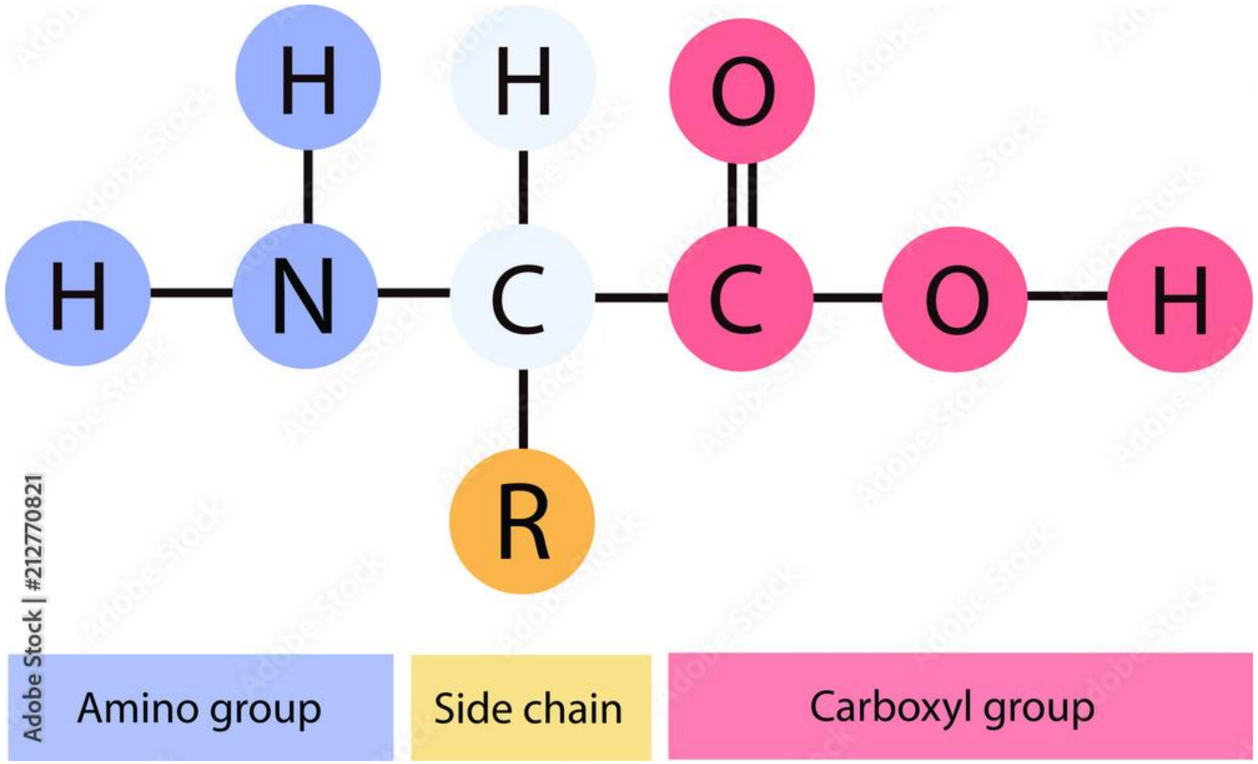
Motif and Domain Discovery

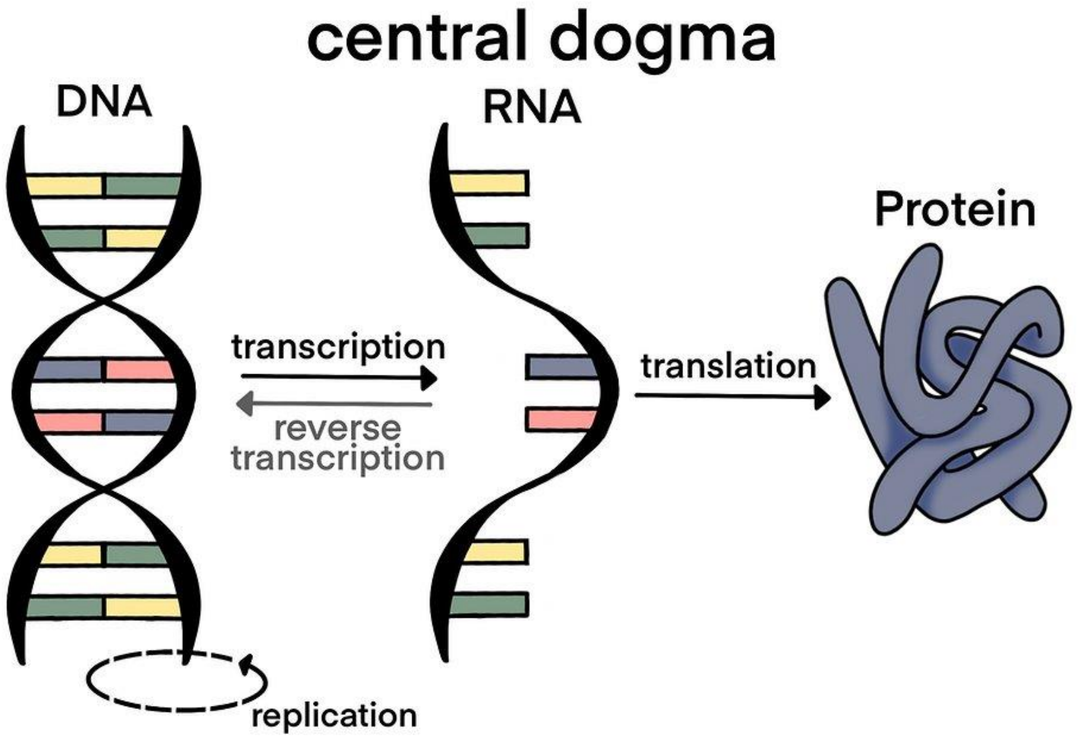
Rohan Babaria and Joely Swanson

What are Proteins?



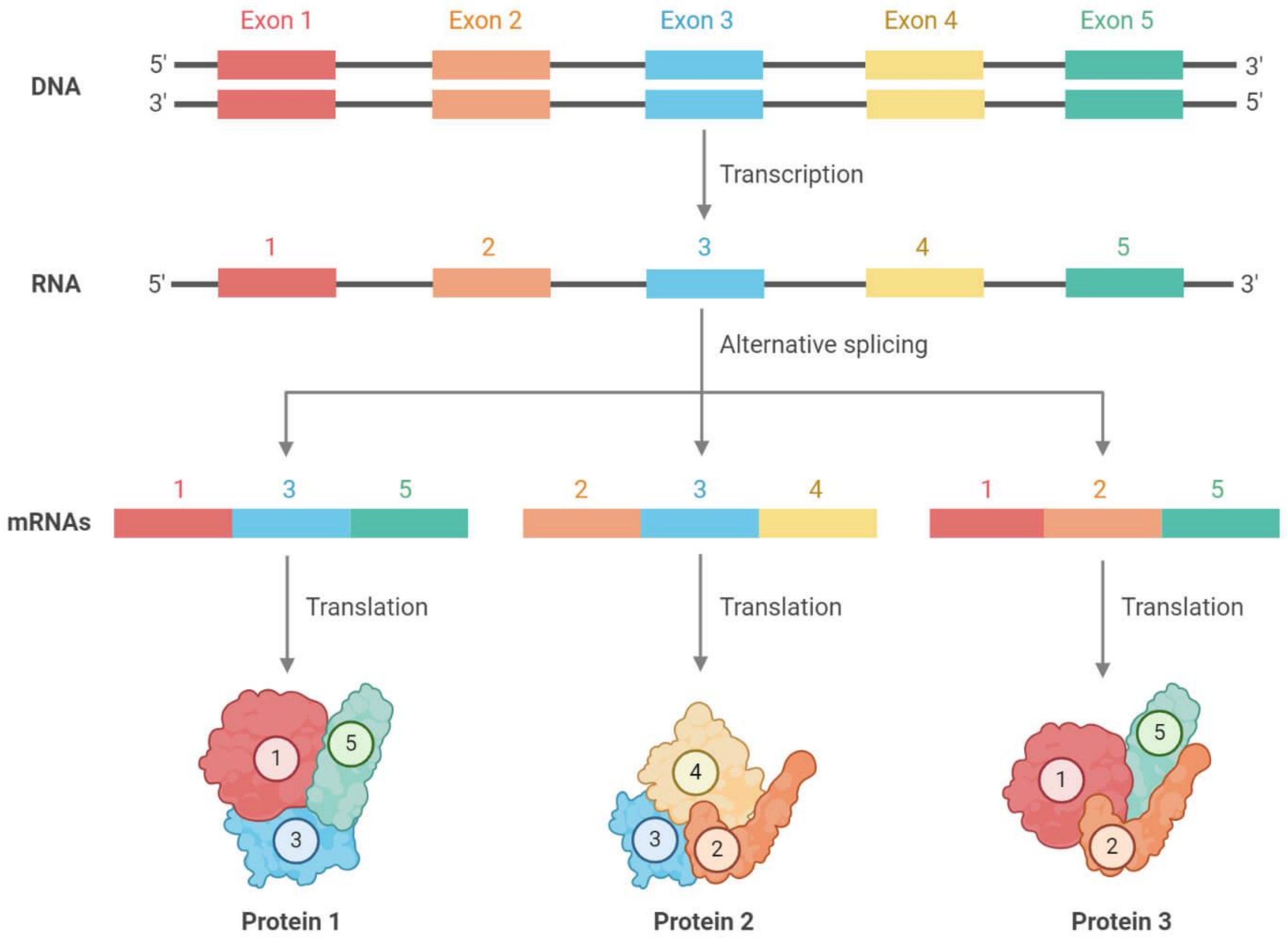
Proteins are polypeptides, chains of amino acids, that act as the functional unit of life.

How are proteins made?



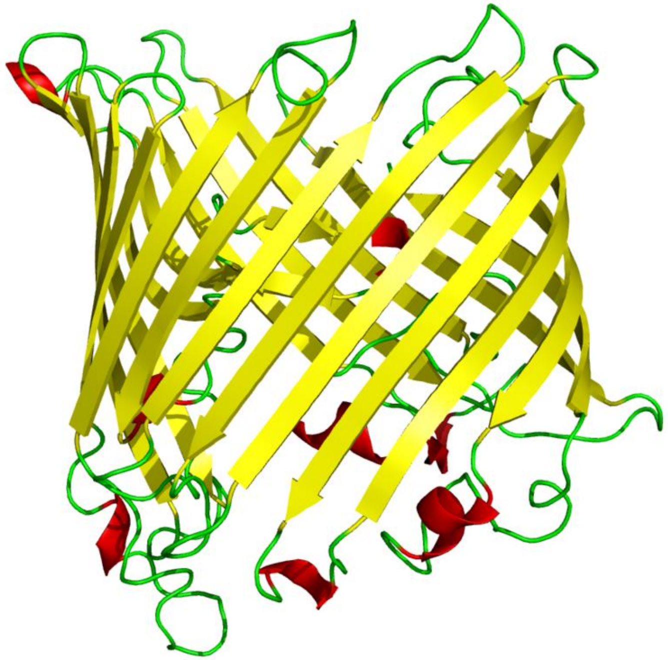
Proteins are translated from mRNA transcripts derived from DNA.

How are multiple protein variants produced?



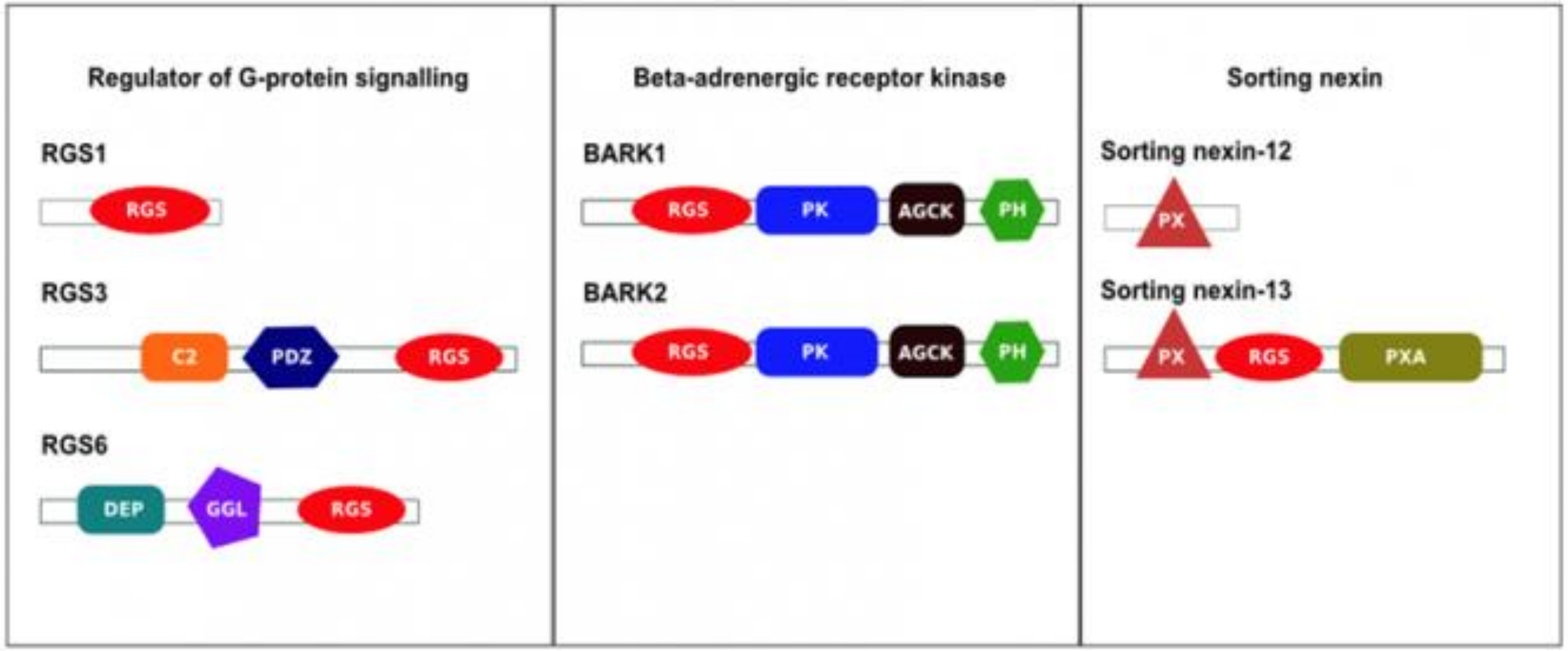
What are the 4 structure types?

What are Protein Motifs?



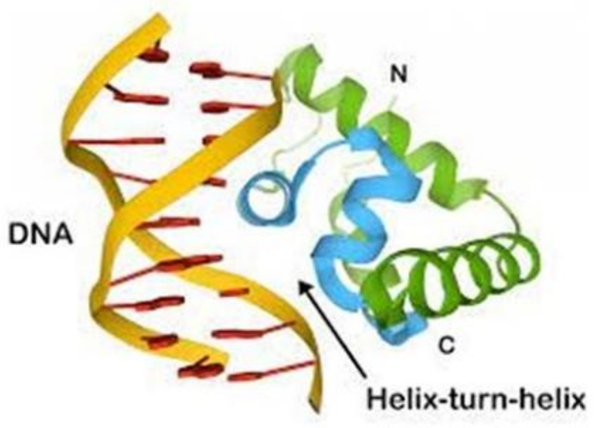
Small region of protein with a common three-dimensional structure/sequence shared among proteins.

What are Protein Domains?

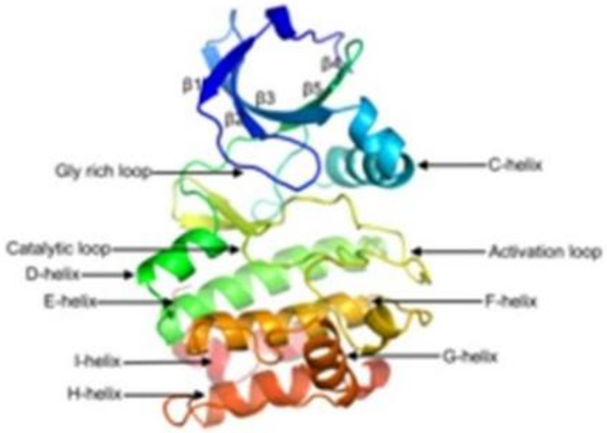


A conserved sequence pattern that acts as an independent functional and structural unit.

