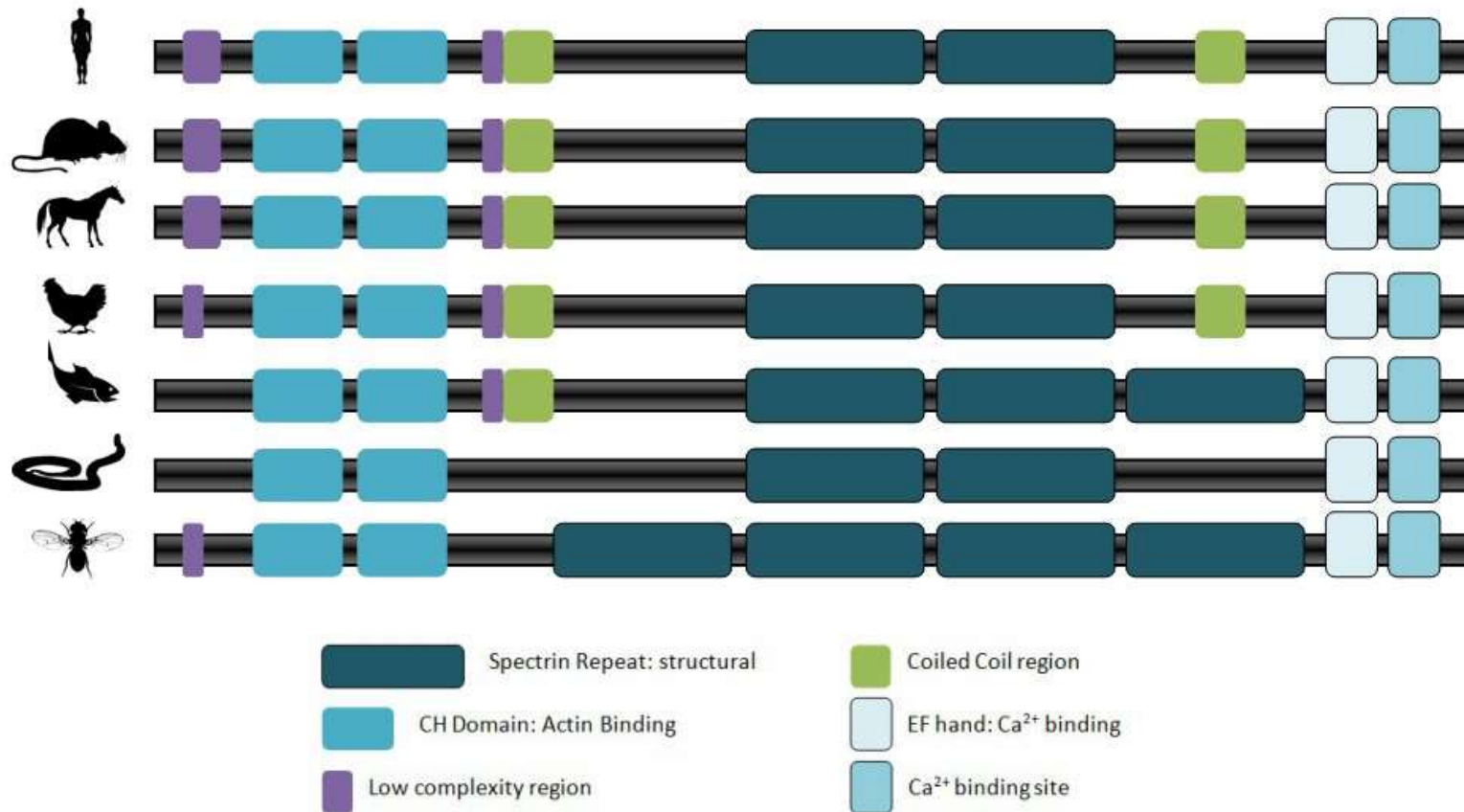
# What is a Protein Domain?



**Conserved structural or functional units of protein**



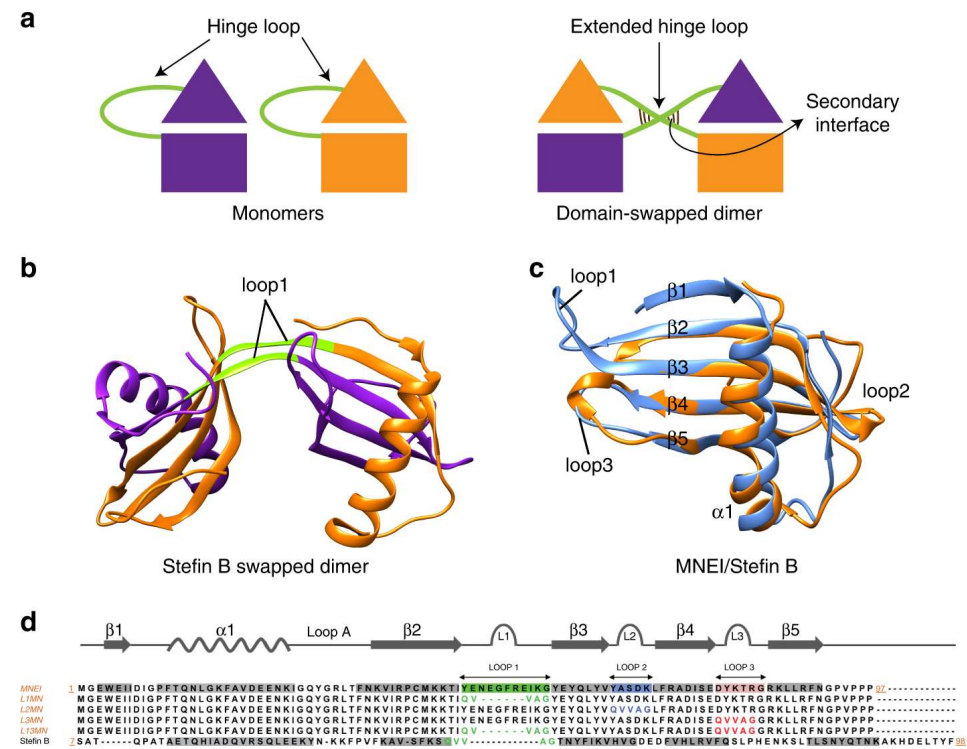
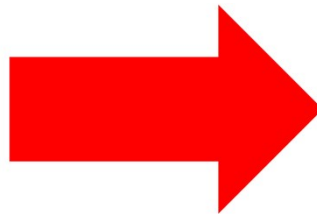
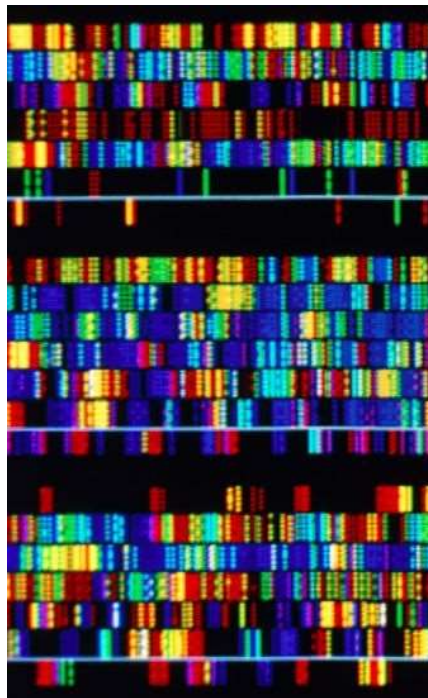
# How do we visualize conserved domains?



# What is the difference between Motifs and Domains?

	Motif	Domain
DEFINITION	Motif is an arrangement of secondary structures of the protein molecule.	Domain is the three dimensional fundamental and functional unit of the protein.
STABILITY	Not stable	Stable by itself
FUNCTIONAL ROLE	Does not depict a functional role.	It is the functional unit of the protein.

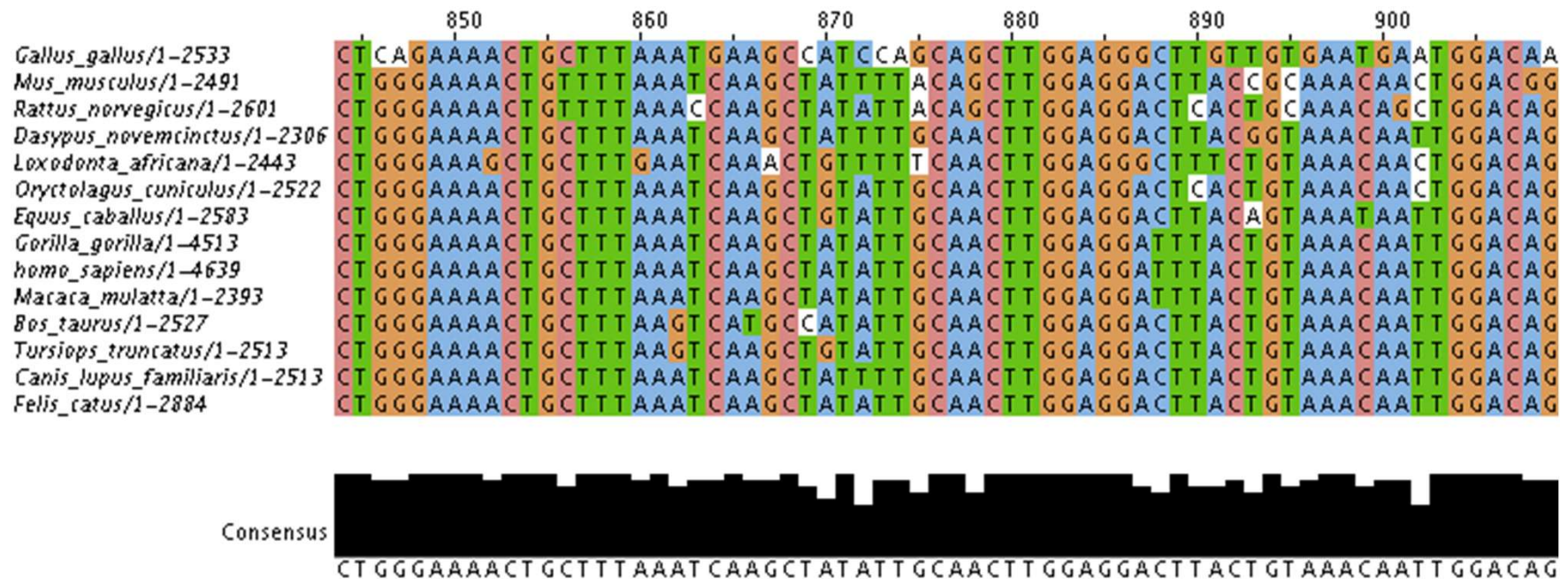
# Why are they important to research?



They allow us to computationally classify protein classification and determine biological function

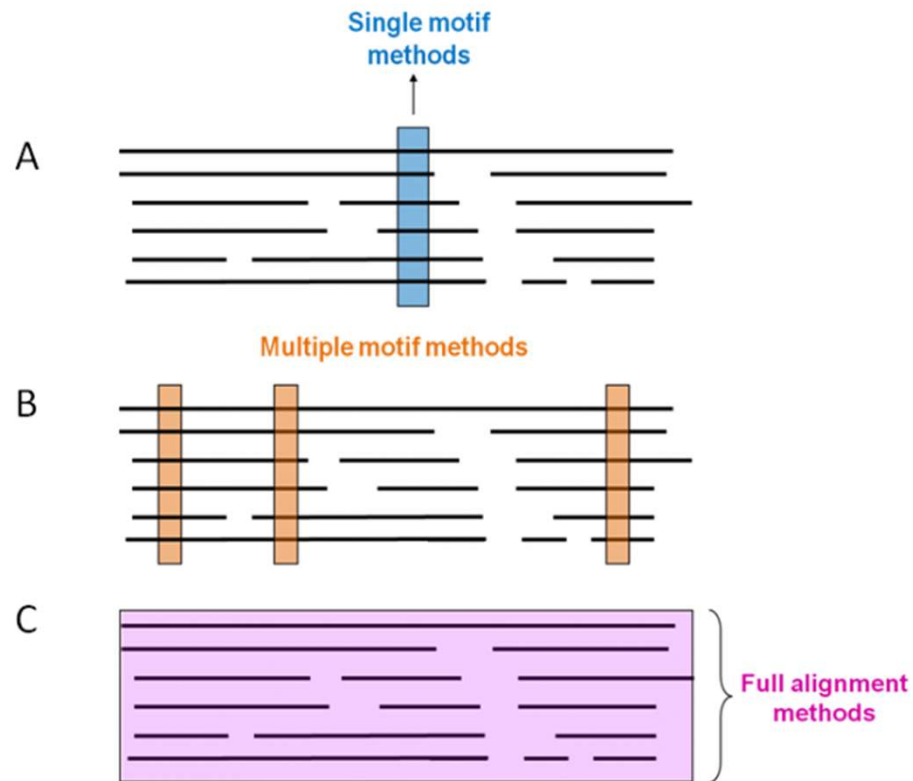


# How do we determine domains?



**Multiple Sequence Alignment using Multiple Sequence Comparison Log-Expectation (MUSCLE)**

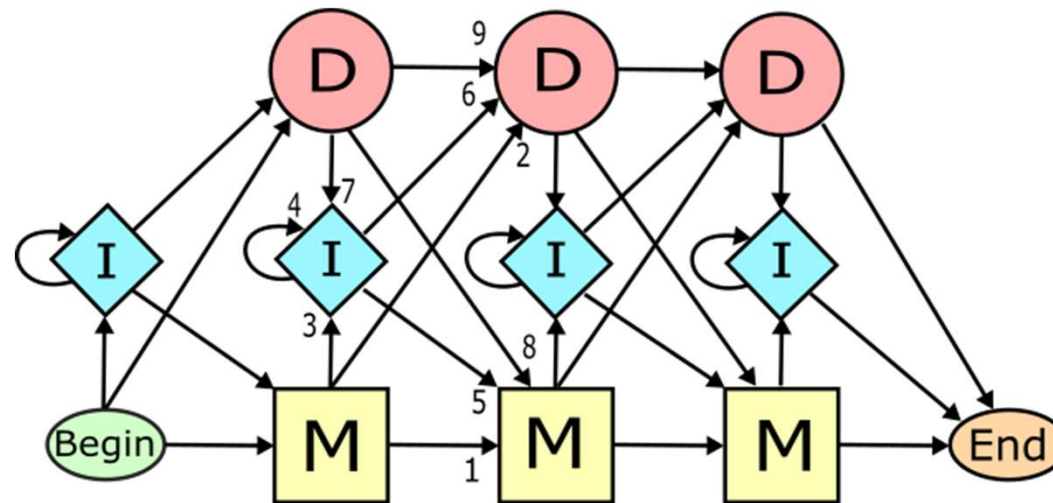
# Why multiple sequence alignment?



