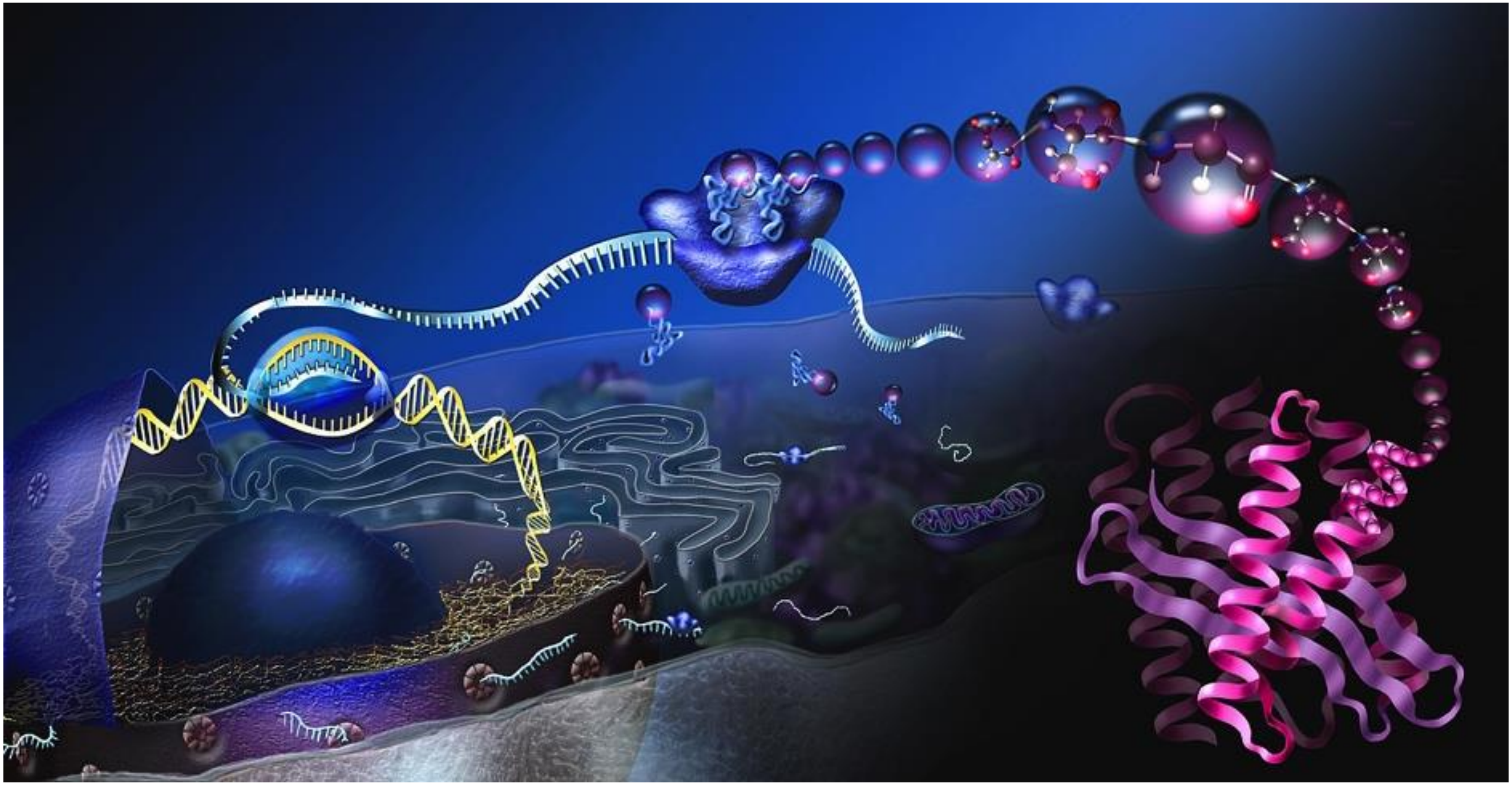
LOSCHMIDT
LABORATORIES

Bioinformatics protein sequences and databases

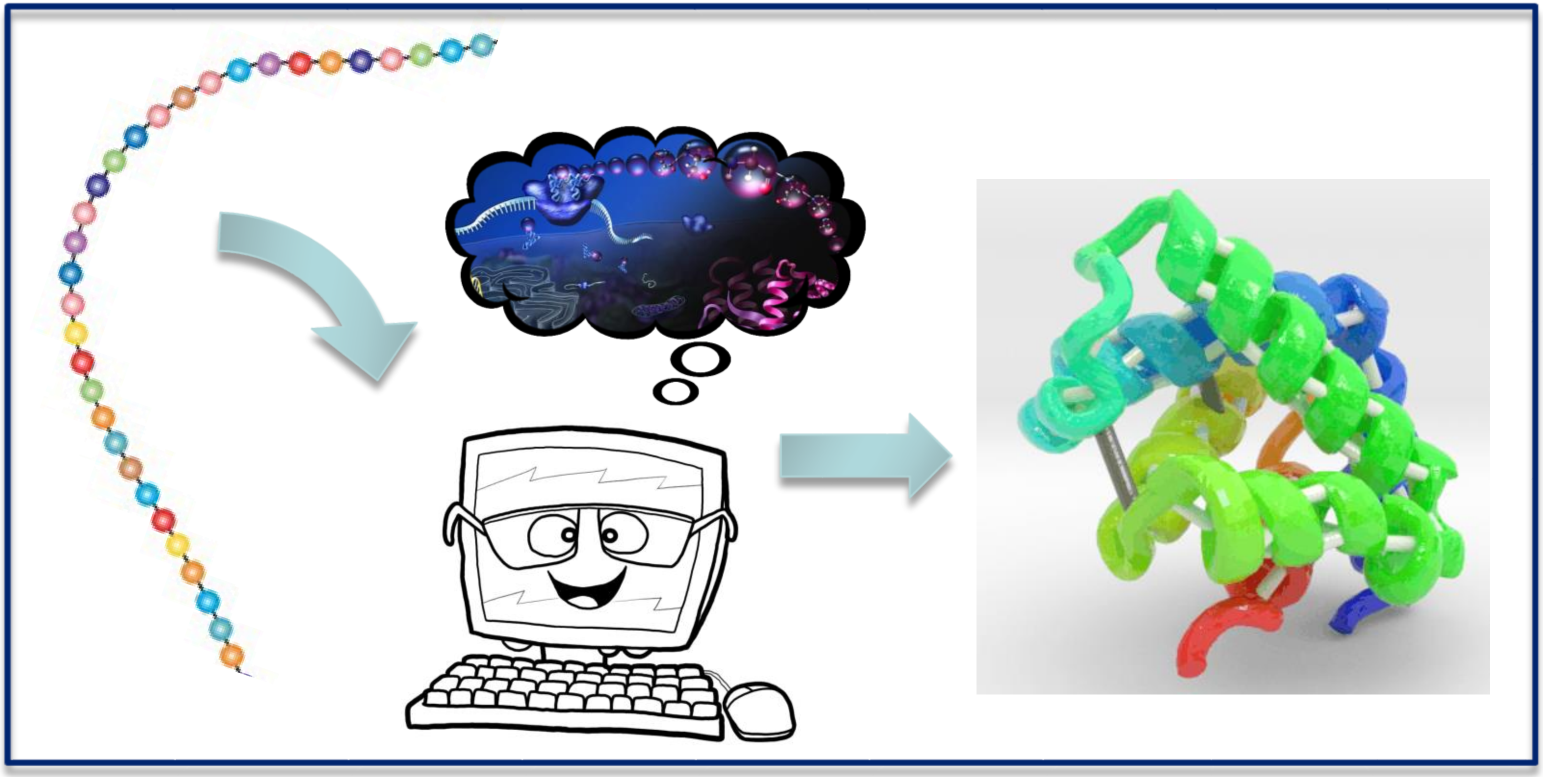
Outline

- Introduction
- Primary sequence of proteins
- Protein sequence databases
- Sequence alignments
    - evolution of proteins
    - Sequence-structure-function paradigm
    - Alignment of sequences
- Prediction of protein properties from sequence

Proteins: a quick overview

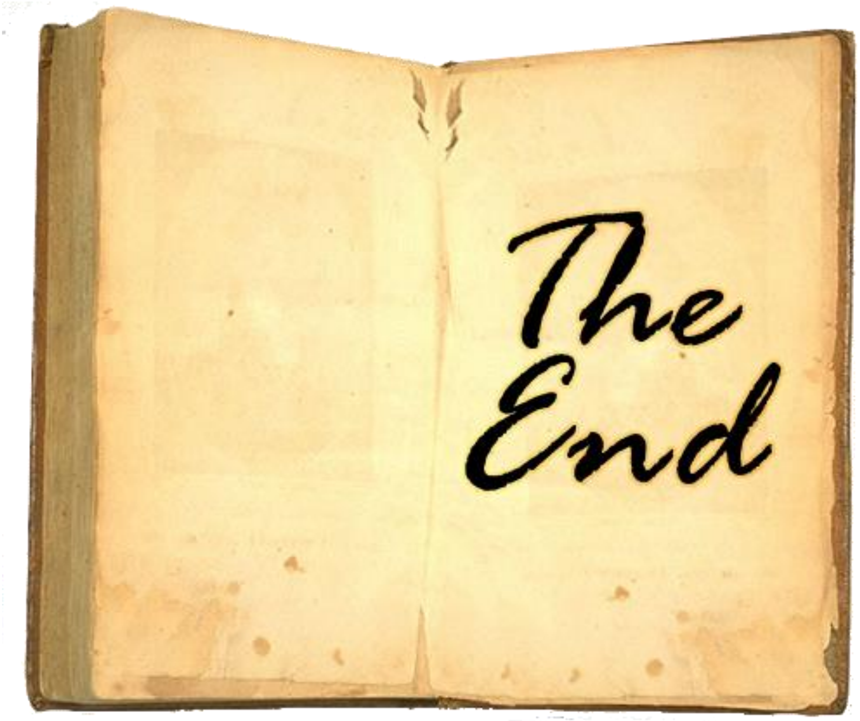


Structure prediction



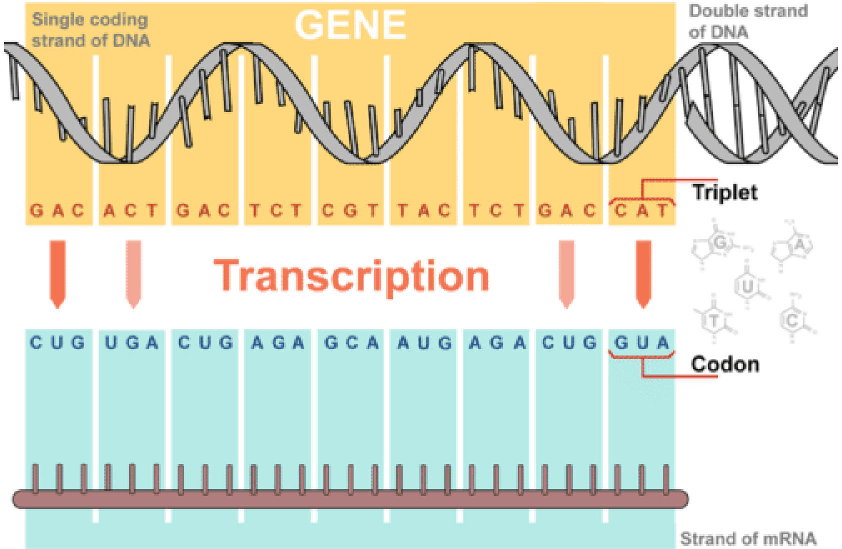
Google DeepMind

Structure prediction



Let’s start from the beginning…

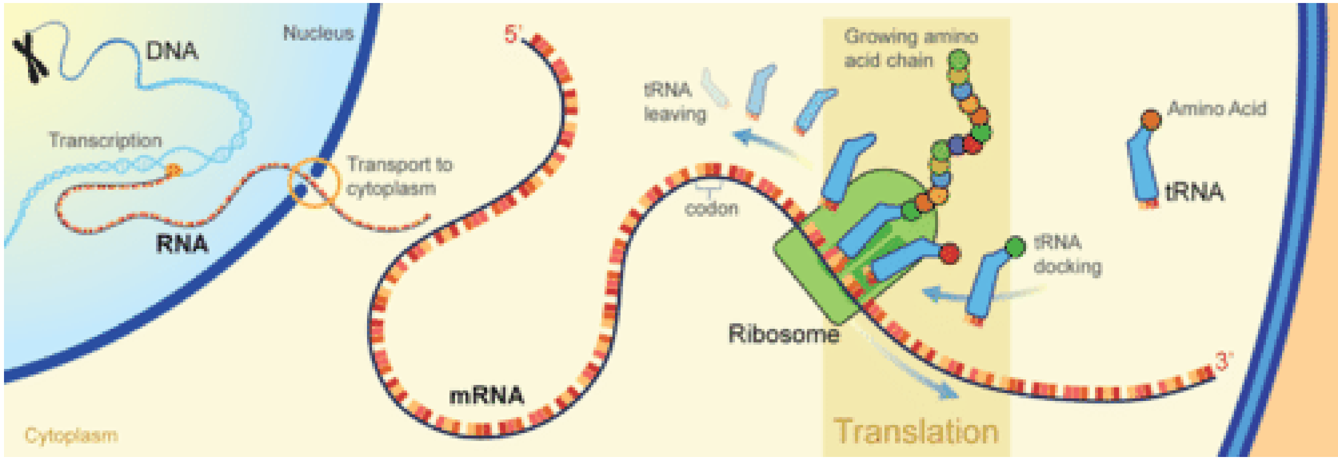
Protein synthesis


Single coding strand of DNA
GENE
Double strand of DNA
GAC ACT GAC TCT CGT TAC TCT GAC CAT
Triplet
Transcription
G
A
U
T
C
CUG UGA CUG AGA GCA AUG AGA CUG GUA
Codon
Strand of mRNA

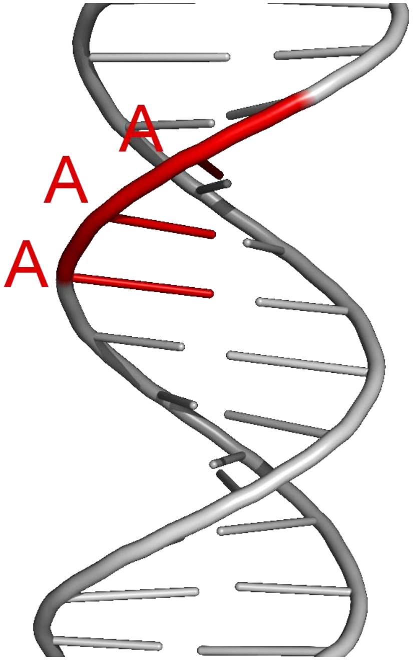
Protein synthesis occurs in two steps:

- Transcription: DNA -> RNA
- Splicing: RNA -> mRNA
- Translation: mRNA -> Protein
- Post-translational modifications: protein → mature protein