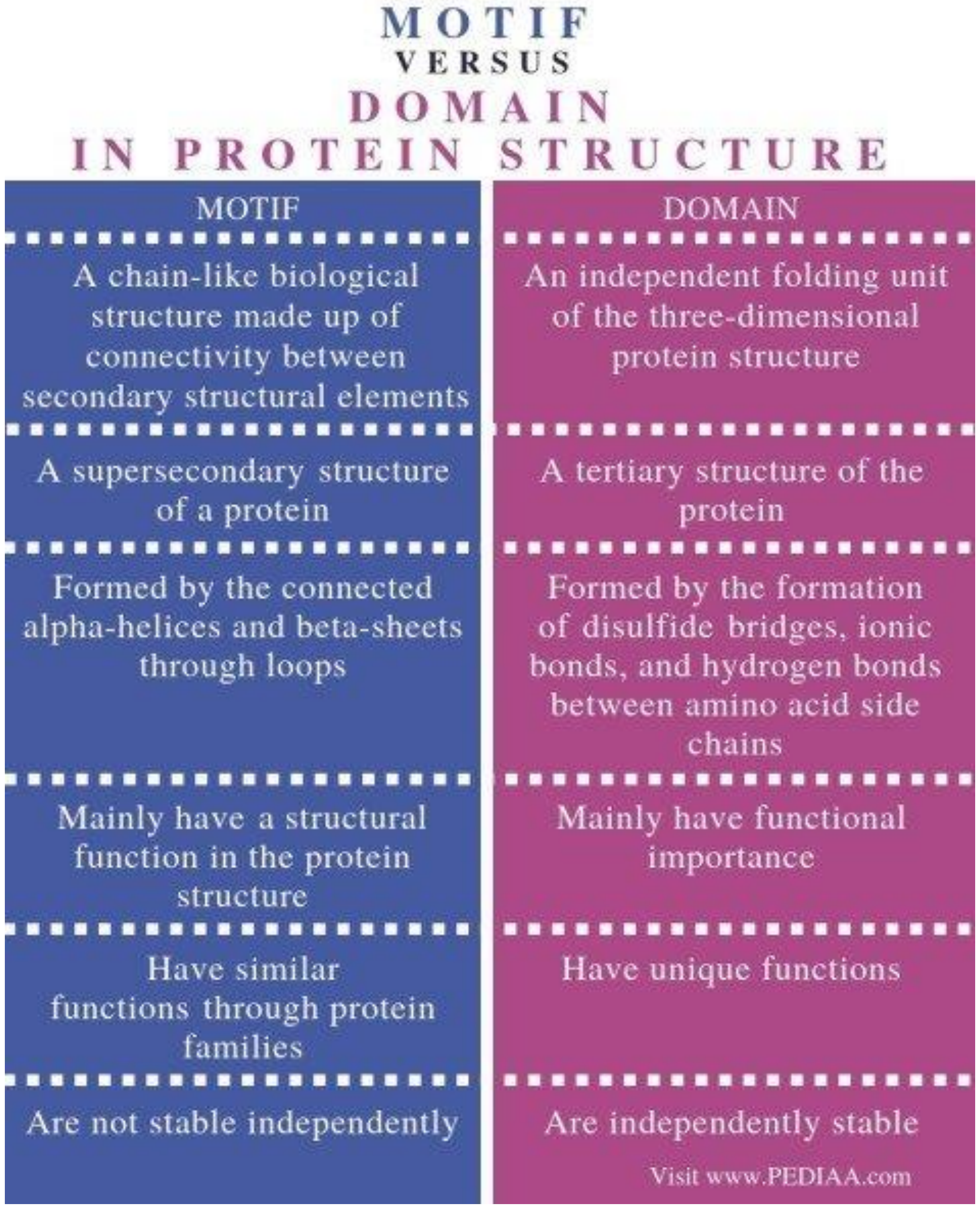
What are the differences?



Motif



Domain

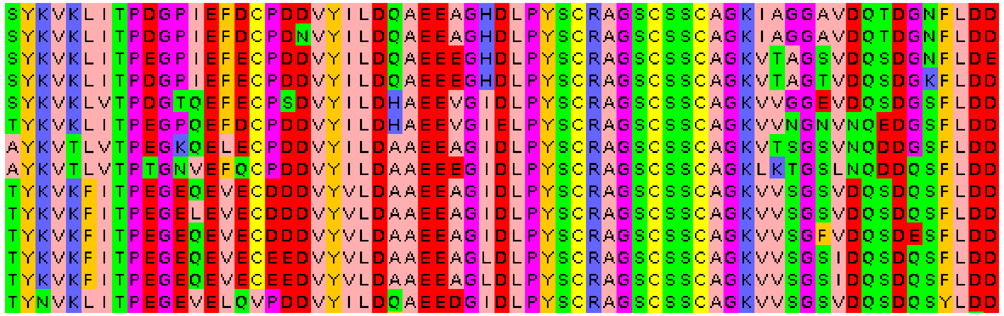


MOTIF VERSUS DOMAIN IN PROTEIN STRUCTURE

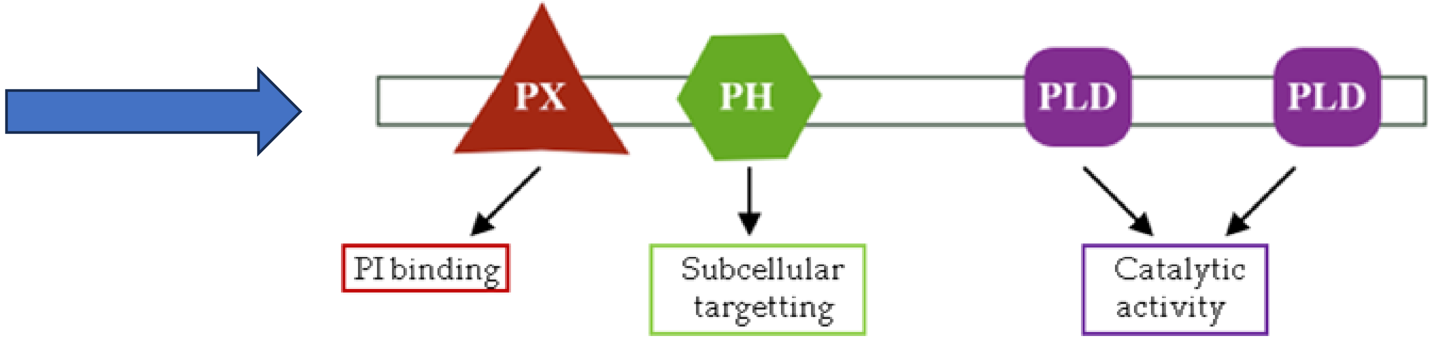
MOTIF	DOMAIN
A chain-like biological structure made up of connectivity between secondary structural elements	An independent folding unit of the three-dimensional protein structure
A supersecondary structure of a protein	A tertiary structure of the protein
Formed by the connected alpha-helices and beta-sheets through loops	Formed by the formation of disulfide bridges, ionic bonds, and hydrogen bonds between amino acid side chains
Mainly have a structural function in the protein structure	Mainly have functional importance
Have similar functions through protein families	Have unique functions
Are not stable independently	Are independently stable

Visit www.PEDIAA.com

Why do we care?



```
SYKVKLITPDGPIEFDCPDDVYILDQAEEAGHDLPYSCRAGSCSSCAGKIAGGAVDQTDGNFLDD
SYKVKLITPDGPIEFDCPDNVYILDQAEEAGHDLPYSCRAGSCSSCAGKIAGGAVDQTDGNFLDD
SYKVKLITPEGPIEFECPDDVYILDQAEEEGHDLPYSCRAGSCSSCAGKVTAGSVDQSDGNFLDE
SYKVKLITPDGPIEFECPDDVYILDQAEEEGHDLPYSCRAGSCSSCAGKVTAGTVDQSDGKFLDD
SYKVKLVTPDGTQEFECPSDVYILDHAEEVGIDLPYSCRAGSCSSCAGKVVGGEVDQSDGSFLDD
TYKVKLITPEGPQEFDCPDDVYILDHAEEVGIELPYSCRAGSCSSCAGKVVNGNVNQEDGSFLDD
AYKVTLVTPEGKQELECPDDVYILDAAEEAGIDLPYSCRAGSCSSCAGKVTSGSVNQDDGSFLDD
AYKVTLVTPTGNVEFQCPDDVYILDAAEEEGIDLPYSCRAGSCSSCAGKLKTGSLNQDDQSFLDD
TYKVKFITPEGEQEVECDDDVYVLDAAEEAGIDLPYSCRAGSCSSCAGKVVSGSVDQSDQSFLDD
TYKVKFITPEGELEVECDDDVYVLDAAEEAGIDLPYSCRAGSCSSCAGKVVSGSVDQSDQSFLDD
TYKVKFITPEGEQEVECDDDVYVLDAAEEAGIDLPYSCRAGSCSSCAGKVVSGFVDQSDESFLDD
TYKVKFITPEGEQEVECEEDVYVLDAAEEAGLDLPYSCRAGSCSSCAGKVVSGSIDQSDQSFLDD
TYKVKFITPEGEQEVECEEDVYVLDAAEEAGLDLPYSCRAGSCSSCAGKVVSGSIDQSDQSFLDD
TYNVKLITPEGEVELQVPDDVYILDQAEEDGIDLPYSCRAGSCSSCAGKVVSGSVDQSDQSYLDD
```



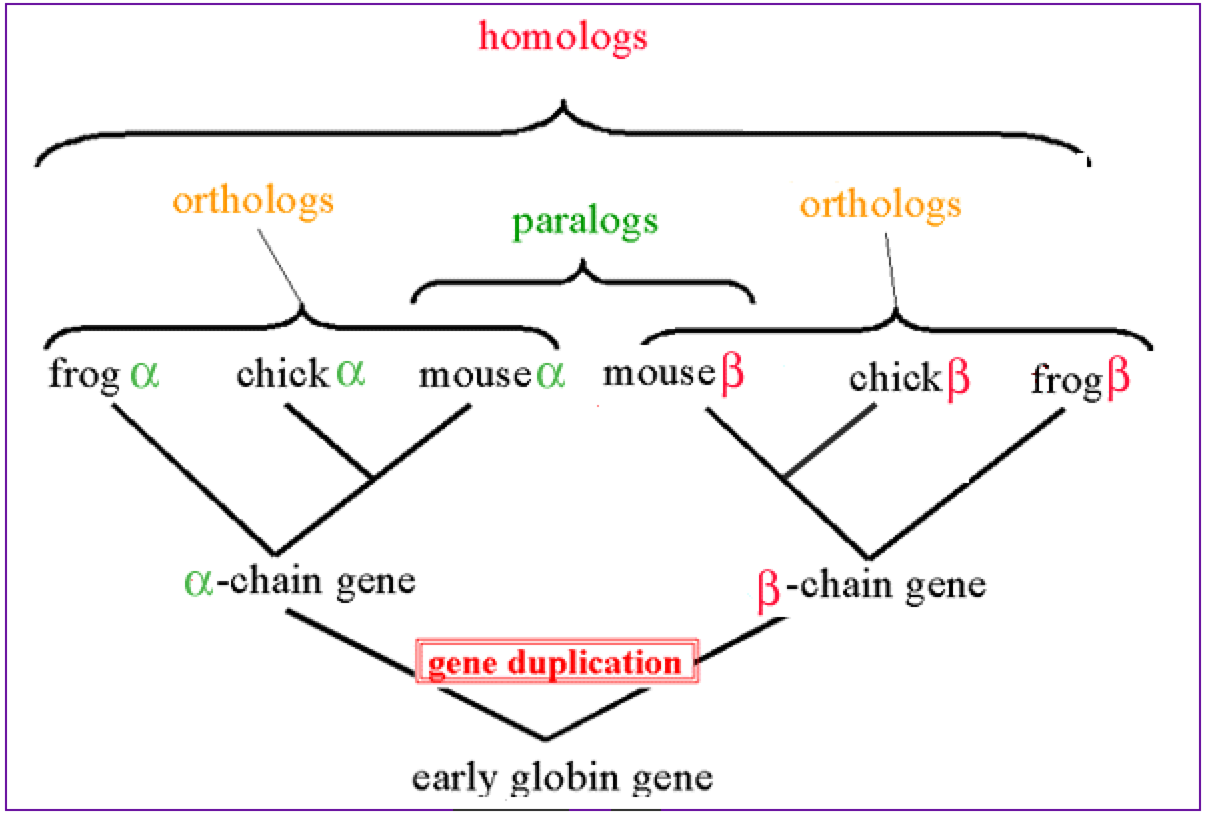
PI binding

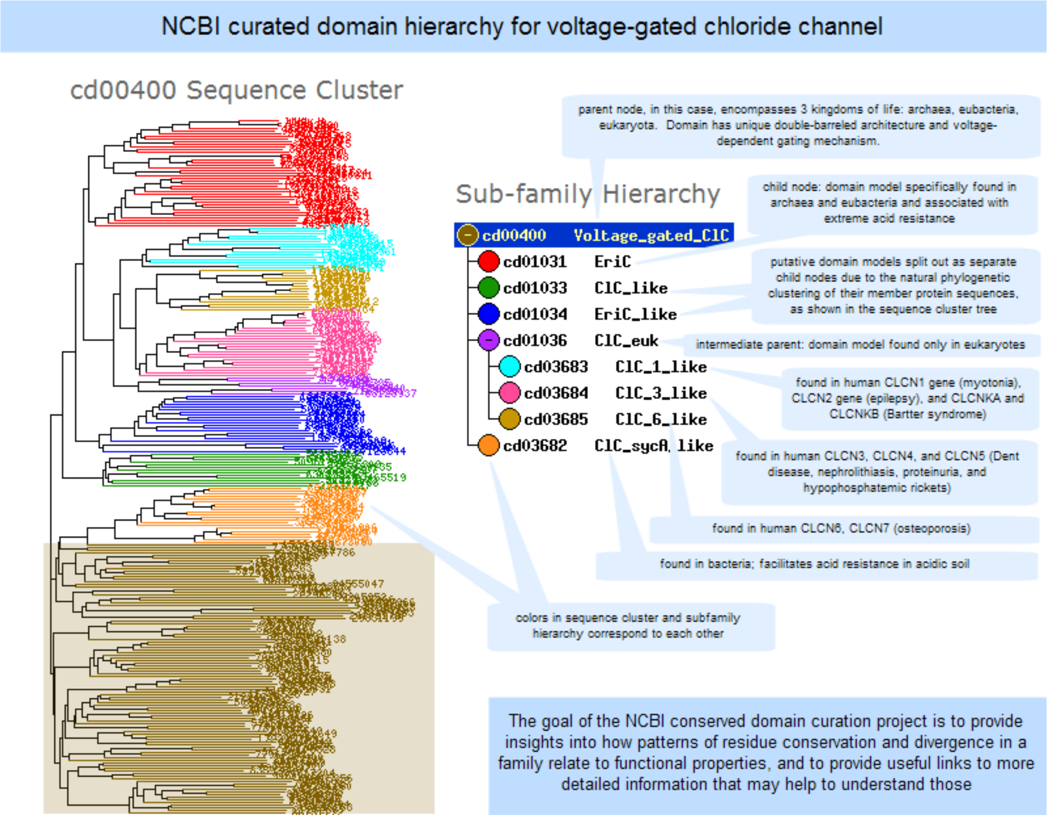
Subcellular targetting

Catalytic activity

Conserved domain sequences give us insights into protein function and history.

How do we characterize proteins through homology?




NCBI curated domain hierarchy for voltage-gated chloride channel

parent node, in this case, encompasses 3 kingdoms of life: archaea, eubacteria, eukaryota. Domain has unique double-barreled architecture and voltage-dependent gating mechanism.

child node: domain model specifically found in archaea and eubacteria and associated with extreme acid resistance

putative domain models split out as separate child nodes due to the natural phylogenetic clustering of their member protein sequences, as shown in the sequence cluster tree

intermediate parent: domain model found only in eukaryotes

found in human CLCN1 gene (myotonia), CLCN2 gene (epilepsy), and CLCNKA and CLCNKB (Bartter syndrome)

found in human CLCN3, CLCN4, and CLCN5 (Dent disease, nephrolithiasis, proteinuria, and hypophosphatemic rickets)

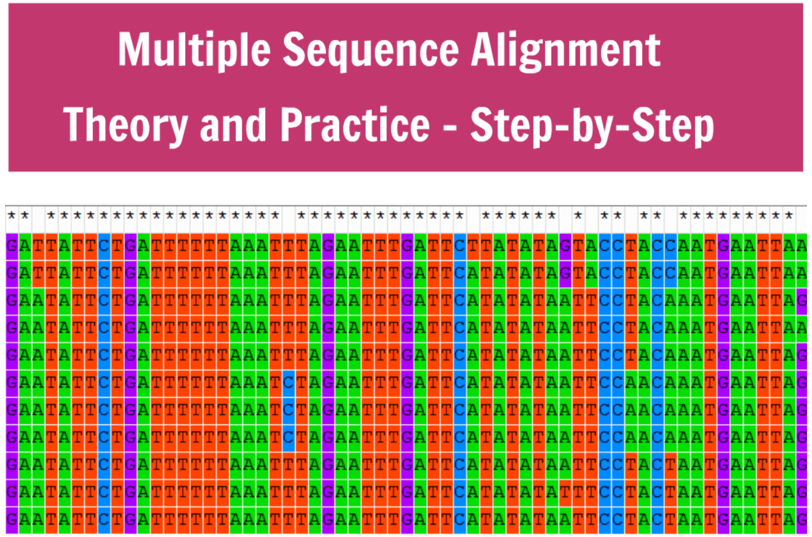
found in human CLCN6, CLCN7 (osteoporosis)

found in bacteria; facilitates acid resistance in acidic soil

colors in sequence cluster and subfamily hierarchy correspond to each other

The goal of the NCBI conserved domain curation project is to provide insights into how patterns of residue conservation and divergence in a family relate to functional properties, and to provide useful links to more detailed information that may help to understand those

How can we identify conserved domains?