**By using this technique, we can observe patterns in the data**



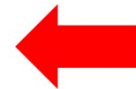
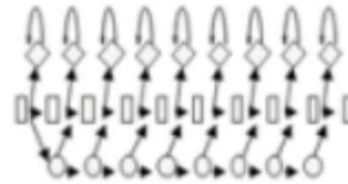
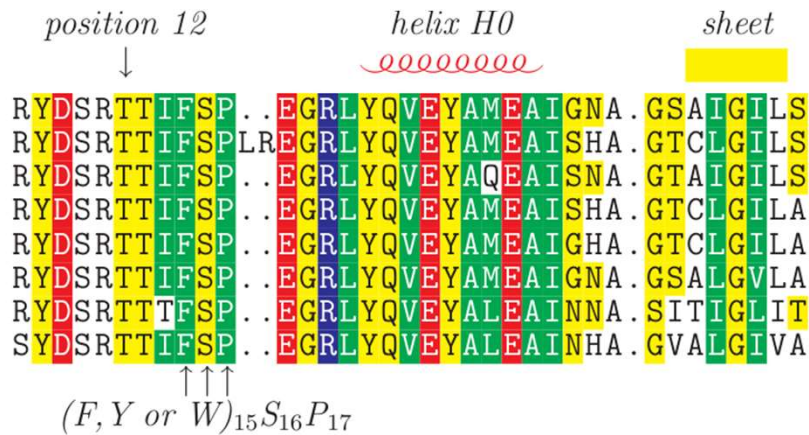
What are other methods we can use?



### Hidden Markov Method

Statistical method that quantifies the position of amino acids at a particular position

# So how do we model domains?



Mature Model

Protein databases refine the model by quantifying the level of amino acid conservation

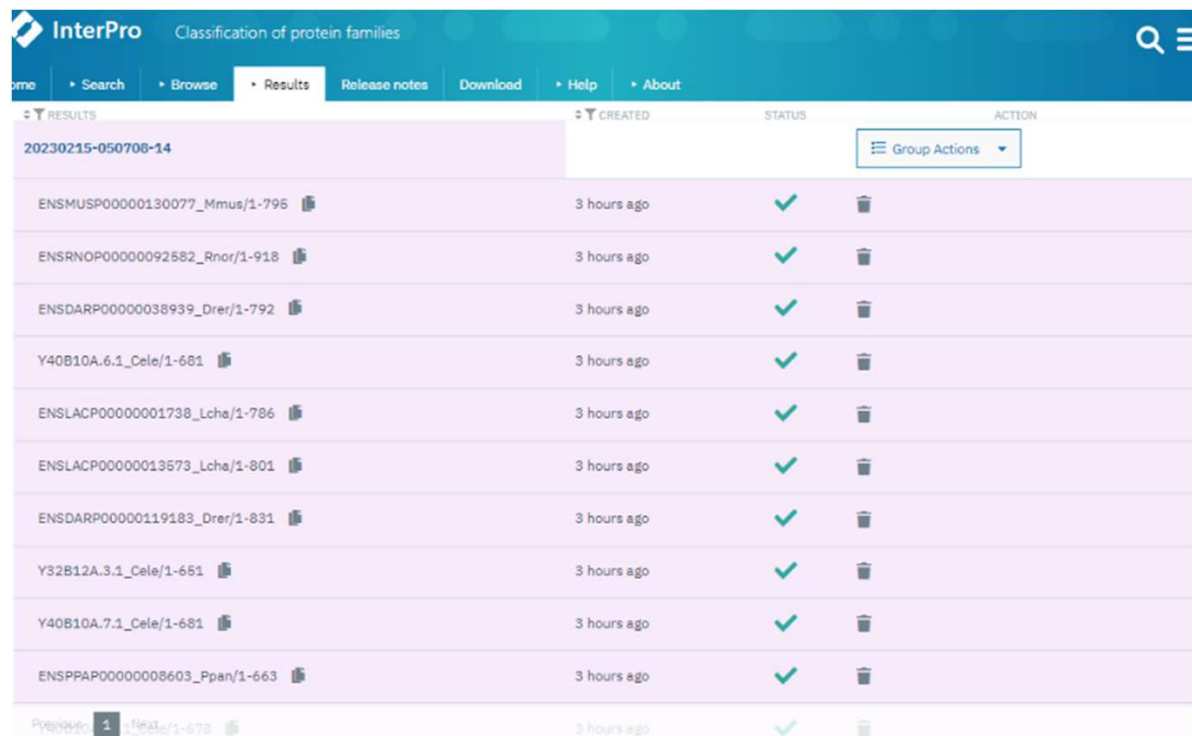
How do we analyze domains?



**Pfam**

**Database of protein families that includes their multiple sequence alignments generated using HMM**

# What results do we get from Pfam?



The screenshot shows the InterPro website interface. The header includes the InterPro logo and the text "Classification of protein families". Below the header is a navigation bar with links: Home, Search, Browse, Results (active), Release notes, Download, Help, and About. The main content area displays a table of protein families. The table has columns for RESULTS, CREATED, STATUS, and ACTION. The RESULTS column lists protein families with their accession numbers and domain names. The CREATED column shows the time since creation (all are "3 hours ago"). The STATUS column shows a green checkmark for each entry. The ACTION column shows a trash icon for each entry. A "Group Actions" dropdown menu is visible in the top right of the table area.

RESULTS	CREATED	STATUS	ACTION
20230215-050708-14			Group Actions
ENSMUSP00000130077_Mmus/1-795	3 hours ago	✓	✖
ENSRNOP00000092582_Rnor/1-918	3 hours ago	✓	✖
ENSDARP00000038939_Drer/1-792	3 hours ago	✓	✖
Y40B10A.6.1_Cele/1-681	3 hours ago	✓	✖
ENSLACP00000001738_Lcha/1-786	3 hours ago	✓	✖
ENSLACP000000013573_Lcha/1-801	3 hours ago	✓	✖
ENSDARP000000119183_Drer/1-831	3 hours ago	✓	✖
Y32B12A.3.1_Cele/1-651	3 hours ago	✓	✖
Y40B10A.7.1_Cele/1-681	3 hours ago	✓	✖
ENSPPAP00000008603_Ppan/1-663	3 hours ago	✓	✖
Previous 1 Next 1-678	3 hours ago	✓	✖

**Note: Pfam has been absorbed into InterPro. True Pfam results no longer available**

# Interpro results, continued

**Pfam** Arc MA domain PF19284

1 - 20 of 540 proteins			
<div><div></div><div>Search</div><div>Export</div></div>			
ACCESSION	NAME	SPECIES	MATCHES
A0A087R736	ARC protein	Aptenodytes forsteri (Emperor penguin)	<div><div></div><div>100200300</div></div>
A0A087V0K3	ARC protein	Balearica regulorum gibbericeps (East African grey crowned-crane)	<div><div></div><div>100200300</div></div>
A0A091D3X6	Arc_MA domain-containing protein	Fukomys damarensis (Damaraland mole rat)	<div><div></div><div>5001000</div></div>
A0A091GC14	ARC protein	Cuculus canorus (common cuckoo)	<div><div></div><div>100200300</div></div>
A0A091HTL1	ARC protein	Buceros rhinoceros silvestris	<div><div></div><div>100200300</div></div>
A0A091I830	ARC protein	Calypte anna (Anna's hummingbird)	<div><div></div><div>100200300</div></div>
A0A091JB13	ARC protein	Egretta garzetta (Little egret)	<div><div></div><div>100200300</div></div>
A0A091KC23	ARC protein	Colius striatus (Speckled mousebird)	<div><div></div><div>100200300</div></div>
Show 20 results			<div>PreviousNext</div>

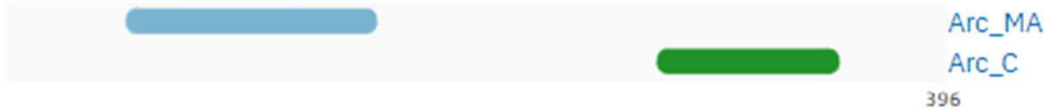
# Interpro results, continued

**Pfam** Arc MA domain PF19284

## 3 domain architectures found.

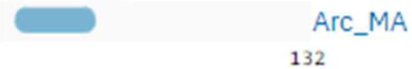
There are 528 proteins with this architecture (represented by Q63053):

PF19284 - PF18162



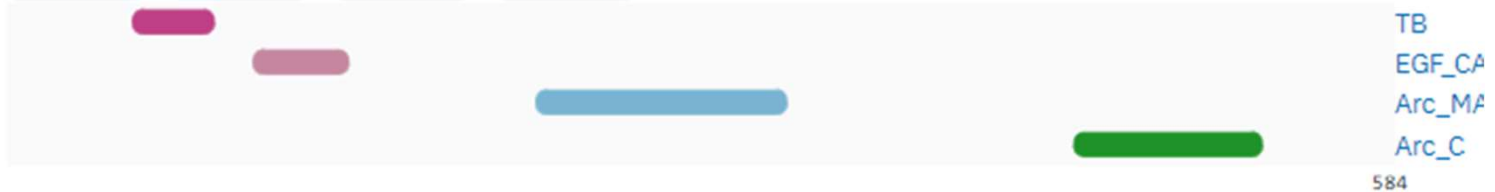
There are 11 proteins with this architecture (represented by Q5RZZ8):

PF19284



There is 1 protein with this architecture (represented by A0A6I9YW12):

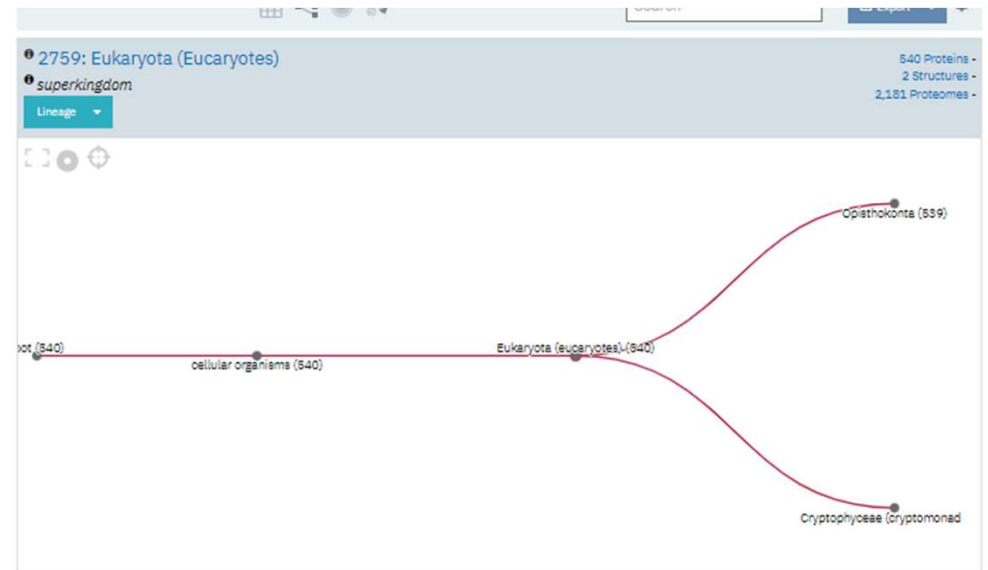
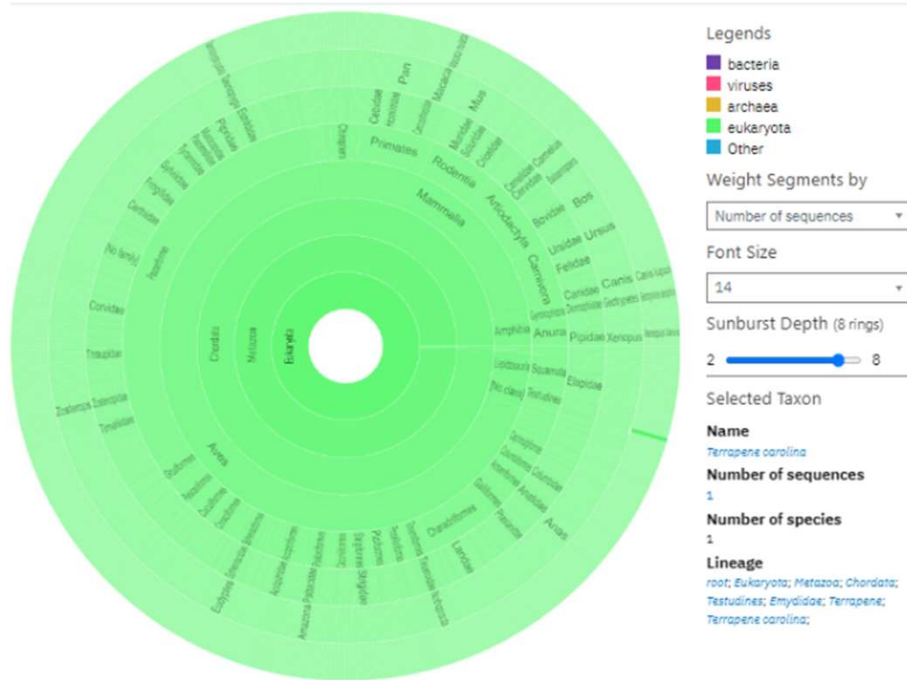
PF00683 - PF07645 - PF19284 - PF18162







## Interpro results, continued

**Pfam** Arc MA domain PF19284




# What did Pfam use to give?

EMBL-EBI  [HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#) | [ABOUT](#) 

### Sequence search results

[Show](#) the detailed description of this results page.  
We found **4** Pfam-A matches to your search sequence (**all** significant)


 DNA\_pol\_A

[Show](#) the search options and sequence that you submitted.  
[Return](#) to the search form to look for Pfam domains on a new sequence.