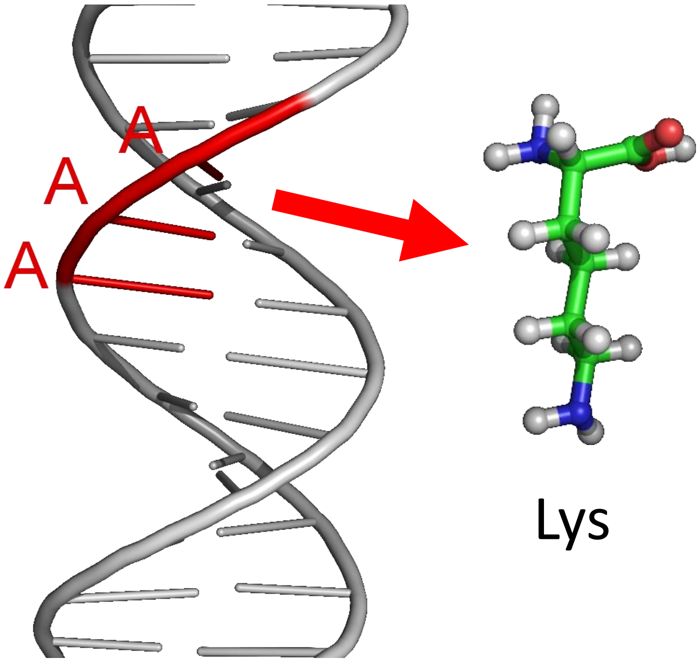
Translation



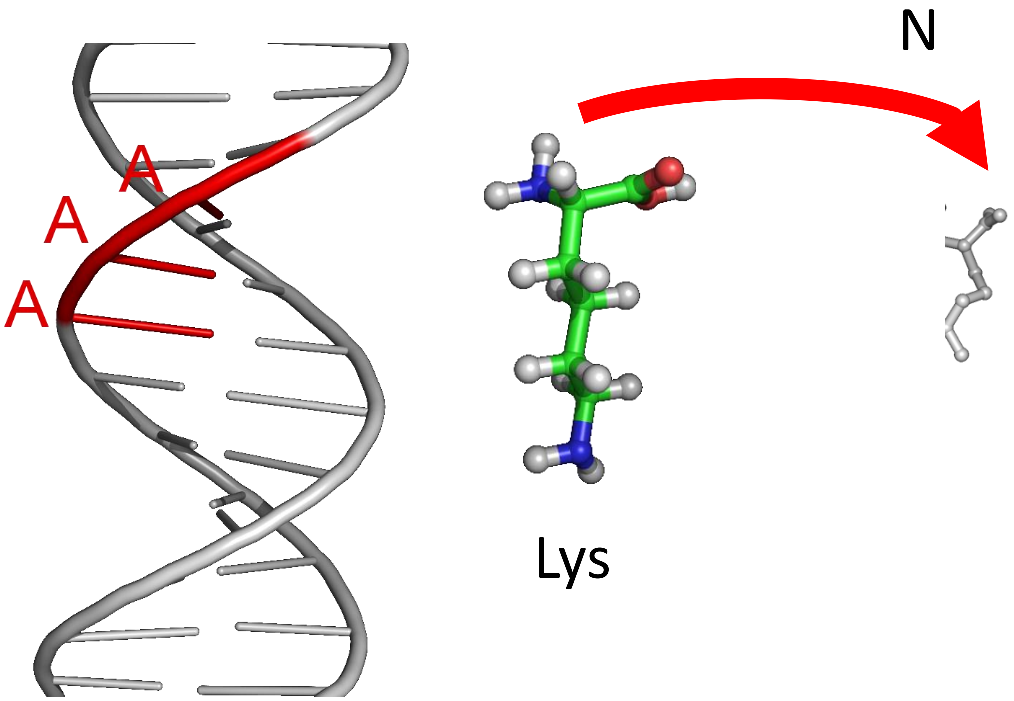
Protein synthesis



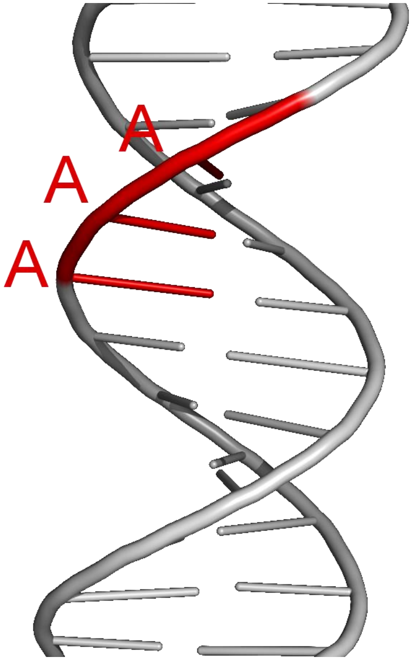
Protein synthesis



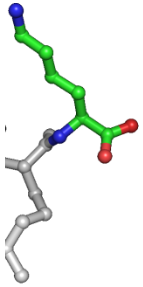
Protein synthesis



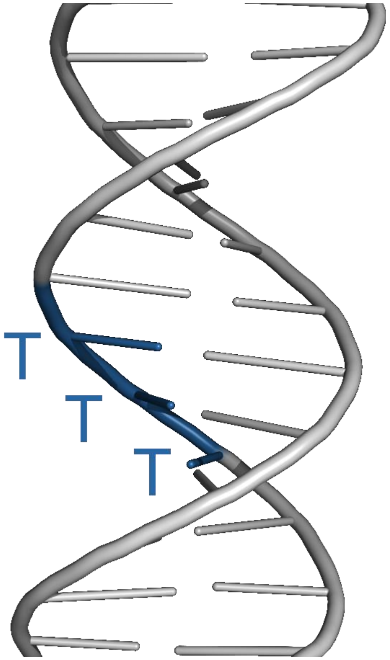
Protein synthesis

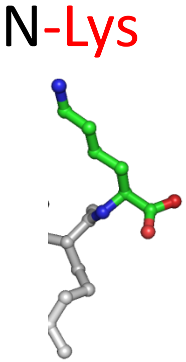


N-Lys

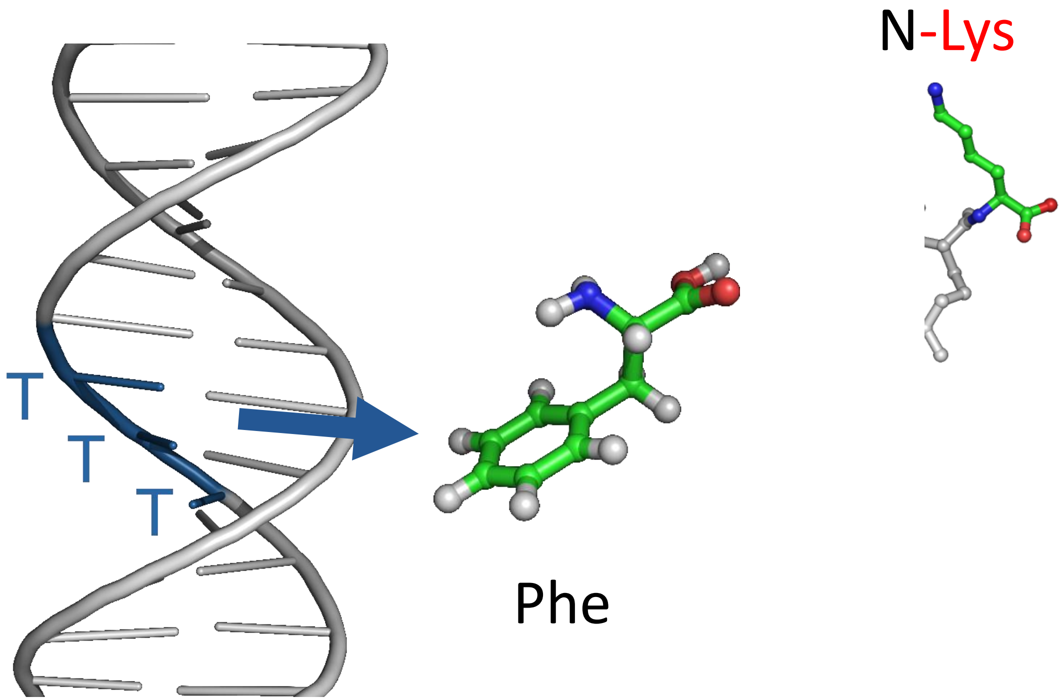


Protein synthesis

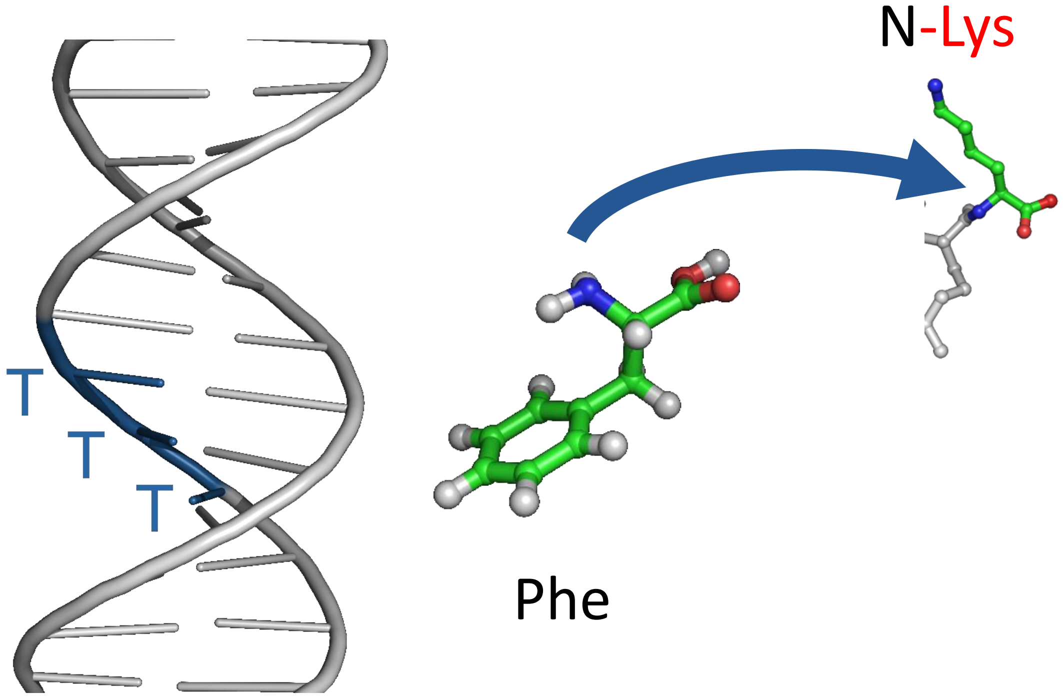




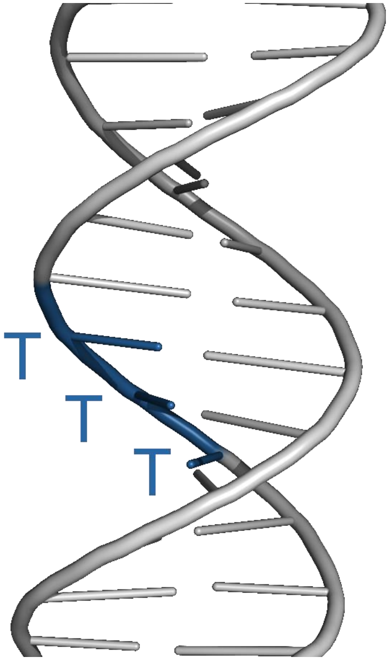
Protein synthesis



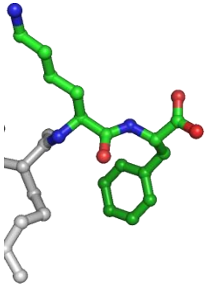
Protein synthesis



Protein synthesis

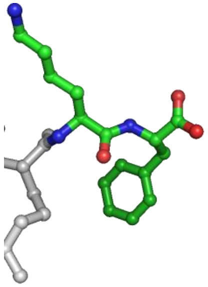


N-Lys-Phe

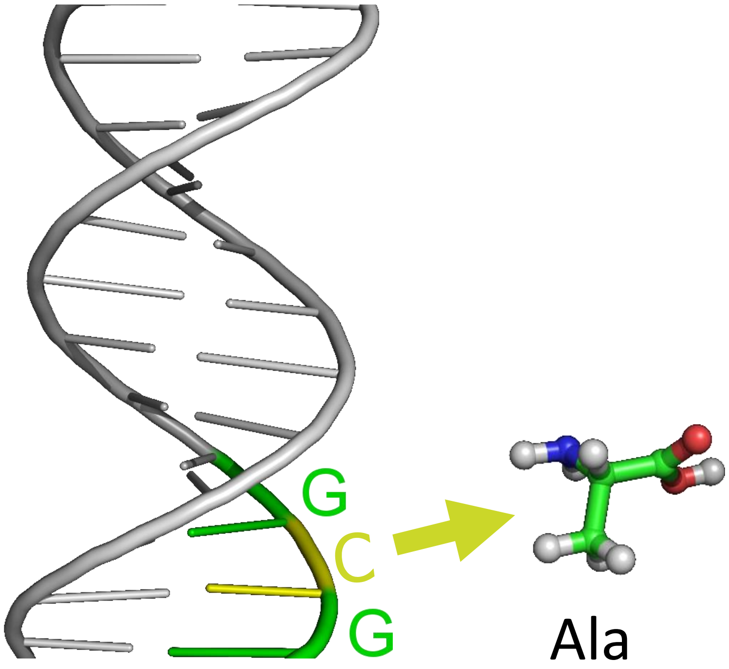


Protein synthesis

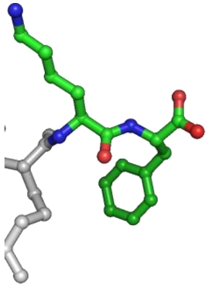
N-Lys-Phe



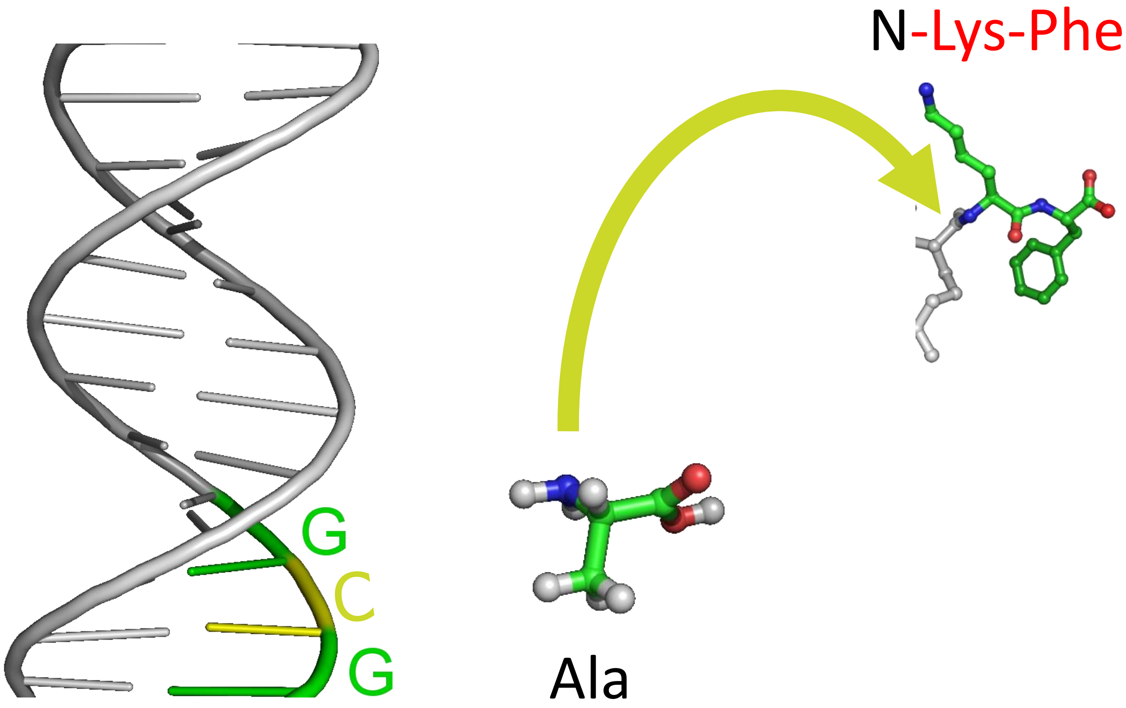
Protein synthesis



N-Lys-Phe



Protein synthesis



N-Lys-Phe

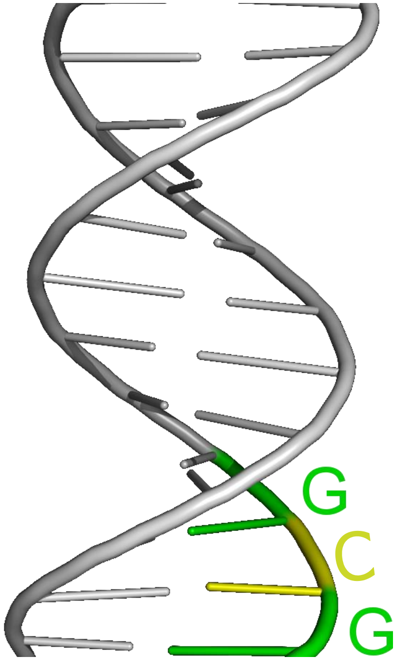
G

C

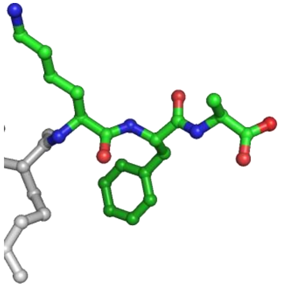
G

Ala

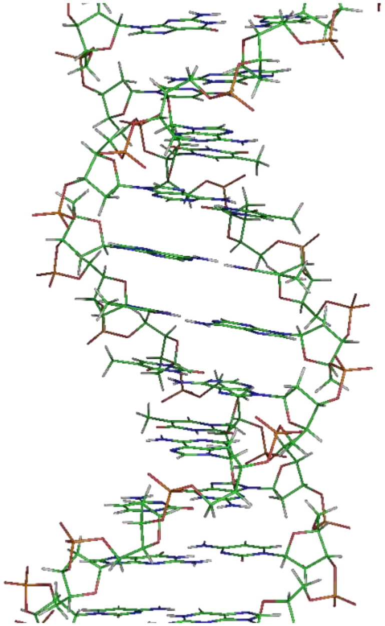
Protein synthesis



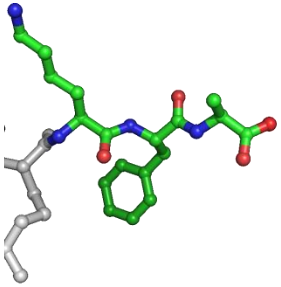
N-Lys-Phe-Ala



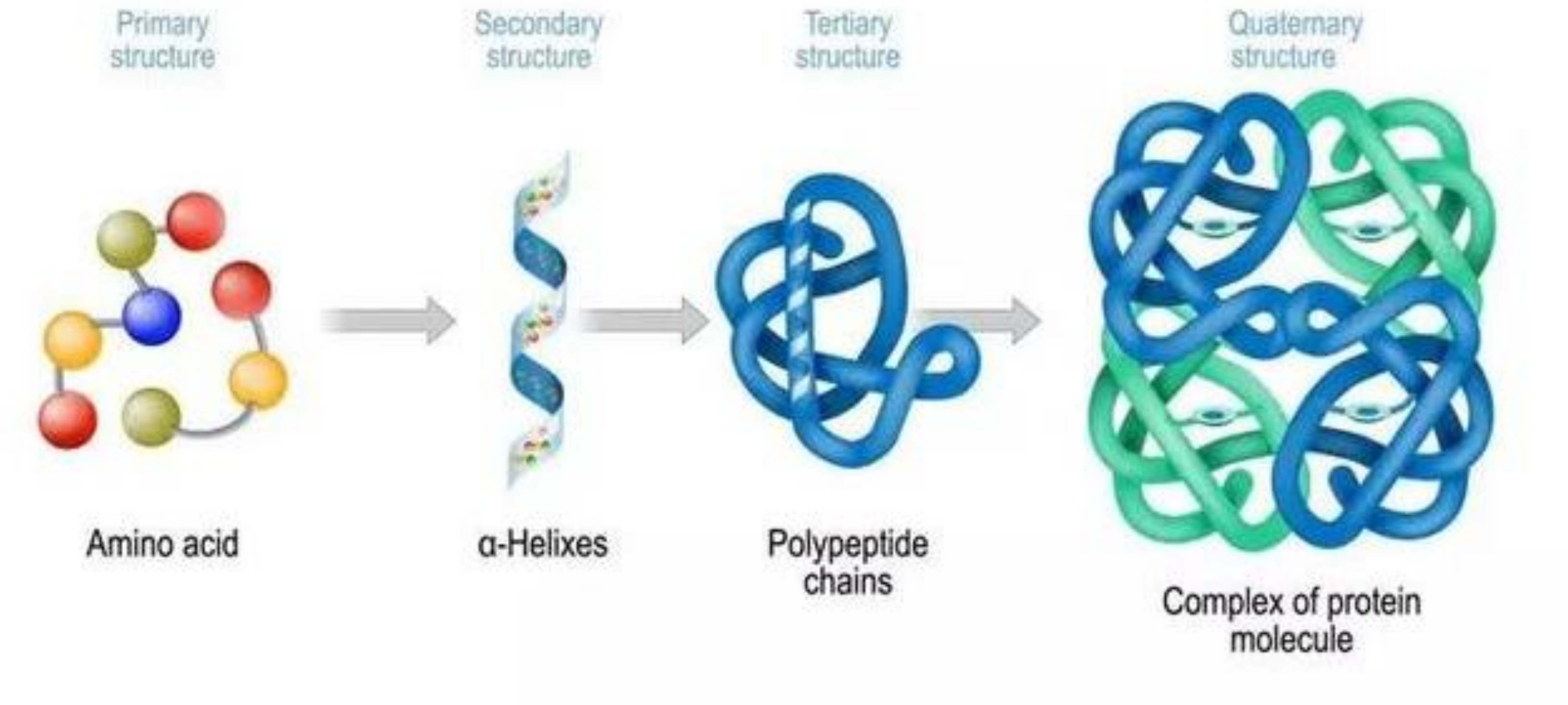
5’-NCG-AAA-TTT-GCG-3’



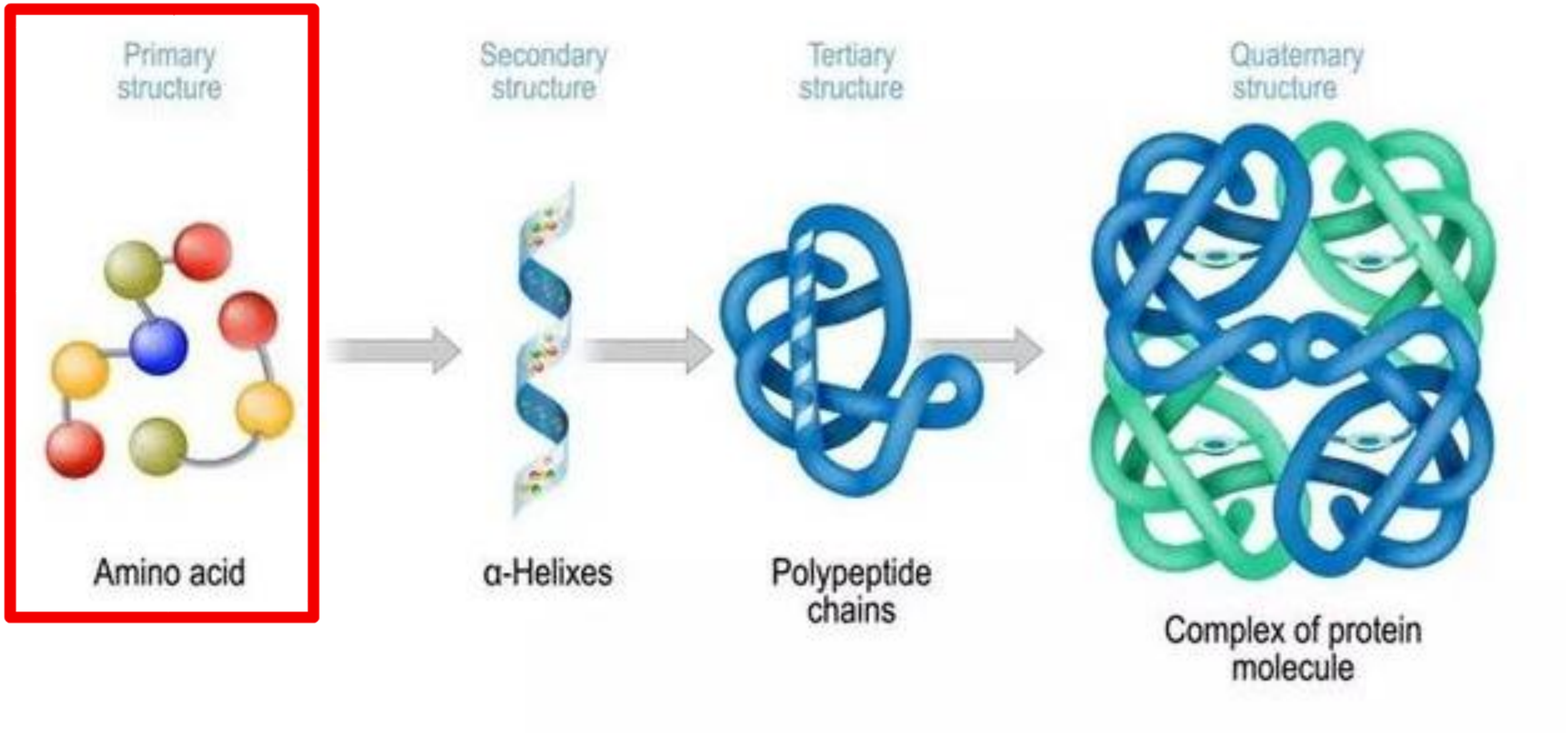
N-Lys-Phe-Ala



Levels of protein structure



Levels of protein structure



Sources of protein sequences

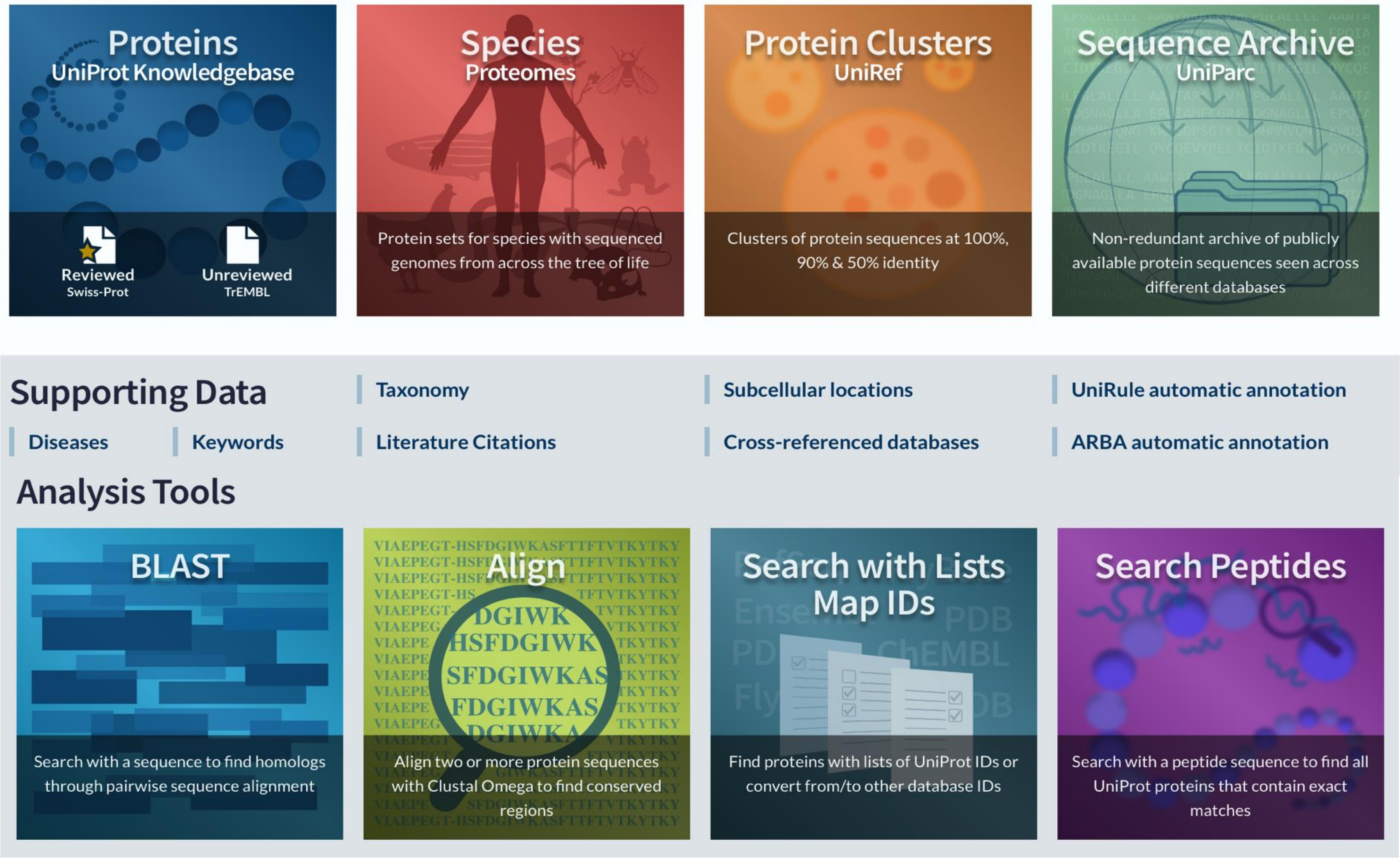
- Multiple databases available:
  - With different scope focus:
    - **Generalist**: sequences from any source (UniProtKB)
    - **Specialist**: sequences focusing on one more specific condition(s) (i.e. biologic pathway, disease, organism) (WormBase)
  - With different types of sequence content:
    - **Primary sequence** of proteins, and annotations and cross-references to that sequence (UniProtKB)
    - **Motifs or profiles databases**: contain information derived from the primary sequence, in the form of abstractions (patterns) that distil the most conserved features among related proteins (PFam)

Sources of protein sequences

- Multiple databases available

- UniProtKB
  - Collaboration between EBI, Swiss Institute of Bioinformatics and Protein Information