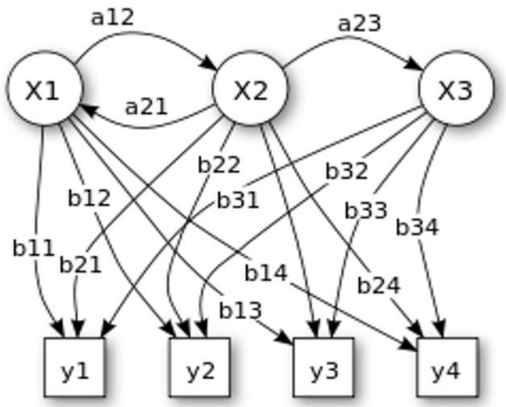


Multiple Sequence Alignment

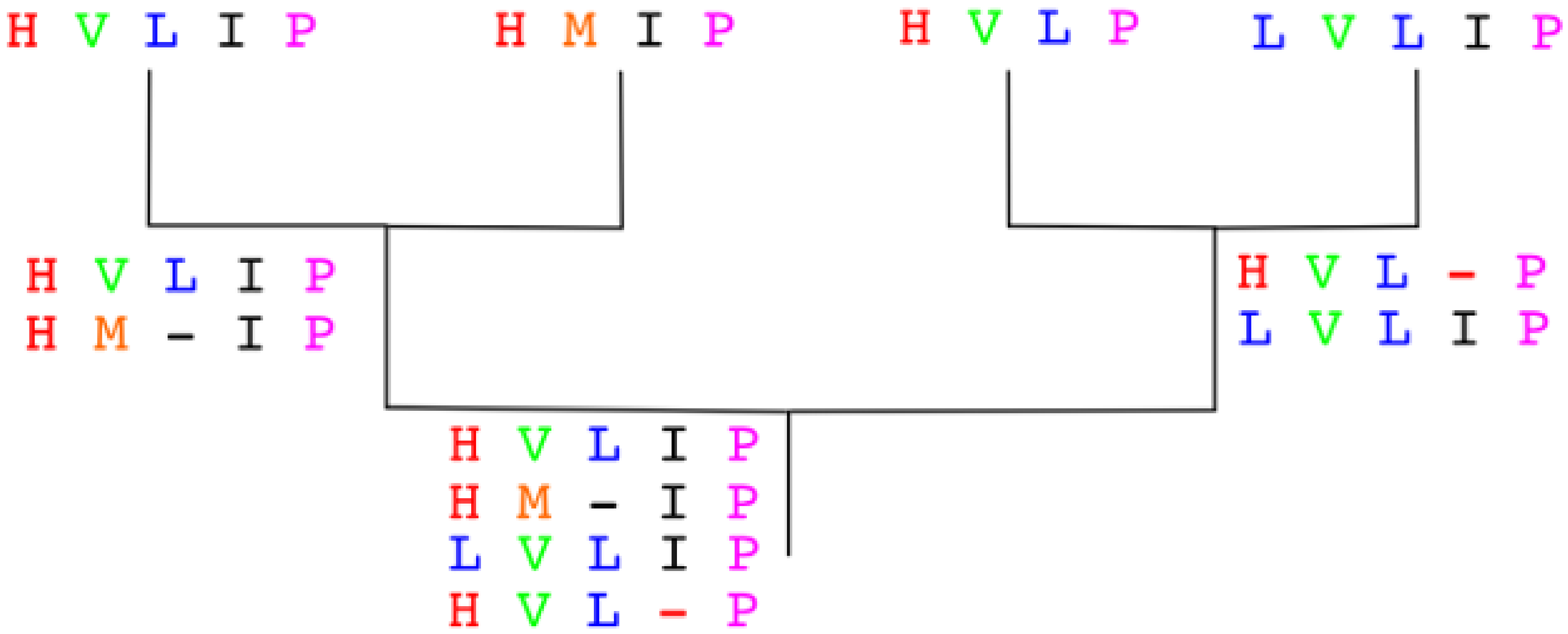
Theory and Practice - Step-by-Step

```
** ****************** ************* ****** * ** ** ********* 
GATTATTCTGATTTTTTAAATTTAGAATTTGATTCTTATATAGTACCTACCAATGAATTAA
GATTATTCTGATTTTTTAAATTTAGAATTTGATTCATATATAGTACCTACCAATGAATTAA
GAATATTCTGATTTTTTAAATTTAGAATTTGATTCATATATAATTCCTACAAATGAATTAG
GAATATTCTGATTTTTTAAATTTAGAATTTGATTCATATATAATTCCTACAAATGAATTAA
GAATATTCTGATTTTTTAAATTTAGAATTTGATTCATATATAATTCCTACAAATGAATTAG
GAATATTCTGATTTTTTAAATCTAGAATTTGATTCATATATAATTCCAACAAATGAATTAG
GAATATTCTGATTTTTTAAATCTAGAATTTGATTCATATATAATTCCAACAAATGAATTAG
GAATATTCTGATTTTTTAAATCTAGAATTTGATTCATATATAATTCCAACAAATGAATTAG
GAATATTCTGATTTTTTAAATTTAGAATTTGATTCATATATAATTCCTACTAATGAATTAG
GAATATTCTGATTTTTTAAATTTAGAATTTGATTCATATATATTTCCTACTAATGAATTAG
GAATATTCTGATTTTTTAAATTTAGAATTTGATTCATATATAATTCCTACTAATGAATTAG
```

Hidden Markov Model



What is MUSCLE?



H V L I P

H M I P

H V L P

L V L I P

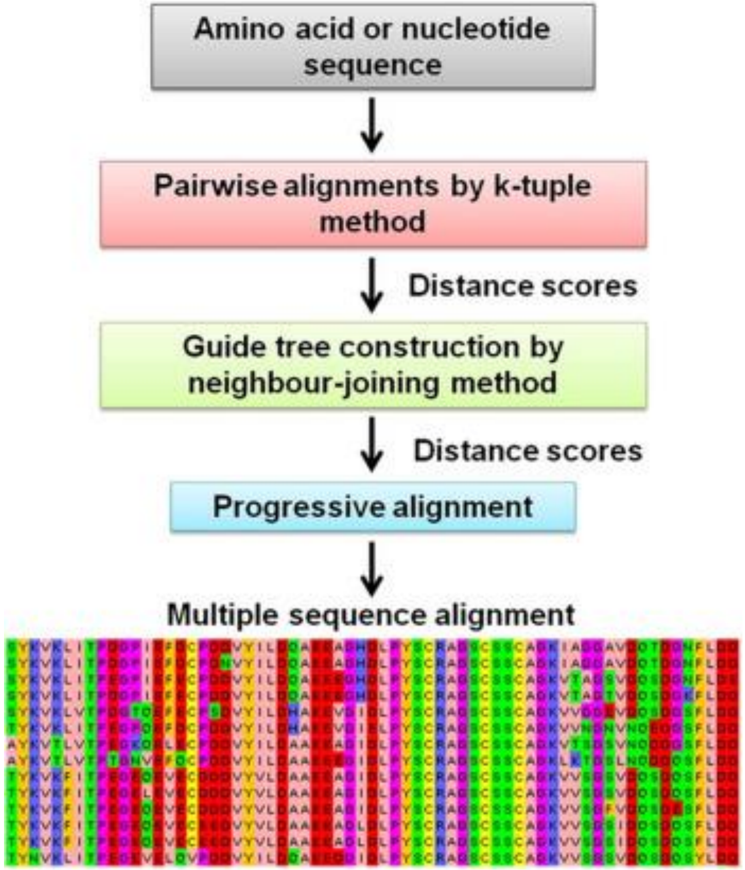
H V L I P
H M - I P

H V L - P
L V L I P

H V L I P
H M - I P
L V L I P
H V L - P

Multiple Sequence Alignment using Multiple Sequence Comparison Log-Expectation

How does MUSCLE work?



Amino acid or nucleotide sequence

Pairwise alignments by k-tuple method

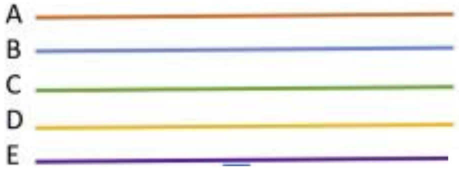
Distance scores

Guide tree construction by neighbour-joining method

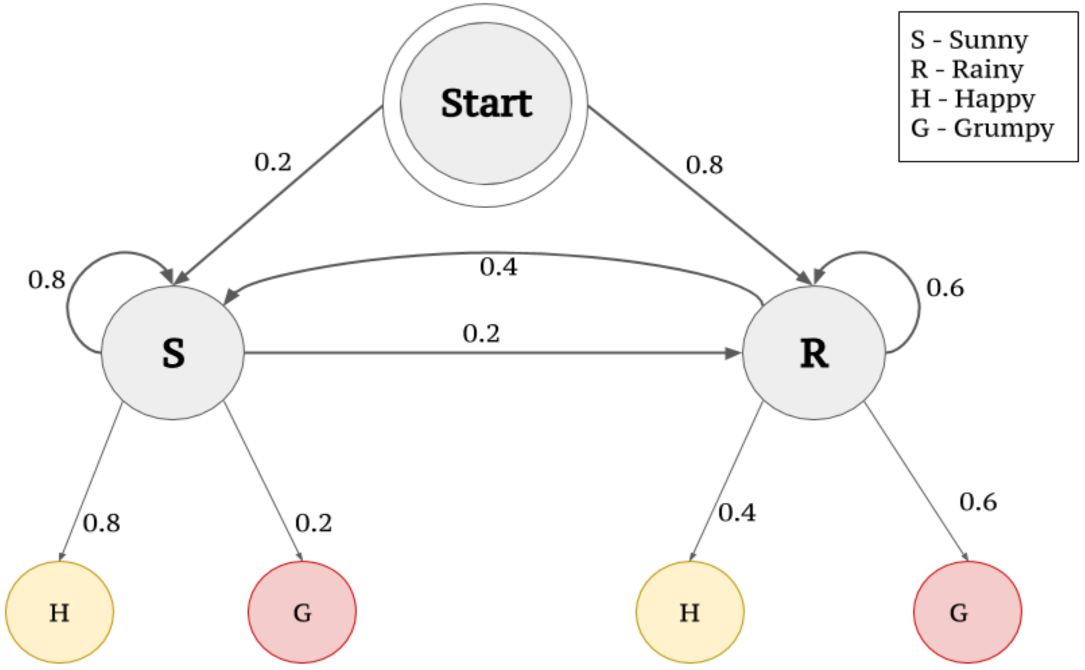
Distance scores

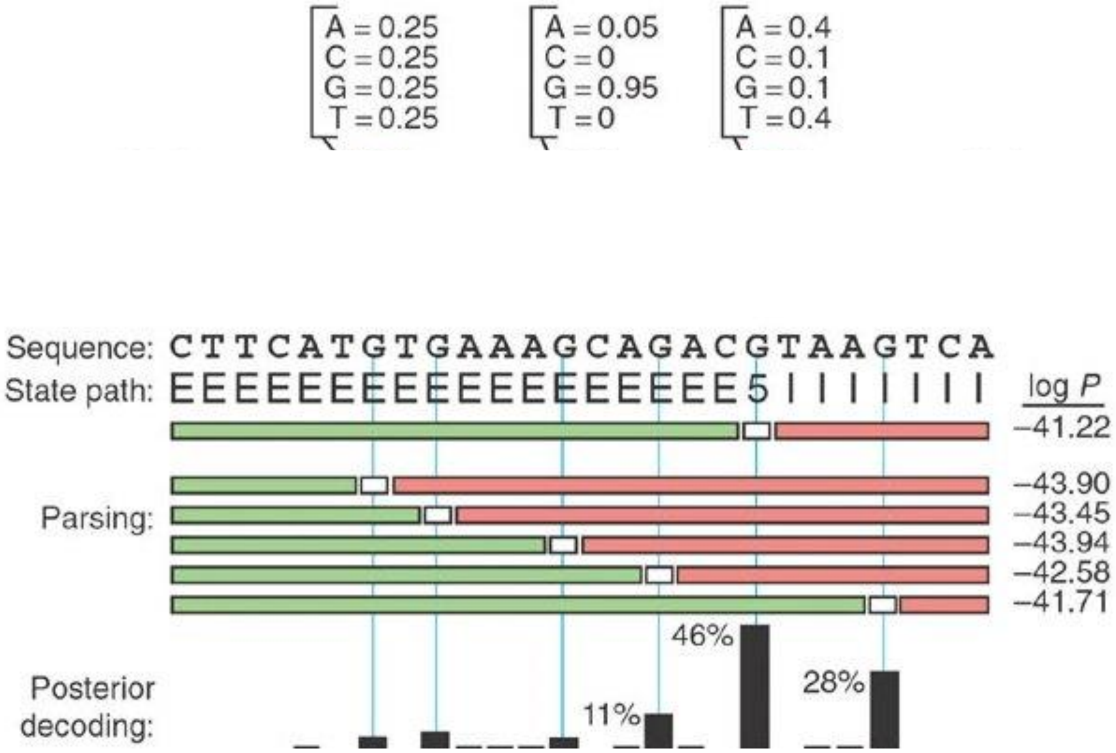
Progressive alignment

Multiple sequence alignment



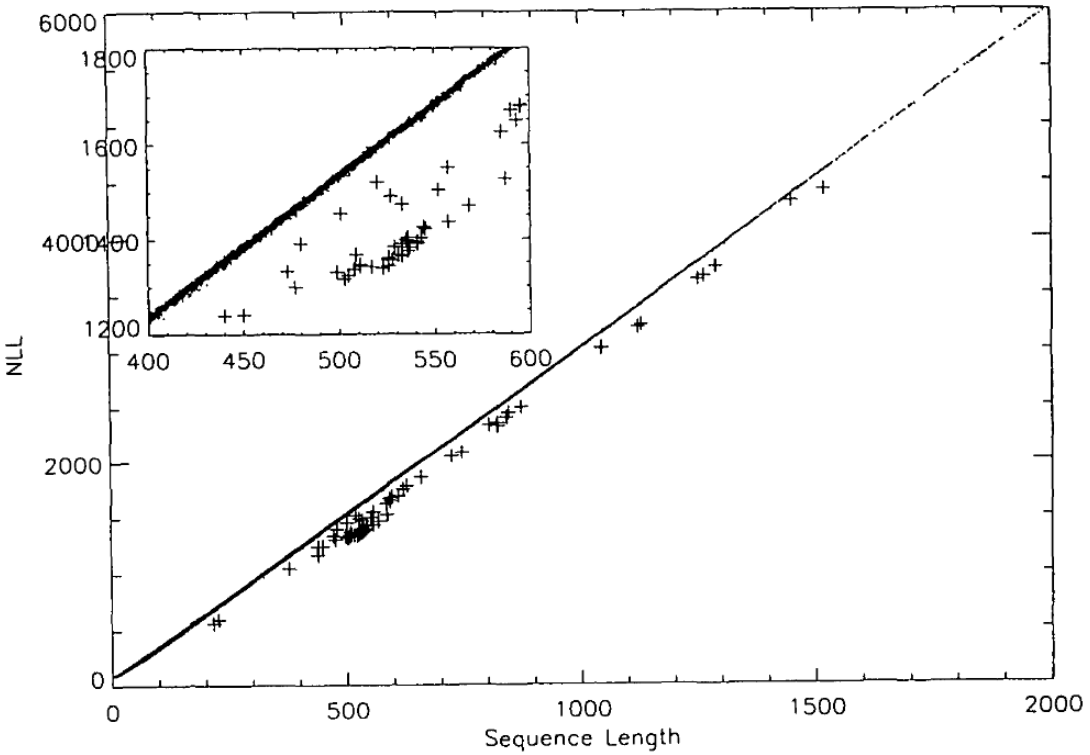
What is HMM and how is it used?

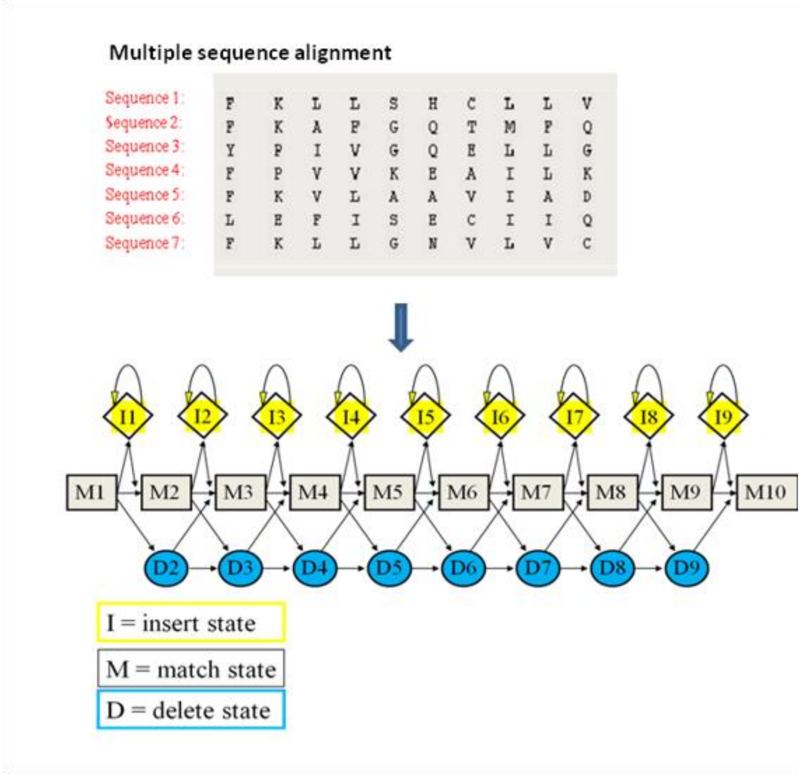




Hidden Markov Model

How is HMM utilized?