### Significant Pfam-A Matches

[Show](#) or [hide](#) all alignments.

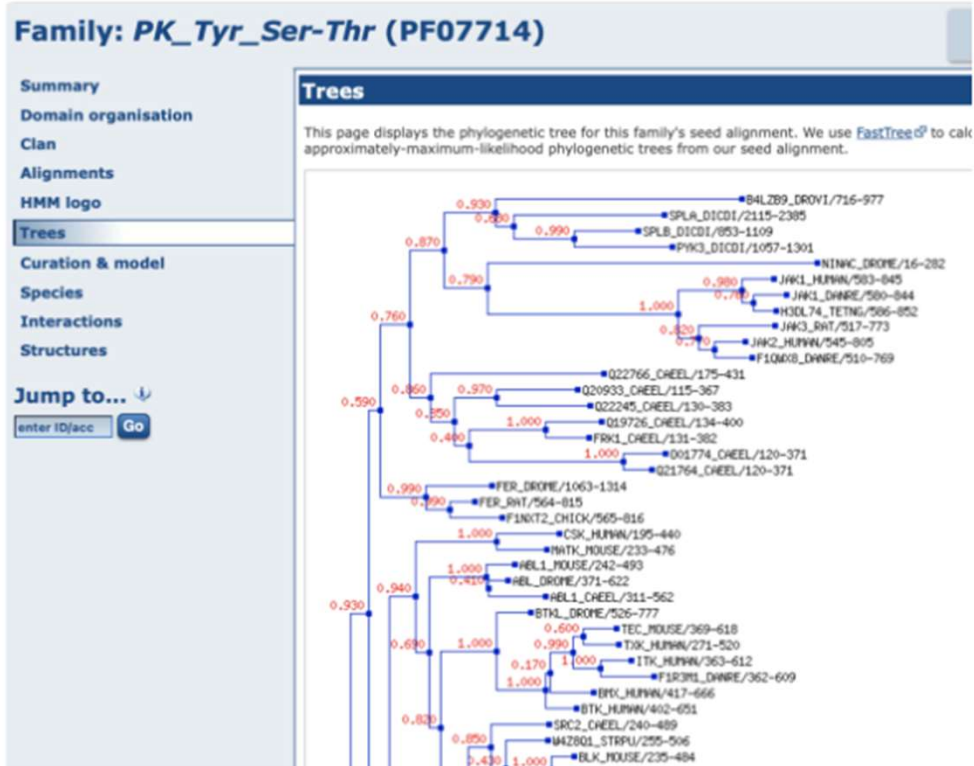
Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
<a href="#">5_3_exonuc_N</a>	5'-3' exonuclease, N-terminal resolvase- ...	Domain	<a href="#">CL0280</a>	13	174	14	174	<b>2</b>	159	159	169.1	6.4e-50	n/a	<a href="#">Show</a>
<a href="#">5_3_exonuc</a>	5'-3' exonuclease, C-terminal SAM fold	Domain	<a href="#">CL0464</a>	175	269	176	262	<b>2</b>	<b>88</b>	97	103.1	8.7e-30	n/a	<a href="#">Show</a>
<a href="#">Taq-exonuc</a>	Taq polymerase, exonuclease	Domain	<a href="#">CL0219</a>	296	422	297	422	<b>2</b>	129	129	162.3	4.7e-48	n/a	<a href="#">Show</a>
<a href="#">DNA_pol_A</a>	DNA polymerase family A	Family	n/a	455	831	457	830	<b>3</b>	<b>375</b>	376	498.1	1.4e-149	n/a	<a href="#">Show</a>

 **Pfam is part of the ELIXIR infrastructure**  
Pfam is an Elixir service [Read more](#)

Comments or questions on the site? Send a mail to [pfam-help@ebi.ac.uk](mailto:pfam-help@ebi.ac.uk).  
**European Molecular Biology Laboratory**

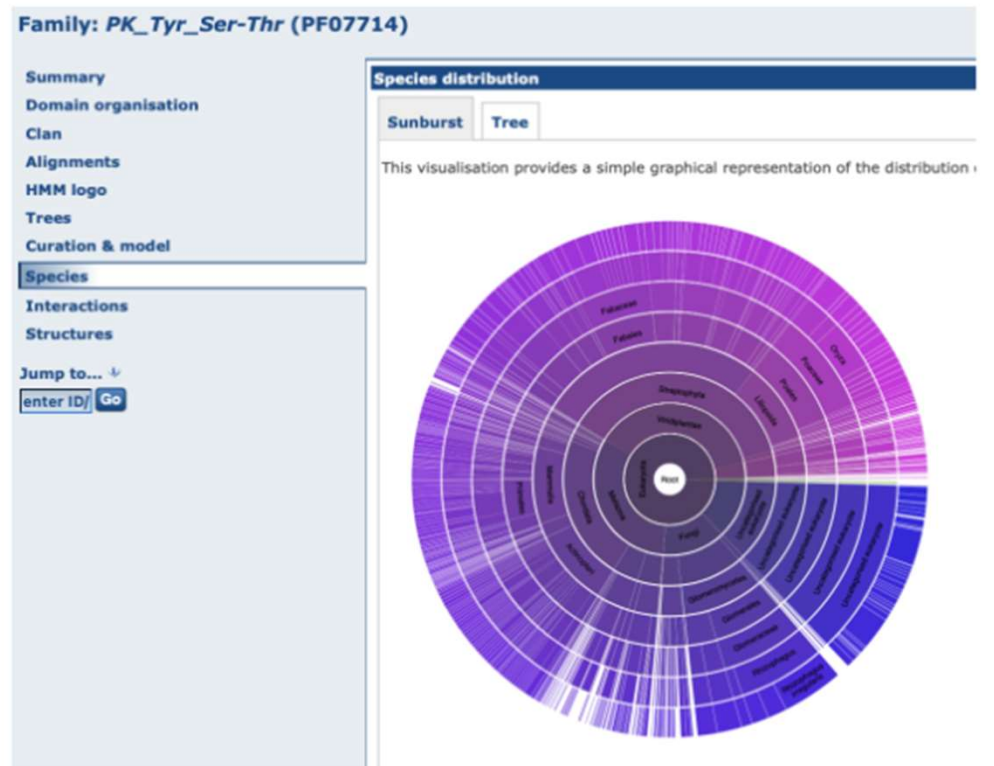
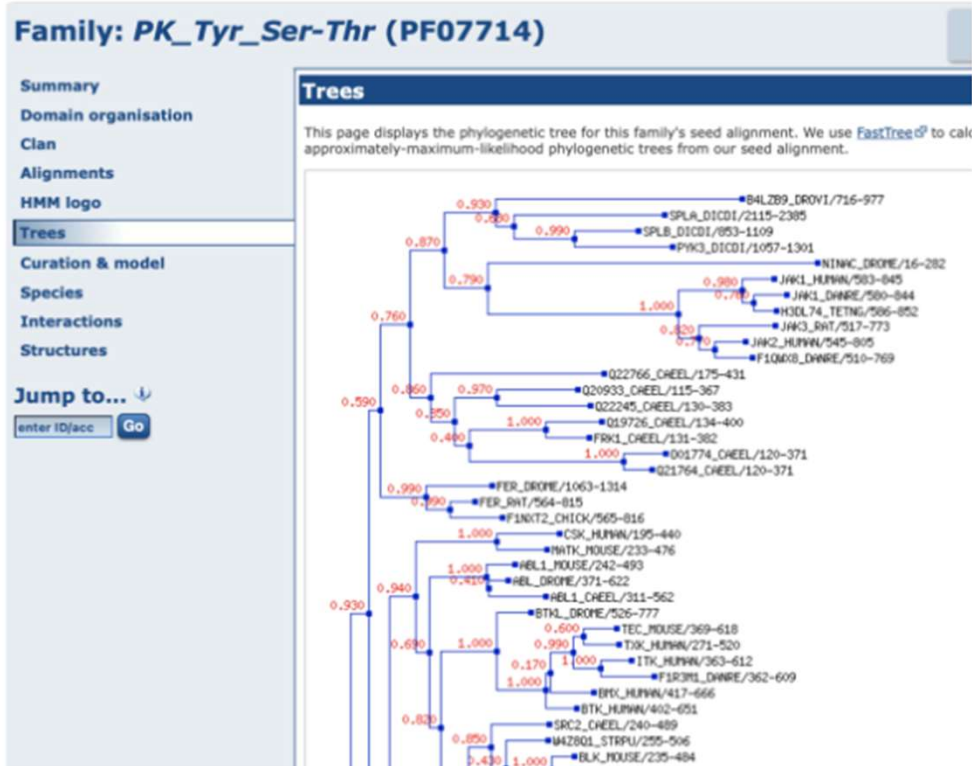
## Defined Protein Domains

# Continued



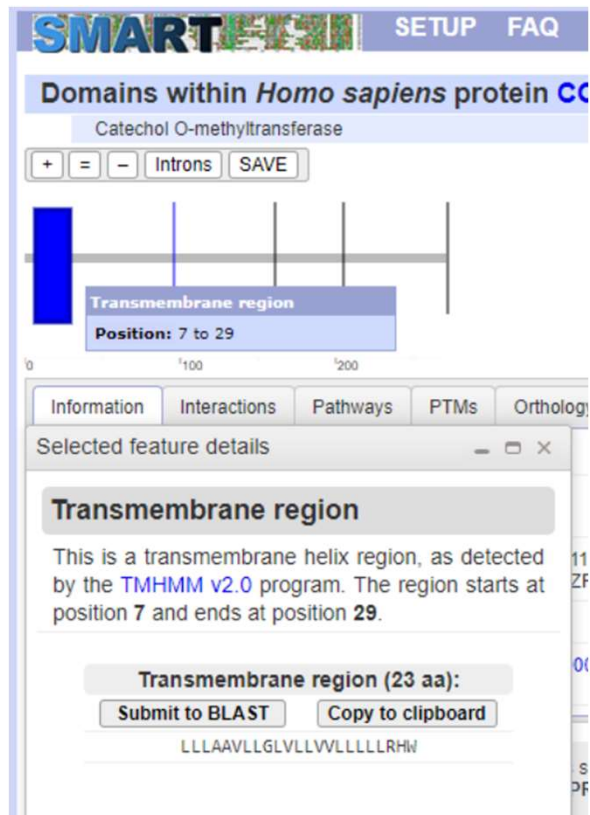
## Evolutionary Relationships & Species Distribution

# Continued



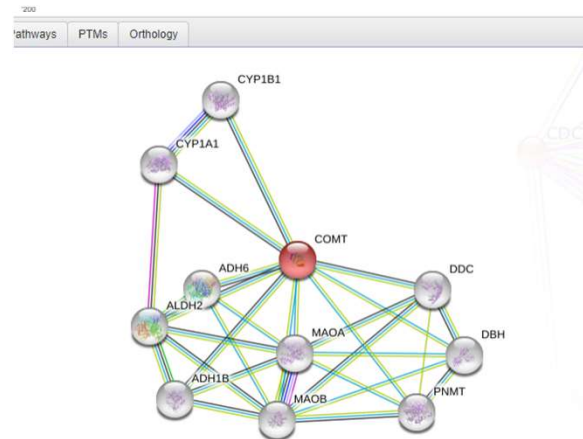
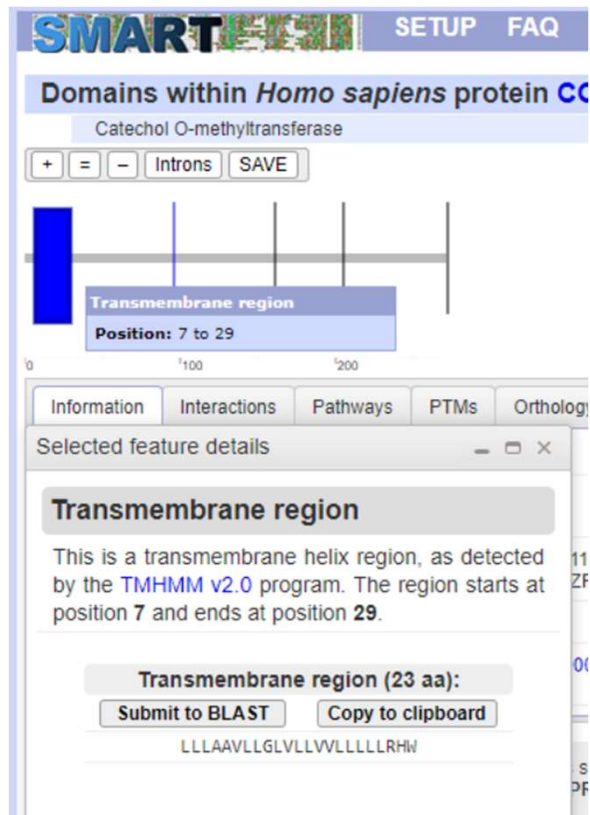
## Evolutionary Relationships & Species Distribution

# What results do we get from SMART?



Simple **M**odular **A**rchitecture **R**esearch **T**ool

# What results do we get from SMART?



Simple **M**odular **A**rchitecture **R**esearch **T**ool



# What results do we get from SMART?

**SMART** SETUP FAQ

Domains within *Homo sapiens* protein CC

Catechol O-methyltransferase

+ = - Introns SAVE

Transmembrane region  
Position: 7 to 29

Information Interactions Pathways PTMs Orthology

Selected feature details

**Transmembrane region**

This is a transmembrane helix region, as detected by the [TMHMM v2.0](#) program. The region starts at position 7 and ends at position 29.

**Transmembrane region (23 aa):**

Submit to BLAST Copy to clipboard

LLLA AVL LGLV L L V L L L L R H W

Pathways PTMs Orthology

Information Interactions Pathways PTMs Orthology

**Posttranslational modifications**

PTM annotation is taken from [PTMcode](#), a resource of known and predicted functional associations between protein posttranslational modifications (PTMs). There are 20 PTMs annotated in this protein:

PTM	Count
Ph Phosphorylation	9
Ub Ubiquitination	9
NT Nitrosylation	2

To see the full details, including possible functional associations between the PTMs, please visit the [PTMcode annotation page for protein COMT](#).