  - Central repository of protein sequences and functional information
  - Quality annotations - information on protein function and individual amino acids, experimental information, biological ontologies, classification, links to other databases
  - Quality level of the annotation (manual vs. automatic)

UniProt KB



Proteins
UniProt Knowledgebase

Reviewed
Swiss-Prot

Unreviewed
TrEMBL

Species
Proteomes

Protein sets for species with sequenced genomes from across the tree of life

Protein Clusters
UniRef

Clusters of protein sequences at 100%, 90% & 50% identity

Sequence Archive
UniParc

Non-redundant archive of publicly available protein sequences seen across different databases

Supporting Data

- Diseases
- Keywords
- Taxonomy
- Literature Citations
- Subcellular locations
- Cross-referenced databases
- UniRule automatic annotation
- ARBA automatic annotation

Analysis Tools

BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

Align two or more protein sequences with Clustal Omega to find conserved regions

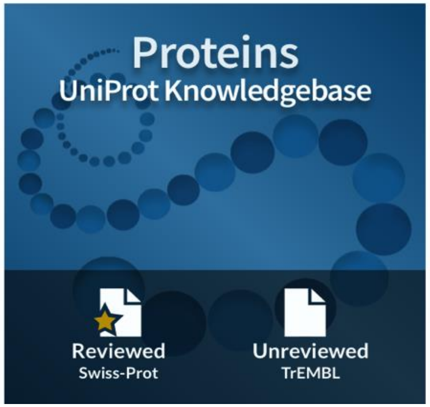
Search with Lists
Map IDs

Find proteins with lists of UniProt IDs or convert from/to other database IDs

Search Peptides

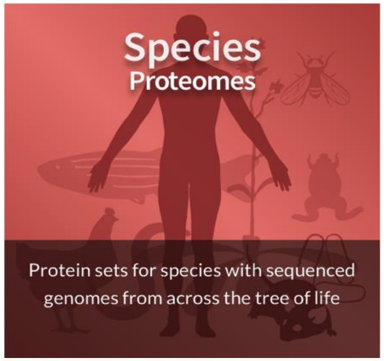
Search with a peptide sequence to find all UniProt proteins that contain exact matches

UniProt KB



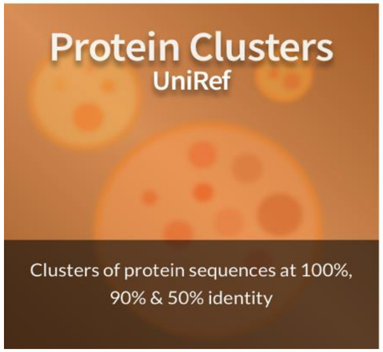
- ❑ Main component of the database
- ❑ Reviewed protein entries (SwissProt):
  - High quality manual annotations
  - ☺ Manual annotations → reliable info
  - ☹ >570,000 protein records (2024)
- ❑ Automatic protein entries (TrEMBL):
  - Automatic translation of protein sequences from EMBL data bank
  - ☹ Automatic annotations → lower quality, chance for errors.
  - ☺ ~250,000,000 protein records (2024) (400x info ammount)

UniProt KB



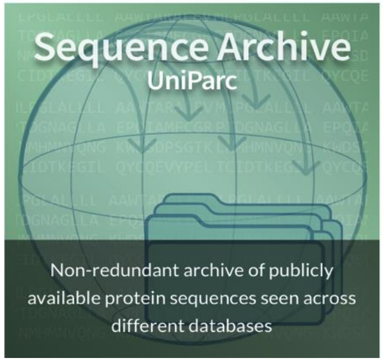
Species
Proteomes

Protein sets for species with sequenced genomes from across the tree of life



Protein Clusters
UniRef

Clusters of protein sequences at 100%, 90% & 50% identity



Sequence Archive
UniParc

Non-redundant archive of publicly available protein sequences seen across different databases