Multiple sequence alignment

Sequence 1:	F	K	L	L	S	H	C	L	L	V
Sequence 2:	F	K	A	F	G	Q	T	M	F	Q
Sequence 3:	Y	P	I	V	G	Q	E	L	L	G
Sequence 4:	F	P	V	V	K	E	A	I	L	K
Sequence 5:	F	K	V	L	A	A	V	I	A	D
Sequence 6:	L	E	F	I	S	E	C	I	I	Q
Sequence 7:	F	K	L	L	G	N	V	L	V	C

I = insert state

M = match state

D = delete state

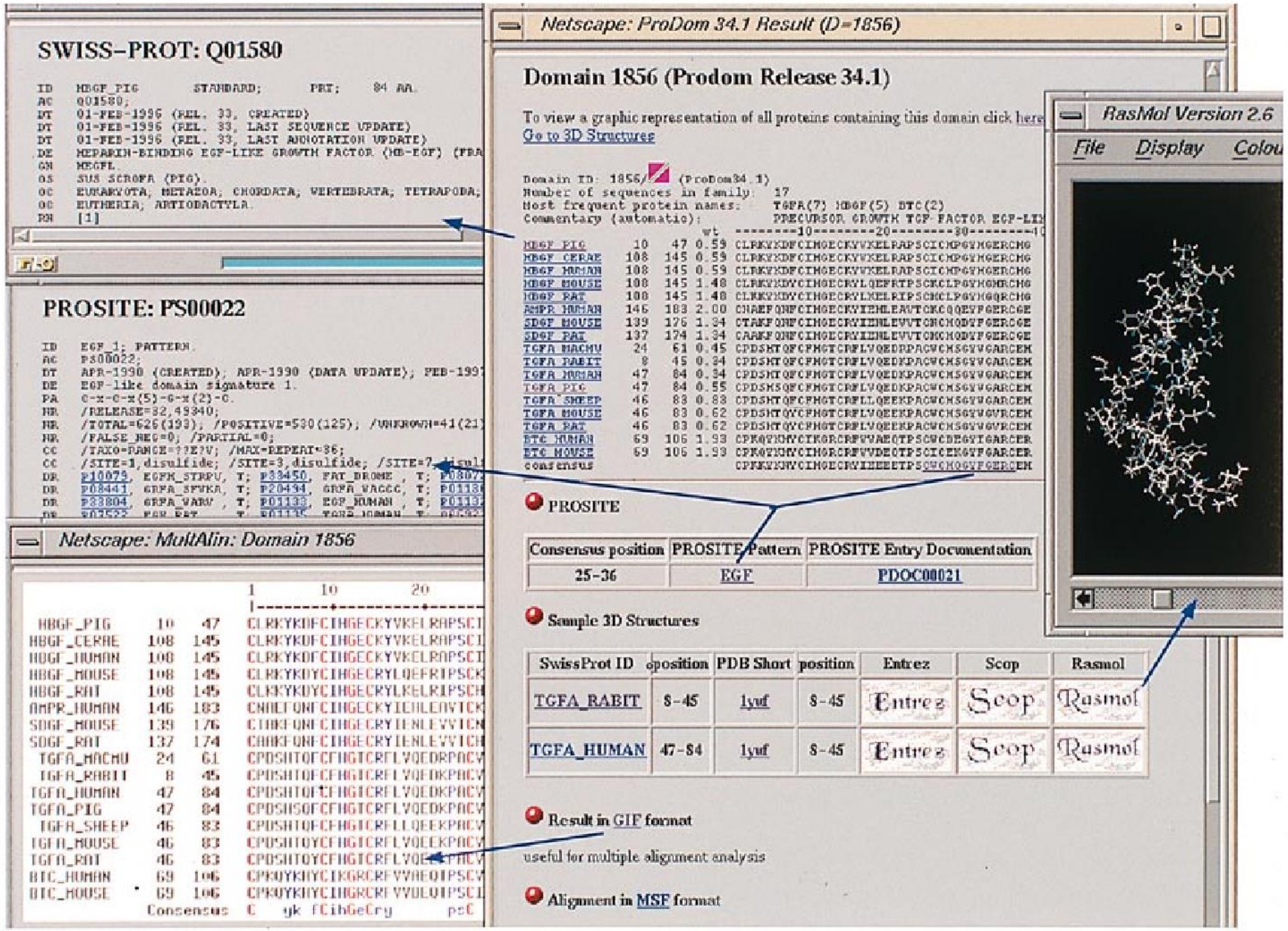
Searching through sequences and identifying protein families and generating profiles from MUSCLE.

What are Family Based Resource Groups?

Techniques that group together protein sequences or protein domains into evolutionary families.

Some utilize techniques like sequence clustering, while others use significant manual curation.

Ex. PFAM, ProDom, PANTHER, SMART, PhyloFacts



SWISS-PROT: Q01580

```
ID   HBGF_PIG        STANDARD;      PRT;    84 AA.
AC   Q01580;
DT   01-FEB-1996 (REL. 33, CREATED)
DT   01-FEB-1996 (REL. 33, LAST SEQUENCE UPDATE)
DT   01-FEB-1996 (REL. 33, LAST ANNOTATION UPDATE)
DE   HEPARIN-BINDING EGF-LIKE GROWTH FACTOR (HB-EGF) (FRA
GN   HEGFL.
OS   SUS SCROFA (PIG).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA;
OC   EUTHERIA; ARTIODACTYLA.
RN   [1]
```

PROSITE: PS00022

```
ID   EGF_1; PATTERN.
AC   PS00022;
DT   APR-1990 (CREATED); APR-1990 (DATA UPDATE); FEB-1997
DE   EGF-like domain signature 1.
PA   C-x-C-x(5)-G-x(2)-C.
NR   /RELEASE=32,49340;
NR   /TOTAL=626(193); /POSITIVE=530(125); /UNKNOWN=41(21)
NR   /FALSE_NEG=0; /PARTIAL=0;
CC   /TAXO-RANGE=??E?V; /MAX-REPEAT=36;
CC   /SITE=1,disulfide; /SITE=3,disulfide; /SITE=7,disulf
DR   P10079, EGFH_STRPU, T; P33450, FAT_DROME , T; P0807
DR   P08441, GRFA_SFVKA, T; P20494, GRFA_VACCC, T; P0113
DR   P33804, GRFA_VARV , T; P01133, EGF_HUMAN , T; P0113
```

Netscape: MultAlin: Domain 1856

```
                            1         10        20
                            |--------+---------+-----
 HBGF_PIG       10   47     CLRKYKDFCIHGECKYVKELRAPSCI
HBGF_CERAE     108  145     CLRKYKDFCIHGECKYVKELRAPSCI
HBGF_HUMAN     108  145     CLRKYKDFCIHGECKYVKELRAPSCI
HBGF_MOUSE     108  145     CLRKYKDYCIHGECRYLQEFRIPSCK
HBGF_RAT       108  145     CLKKYKDYCIHGECRYLKELRIPSCH
AMPR_HUMAN     146  183     CNAEFQNFCIHGECKYIEHLEAVTCK
SDGF_MOUSE     139  176     CTAKFQNFCIHGECRYIENLEVVTCN
SDGF_RAT       137  174     CAAKFQNFCIHGECRYIENLEVVTCH
 TGFA_MACMU     24   61     CPDSHTQFCFHGTCRFLVQEDRPACV
 TGFA_RABIT      8   45     CPDSHTQFCFHGTCRFLVQEDKPACV
TGFA_HUMAN      47   84     CPDSHTQFCFHGTCRFLVQEDKPACV
TGFA_PIG        47   84     CPDSHSQFCFHGTCRFLVQEDKPACV
 TGFA_SHEEP     46   83     CPDSHTQFCFHGTCRFLLQEEKPACV
TGFA_MOUSE      46   83     CPDSHTQYCFHGTCRFLVQEEKPACV
TGFA_RAT        46   83     CPDSHTQYCFHGTCRFLVQEEKPACV
BTC_HUMAN       69  106     CPKQYKHYCIKGRCRFVVAEQIPSCV
BTC_MOUSE       69  106     CPKQYKHYCIHGRCRFVVDEQIPSCI
               Consensus    C   yk fCihGeCry        psC
```

Netscape: ProDom 34.1 Result (D=1856)

Domain 1856 (Prodom Release 34.1)

To view a graphic representation of all proteins containing this domain click here
Go to 3D Structures

```
Domain ID: 1856/    (ProDom34.1)
Number of sequences in family:   17
Most frequent protein names:     TGFA(7) HBGF(5) BTC(2)
Commentary (automatic):          PRECURSOR GROWTH TGF FACTOR EGF-LIK
                          wt  ---------10---------20---------30---------4
HBGF_PIG      10   47  0.59  CLRKYKDFCIHGECKYVKELRAPSCICHPGYHGERCHG
HBGF_CERAE   108  145  0.59  CLRKYKDFCIHGECKYVKELRAPSCICHPGYHGERCHG
HBGF_HUMAN   108  145  0.59  CLRKYKDFCIHGECKYVKELRAPSCICHPGYHGERCHG
HBGF_MOUSE   108  145  1.48  CLRKYKDYCIHGECRYLQEFRTPSCKCLPGYHGHRCHG
HBGF_RAT     108  145  1.48  CLKKYKDYCIHGECRYLKELRIPSCHCLPGYHGQRCHG
AMPR_HUMAN   146  183  2.00  CNAEFQNFCIHGECKYIEHLEAVTCKCQQEYFGERCGE
SDGF_MOUSE   139  176  1.34  CTAKFQNFCIHGECRYIENLEVVTCHCHQDYFGERCGE
SDGF_RAT     137  174  1.34  CAAKFQNFCIHGECRYIENLEVVTCHCHQDYFGERCGE
TGFA_MACMU    24   61  0.45  CPDSHTQFCFHGTCRFLVQEDRPACVCHSGYVGARCEH
TGFA_RABIT     8   45  0.34  CPDSHTQFCFHGTCRFLVQEDKPACVCHSGYVGARCEH
TGFA_HUMAN    47   84  0.34  CPDSHTQFCFHGTCRFLVQEDKPACVCHSGYVGARCEH
TGFA_PIG      47   84  0.55  CPDSHSQFCFHGTCRFLVQEDKPACVCHSGYVGARCEH
TGFA_SHEEP    46   83  0.83  CPDSHTQFCFHGTCRFLLQEEKPACVCHSGYVGARCEH
TGFA_MOUSE    46   83  0.62  CPDSHTQYCFHGTCRFLVQEEKPACVCHSGYVGVRCEH
TGFA_RAT      46   83  0.62  CPDSHTQYCFHGTCRFLVQEEKPACVCHSGYVGVRCEH
BTC_HUMAN     69  106  1.93  CPKQYKHYCIKGRCRFVVAEQTPSCVCDEGYIGARCER
BTC_MOUSE     69  106  1.93  CPKQYKHYCIHGRCRFVVDEQTPSCICEKGYFGARCER
consensus                    CPKKYKHYCIHGECRYIEEEETPSCVCHQGYFGERCEH
```

● PROSITE

Consensus position	PROSITE Pattern	PROSITE Entry Documentation
25-36	EGF	PDOC00021

● Sample 3D Structures

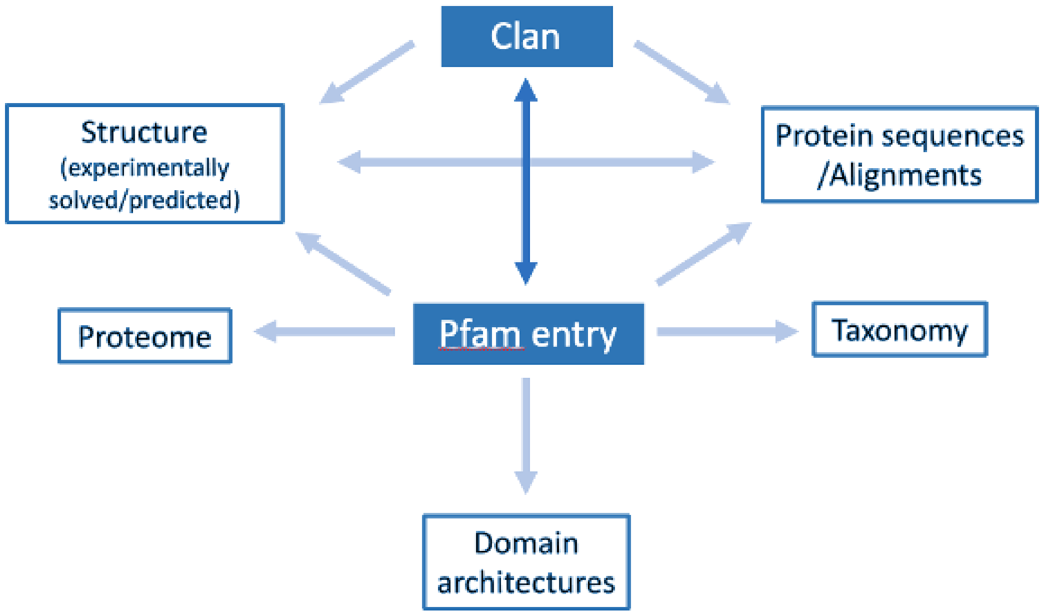
SwissProt ID	position	PDB Short	position	Entrez	Scop	Rasmol
TGFA_RABIT	8-45	1yuf	8-45	Entrez	Scop	Rasmol
TGFA_HUMAN	47-84	1yuf	8-45	Entrez	Scop	Rasmol

● Result in GIF format

useful for multiple alignment analysis

● Alignment in MSF format

What is Pfam


Clan
Structure (experimentally solved/predicted)
Protein sequences /Alignments
Proteome
Pfam entry
Taxonomy
Domain architectures

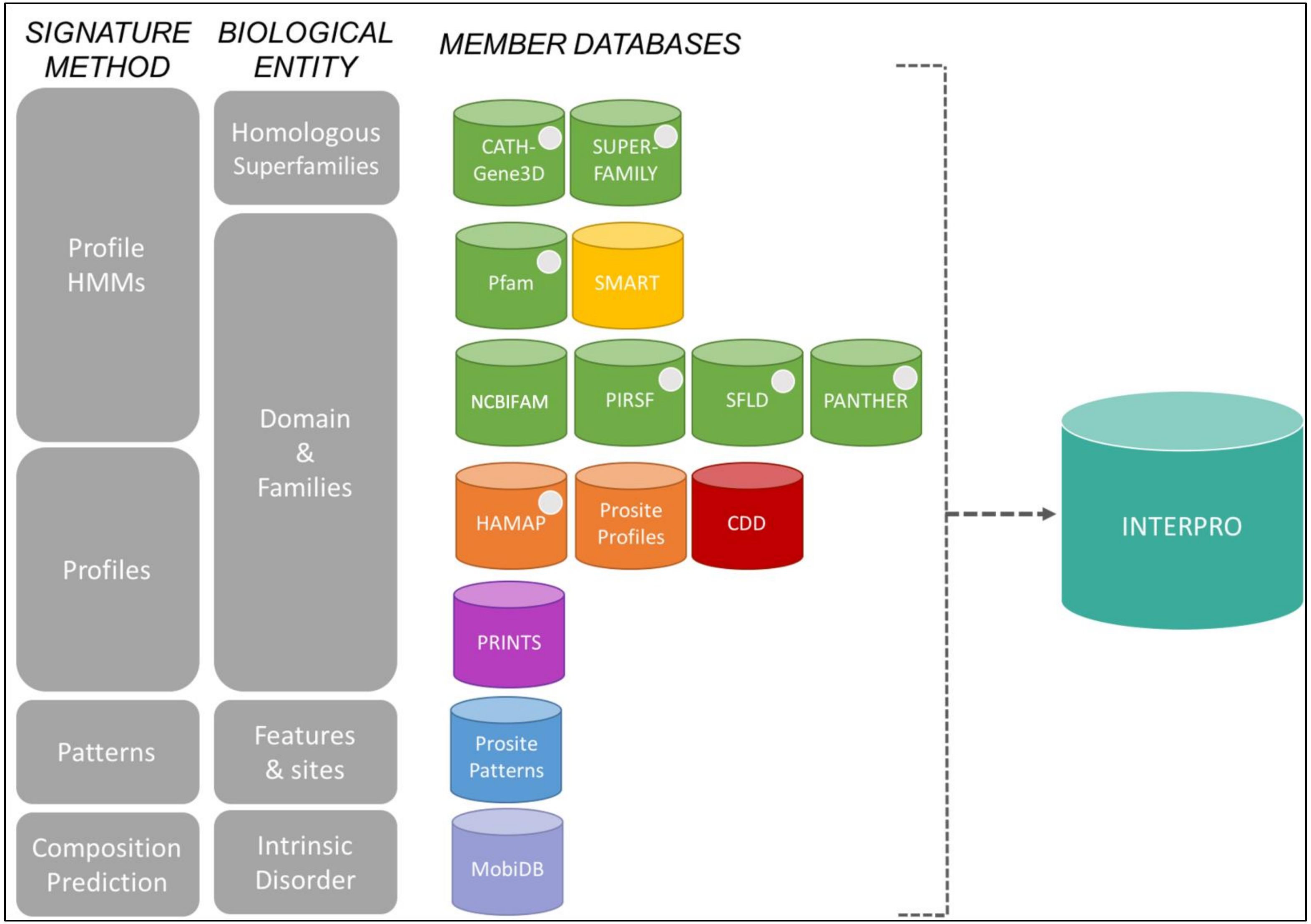
PFAM is a database that allows you to analyze proteins and find families using HMM to detect homology.

What can you do with PFAM?