Simple **Modular Architecture Research Tool**

# Why would we use one or the other?

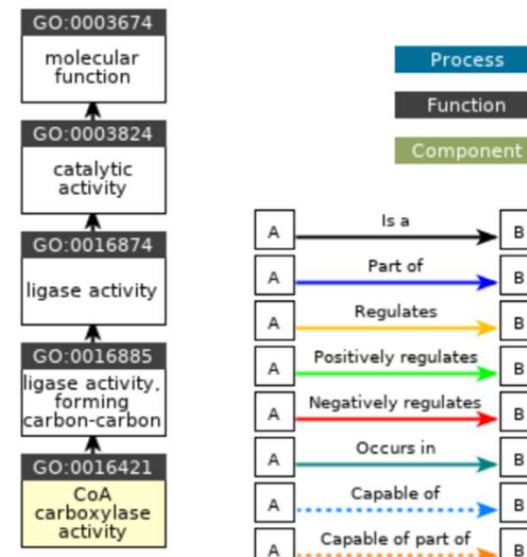
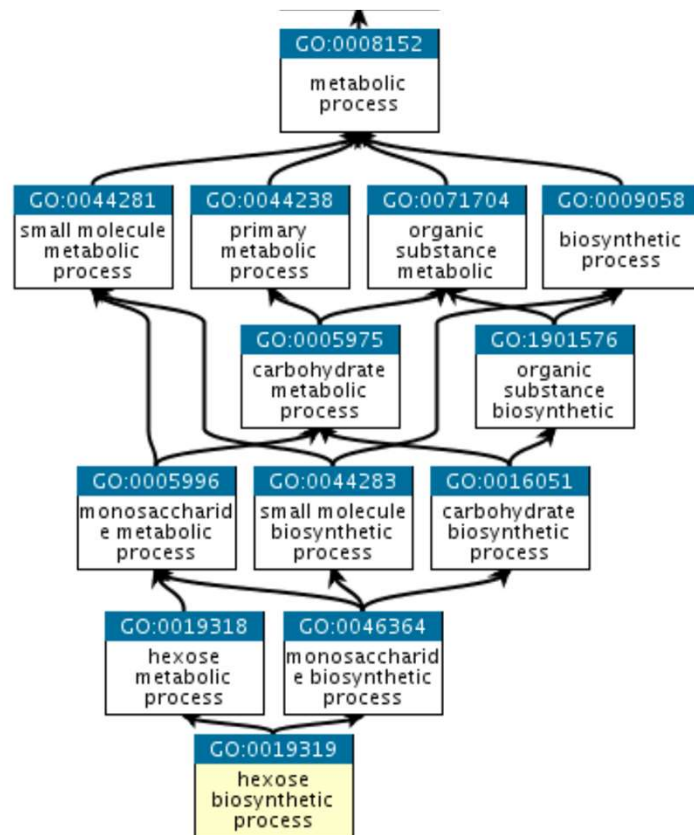
## SMART

More accurate domain  
identification  
Domains exclusively  
annotated  
200M+ proteins in  
database  
Less comprehensive

## Pfam

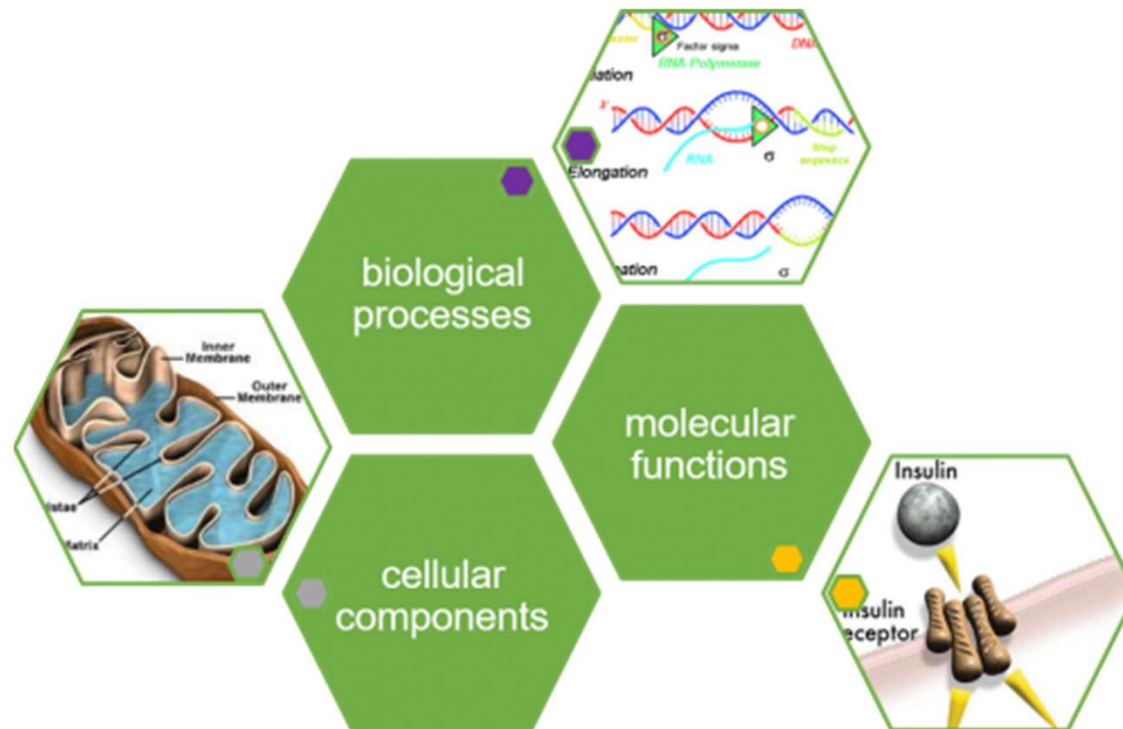
Classifies novel  
sequences into protein  
domain families  
Most comprehensive  
(16K+ families)  
No longer exists

# How do we make sure ALL biologists can understand?

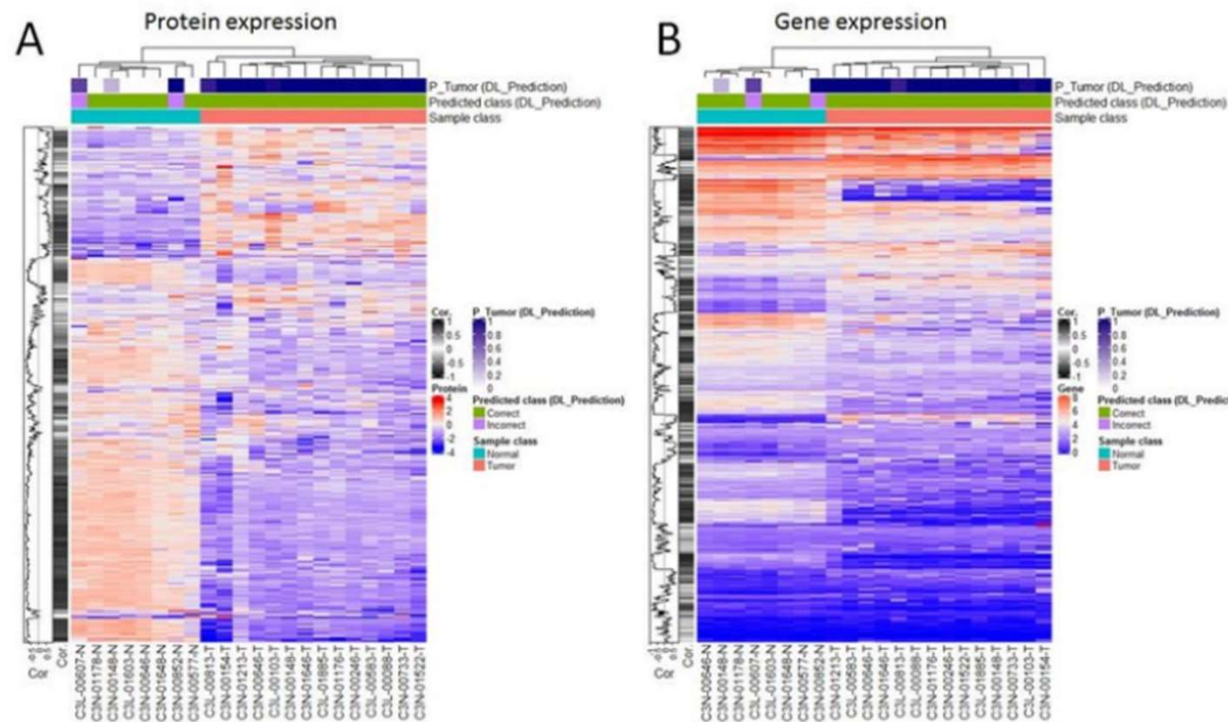


**Gene Ontology**  
logic-based organization  
structure for knowledge

# What are the three Ontologies used?



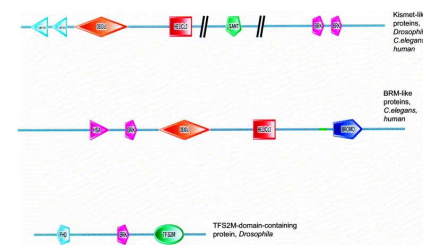
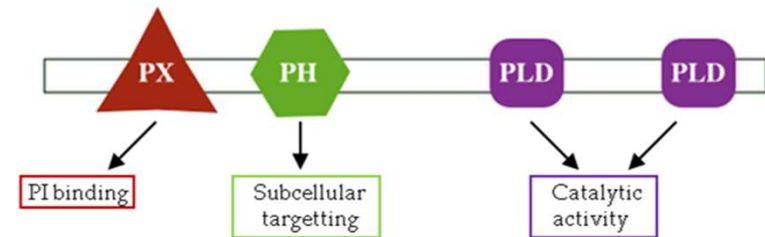
# How do we validate our computational data?



**Proteins that interact should be expressed in the same cell types or under similar conditions**

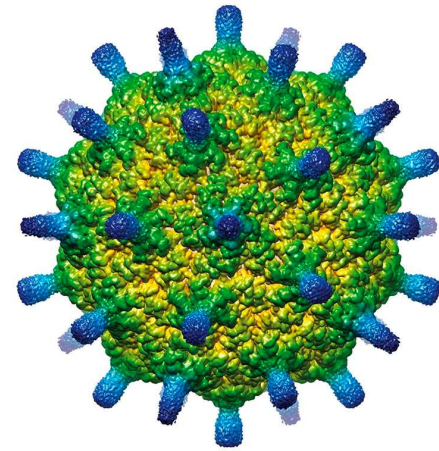
# Summary

1. Domains are conserved structural/functional units of protein
2. Domains can be discovered and analyzed by using bioinformatic approaches
3. Domain analysis helps us classify proteins and impact genetic research