Proteomes for 25,000 model organisms available

Different degrees of coverage (other 160,000 available)

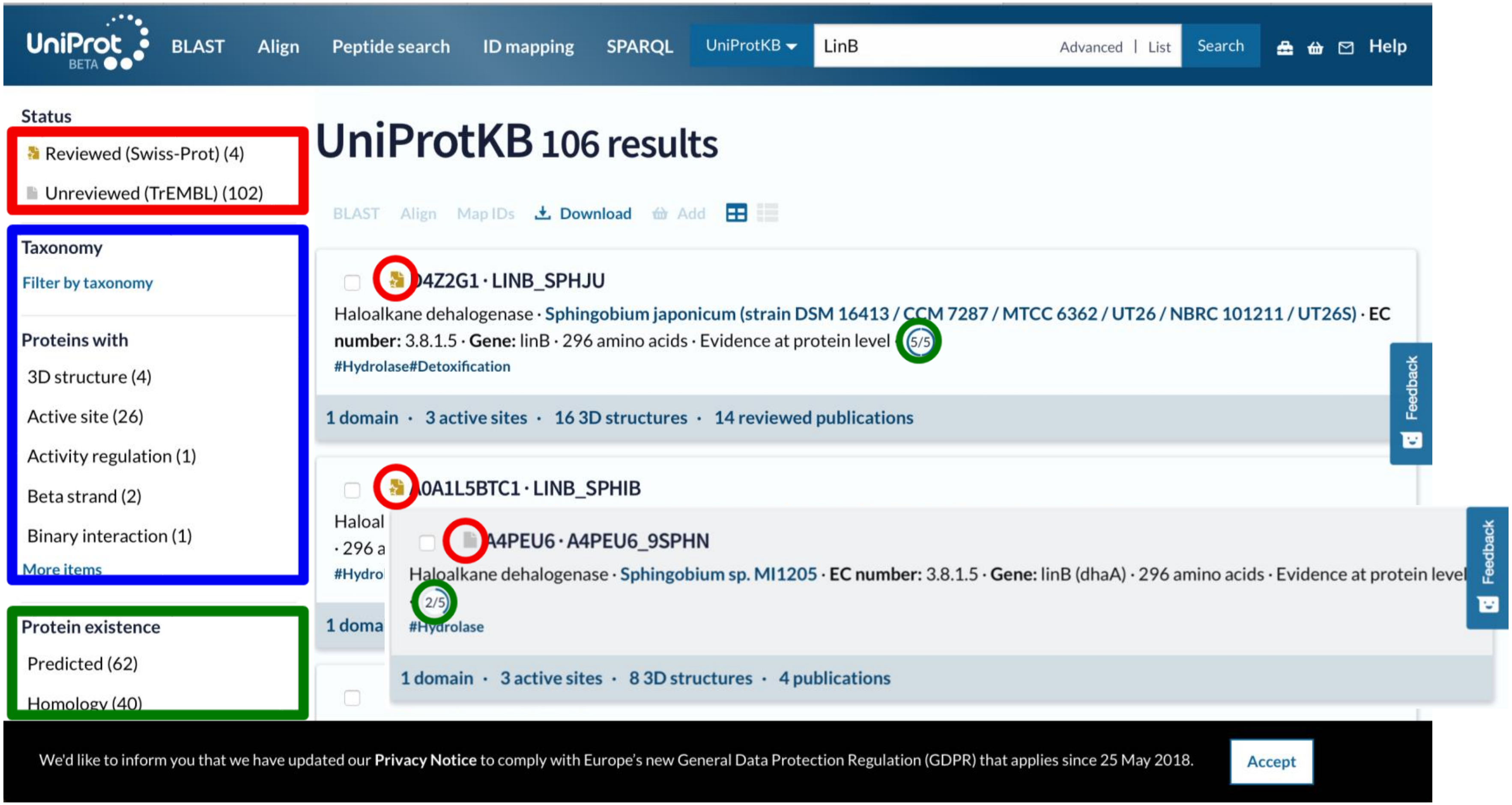
Clusters of proteins at 100%, 90%, and 50% seq. ID

Groups of similar proteins where to sample from

Stable identifier repository

Cross-references to a wealth of 40 external different databases (generalist and specialist)

UniProt KB



Status

Reviewed (Swiss-Prot) (4)

Unreviewed (TrEMBL) (102)

Taxonomy

Filter by taxonomy

Proteins with

3D structure (4)

Active site (26)

Activity regulation (1)

Beta strand (2)

Binary interaction (1)

More items

Protein existence

Predicted (62)

Homology (40)

UniProtKB 106 results

□ D4Z2G1 · LINB_SPHJU

Haloalkane dehalogenase · Sphingobium japonicum (strain DSM 16413 / CCM 7287 / MTCC 6362 / UT26 / NBRC 101211 / UT26S) · EC number: 3.8.1.5 · Gene: linB · 296 amino acids · Evidence at protein level 5/5

#Hydrolase#Detoxification

1 domain · 3 active sites · 16 3D structures · 14 reviewed publications

□ A0A1L5BTC1 · LINB_SPHIB

Haloal

· 296 a

#Hydro

1 doma

□ A4PEU6 · A4PEU6_9SPHN

Haloalkane dehalogenase · Sphingobium sp. MI1205 · EC number: 3.8.1.5 · Gene: linB (dhaA) · 296 amino acids · Evidence at protein level 2/5

#Hydrolase

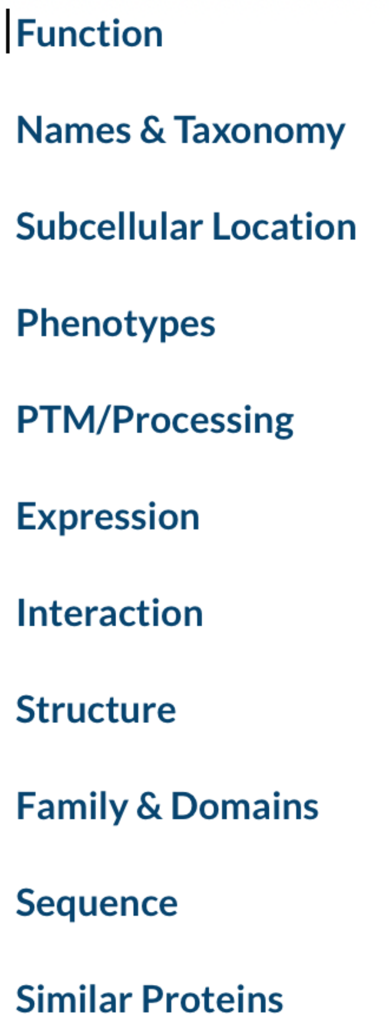
1 domain · 3 active sites · 8 3D structures · 4 publications

We'd like to inform you that we have updated our **Privacy Notice** to comply with Europe's new General Data Protection Regulation (GDPR) that applies since 25 May 2018.

Quality **Info:** Name/Organism source/EC activity/gene name/length.

Filters **Protein evidence** **+Info**: Domain/3D structure/active site/pubs.

UniProt KB



Function

Names & Taxonomy

Subcellular Location

Phenotypes

PTM/Processing

Expression

Interaction

Structure

Family & Domains

Sequence

Similar Proteins

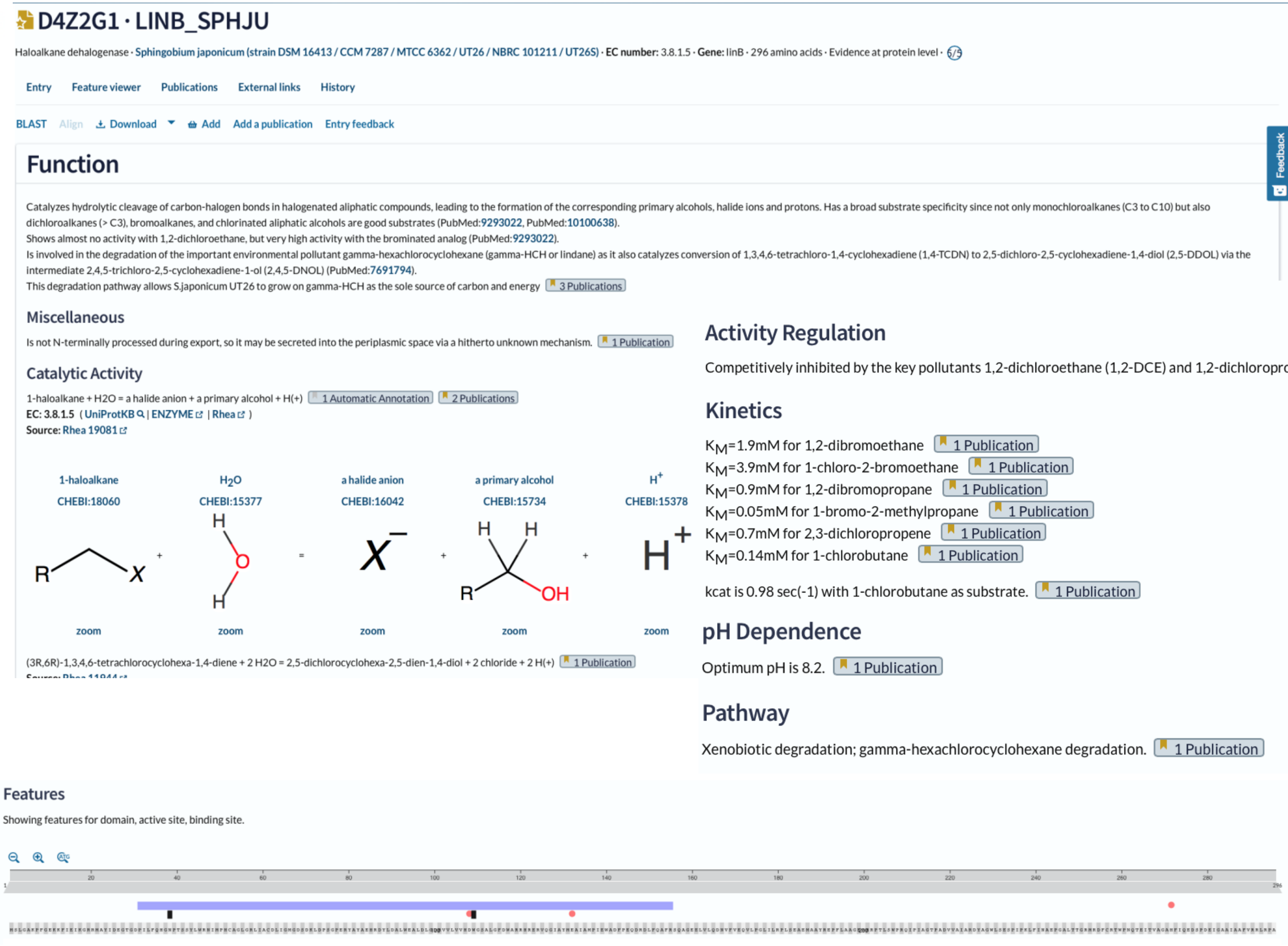
Human readable explanation of the protein function

Wealth of systematically organized information. In the illustrated example:

- **Catalytic activity**: with details of the enzymatic reaction and cross-links to chemical databases
- **Activity regulation**: competitive inhibitors
- **Kinetics**: experimental measurements towards n substrates
- **Optimal pH**
- Implication in **biological pathways**
- **Catalytic and Key Residues** (active/binding sites)
- **Gene Ontology (GO) annotations** (enrichment values)
- **Enzyme/Pathways and Protein Family DBs**
- **Keywords**

UniProt KB

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins



D4Z2G1 · LINB_SPHJU

Haloalkane dehalogenase · Sphingobium japonicum (strain DSM 16413 / CCM 7287 / MTCC 6362 / UT26 / NBRC 101211 / UT26S) · **EC number:** 3.8.1.5 · **Gene:** linB · 296 amino acids · Evidence at protein level · 5/5

Function

Catalyzes hydrolytic cleavage of carbon-halogen bonds in halogenated aliphatic compounds, leading to the formation of the corresponding primary alcohols, halide ions and protons. Has a broad substrate specificity since not only monochloroalkanes (C3 to C10) but also dichloroalkanes (> C3), bromoalkanes, and chlorinated aliphatic alcohols are good substrates (PubMed:9293022, PubMed:10100638).
Shows almost no activity with 1,2-dichloroethane, but very high activity with the brominated analog (PubMed:9293022).
Is involved in the degradation of the important environmental pollutant gamma-hexachlorocyclohexane (gamma-HCH or lindane) as it also catalyzes conversion of 1,3,4,6-tetrachloro-1,4-cyclohexadiene (1,4-TCDN) to 2,5-dichloro-2,5-cyclohexadiene-1,4-diol (2,5-DDOL) via the intermediate 2,4,5-trichloro-2,5-cyclohexadiene-1-ol (2,4,5-DNOL) (PubMed:7691794).
This degradation pathway allows S.japonicum UT26 to grow on gamma-HCH as the sole source of carbon and energy 3 Publications

Miscellaneous

Is not N-terminally processed during export, so it may be secreted into the periplasmic space via a hitherto unknown mechanism. 1 Publication

Catalytic Activity

1-haloalkane + H2O = a halide anion + a primary alcohol + H(+) 1 Automatic Annotation 2 Publications
EC: 3.8.1.5 (UniProtKB | ENZYME | Rhea)
Source: Rhea 19081

(3R,6R)-1,3,4,6-tetrachlorocyclohexa-1,4-diene + 2 H2O = 2,5-dichlorocyclohexa-2,5-dien-1,4-diol + 2 chloride + 2 H(+) 1 Publication
Source: Rhea 11944

Activity Regulation

Competitively inhibited by the key pollutants 1,2-dichloroethane (1,2-DCE) and 1,2-dichloropro

Kinetics

K_M=1.9mM for 1,2-dibromoethane 1 Publication
K_M=3.9mM for 1-chloro-2-bromoethane 1 Publication
K_M=0.9mM for 1,2-dibromopropane 1 Publication
K_M=0.05mM for 1-bromo-2-methylpropane 1 Publication
K_M=0.7mM for 2,3-dichloropropene 1 Publication
K_M=0.14mM for 1-chlorobutane 1 Publication

kcat is 0.98 sec(-1) with 1-chlorobutane as substrate. 1 Publication

pH Dependence

Optimum pH is 8.2. 1 Publication

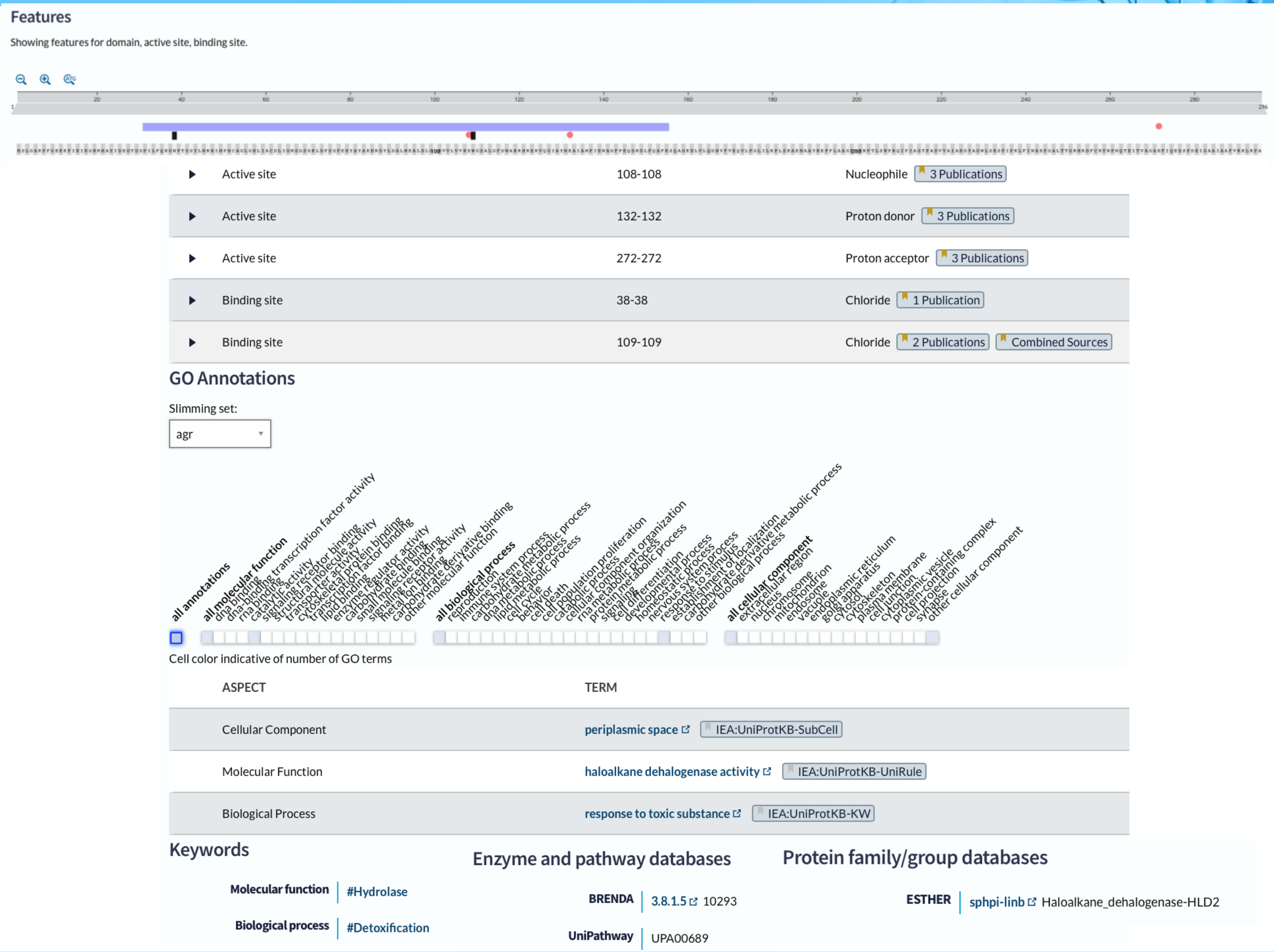
Pathway

Xenobiotic degradation; gamma-hexachlorocyclohexane degradation. 1 Publication

Features

Showing features for domain, active site, binding site.