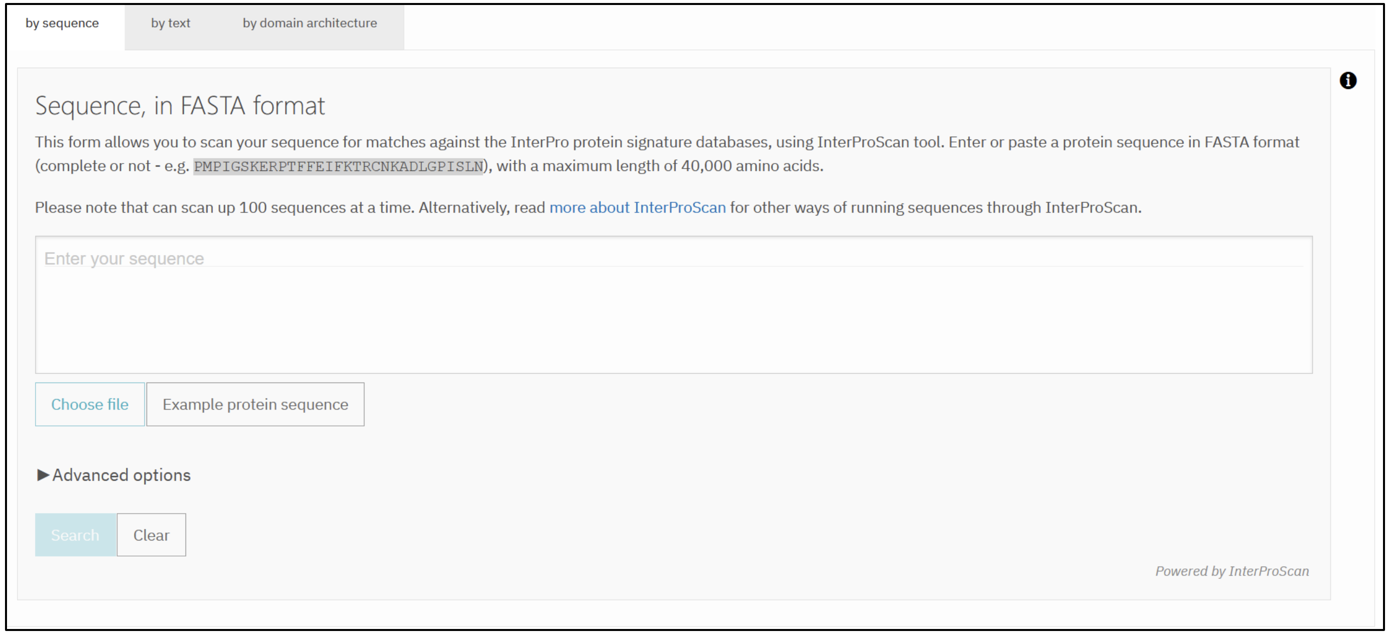
What can you do with PFAM continued?

What is InterPro?

Interpro is a large secondary protein database that incorporates many different databases



How to use Interpro?



Sequence, in FASTA format

This form allows you to scan your sequence for matches against the InterPro protein signature databases, using InterProScan tool. Enter or paste a protein sequence in FASTA format (complete or not - e.g. `PMPIGSKERPTFFEIFKTRCNKADLGPISLN`), with a maximum length of 40,000 amino acids.

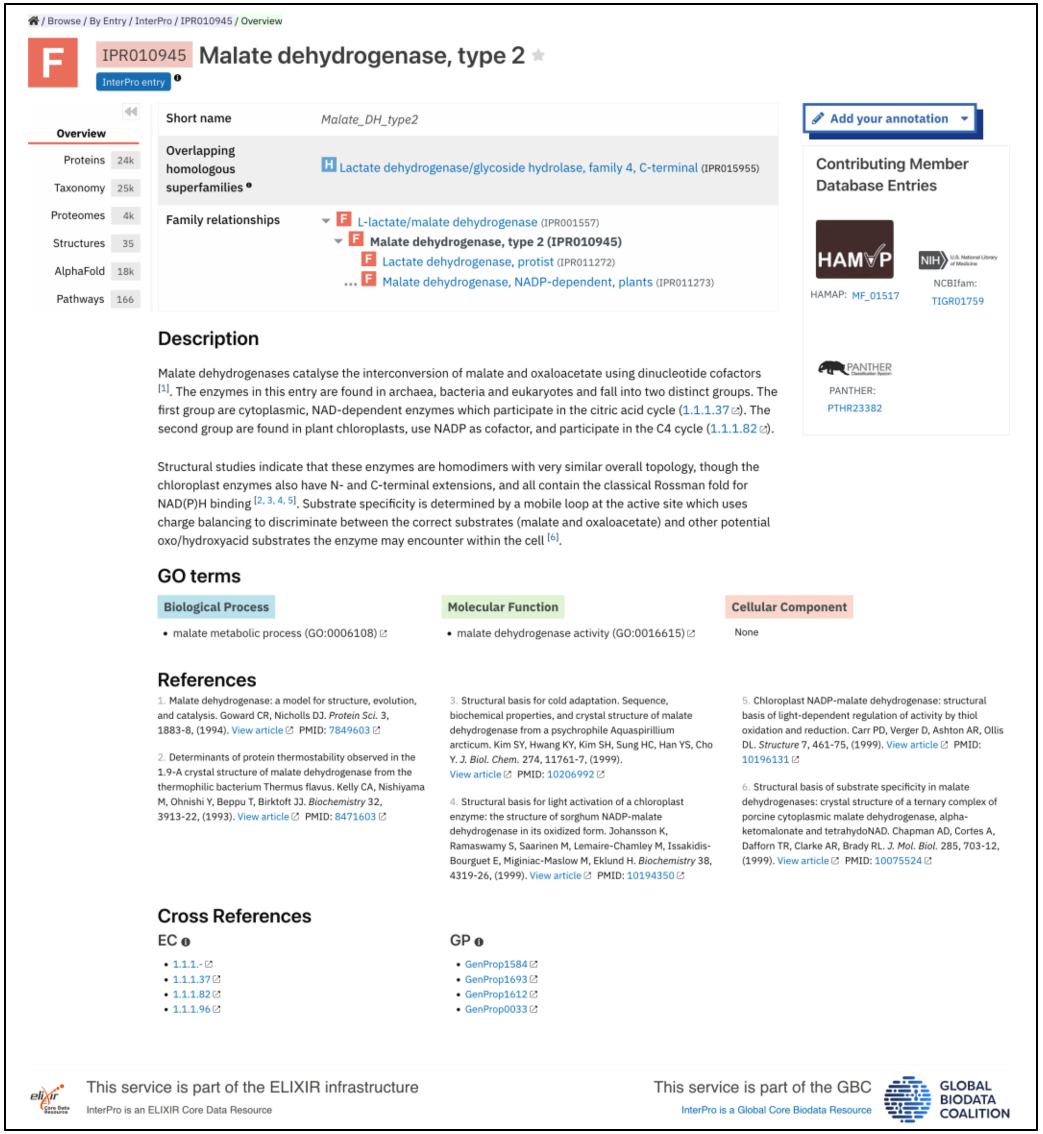
Please note that can scan up 100 sequences at a time. Alternatively, read more about InterProScan for other ways of running sequences through InterProScan.

Enter your sequence

▶Advanced options

Powered by InterProScan

Search by sequence or Ascension -> Interpro entry



IPR010945 Malate dehydrogenase, type 2

Short name: *Malate_DH_type2*

Overlapping homologous superfamilies: H Lactate dehydrogenase/glycoside hydrolase, family 4, C-terminal (IPR015955)

Family relationships:
- F L-lactate/malate dehydrogenase (IPR001557)
  - F Malate dehydrogenase, type 2 (IPR010945)
    - F Lactate dehydrogenase, protist (IPR011272)
    - … F Malate dehydrogenase, NADP-dependent, plants (IPR011273)

Contributing Member Database Entries

HAMAP: MF_01517

NCBIfam: TIGR01759

PANTHER: PTHR23382

Description

Malate dehydrogenases catalyse the interconversion of malate and oxaloacetate using dinucleotide cofactors [1]. The enzymes in this entry are found in archaea, bacteria and eukaryotes and fall into two distinct groups. The first group are cytoplasmic, NAD-dependent enzymes which participate in the citric acid cycle (1.1.1.37). The second group are found in plant chloroplasts, use NADP as cofactor, and participate in the C4 cycle (1.1.1.82).

Structural studies indicate that these enzymes are homodimers with very similar overall topology, though the chloroplast enzymes also have N- and C-terminal extensions, and all contain the classical Rossman fold for NAD(P)H binding [2, 3, 4, 5]. Substrate specificity is determined by a mobile loop at the active site which uses charge balancing to discriminate between the correct substrates (malate and oxaloacetate) and other potential oxo/hydroxyacid substrates the enzyme may encounter within the cell [6].

GO terms

Biological Process
- malate metabolic process (GO:0006108)

Molecular Function
- malate dehydrogenase activity (GO:0016615)

Cellular Component

None

References

1. Malate dehydrogenase: a model for structure, evolution, and catalysis. Goward CR, Nicholls DJ. *Protein Sci.* 3, 1883-8, (1994). View article PMID: 7849603

2. Determinants of protein thermostability observed in the 1.9-A crystal structure of malate dehydrogenase from the thermophilic bacterium Thermus flavus. Kelly CA, Nishiyama M, Ohnishi Y, Beppu T, Birktoft JJ. *Biochemistry* 32, 3913-22, (1993). View article PMID: 8471603

3. Structural basis for cold adaptation. Sequence, biochemical properties, and crystal structure of malate dehydrogenase from a psychrophile Aquaspirillium arcticum. Kim SY, Hwang KY, Kim SH, Sung HC, Han YS, Cho Y. *J. Biol. Chem.* 274, 11761-7, (1999). View article PMID: 10206992

4. Structural basis for light activation of a chloroplast enzyme: the structure of sorghum NADP-malate dehydrogenase in its oxidized form. Johansson K, Ramaswamy S, Saarinen M, Lemaire-Chamley M, Issakidis-Bourguet E, Miginiac-Maslow M, Eklund H. *Biochemistry* 38, 4319-26, (1999). View article PMID: 10194350

5. Chloroplast NADP-malate dehydrogenase: structural basis of light-dependent regulation of activity by thiol oxidation and reduction. Carr PD, Verger D, Ashton AR, Ollis DL. *Structure* 7, 461-75, (1999). View article PMID: 10196131

6. Structural basis of substrate specificity in malate dehydrogenases: crystal structure of a ternary complex of porcine cytoplasmic malate dehydrogenase, alpha-ketomalonate and tetrahydoNAD. Chapman AD, Cortes A, Dafforn TR, Clarke AR, Brady RL. *J. Mol. Biol.* 285, 703-12, (1999). View article PMID: 10075524

Cross References

EC
- 1.1.1.-
- 1.1.1.37
- 1.1.1.82
- 1.1.1.96

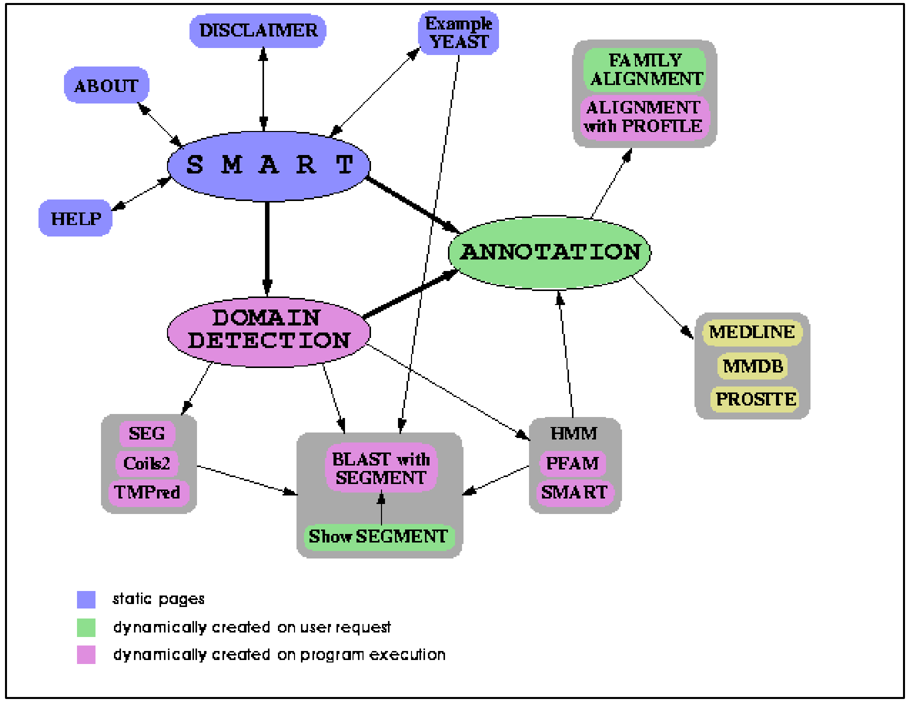
GP
- GenProp1584
- GenProp1693
- GenProp1612
- GenProp0033

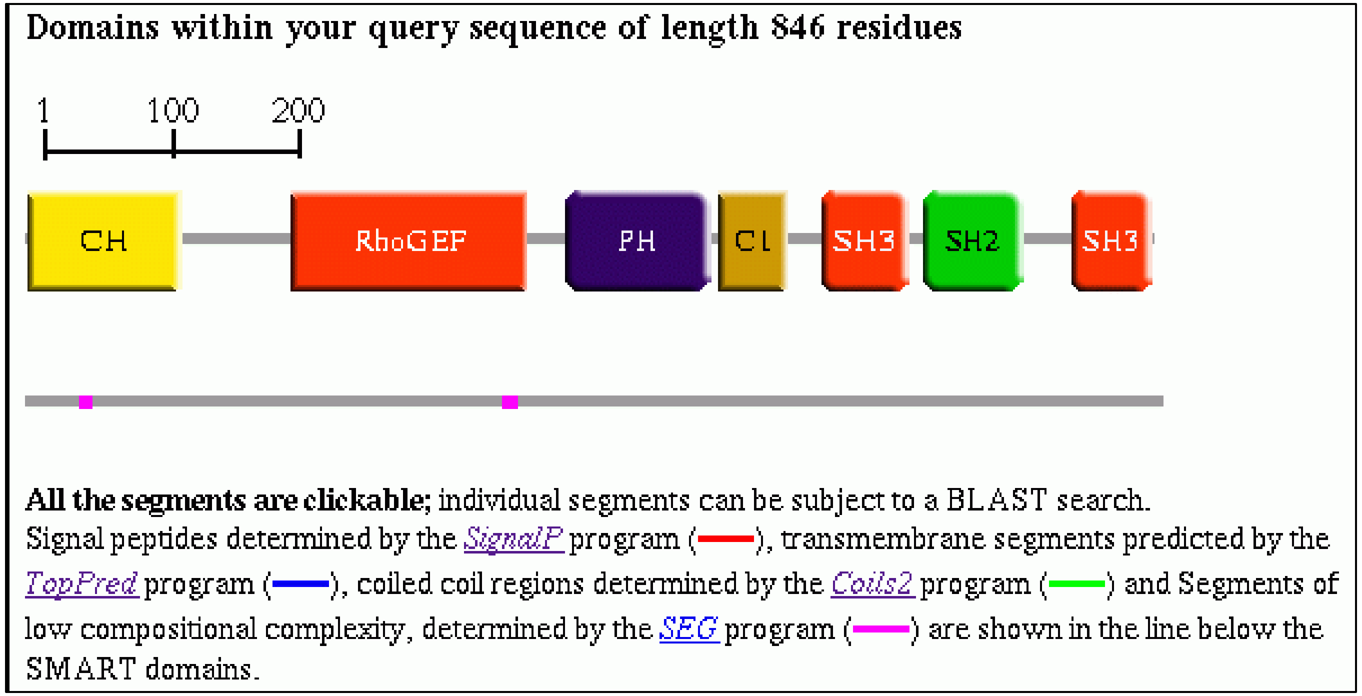
This service is part of the ELIXIR infrastructure
InterPro is an ELIXIR Core Data Resource

This service is part of the GBC
InterPro is a Global Core Biodata Resource

GLOBAL BIODATA COALITION

How can SMART be used?




Domains within your query sequence of length 846 residues

All the segments are clickable; individual segments can be subject to a BLAST search. Signal peptides determined by the *SignalP* program (🟥), transmembrane segments predicted by the *TopPred* program (🟦), coiled coil regions determined by the *Coils2* program (🟩) and Segments of low compositional complexity, determined by the *SEG* program (🟪) are shown in the line below the SMART domains.

SMART is a database that is used in the analysis of protein domains within protein sequences.

Also uses Hidden Markov models to best sort through the database.

SMART vs PFAM- when to use what?

SMART

- Specializes in extracellular, signaling, and chromatin associated domains
- Significant manual curation (>1200 manually curated models)
- Exclusive annotation

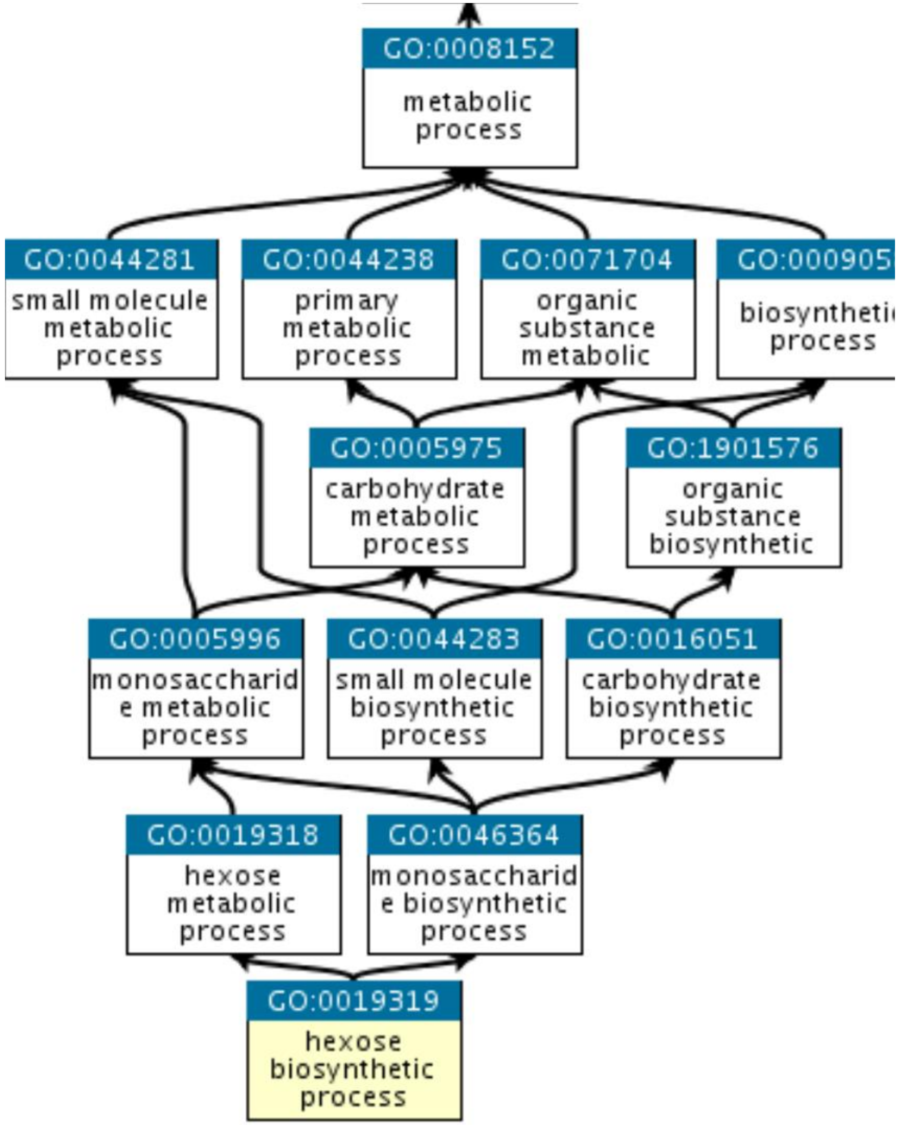
PFAM

- Many more proteins in the database
- Novel sequences are classified into families
- No longer exists (incorporated into InterPro)

What is Gene Ontology?