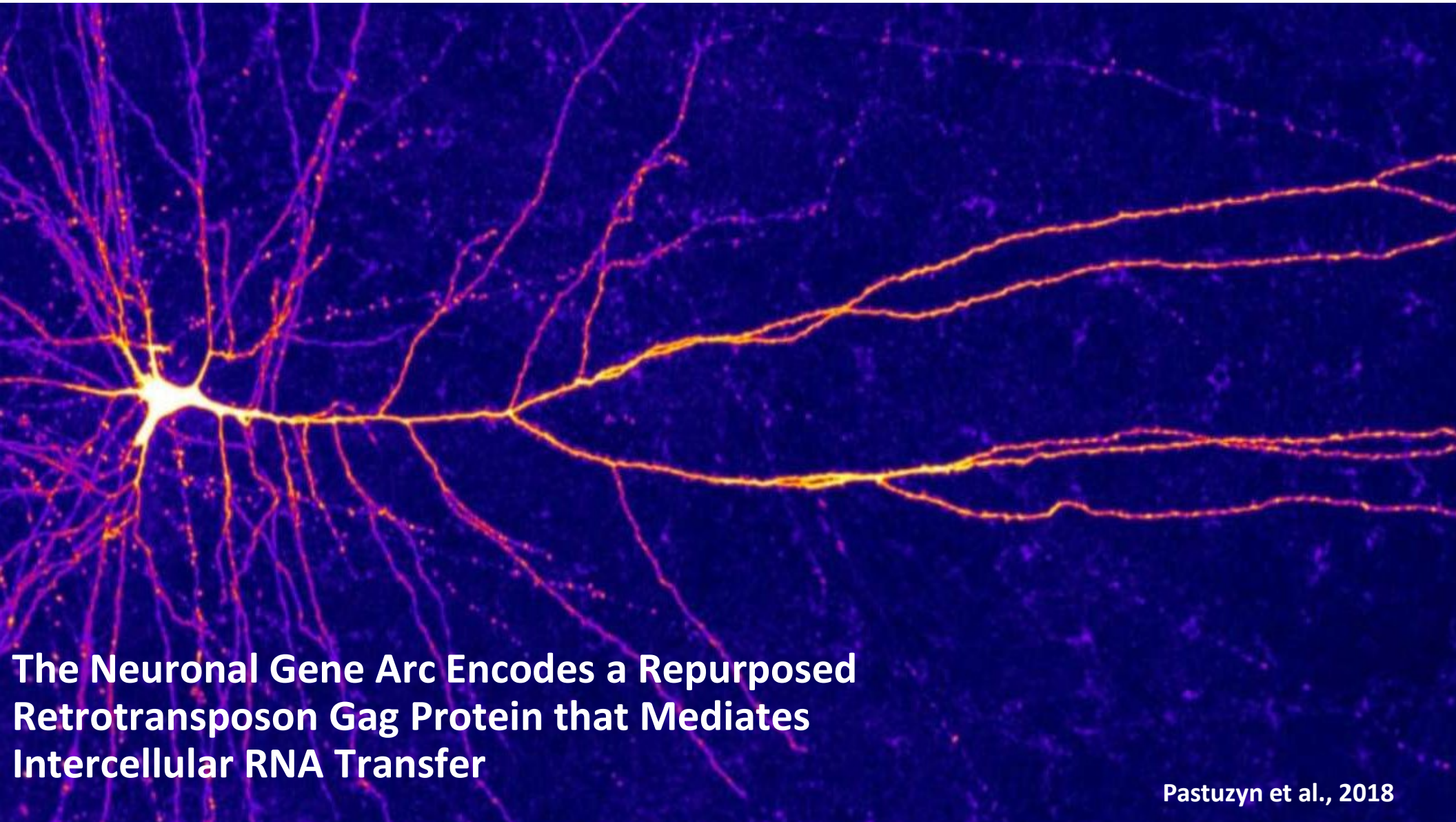
# Dr. Jason Shephard & the Arc protein



**University of Utah, School of Medicine**

Molecular function of the Arc protein in long-term memory formation, 2018



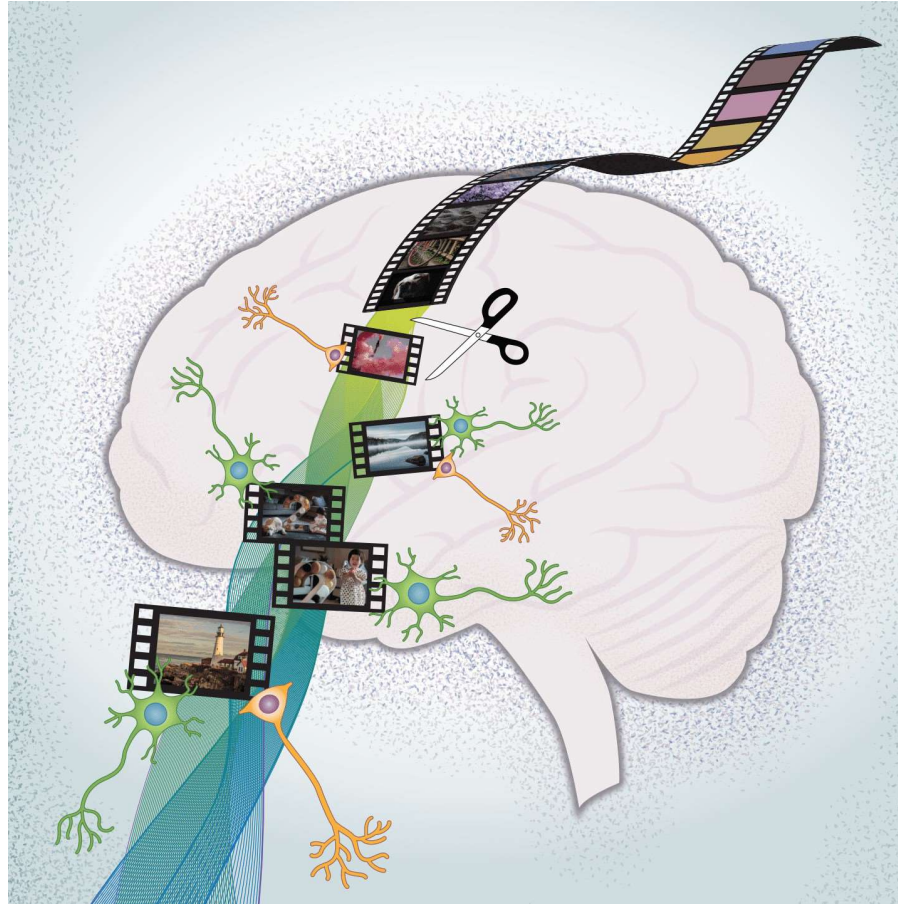


**The Neuronal Gene Arc Encodes a Repurposed  
Retrotransposon Gag Protein that Mediates  
Intercellular RNA Transfer**

Pastuzyn et al., 2018

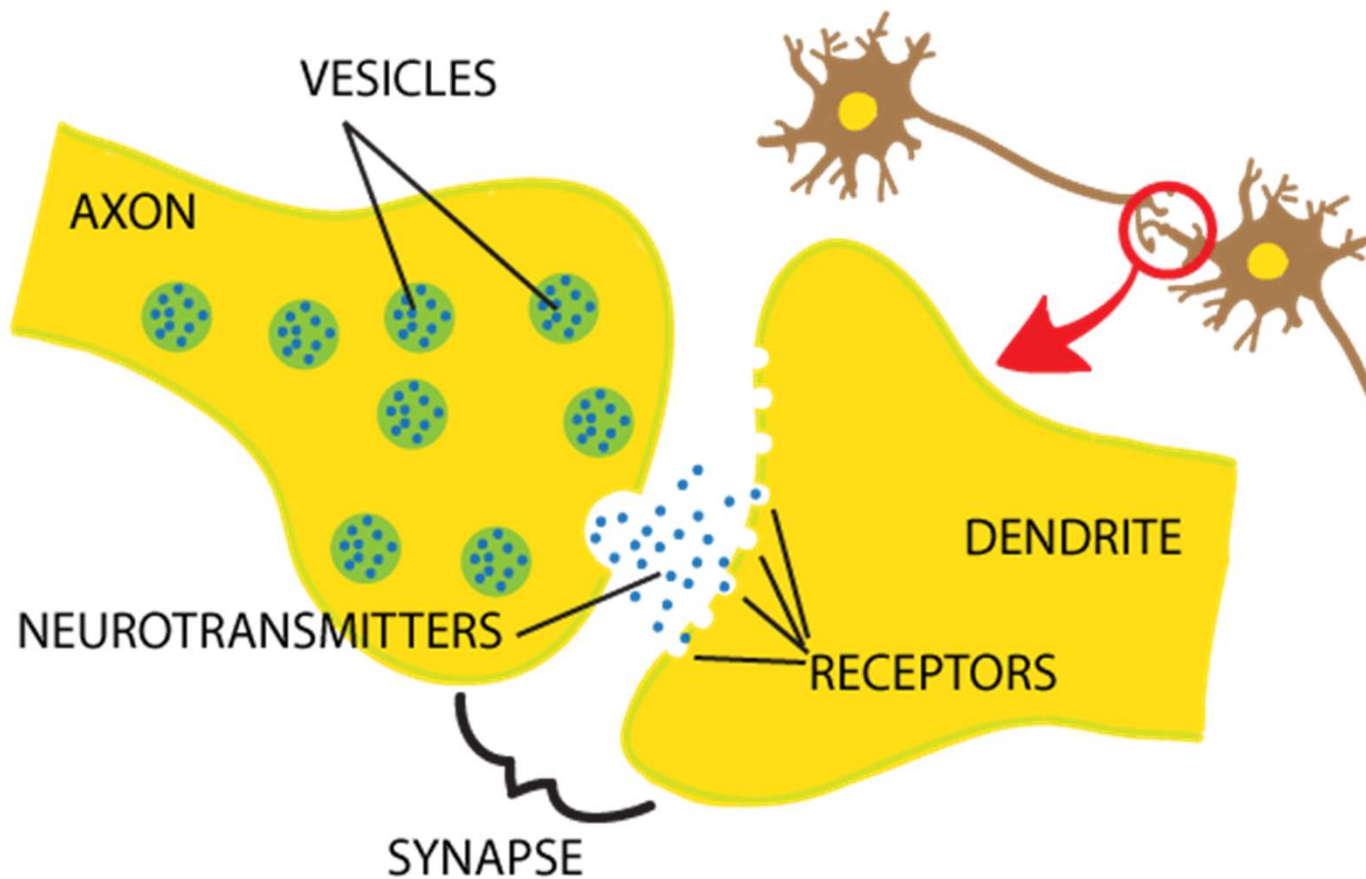


# How does the brain store information?



**Through synaptic connections between interconnected networks of neurons**

# Synapses move information

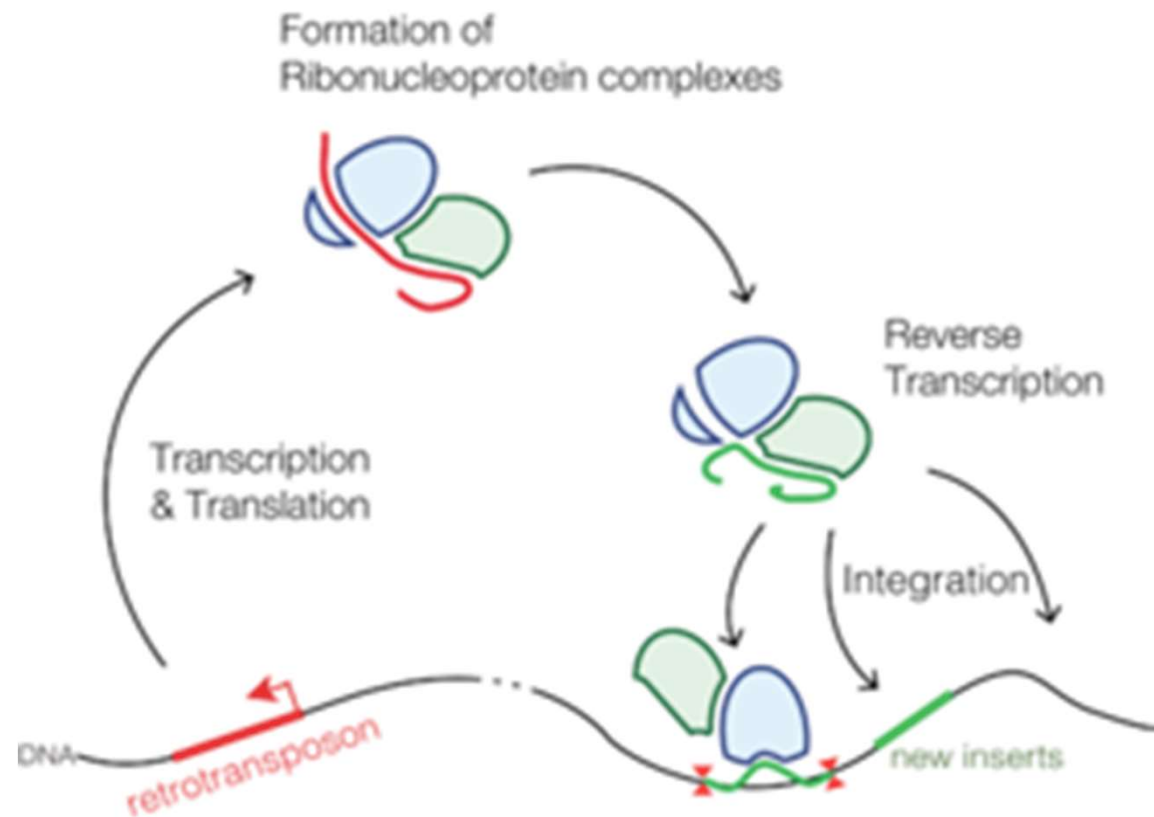


**It is unknown how memories are moved around**



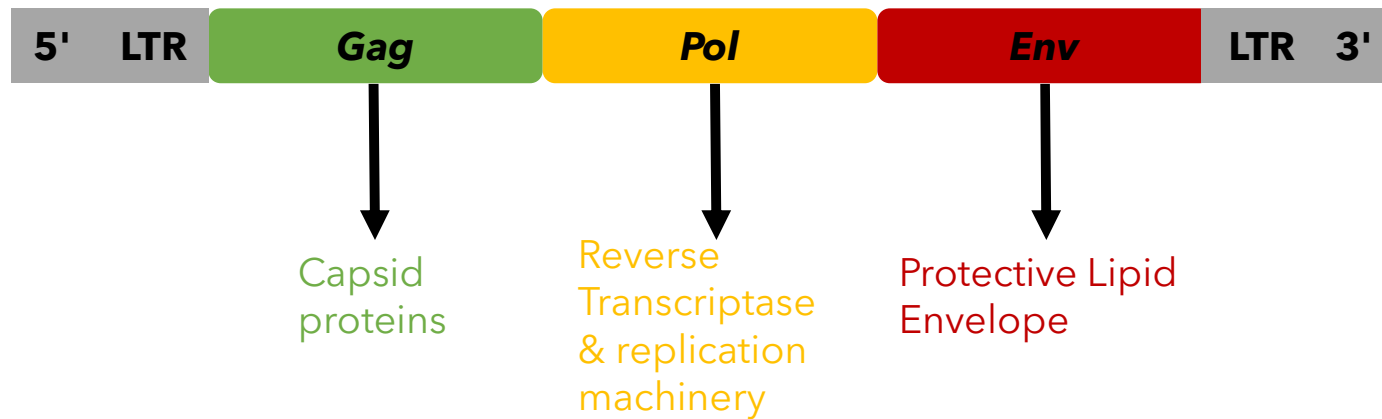


# Memories are thought to move via **retrotransposons**



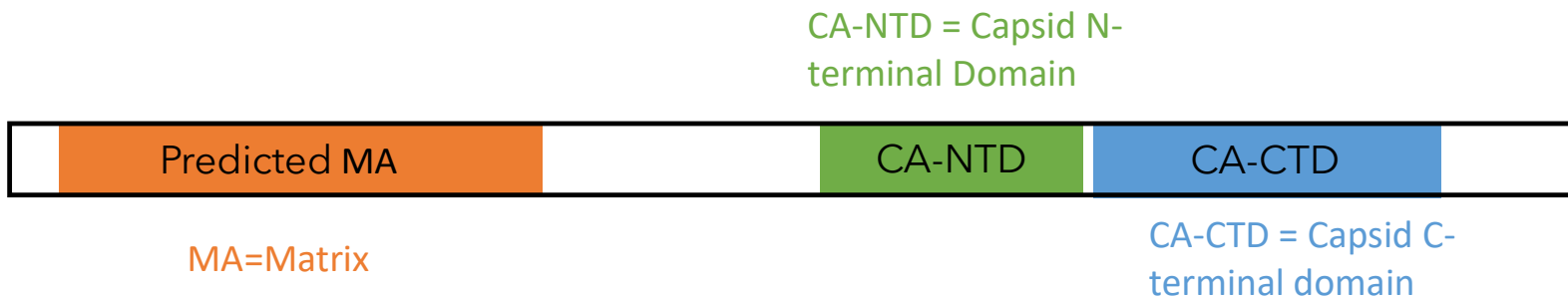
**Genetic components that can insert themselves elsewhere in the genome**