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins



Features

Showing features for domain, active site, binding site.

▶	Active site	108-108	Nucleophile 3 Publications
▶	Active site	132-132	Proton donor 3 Publications
▶	Active site	272-272	Proton acceptor 3 Publications
▶	Binding site	38-38	Chloride 1 Publication
▶	Binding site	109-109	Chloride 2 Publications Combined Sources

GO Annotations

Slimming set:

agr

Cell color indicative of number of GO terms

ASPECT	TERM
Cellular Component	periplasmic space IEA:UniProtKB-SubCell
Molecular Function	haloalkane dehalogenase activity IEA:UniProtKB-UniRule
Biological Process	response to toxic substance IEA:UniProtKB-KW

Keywords

Molecular function	#Hydrolase
Biological process	#Detoxification

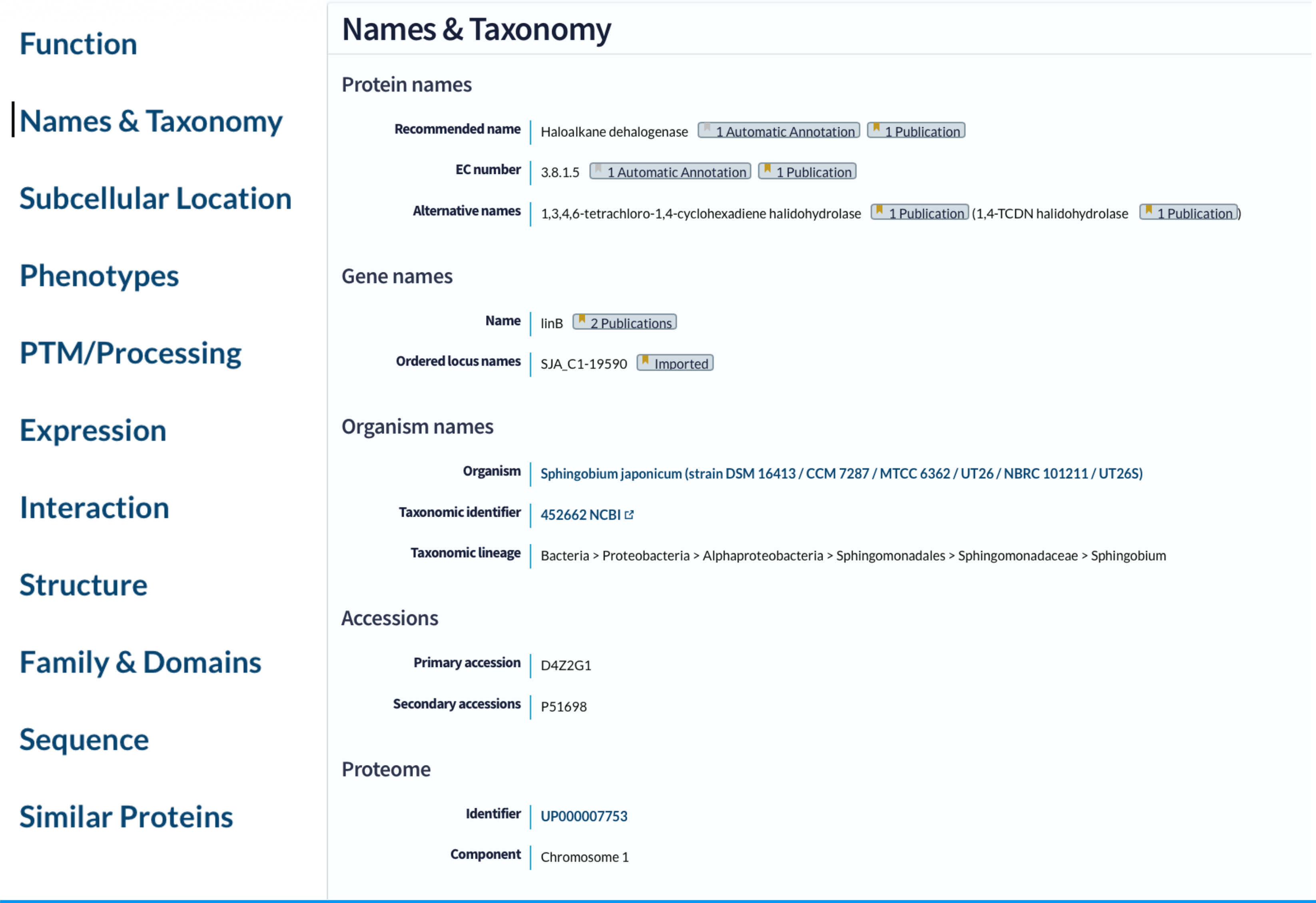
Enzyme and pathway databases

BRENDA	3.8.1.5 10293
UniPathway	UPA00689

Protein family/group databases

ESTHER	sphpi-linb Haloalkane_dehalogenase-HLD2

UniProt KB



- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins

Names & Taxonomy

Protein names

Recommended name	Haloalkane dehalogenase 1 Automatic Annotation 1 Publication
EC number	3.8.1.5 1 Automatic Annotation 1 Publication
Alternative names	1,3,4,6-tetrachloro-1,4-cyclohexadiene halidohydrolase 1 Publication (1,4-TCDN halidohydrolase 1 Publication)

Gene names

Name	linB 2 Publications
Ordered locus names	SJA_C1-19590 Imported

Organism names

Organism	Sphingobium japonicum (strain DSM 16413 / CCM 7287 / MTCC 6362 / UT26 / NBRC 101211 / UT26S)
Taxonomic identifier	452662 NCBI
Taxonomic lineage	Bacteria > Proteobacteria > Alphaproteobacteria > Sphingomonadales > Sphingomonadaceae > Sphingobium

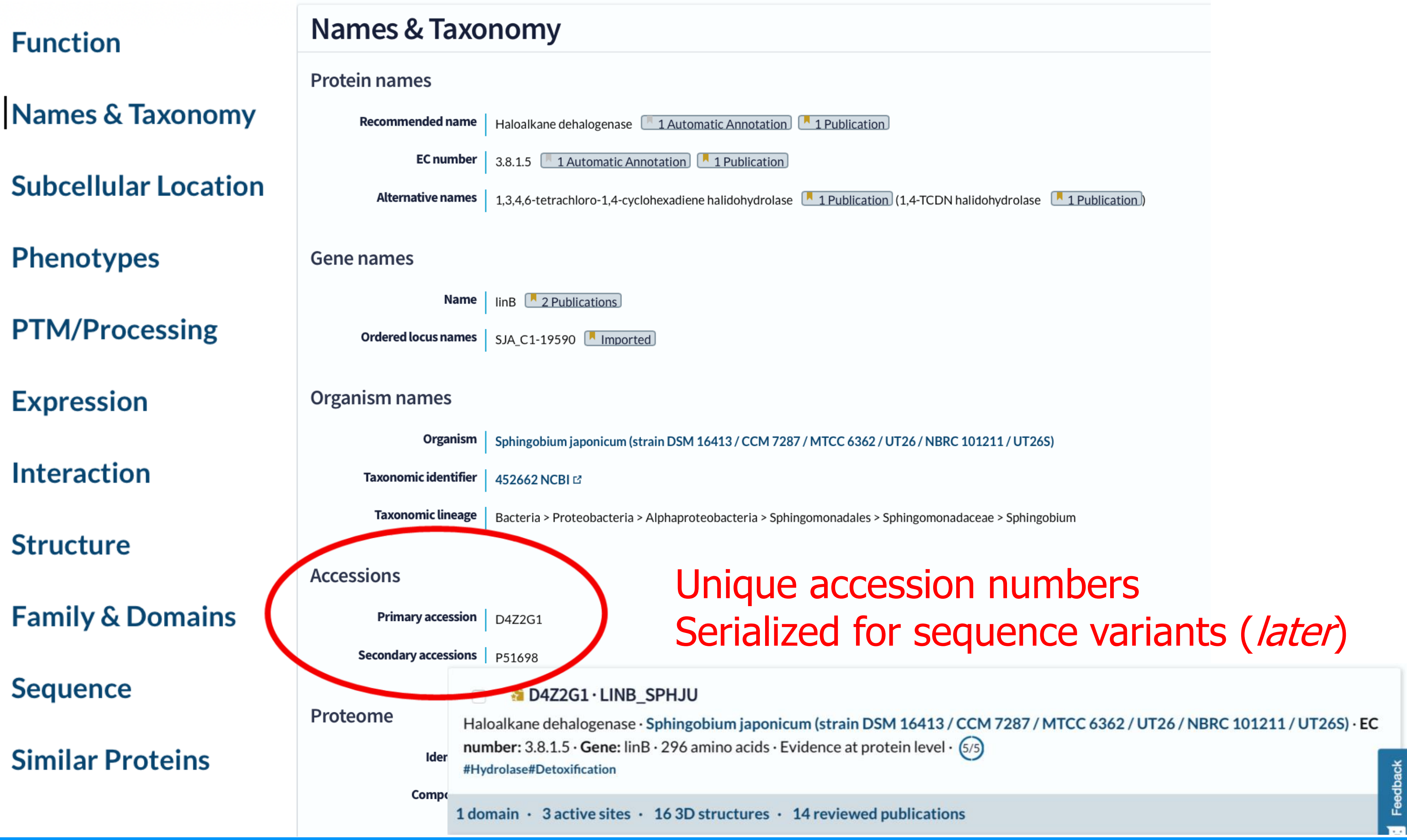
Accessions

Primary accession	D4Z2G1
Secondary accessions	P51698

Proteome

Identifier	UP000007753
Component	Chromosome 1

UniProt KB



Function

Names & Taxonomy

Subcellular Location

Phenotypes

PTM/Processing

Expression

Interaction

Structure

Family & Domains

Sequence

Similar Proteins

Names & Taxonomy

Protein names

Recommended name	Haloalkane dehalogenase 1 Automatic Annotation 1 Publication
EC number	3.8.1.5 1 Automatic Annotation 1 Publication
Alternative names	1,3,4,6-tetrachloro-1,4-cyclohexadiene halidohydrolase 1 Publication (1,4-TCDN halidohydrolase 1 Publication)

Gene names

Name	linB 2 Publications
Ordered locus names	SJA_C1-19590 Imported

Organism names

Organism	Sphingobium japonicum (strain DSM 16413 / CCM 7287 / MTCC 6362 / UT26 / NBRC 101211 / UT26S)
Taxonomic identifier	452662 NCBI
Taxonomic lineage	Bacteria > Proteobacteria > Alphaproteobacteria > Sphingomonadales > Sphingomonadaceae > Sphingobium

Accessions

Primary accession	D4Z2G1
Secondary accessions	P51698

Proteome

Ider

Compo

D4Z2G1 · LINB_SPHJU

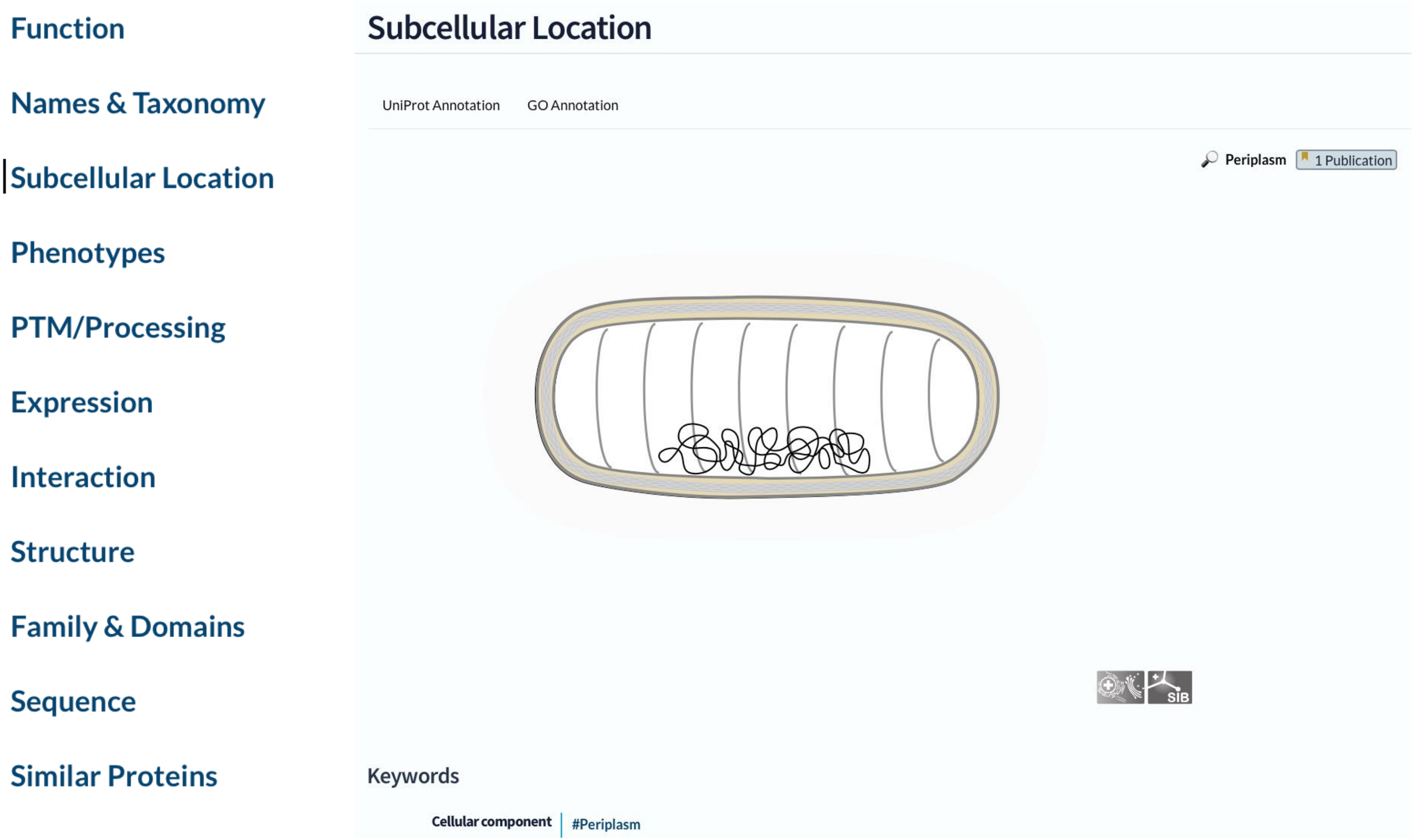
Haloalkane dehalogenase · Sphingobium japonicum (strain DSM 16413 / CCM 7287 / MTCC 6362 / UT26 / NBRC 101211 / UT26S) · **EC number:** 3.8.1.5 · **Gene:** linB · 296 amino acids · Evidence at protein level · 5/5