Ontology: A set of well-defined terms with well-defined relationships

Gene ontology is a way of categorizing and organizing data to facilitate data retrieval.

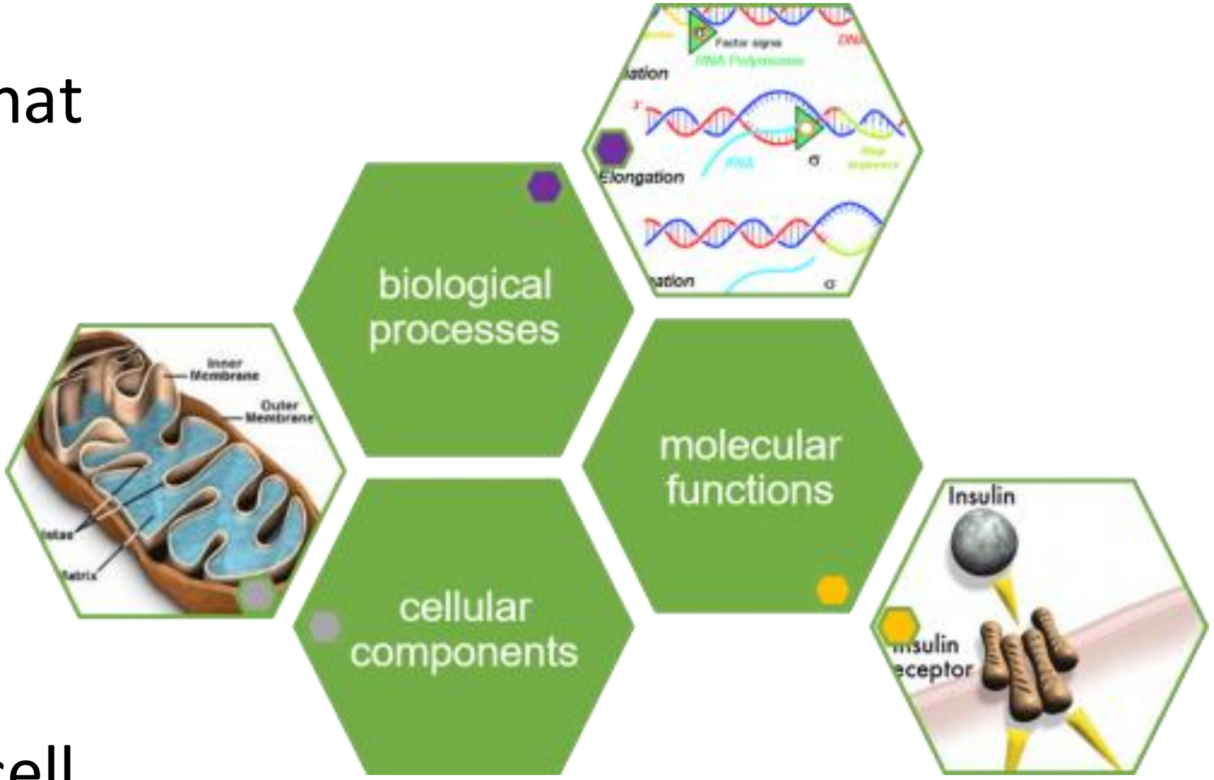

GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:000905
biosyntheti
process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharide metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0019318
hexose metabolic process
GO:0046364
monosaccharide biosynthetic process
GO:0019319
hexose biosynthetic process

Categories of Gene ontology

Biological process: the objective that the gene/protein contributes to.

Molecular function: The role/biochemical activity of a gene product.

Cellular component: Place in the cell where a gene/protein is active.



Summary

1) Domains are conserved, functional units of protein

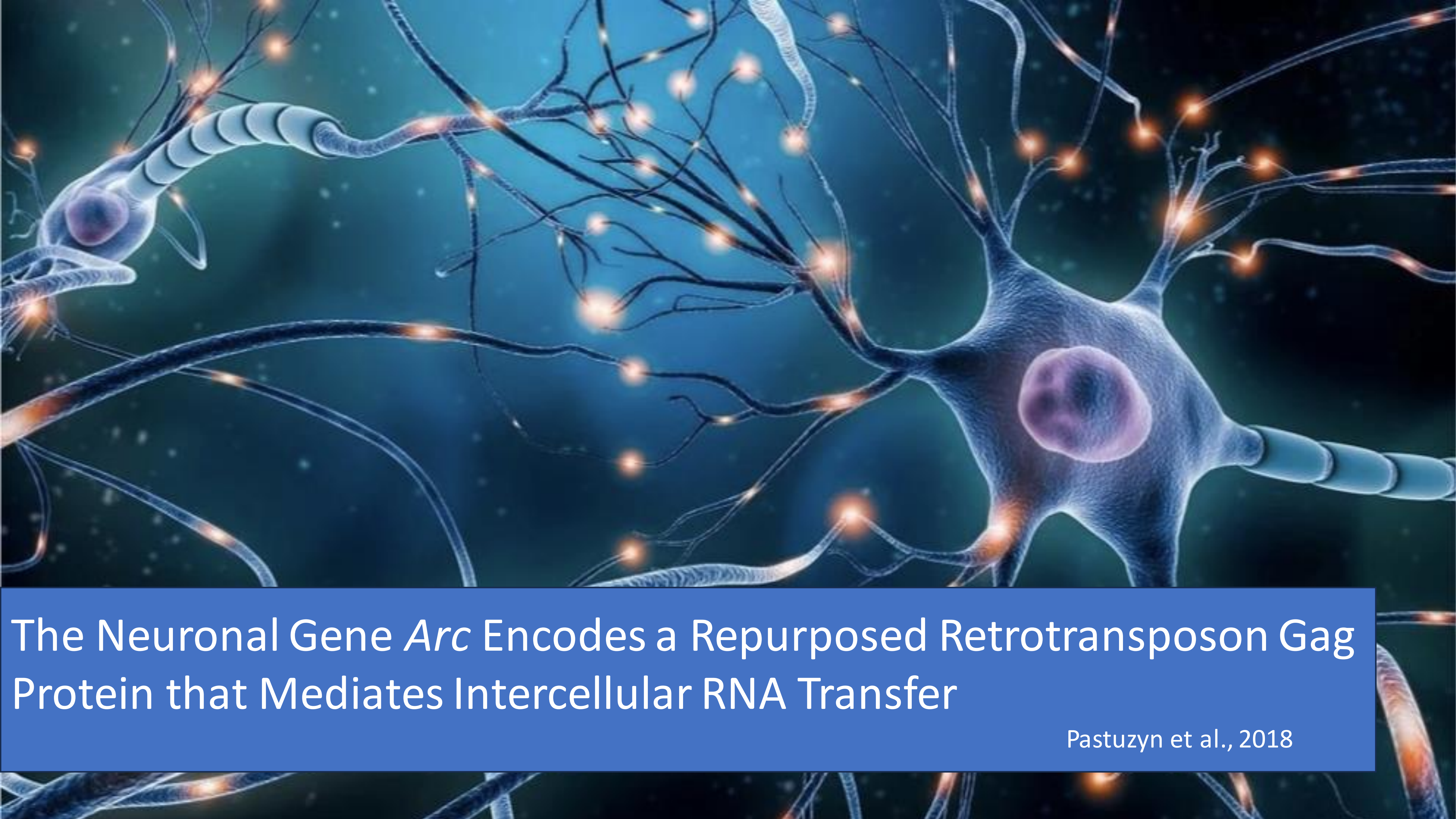
2) Domains can be uncovered and analyzed by various homology and non-homology directed methods.

3) Domain analysis allows us to infer protein function and better classify protein families.

Dr. Jason Shepherd & the Arc protein

Molecular function of Arc protein in long-term memory formation

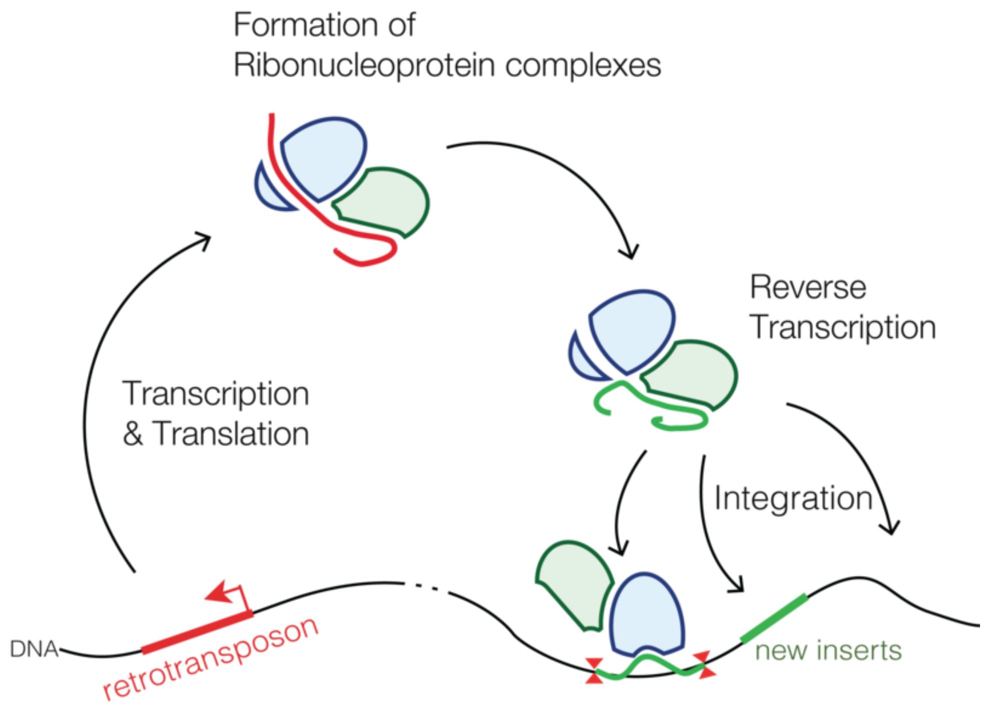
Questions?



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

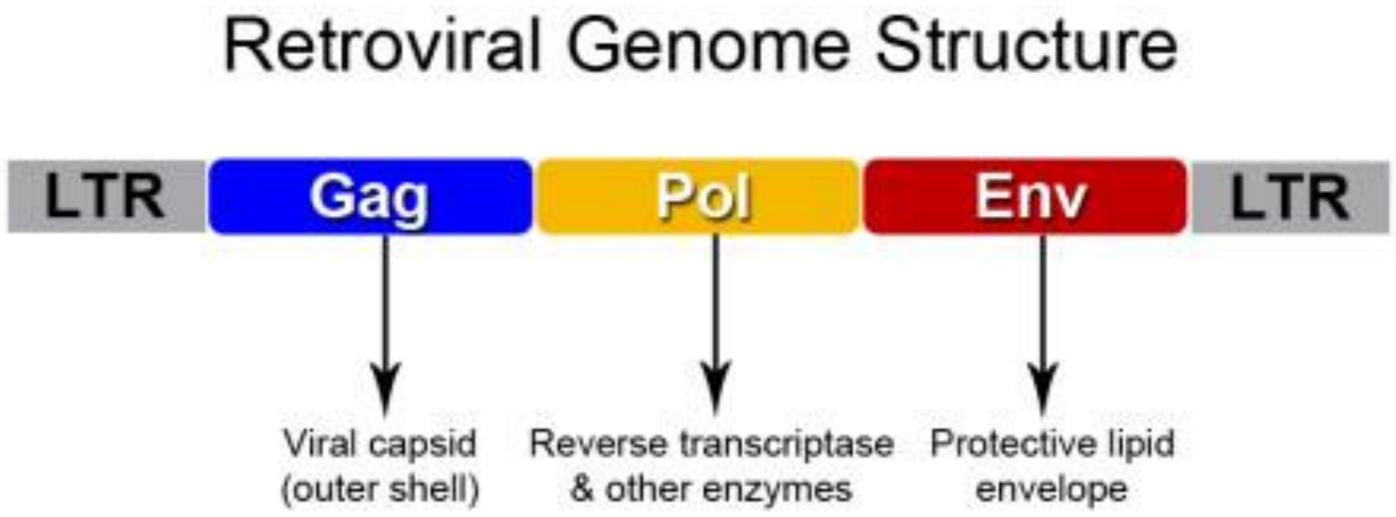
Pastuzyn et al., 2018

How is information stored in the brain?



Retrotransposons store long-lasting information along with genetic memory.

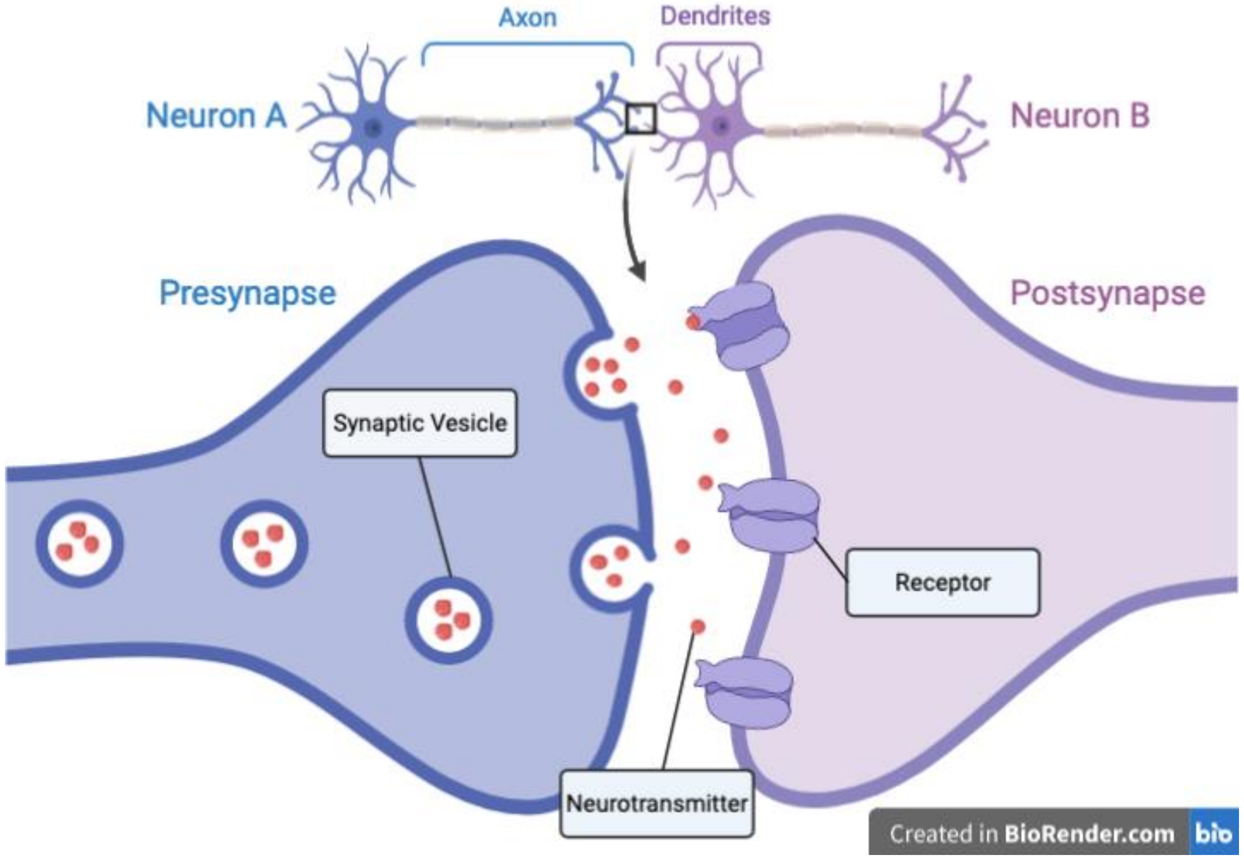
What is a retrovirus?


Retroviral Genome Structure
LTR
Gag
Pol
Env
LTR
Viral capsid (outer shell)
Reverse transcriptase & other enzymes
Protective lipid envelope

Gag and Env: form capsid shell

Pol: encodes enzymes

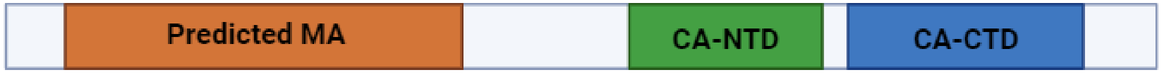
What is a synapse?