# What is a retrovirus?



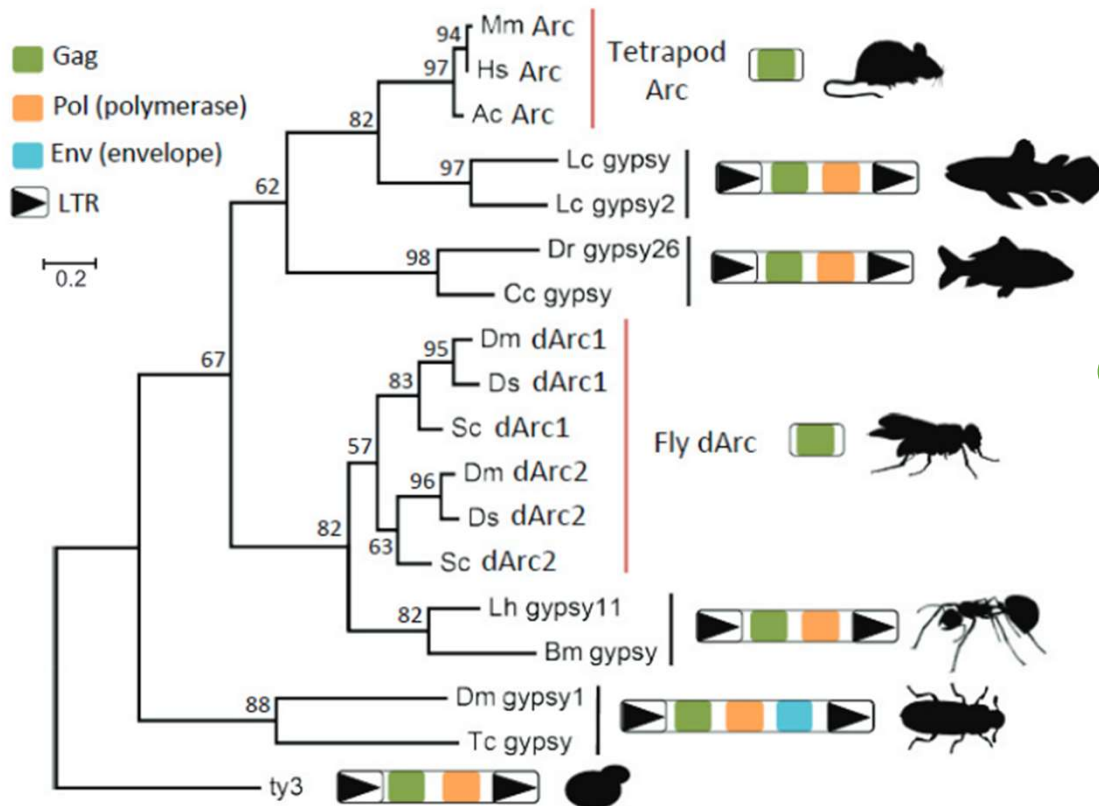
An RNA virus that replicates itself into a host genome via **capsids**

# Arc is a protein important for memory



**Arc plays an important role in synaptic plasticity**

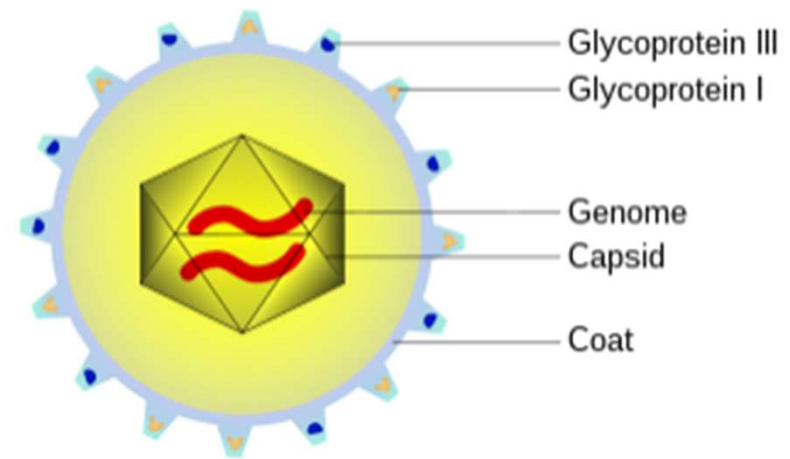
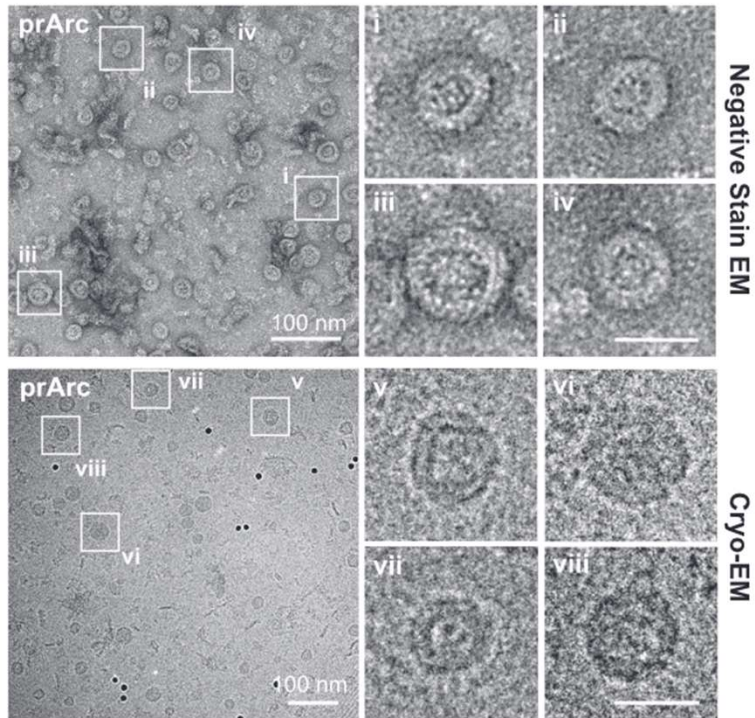
# What are the evolutionary origins of Arc?



Conserved retroviral Gag domain

**Ty3/gypsy retrotransposon family**

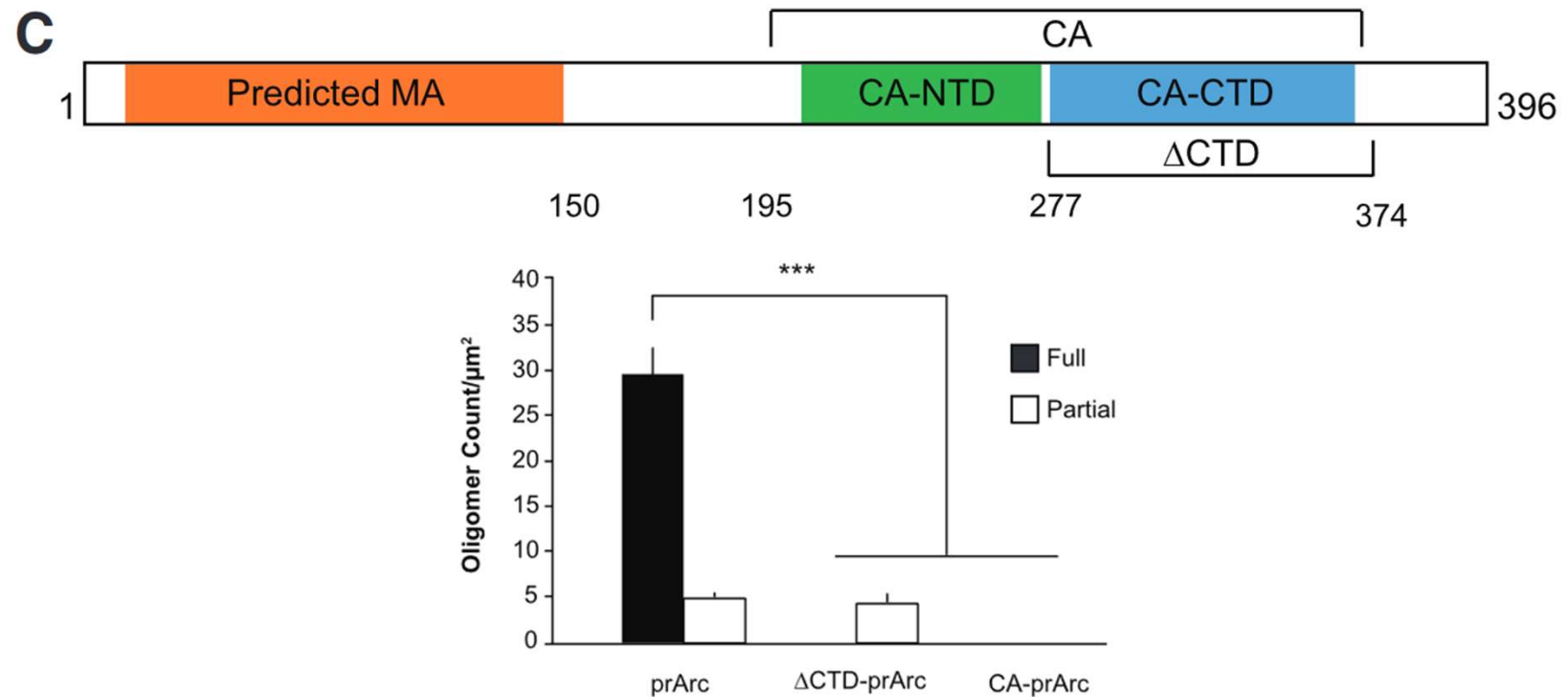
# Does Arc form virus-like capsids?



**Purified Arc resembles retrovirus capsid structure**

Pastuzyn et al., 2018

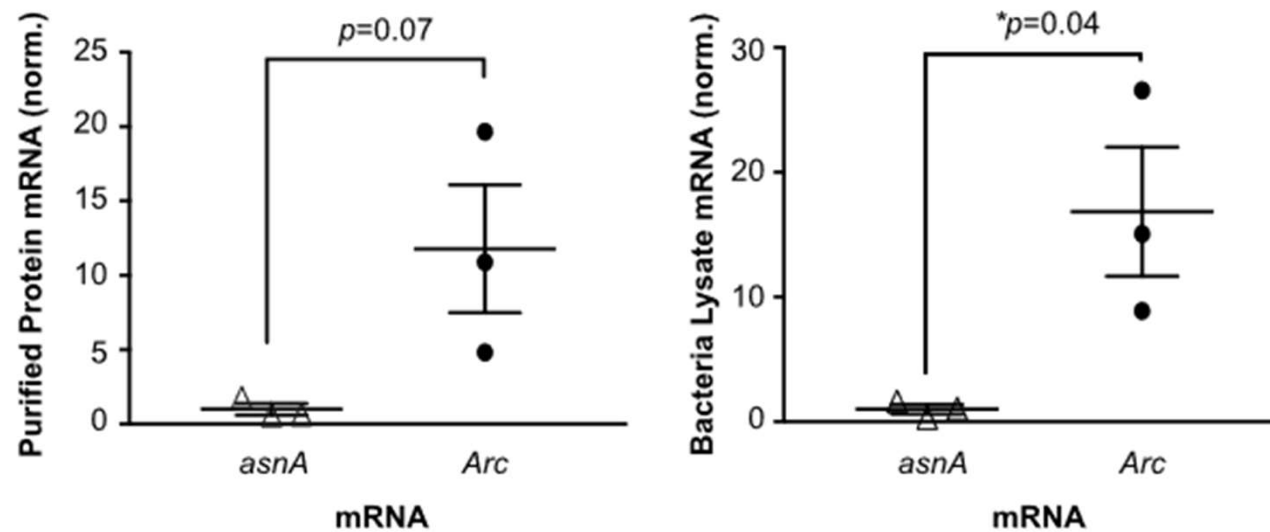
# What regions of Arc are important for self-assembly into capsids?



**Double-shelled protein structure of Arc cannot form when the CTD is deleted**

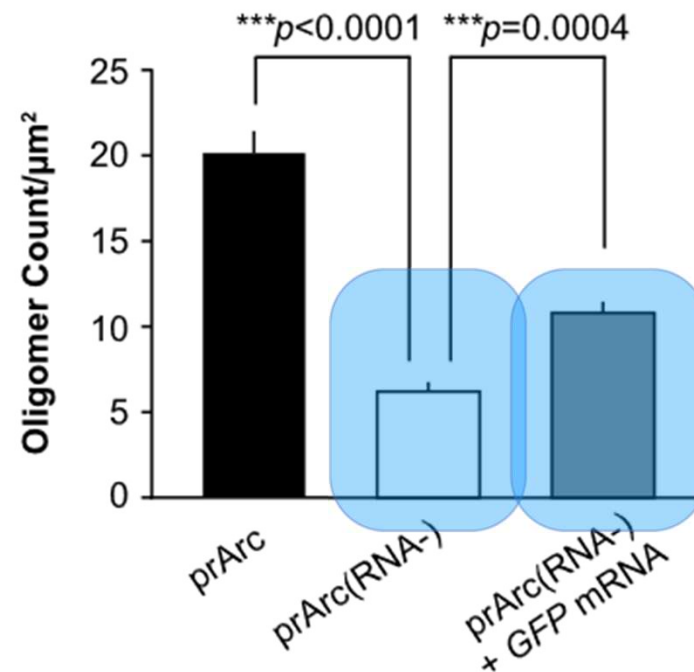


# Do Arc capsids contain mRNA?



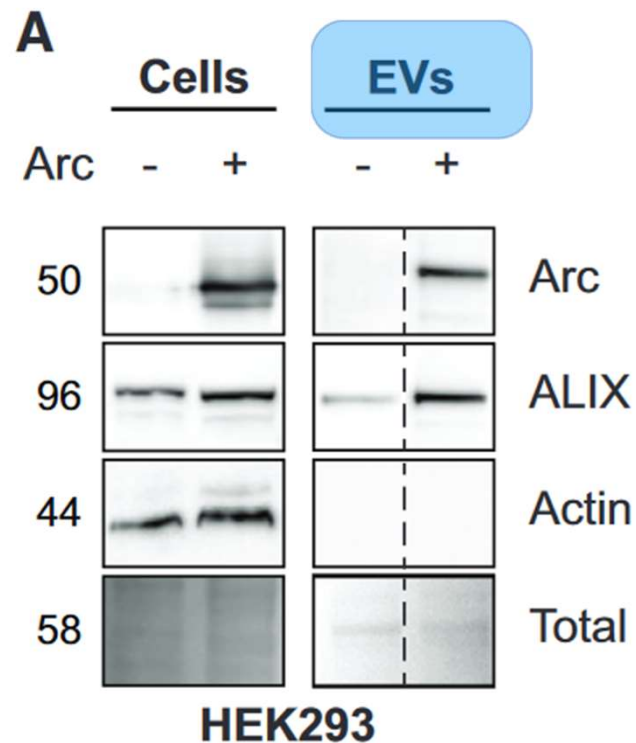
**Interactions between the Arc protein and the Arc mRNA is important for the stability of the mRNA**