#Hydrolase#Detoxification

1 domain · 3 active sites · 16 3D structures · 14 reviewed publications

Unique accession numbers
Serialized for sequence variants (*later*)

UniProt KB



- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins

Subcellular Location

UniProt Annotation GO Annotation

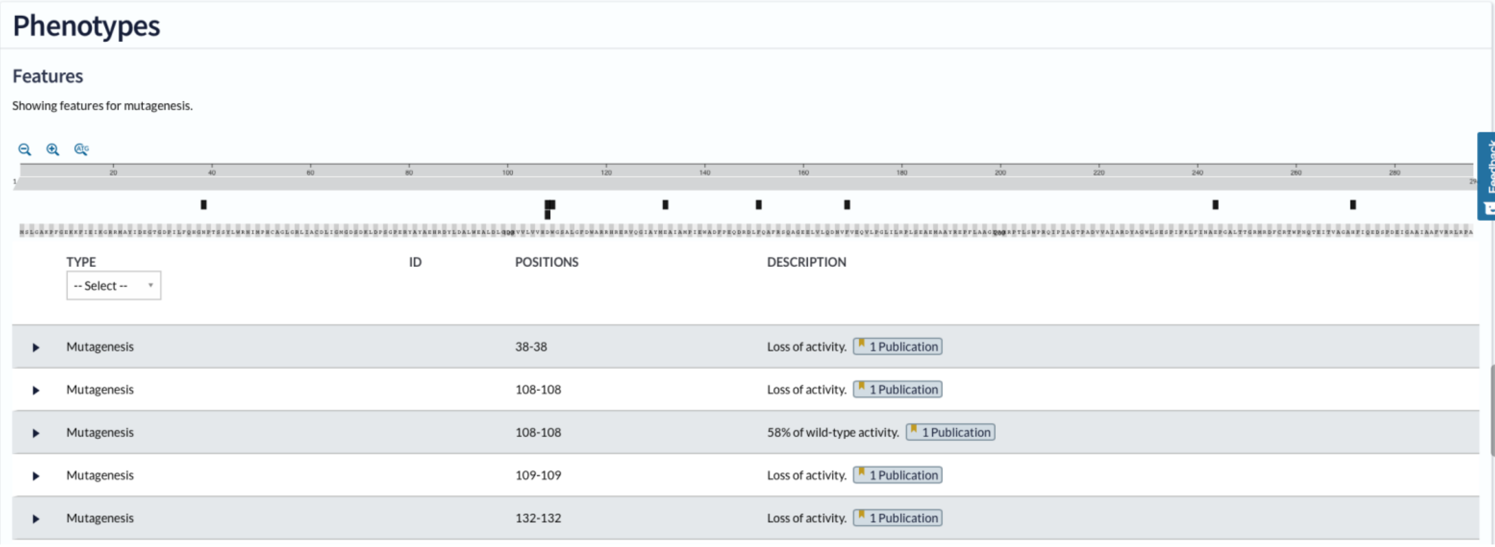
Periplasm 1 Publication

Keywords

Cellular component | #Periplasm

UniProt KB

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins



Phenotypes

Features

Showing features for mutagenesis.

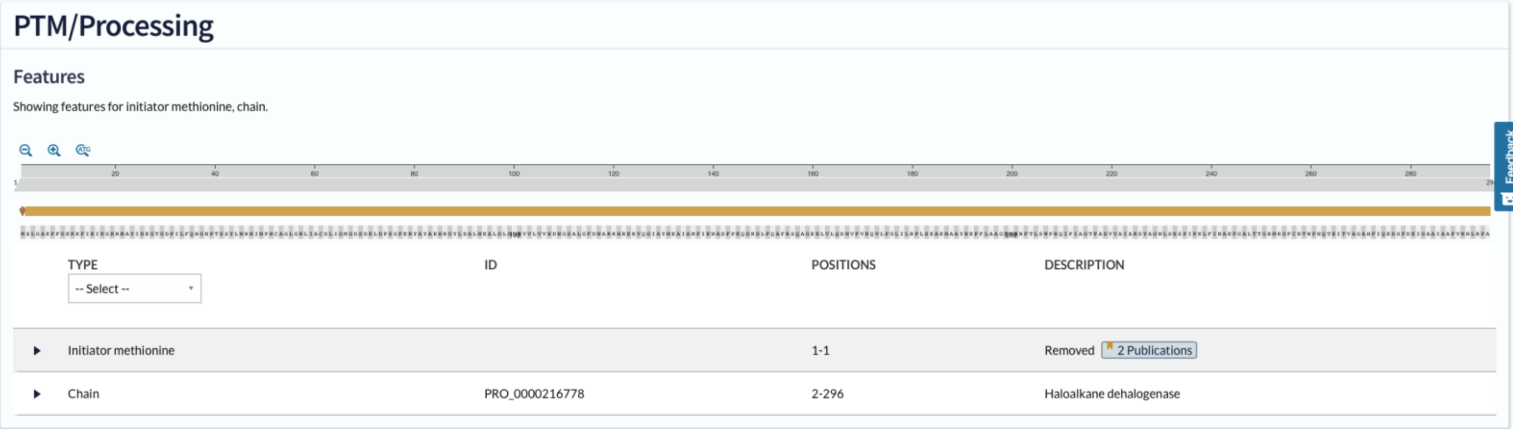
TYPE -- Select --	ID	POSITIONS	DESCRIPTION
Mutagenesis		38-38	Loss of activity. 1 Publication
Mutagenesis		108-108	Loss of activity. 1 Publication
Mutagenesis		108-108	58% of wild-type activity. 1 Publication
Mutagenesis		109-109	Loss of activity. 1 Publication
Mutagenesis		132-132	Loss of activity. 1 Publication

Describe the effect of ***mutations*** in the activity of the protein

Mutations mapped on the protein sequence

UniProt KB

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins



PTM/Processing

Features

Showing features for initiator methionine, chain.

TYPE -- Select --	ID	POSITIONS	DESCRIPTION
▸ Initiator methionine		1-1	Removed 2 Publications
▸ Chain	PRO_0000216778	2-296	Haloalkane dehalogenase

Describe post-translational modifications and other processing of the protein (i.e. cleaving for activation). Positions mapped on the protein sequence.

UniProt KB

Function

Names & Taxonomy

Subcellular Location

Phenotypes

PTM/Processing

Expression

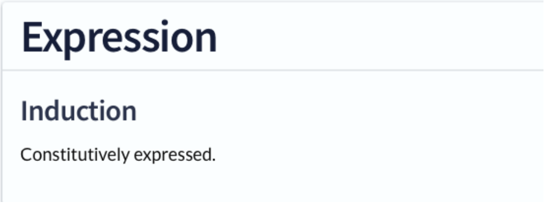
Interaction

Structure

Family & Domains

Sequence

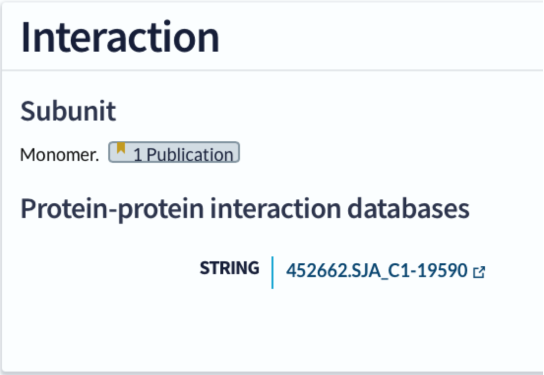
Similar Proteins



Expression

Induction

Constitutively expressed.



Interaction

Subunit

Monomer. 1 Publication

Protein-protein interaction databases

STRING | 452662.SJA_C1-19590

Expression:

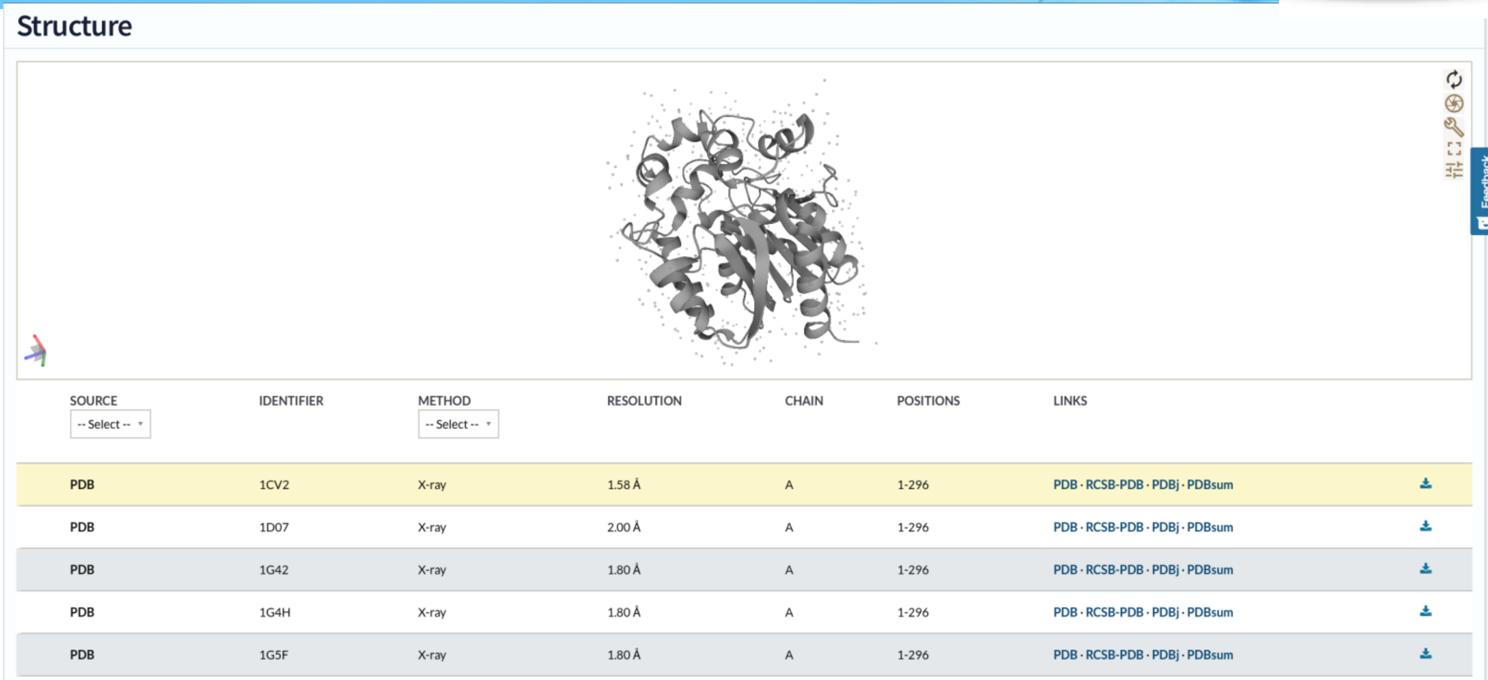
- Describe the expression conditions of the protein

Interaction:

- Refers to the **quaternary structure** of the protein
- Describes its native oligomeric state, and
- Lists interactions with other proteins

UniProt KB

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins



Structure

SOURCE -- Select --	IDENTIFIER	METHOD -- Select --	RESOLUTION	CHAIN	POSITIONS	LINKS
PDB	1CV2	X-ray	1.58 Å	A	1-296	PDB · RCSB-PDB · PDBj · PDBsum
PDB	1D07	X-ray	2.00 Å	A	1-296	PDB · RCSB-PDB · PDBj · PDBsum
PDB	1G42	X-ray	1.80 Å	A	1-296	PDB · RCSB-PDB · PDBj · PDBsum
PDB	1G4H	X-ray	1.80 Å	A	1-296	PDB · RCSB-PDB · PDBj · PDBsum
PDB	1G5F	X-ray	1.80 Å	A	1-296	PDB · RCSB-PDB · PDBj · PDBsum

Displays available **tertiary structures** (experimentally determined) for the protein.

Links to *AlphaFold* predictions if available (*cover later*)

Describes **secondary structure** content mapped to seq.

Links to databases with 3D structure models

UniProt KB

Function

Names & Taxonomy

Subcellular Location

Phenotypes

PTM/Processing

Expression

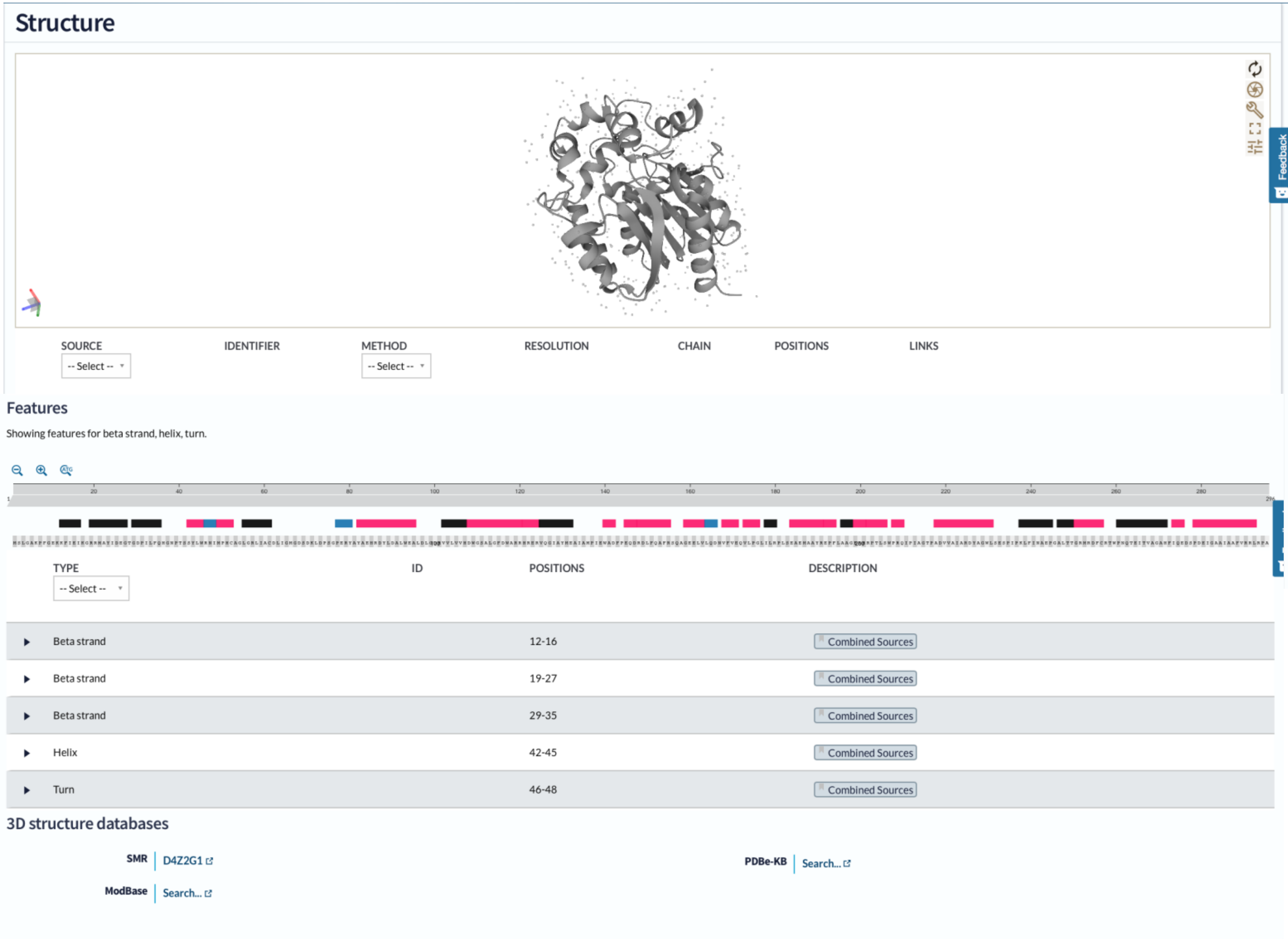
Interaction

Structure

Family & Domains

Sequence

Similar Proteins



Structure

SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
-- Select --		-- Select --				

Features

Showing features for beta strand, helix, turn.