Created in BioRender.com

A synapse is the gap between neurons that allow movement of information in the brain.

What is Arc?



MA = Matrix

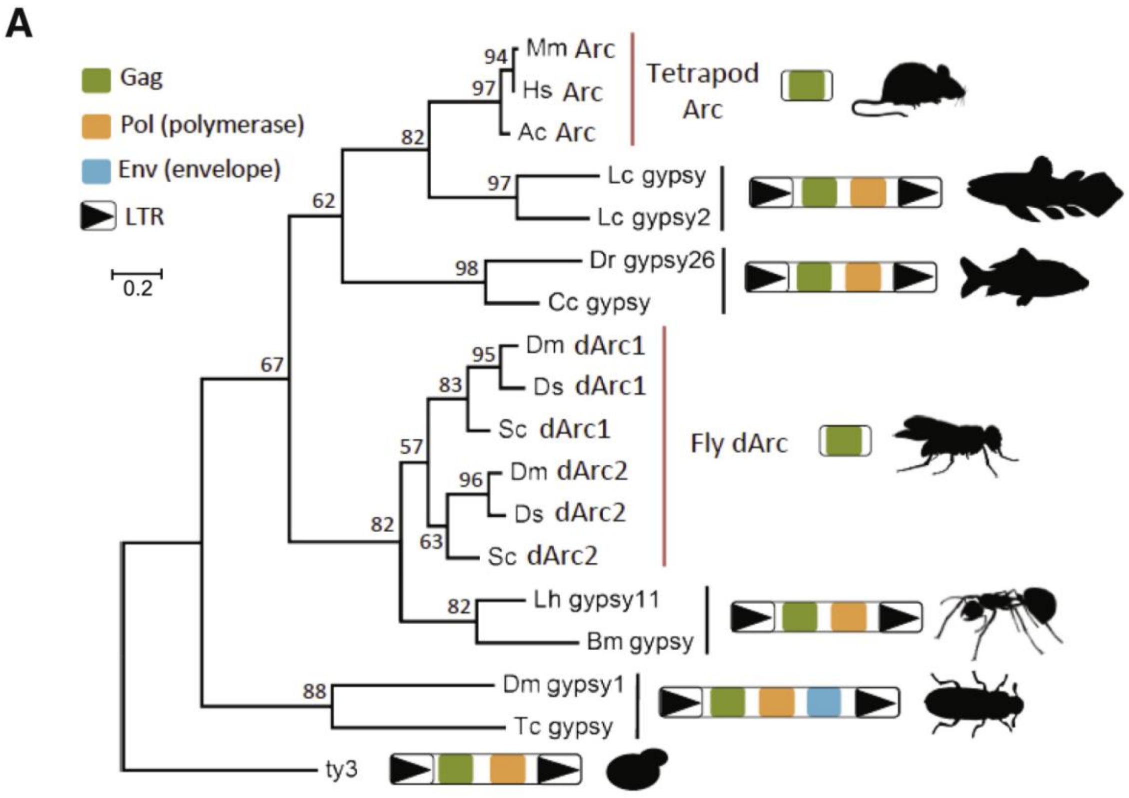
CA-NTD = Capsid N-terminal domain

CA-CTD = Capsid C-terminal domain

Created in BioRender.com bio

Arc is a neuronal gene that is important for memory and synaptic plasticity regulation.

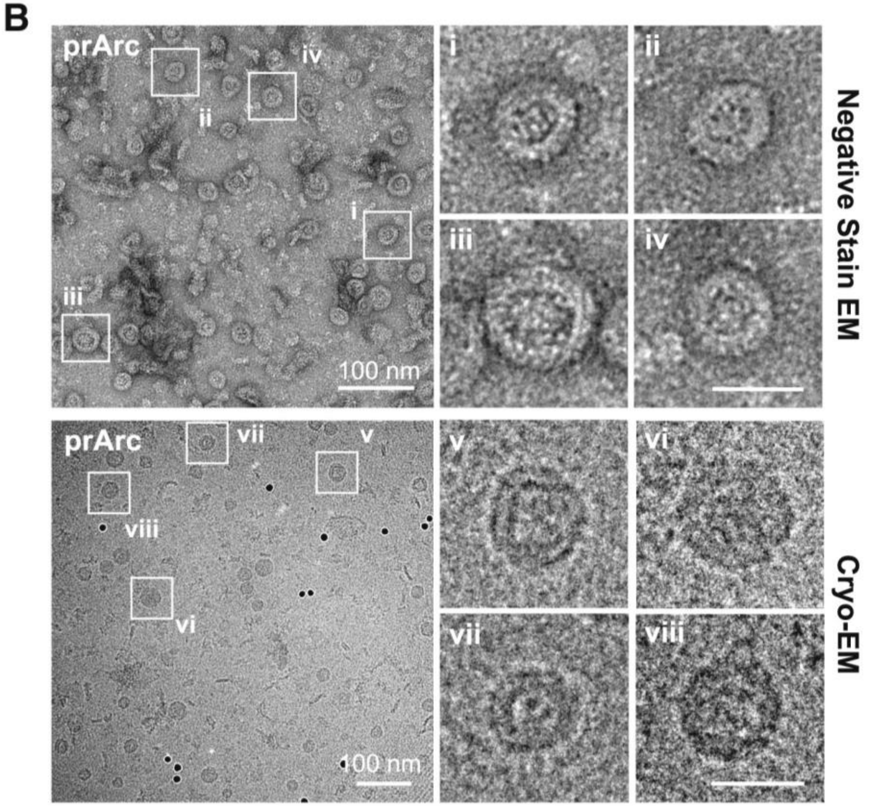
What are the evolutionary origins of Arc?



Retroviral Gag domain

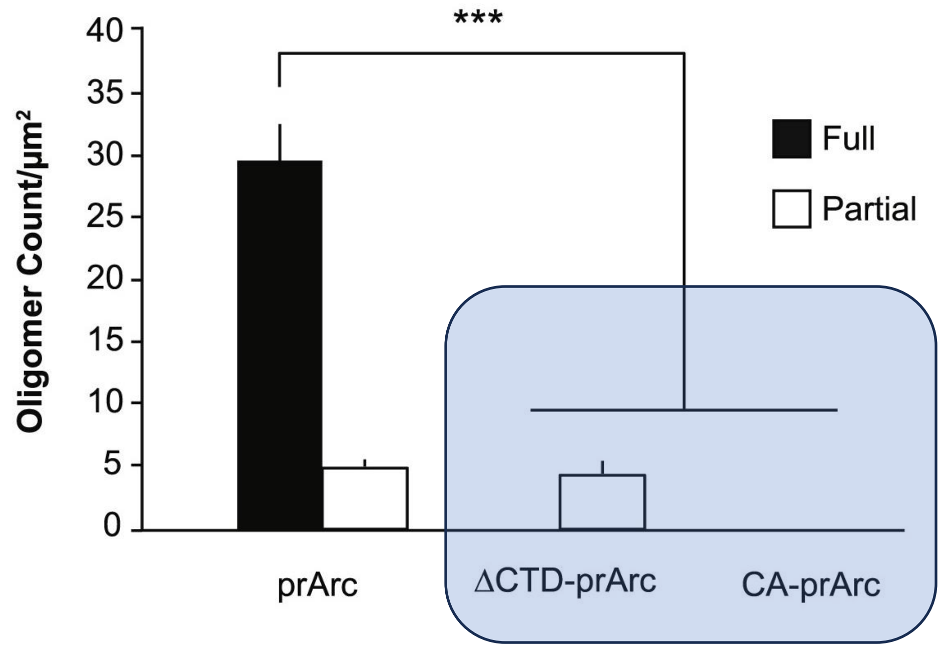
Ty3/gypsy transposons

Does the Arc protein form capsids?



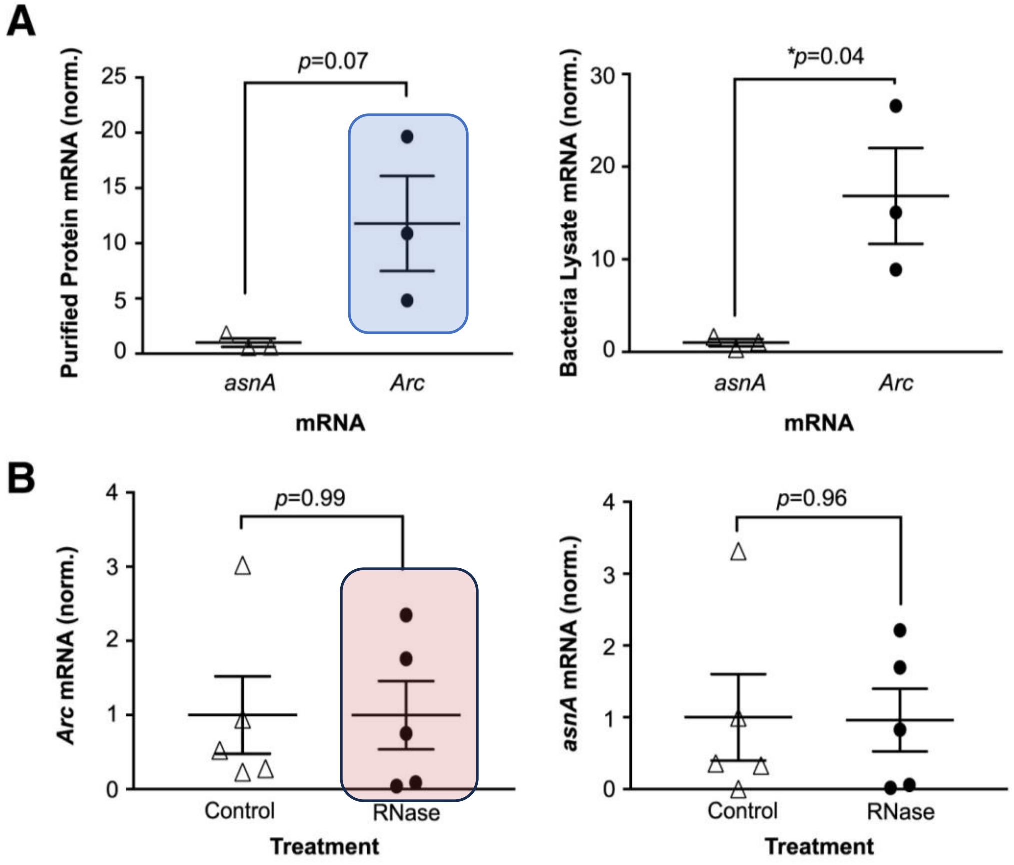
Arc proteins self-assemble into virus-like capsids.

Is CTD required for capsid formation?



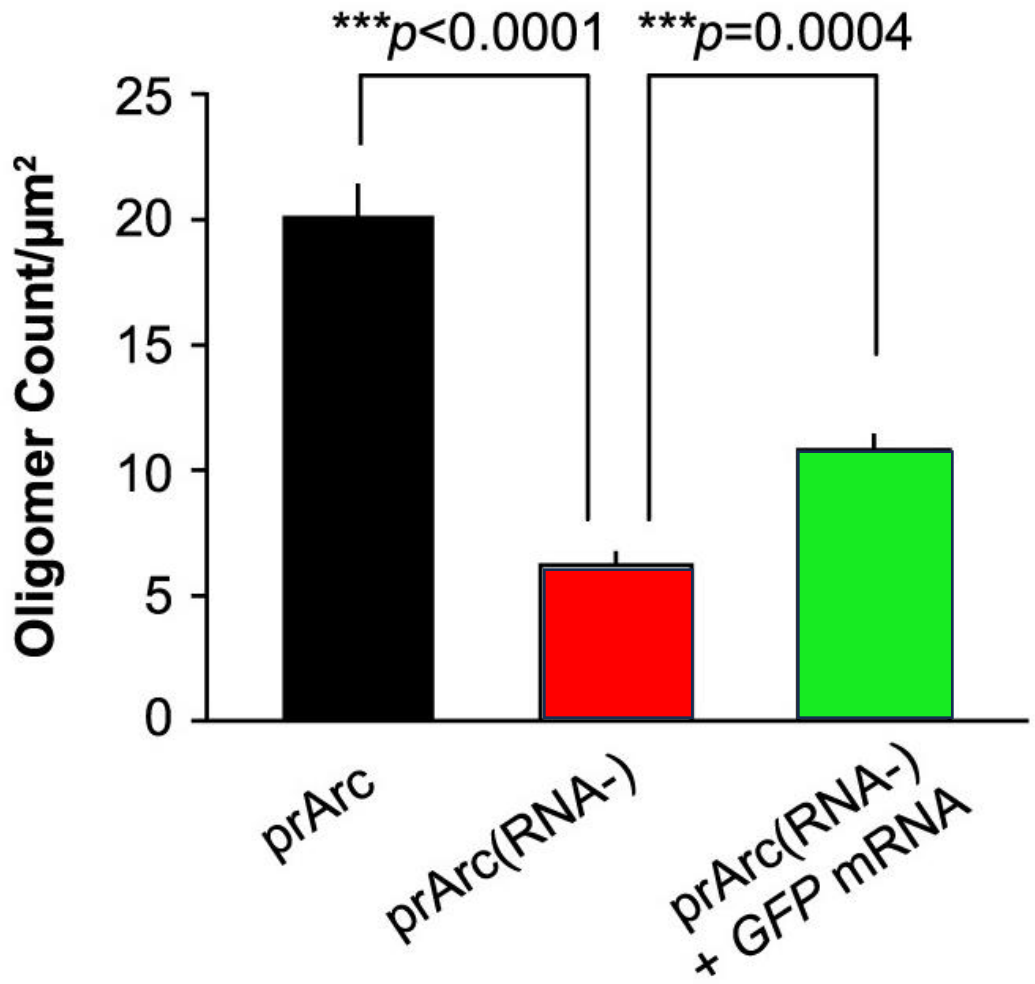
Capsids are not formed without CTD.

Does Arc bind and encapsulate mRNA?



Arc binds and encapsulates mRNA, protecting it from degradation.

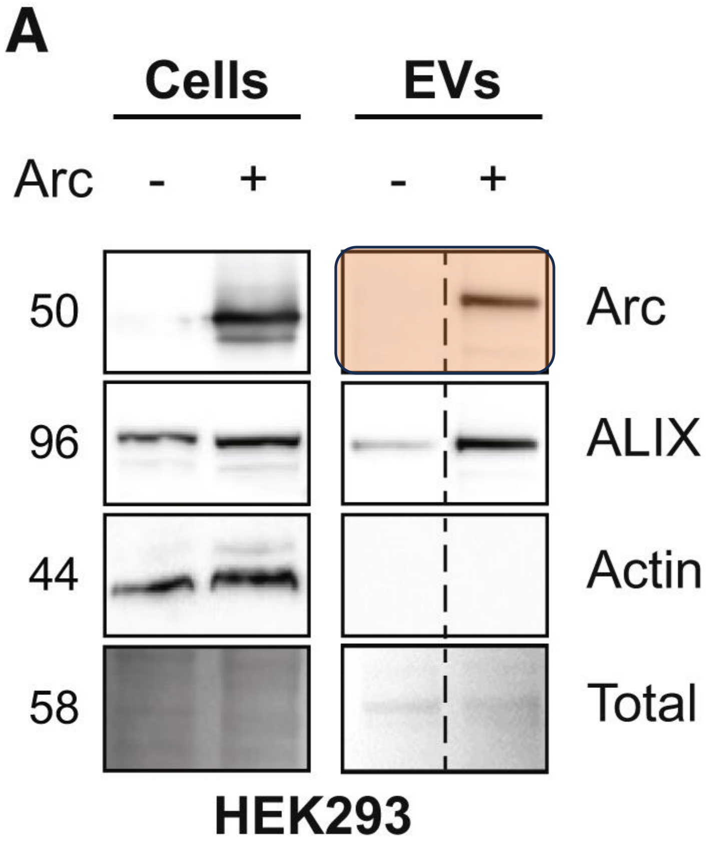
Is RNA required for proper capsid assembly?



Removing RNA bases decreased proper capsid formation.