# Does Arc require mRNA for capsid formation?



**Removal of RNA bases decreases capsid formation**  
**Introduction of mRNA increases capsid formation**

# Where is Arc mRNA found in relation to neurons?

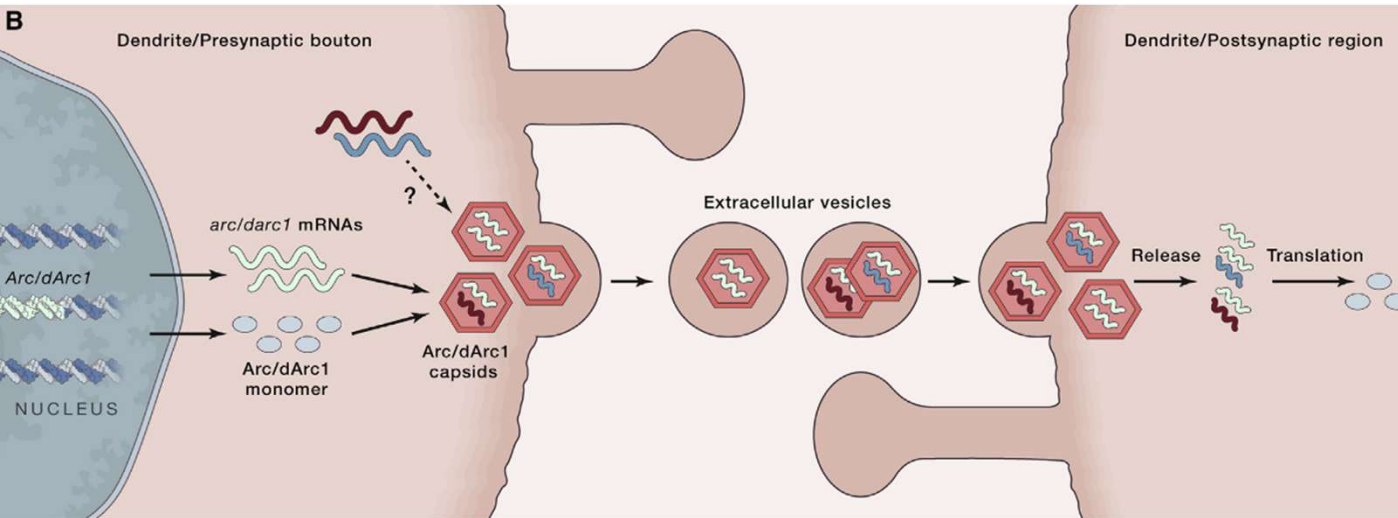


Arc mRNA is found in **Extracellular Vesicles (EVs)** outside of neurons

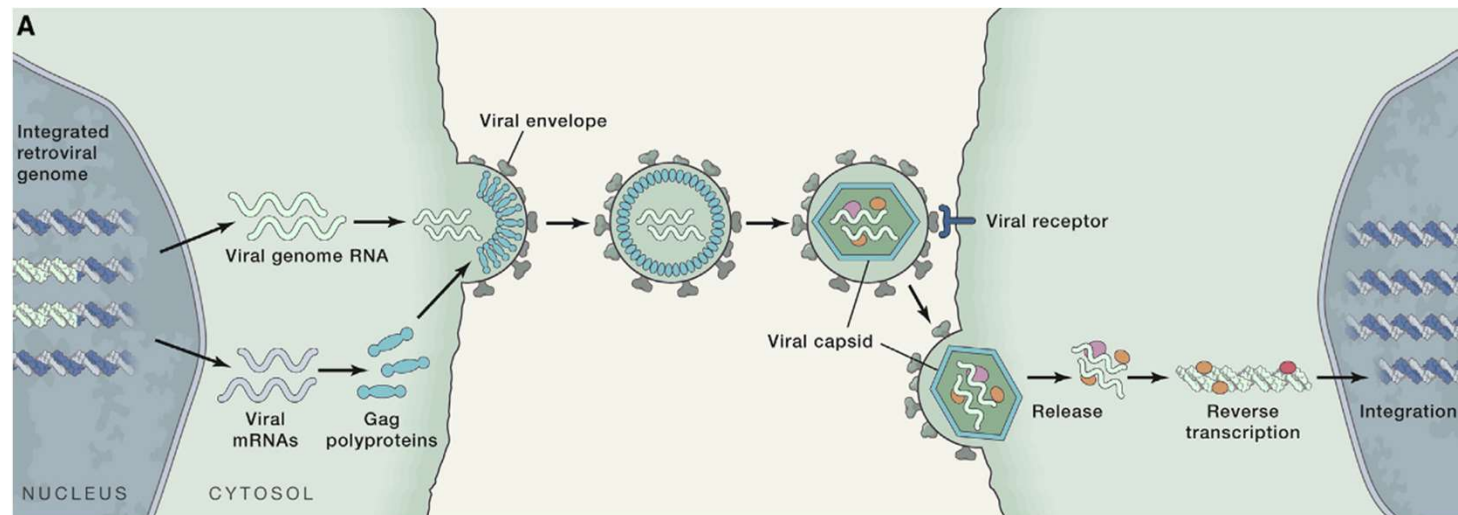
Pastuzyn et al., 2018

# What are extracellular vesicles?

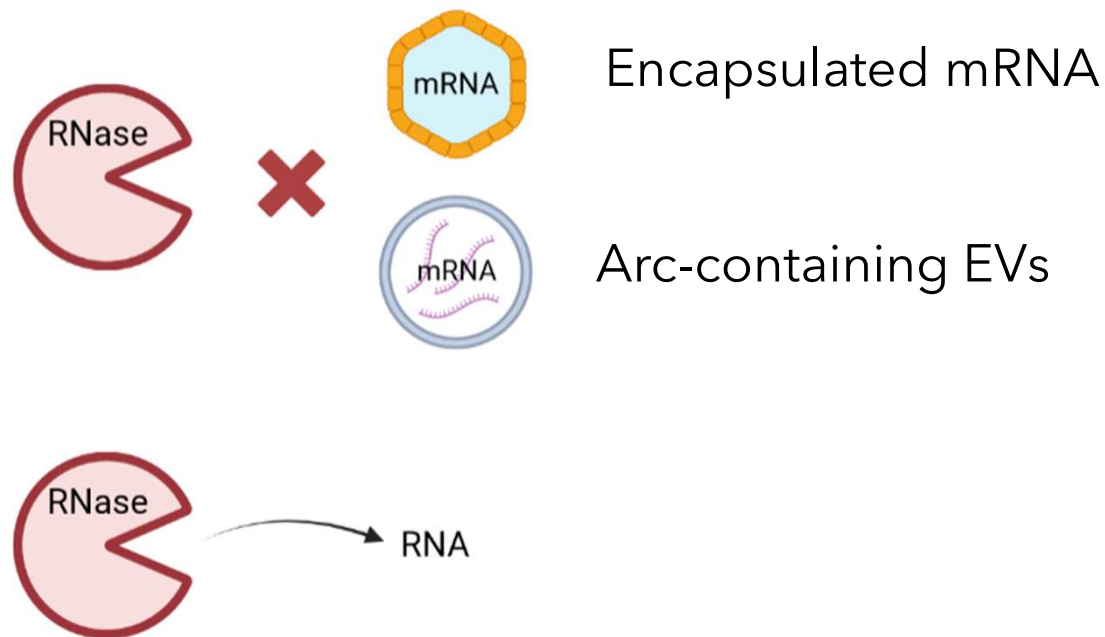
Parrish et al., 2018



**Vs.**

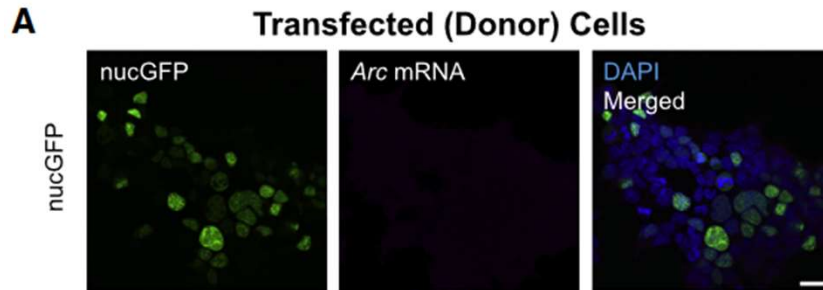


# Do EVs degrade outside of neurons?



**mRNA within EVs or encapsulated are stable!**

# Do EVs transfer **Arc protein** and **mRNA** to recipient cells?

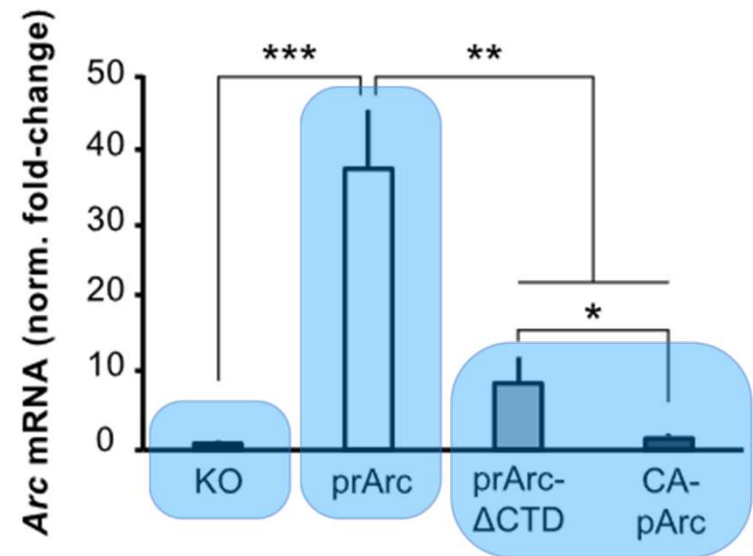
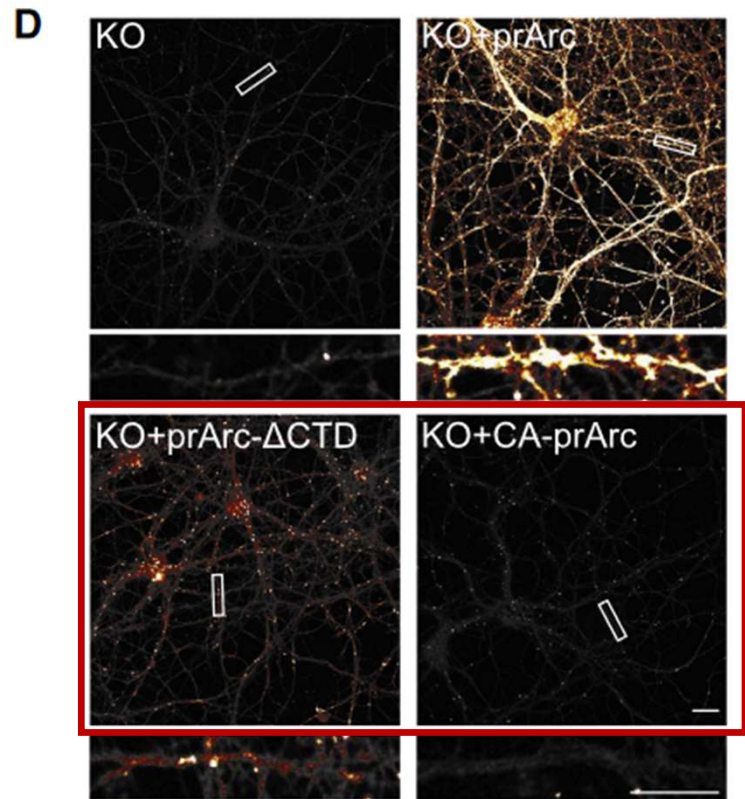


**EVs can transfer **protein** and **mRNA** to recipient cells**

Pastuzyn et al., 2018



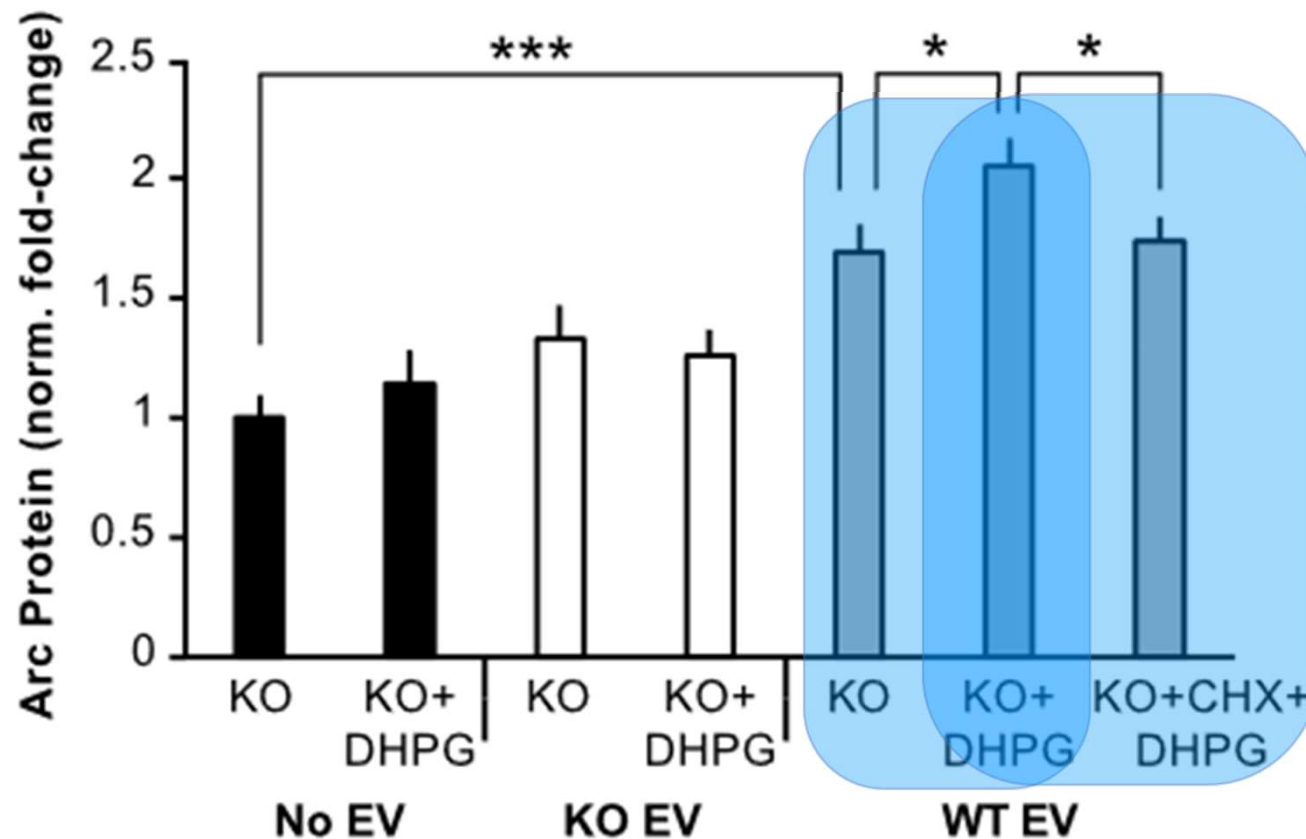
# Are Arc capsids required for neuron uptake?