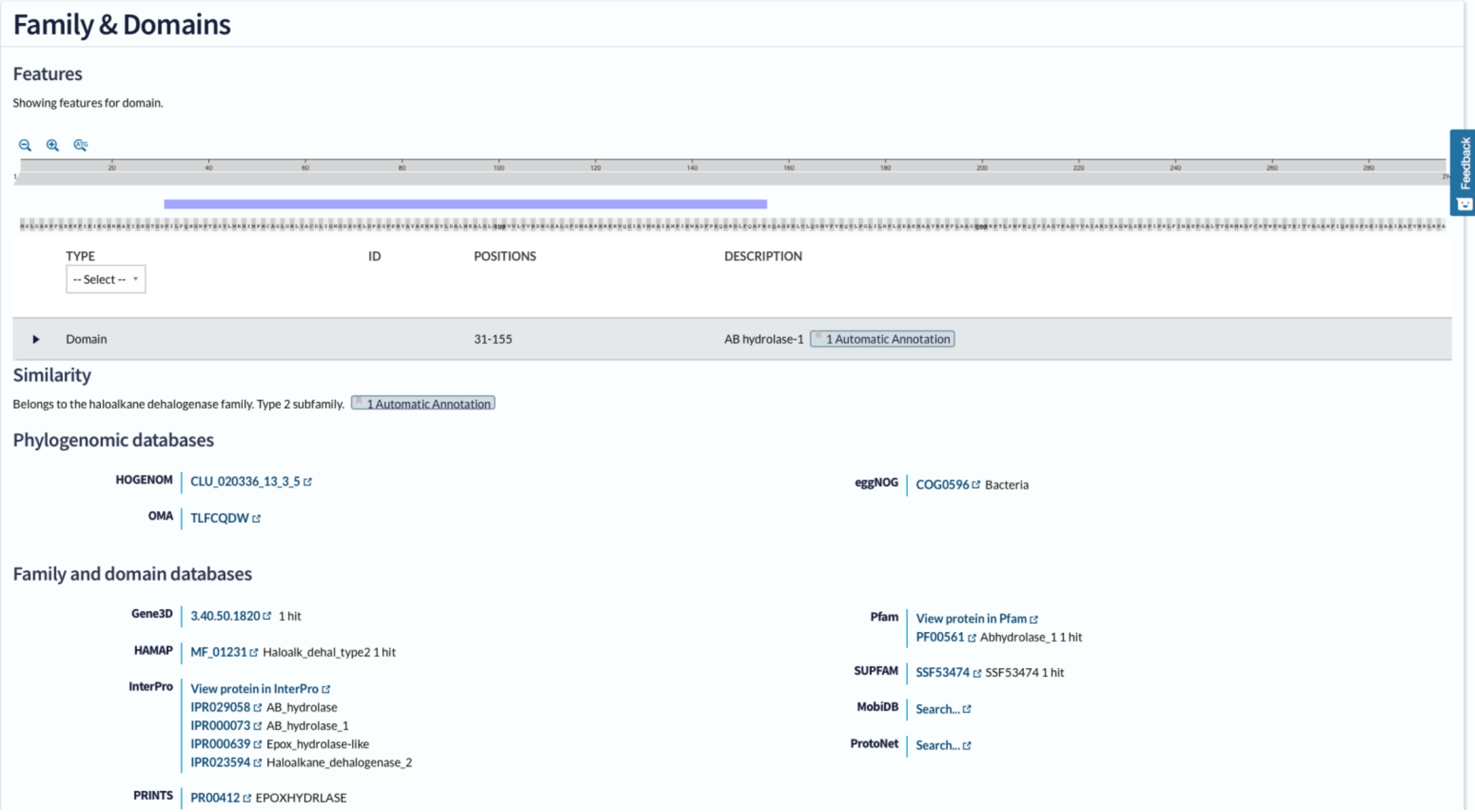
TYPE	ID	POSITIONS	DESCRIPTION
-- Select --			
▸ Beta strand		12-16	Combined Sources
▸ Beta strand		19-27	Combined Sources
▸ Beta strand		29-35	Combined Sources
▸ Helix		42-45	Combined Sources
▸ Turn		46-48	Combined Sources

3D structure databases

SMR	D4Z2G1	PDBe-KB	Search...
ModBase	Search...		

UniProt KB

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins



Family & Domains

Features

Showing features for domain.

TYPE	ID	POSITIONS	DESCRIPTION
-- Select --			
▸ Domain		31-155	AB hydrolase-1 1 Automatic Annotation

Similarity

Belongs to the haloalkane dehalogenase family. Type 2 subfamily. 1 Automatic Annotation

Phylogenomic databases

HOGENOM	CLU_020336_13_3_5	eggNOG	COG0596 Bacteria
OMA	TLFCQDW		

Family and domain databases

Gene3D	3.40.50.1820 1 hit	Pfam	View protein in Pfam PF00561 Abhydrolase_1 1 hit
HAMAP	MF_01231 Haloalk_dehal_type2 1 hit	SUPFAM	SSF53474 SSF53474 1 hit
InterPro	View protein in InterPro IPR029058 AB_hydrolase IPR000073 AB_hydrolase_1 IPR000639 Epox_hydrolase-like IPR023594 Haloalkane_dehalogenase_2	MobiDB	Search...
		ProtoNet	Search...
PRINTS	PR00412 EPOXHYDRLASE		

Cross-references to **motifs and profiles databases**

Convenient to find other proteins that share one particular sequence feature.

UniProt KB

Function

Names & Taxonomy

Subcellular Location

Phenotypes

PTM/Processing

Expression

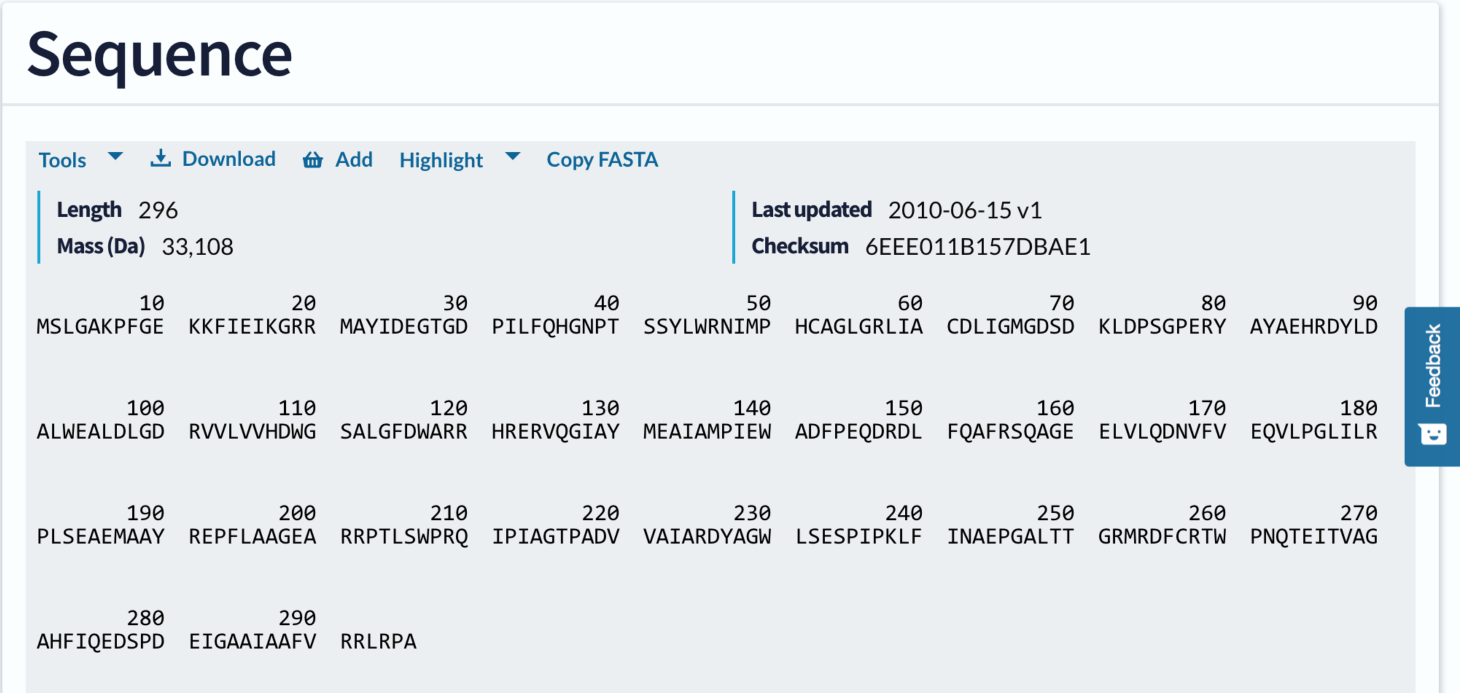
Interaction

Structure

Family & Domains

Sequence

Similar Proteins



Sequence

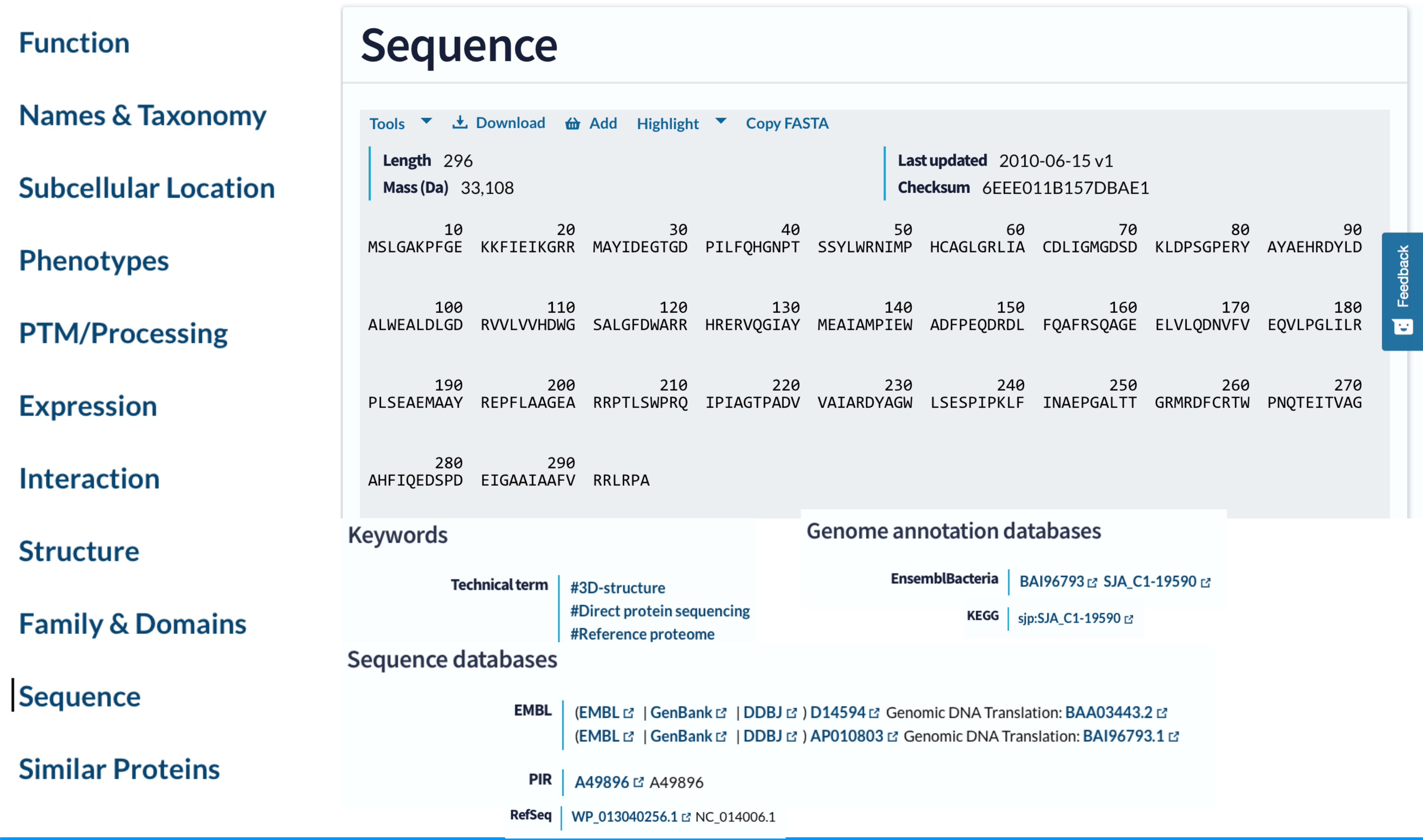
Length 296
Mass (Da) 33,108

Last updated 2010-06-15 v1
Checksum 6EEE011B157DBAE1

```
        10         20         30         40         50         60         70         80         90
MSLGAKPFGE KKFIEIKGRR MAYIDEGTGD PILFQHGNPT SSYLWRNIMP HCAGLGRLIA CDLIGMGDSD KLDPSGPERY AYAEHRDYLD

       100        110        120        130        140        150        160        170        180
ALWEALDLGD RVVLVVHDWG SALGFDWARR HRERVQGIAY MEAIAMPIEW ADFPEQDRDL FQAFRSQAGE ELVLQDNVFV EQVLPGLILR

       190        200        210        220        230        240        250        260        270
PLSEAEMAAY REPFLAAGEA RRPTLSWPRQ IPIAGTPADV VAIARDYAGW LSESPIPKLF INAEPGALTT GRMRDFCRTW PNQTEITVAG

       280        290
AHFIQEDSPD EIGAAIAAFV RRLRPA
```

When multiple ***isoforms*** are avaliable due to ***alternative splicing*** the different sequences are available here, with serialized accession codes (i.e. P21397-**1**, P21397-**2**)

UniProt KB



Function

Names & Taxonomy

Subcellular Location

Phenotypes

PTM/Processing

Expression

Interaction

Structure

Family & Domains

Sequence

Similar Proteins

Sequence

Length 296
Mass (Da) 33,108

Last updated 2010-06-15 v1
Checksum 6EEE011B157DBAE1

```
        10         20         30         40         50         60         70         80         90
MSLGAKPFGE KKFIEIKGRR MAYIDEGTGD PILFQHGNPT SSYLWRNIMP HCAGLGRLIA CDLIGMGDSD KLDPSGPERY AYAEHRDYLD

       100        110        120        130        140        150        160        170        180
ALWEALDLGD RVVLVVHDWG SALGFDWARR HRERVQGIAY MEAIAMPIEW ADFPEQDRDL FQAFRSQAGE ELVLQDNVFV EQVLPGLILR

       190        200        210        220        230        240        250        260        270
PLSEAEMAAY REPFLAAGEA RRPTLSWPRQ IPIAGTPADV VAIARDYAGW LSESPIPKLF INAEPGALTT GRMRDFCRTW PNQTEITVAG

       280        290
AHFIQEDSPD EIGAAIAAFV RRLRPA
```

Keywords

Technical term #3D-structure
#Direct protein sequencing
#Reference proteome

Genome annotation databases

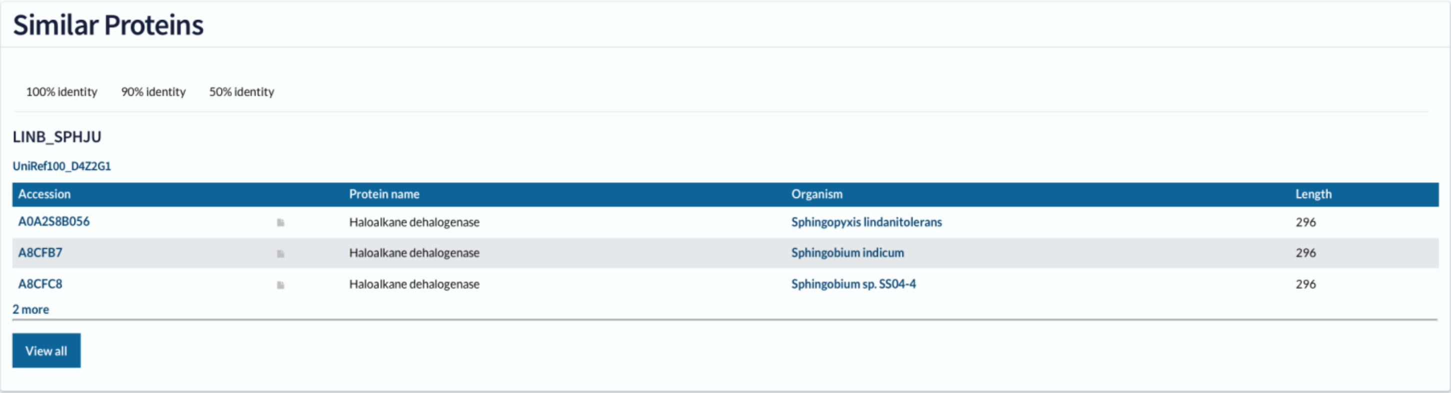
EnsemblBacteria	BAI96793 SJA_C1-19590
KEGG	sjp:SJA_C1-19590