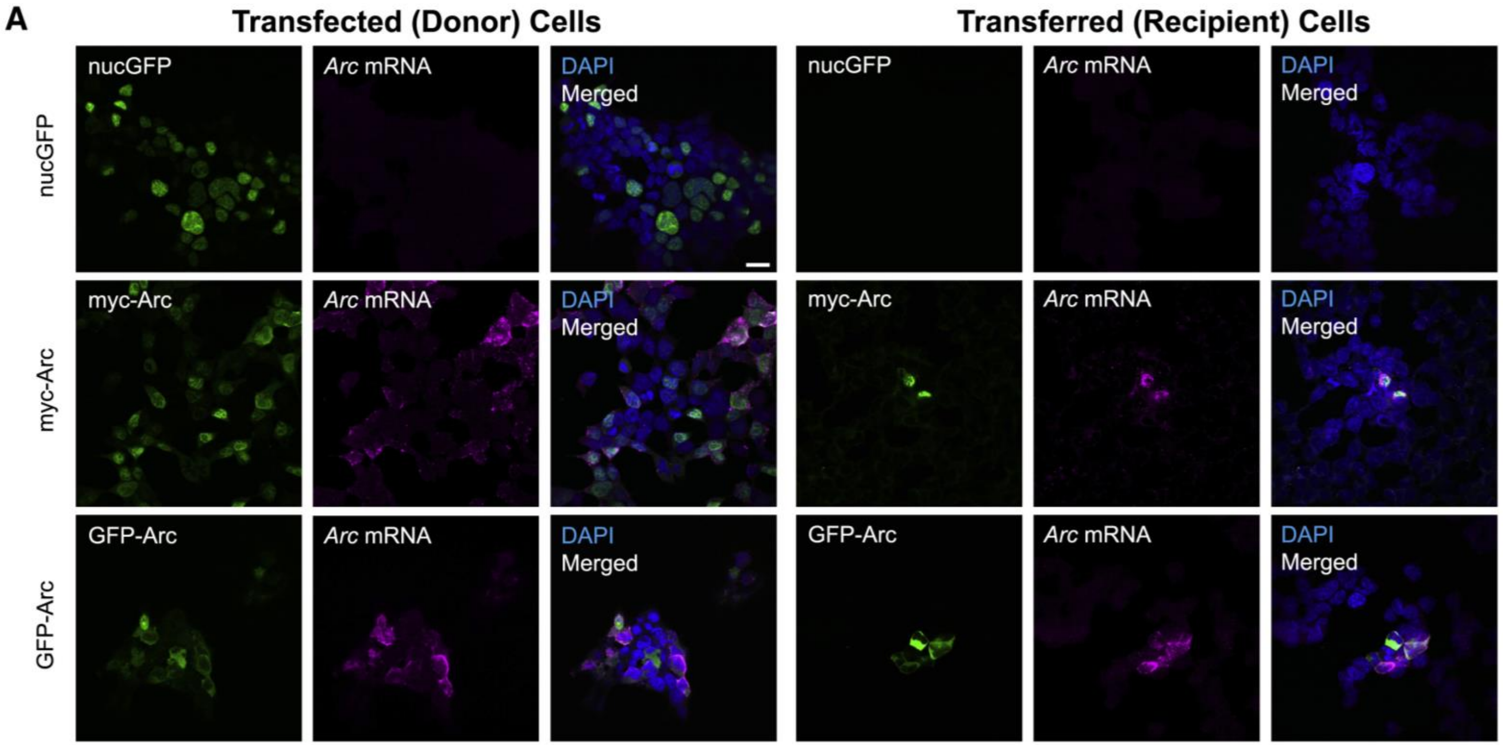
Addition of mRNA increased proper capsid formation.

Is Arc mRNA found in extracellular vesicles (EVs)?

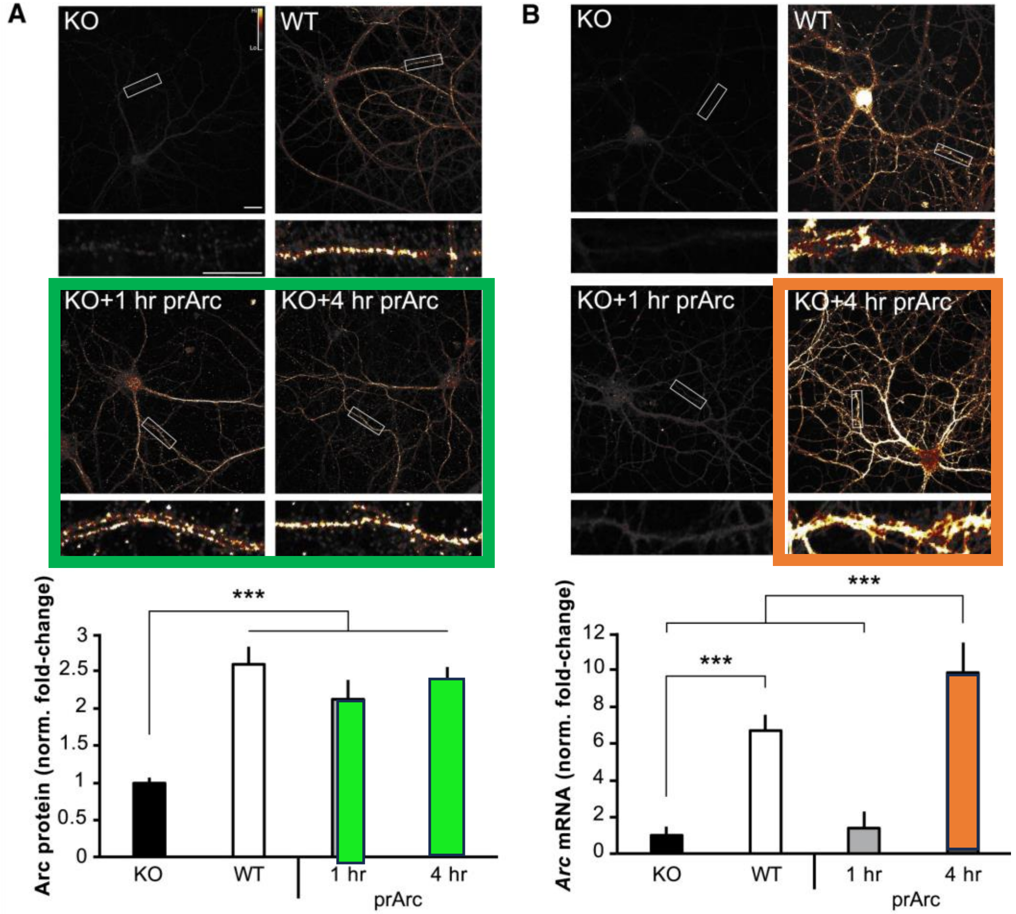


Arc mRNA is found in EVs.

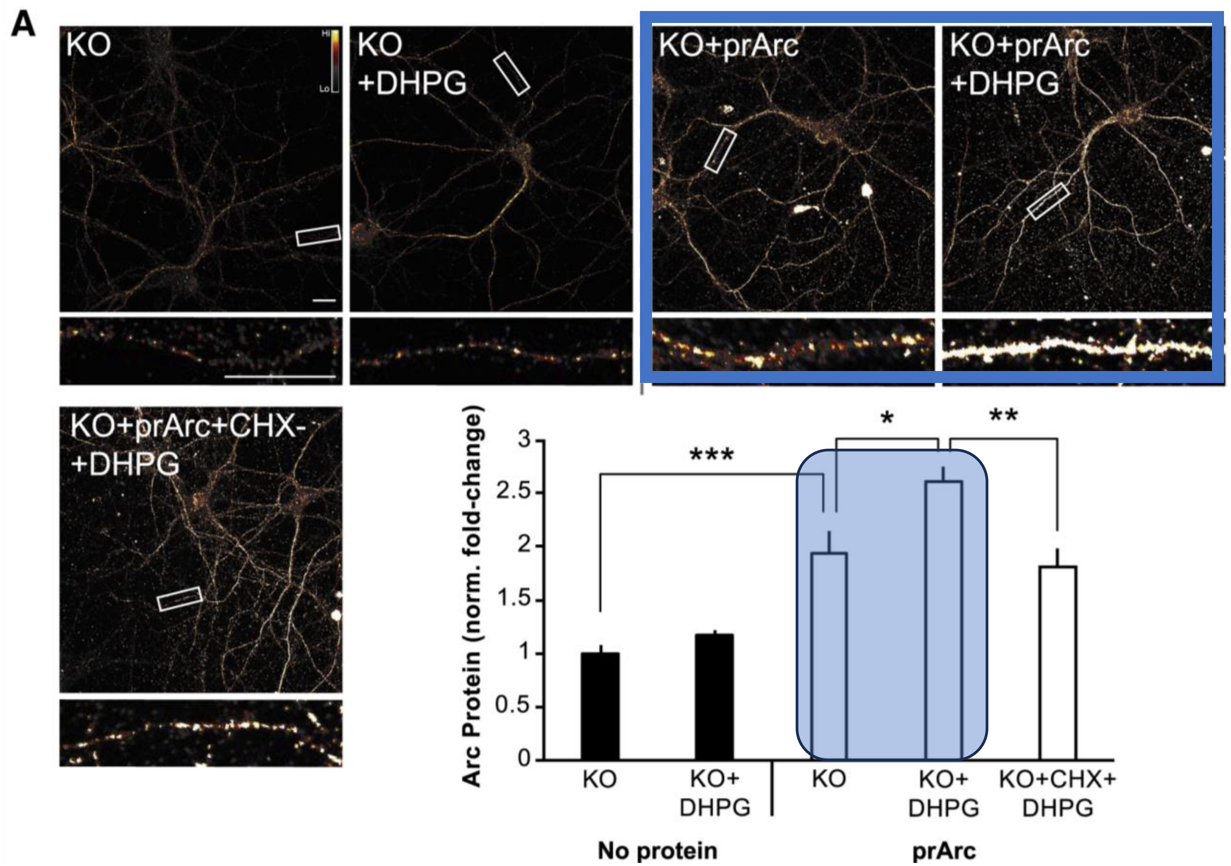
Can Arc transfer mRNA with capsids or EVs?



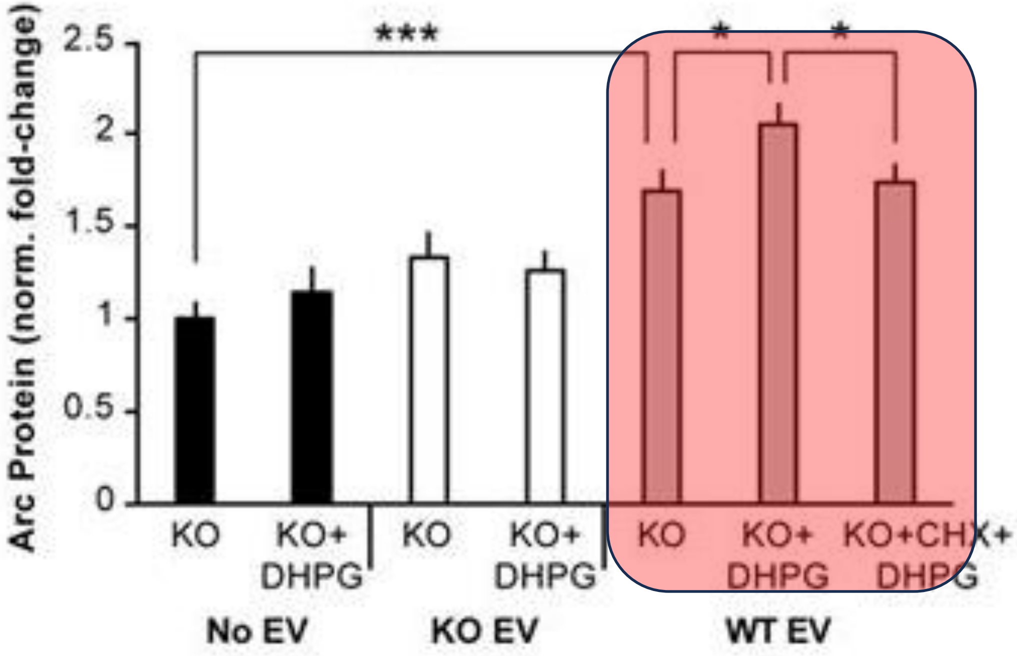
Are capsids required for neuronal uptake?



Is mRNA transferred by Arc available for translation?

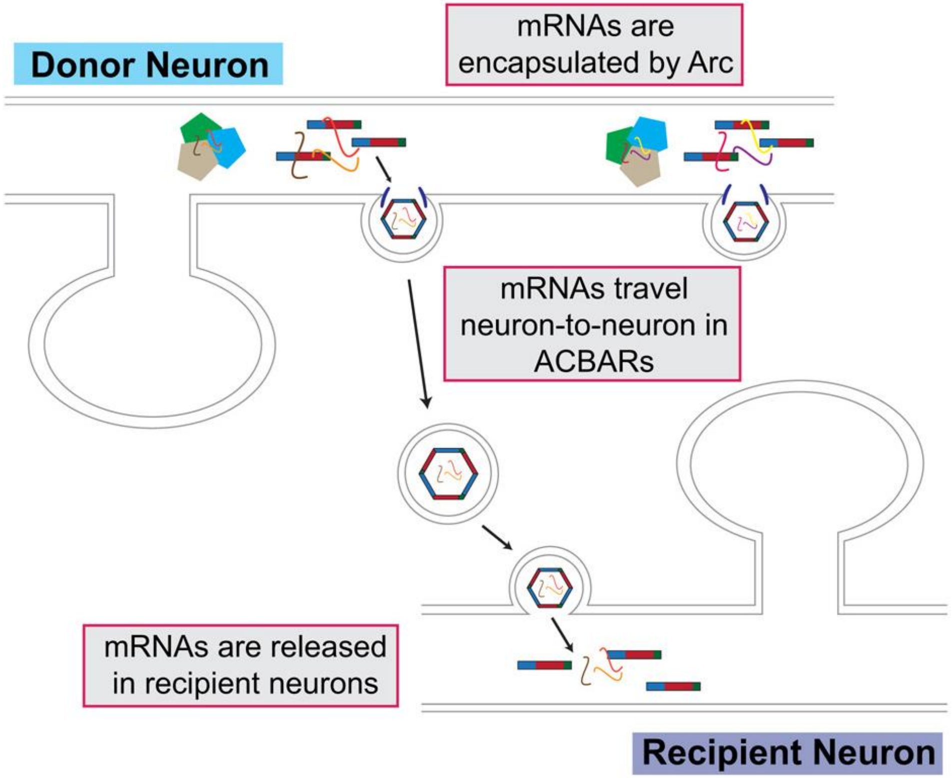


Is mRNA transferred by Arc available for translation?



Translation is occurring after being transferred to neurons.

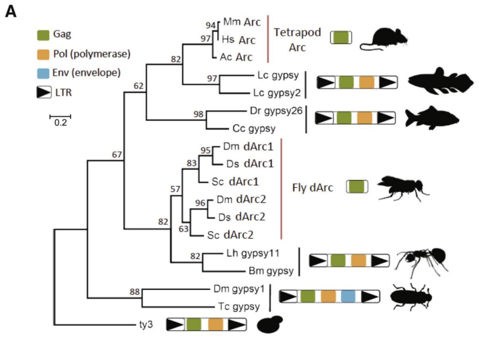
How does Arc work in the brain?



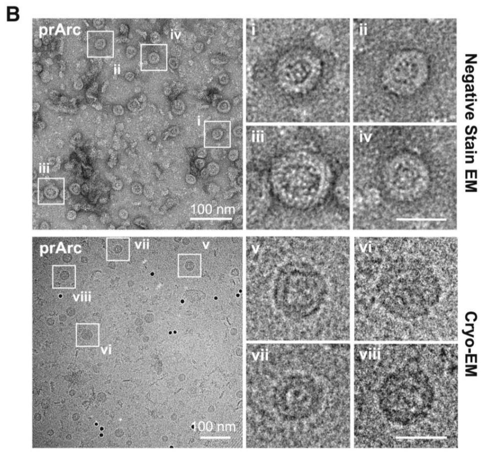
Donor Neuron
mRNAs are encapsulated by Arc
mRNAs travel neuron-to-neuron in ACBARs
mRNAs are released in recipient neurons
Recipient Neuron

Like a retrovirus!

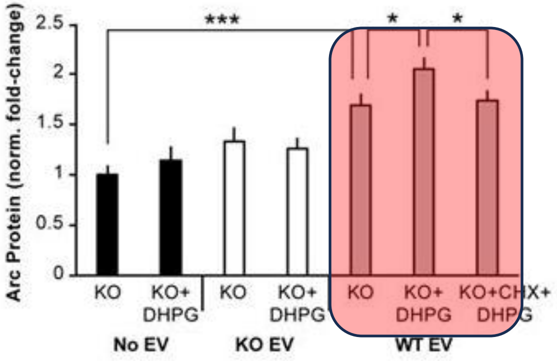
Summary



Arc shares properties of the retroviral Gag protein



Arc can form stable virus-like capsids



These capsid structures allow for mRNA transfer from neuron to neuron.

Questions?

Citations

Xiong J. Protein Motifs and Domain Prediction. In: *Essential Bioinformatics*. Cambridge University Press; 2006:85-94.

Eddy, S. What is a hidden Markov model?. *Nat Biotechnol* **22**, 1315–1316 (2004). https://doi.org/10.1038/nbt1004-1315

Finn, R.D., Mistry, J., Tate, J.G., Coggill, P.C., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R., & Bateman, A. (2007). The Pfam protein families database. Nucleic Acids Research, 38, D211 - D222.

Punta, Marco et al. "The Pfam protein families database." Nucleic Acids Research 40 (2011): D290 - D301.

UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res*. 2008;36(Database issue):D190-D195. doi:10.1093/nar/gkm895

Blum M, Chang HY, Chuguransky S, et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res*. 2021;49(D1):D344-D354. doi:10.1093/nar/gkaa977

Pastuzyn ED, et al., 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.

Image links

- https://static.wixstatic.com/media/309e2a_1fccf0f30277416b8ed317c7548b3401~mv2.png/v1/fill/w_1000,h_667,al_c,q_90,usm_0.66_1.00_0.01/309e2a_1fccf0f30277416b8ed317c7548b3401~mv2.png
- https://microbenotes.com/wp-content/uploads/2020/10/Alternative-Splicing.jpeg
- https://www.ebi.ac.uk/training/online/courses/protein-classification-intro-ebi-resources/wp-content/uploads/sites/96/2020/07/figure7.png
- https://www.ncbi.nlm.nih.gov/Structure/cdd/docs/images/cd00400_hierarchy_tree.png
- https://i.ytimg.com/vi/6OMo9M-nnuk/maxresdefault.jpg
- https://i0.wp.com/pediaa.com/wp-content/uploads/2019/06/Difference-Between-Motif-and-Domain-in-Protein-Structure-Comparison-Summary.jpg?resize=475%2C600&ssl=1
- https://www.ebi.ac.uk/training/online/courses/interpro-functional-and-structural-analysis/wp-content/uploads/sites/32/h5p/content/5/images/image-65b9130f47105.png
- https://www.sciencedirect.com/science/article/pii/S0092867417315040
- https://clarkesworldmagazine.com/koboldt_02_16/
- https://en.wikipedia.org/wiki/Retrotransposon