**Capsid formation is necessary for transfer of mRNA**

Pastuzyn et al., 2018

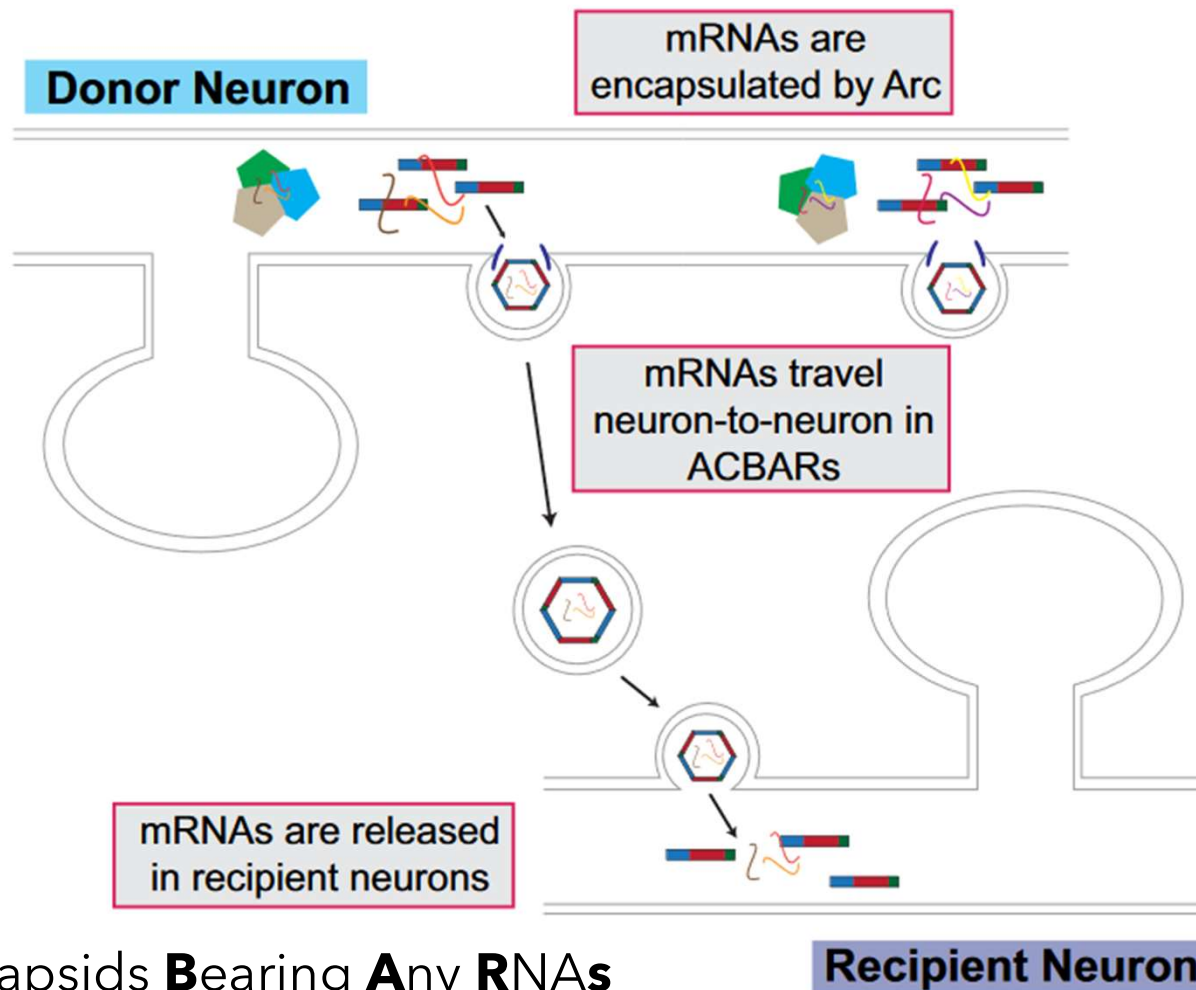
# Does Arc undergo activity-dependent translation in neurons?



Translation of Arc mRNA is increased by DHPG treatment in neurons

Pastuzyn et al., 2018

# Arc works in the brain like a retrovirus



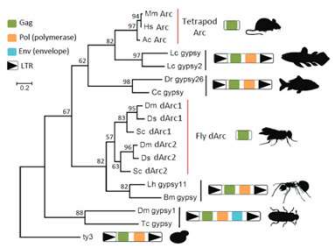
ACBARs=**A**rc **C**apsids **B**earing **A**ny **R**NAs

**Recipient Neuron**

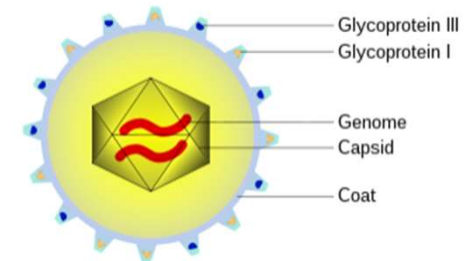
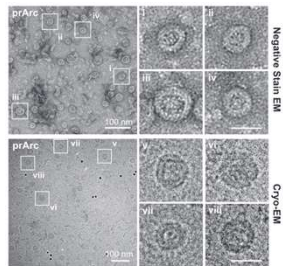
Pastuzyn et al., 2018

# Summary

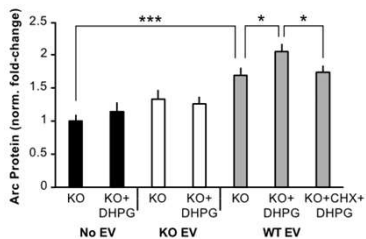
**Arc shares retroviral Gag protein properties**



**Arc forms stable capsid-like structures**

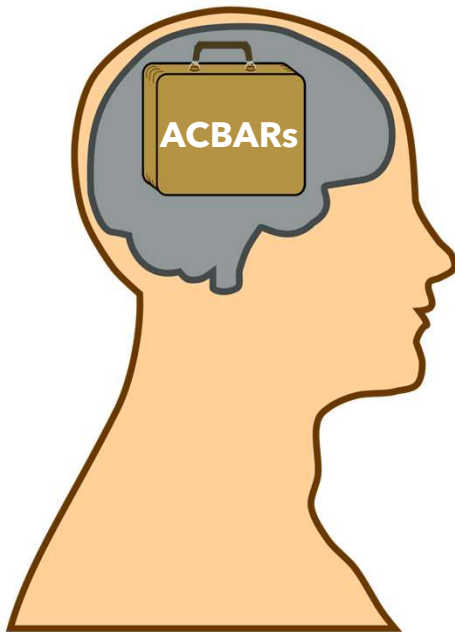


**These structures allow for the transfer of mRNA from neuron to neuron**

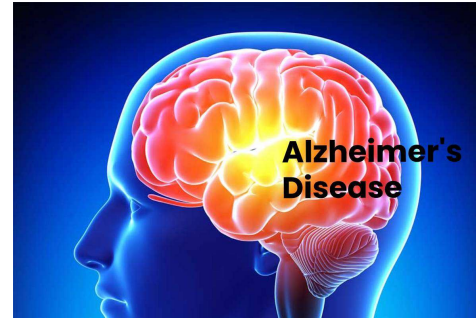


# Future Directions

**What else do ACBARs contain?**



**What else does Arc play a role in?**



**Questions?**





# References

- 1) Pastuzyn ED, Day CE, Kearns RB, Kyrke-Smith M, Taibi AV, McCormick J, Yoder N, Belnap DM, Erlendsson S, Morado DR, Briggs JAG, Feschotte C, Shepherd JD. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell*. 2018 Jan 11;172(1-2):275-288.e18. doi: 10.1016/j.cell.2017.12.024. Erratum in: *Cell*. 2018 Mar 22;173(1):275. PMID: 29328916; PMCID: PMC5884693
- 2) Parrish, N. F., & Tomonaga, K. (2018). A Viral (Arc)hive for Metazoan Memory. *Cell*, 172(1-2), 8–10. doi: 10.1016/j.cell.2017.12.029

Images:

<https://www.resonancescience.org/blog/Neurons-Act-Not-As-Simple-Logic-Gates-But-As-Complex-Multi-Unit-Processing-Systems>

<https://en.wikipedia.org/wiki/Retrotransposon>

<https://clarkesworldmagazine.com/images/commentary/hirv-genome-figure-koboldt.jpg>

<https://en.wikipedia.org/wiki/Capsid>

<https://sarabloggingadventure.wordpress.com/2018/07/21/thoughts-on-podcasts/>

<https://www.wsj.com/articles/what-questions-to-ask-in-a-job-interview-11606259284>

Biorender.com

<https://clipartmag.com/cartoon-brain-clipart>

<https://www.downloadclipart.net/browse/40067/little-tan-suitcase-clipart>

- <https://www.ebi.ac.uk/training/online/courses/goa-and-quickgo-quick-tour/what-is-go/>
- <http://geneontology.org/docs/ontology-documentation/>
- [https://teaching.ncl.ac.uk/bms/seq/cmb2000/info/pages/gifs/protein\\_analysis\\_pfam\\_2\\_2018.png](https://teaching.ncl.ac.uk/bms/seq/cmb2000/info/pages/gifs/protein_analysis_pfam_2_2018.png)
- <http://smart.embl-heidelberg.de/>
- <https://en.wikipedia.org/wiki/Pfam>
- <https://www.uniprot.org/>
- <https://publish.illinois.edu/msaevaluation/>
- <https://towardsdatascience.com/hidden-markov-model-applied-to-biological-sequence-373d2b5a24c>
- <https://www.sciencedirect.com/science/article/pii/S0014579301032896>
- <https://www.tedmed.com/talks/show?id=729641>
- <https://www2.mrc-lmb.cam.ac.uk/structures-of-virus-like-capsids-involved-in-learning-and-memory-formation/>
- <https://www.linkedin.com/pulse/art-asking-questions-bala-pitchandi/>

- [https://en.wikiversity.org/wiki/File:Amino\\_acid\\_structure.png](https://en.wikiversity.org/wiki/File:Amino_acid_structure.png)
- [https://en.m.wikipedia.org/wiki/File:Protein\\_TNKS2\\_PDB\\_3KR7.png](https://en.m.wikipedia.org/wiki/File:Protein_TNKS2_PDB_3KR7.png)
- <https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-rna-and-protein-synthesis/a/the-genetic-code>
- [https://en.wikipedia.org/wiki/Alternative\\_splicing](https://en.wikipedia.org/wiki/Alternative_splicing)
- [https://www.researchgate.net/figure/Cereblon-gene-and-protein-conservation-across-different-vertebrate-species-top\\_fig5\\_338724512](https://www.researchgate.net/figure/Cereblon-gene-and-protein-conservation-across-different-vertebrate-species-top_fig5_338724512)
- <http://libertgen564s15.weebly.com/clustal-omega-alignment.html>
- <https://www.bbc.com/future/article/20121102-will-we-ever-crack-lifes-code>
- <https://www.nature.com/articles/s41467-019-08295-x>
- [https://www.researchgate.net/figure/Motifs-Domains-and-Full-length-proteins-A-Secondary-structure-often-packs-into-motifs\\_fig3\\_344391925](https://www.researchgate.net/figure/Motifs-Domains-and-Full-length-proteins-A-Secondary-structure-often-packs-into-motifs_fig3_344391925)
- <http://kohlmannngen677s13.weebly.com/motifs-and-domains.html>
- [https://www.researchgate.net/figure/Graphical-representation-of-sequence-motifs-detected-in-the-upstream-DNA-sequence-of-TA\\_fig5\\_258830817](https://www.researchgate.net/figure/Graphical-representation-of-sequence-motifs-detected-in-the-upstream-DNA-sequence-of-TA_fig5_258830817)
- [https://www.researchgate.net/figure/DxDxDG-structural-motifs-in-a-variety-of-structural-contextsA-Structures-of-two\\_fig3\\_229555712](https://www.researchgate.net/figure/DxDxDG-structural-motifs-in-a-variety-of-structural-contextsA-Structures-of-two_fig3_229555712)
- [https://en.wikipedia.org/wiki/Solenoid\\_protein\\_domain](https://en.wikipedia.org/wiki/Solenoid_protein_domain)
- <https://www.ebi.ac.uk/training/online/courses/protein-classification-intro-ebi-resources/protein-classification/what-are-protein-domains/>
- <https://www.ebi.ac.uk/training/online/courses/protein-classification-intro-ebi-resources/what-are-protein-signatures/signature-types/>