Sequence databases

EMBL	(EMBL | GenBank | DDBJ) D14594 Genomic DNA Translation: BAA03443.2 (EMBL | GenBank | DDBJ) AP010803 Genomic DNA Translation: BAI96793.1
PIR	A49896 A49896
RefSeq	WP_013040256.1 NC_014006.1

UniProt KB

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins



Similar Proteins

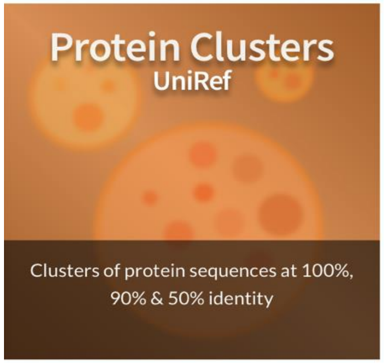
LINB_SPHJU

UniRef100_D4Z2G1

Accession		Protein name	Organism	Length
A0A2S8B056	▫	Haloalkane dehalogenase	Sphingopyxis lindanitolerans	296
A8CFB7	▫	Haloalkane dehalogenase	Sphingobium indicum	296
A8CFC8	▫	Haloalkane dehalogenase	Sphingobium sp. SS04-4	296

2 more

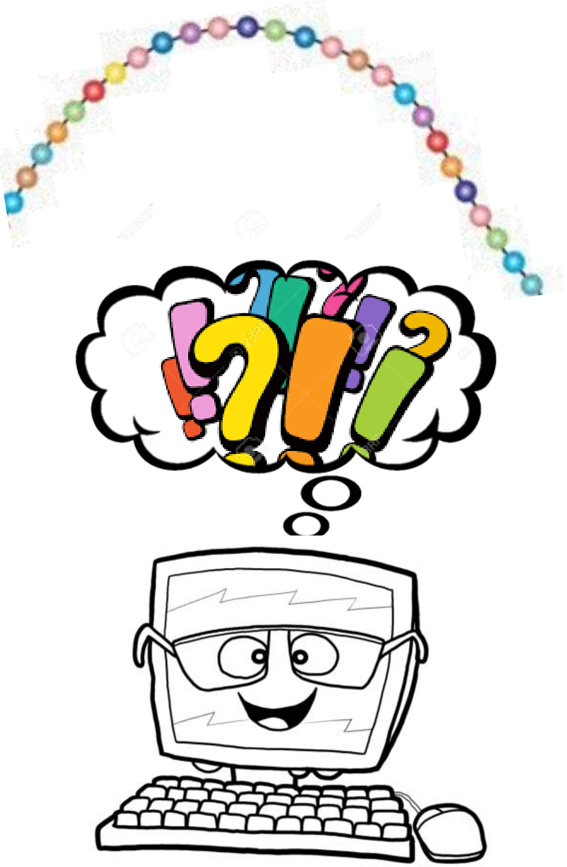
Retrieve groups of proteins that are 100%, up to 90%, or up to 50% identical



Uses for protein sequences

What can we do with protein sequences and computers?

Summary of 1D predictions

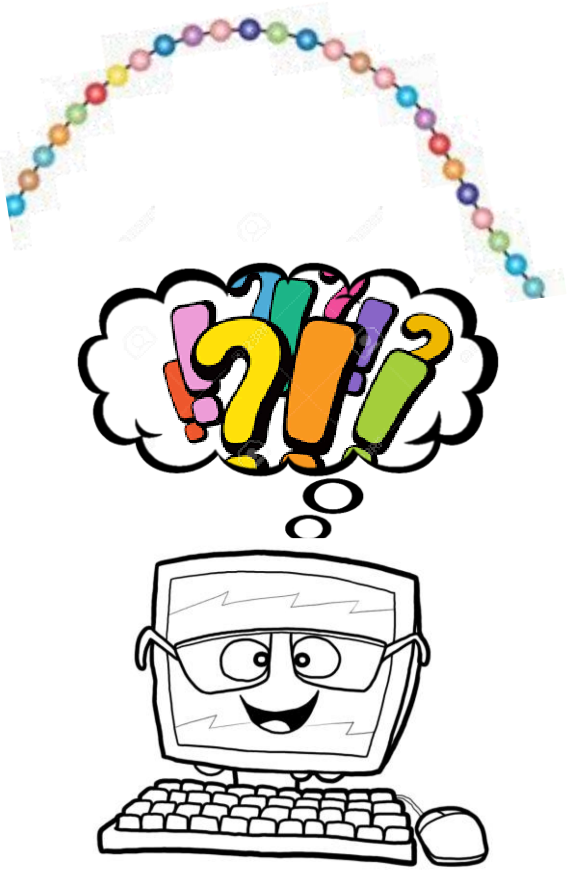


Different protein properties or characteristics can be predicted from its primary sequence:

- Secondary structure
- Solvent accessibility
- Solubility/expressability
- Transmembrane regions

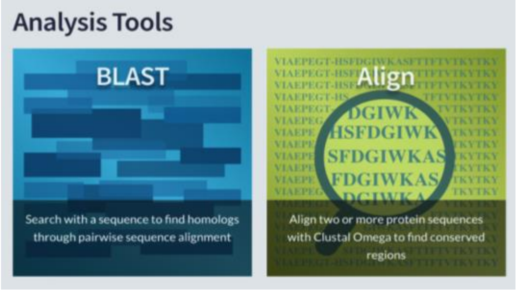
The methods that do such predictions improve if they consider ***evolutionary information***

Introduction to sequence alignment



Protein sequences can also be directly “compared” among them. Their similarities or differences can be assessed..

Alignments are models that aim to pair the most similar parts among different proteins.
If the model considers ***evolutionary information*** (and biologically relevant protein alignments do), evolutionary relationships (***homology***) can be inferred from sequence similarity.



Analysis Tools

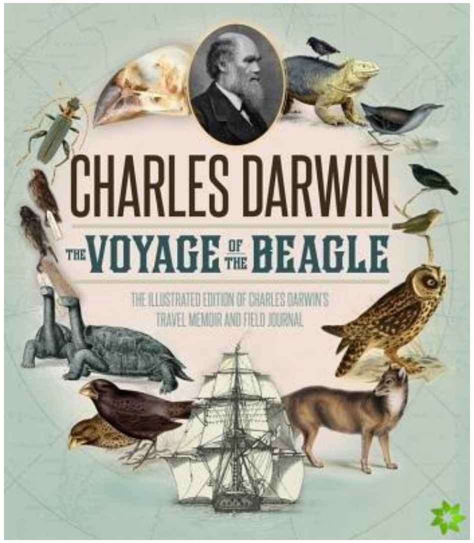
BLAST

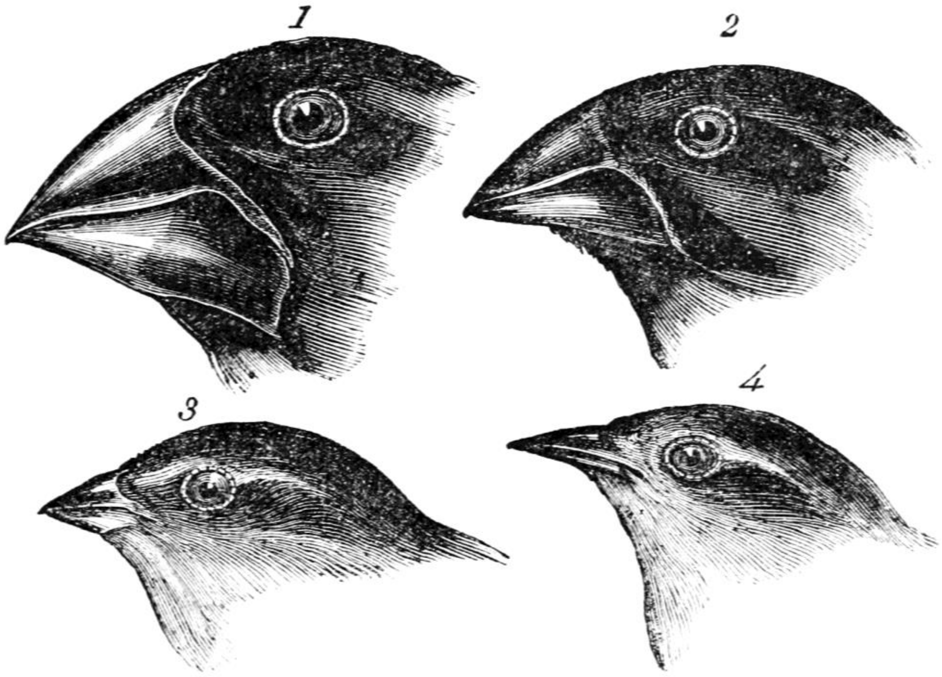
Search with a sequence to find homologs through pairwise sequence alignment

Align

Align two or more protein sequences with Clustal Omega to find conserved regions

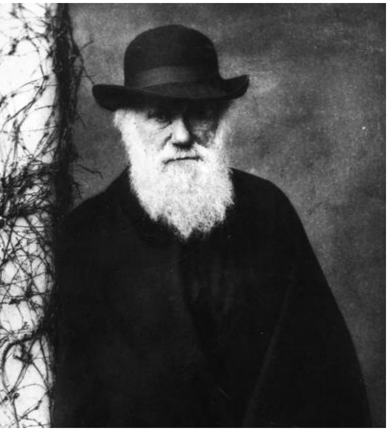
A few words on evolution


CHARLES DARWIN
THE VOYAGE OF THE BEAGLE
THE ILLUSTRATED EDITION OF CHARLES DARWIN'S TRAVEL MEMOIR AND FIELD JOURNAL



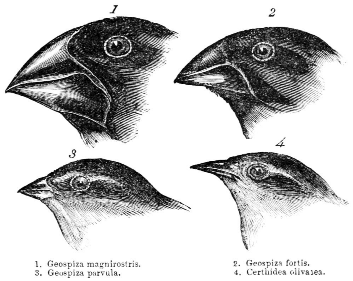
1. Geospiza magnirostris.
3. Geospiza parvula.

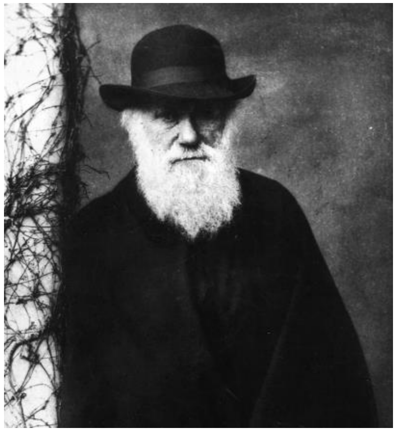
2. Geospiza fortis.
4. Certhidea olivasea.



”[…] one might really fancy that from an original paucity of birds in this archipelago, one species had been taken and modified for different ends."

A few words on evolution





Darwinian ideas on evolution:

All ***species*** of organisms arise and develop through the ***natural selection*** of ***small***, ***inherited variations*** that ***increase*** the ***individual's ability*** to compete, survive, and reproduce (***biological fitness***).

Inter-individual differences need to be:

- Small
- Inheritable

There exists a natural selective pressure.

Variations that make an individual fitter (**improve its functions**) to the conditions of the selective pressure are more likely to be transmitted to next generations.

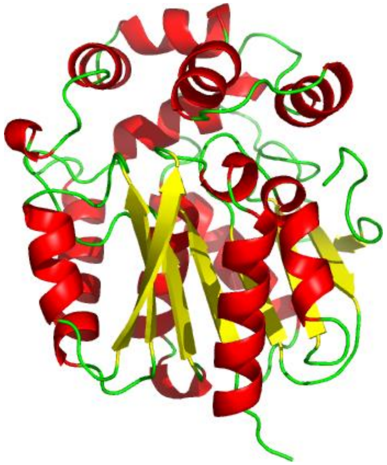
Accumulation of variation causes speciation.

A few words on **molecular** evolution

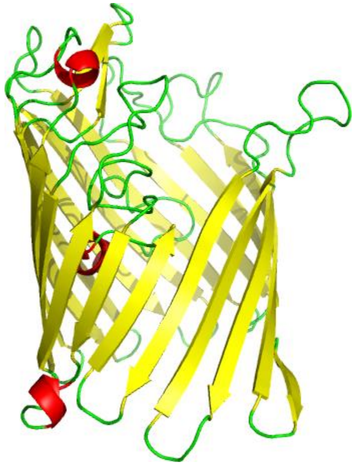
Improved function on a given environment (**adaptation**) is a key concept in evolution.

How does this apply to proteins?

How do proteins function?



Molecular Catalyst
[gift box]



Molecular Pore
[tube]

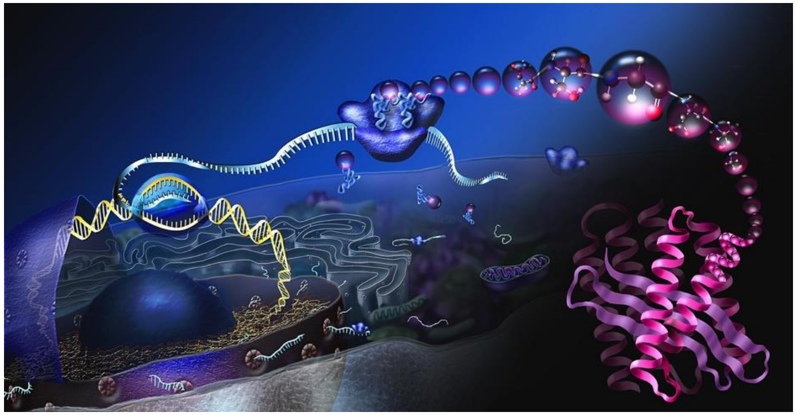
Function is dictated by shape (3D structure)

A few words on **molecular** evolution

Improved function on a given environment (**adaptation**) is a key concept in evolution.

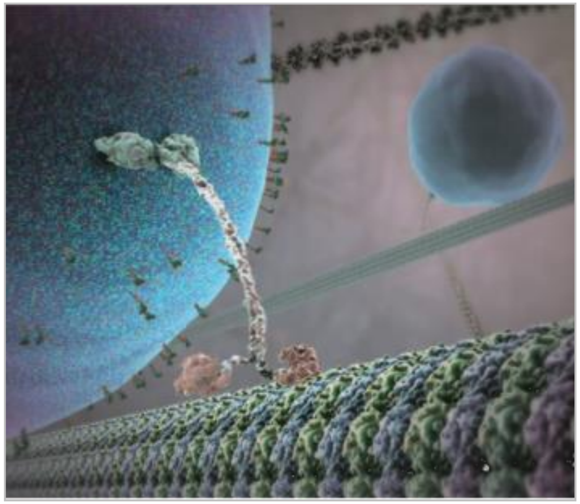
How does this apply to proteins?

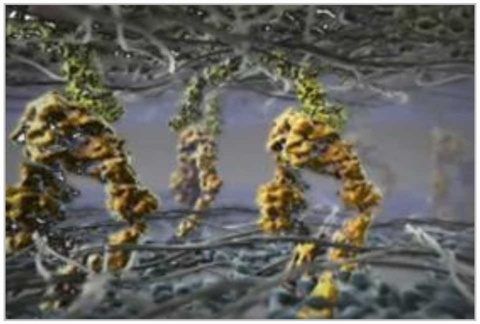
How do proteins function?

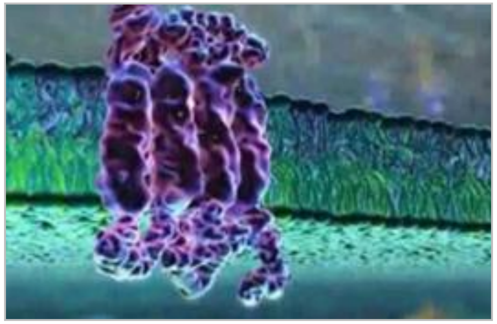


Structure is determined by sequence.

Function is dictated by shape (3D structure)



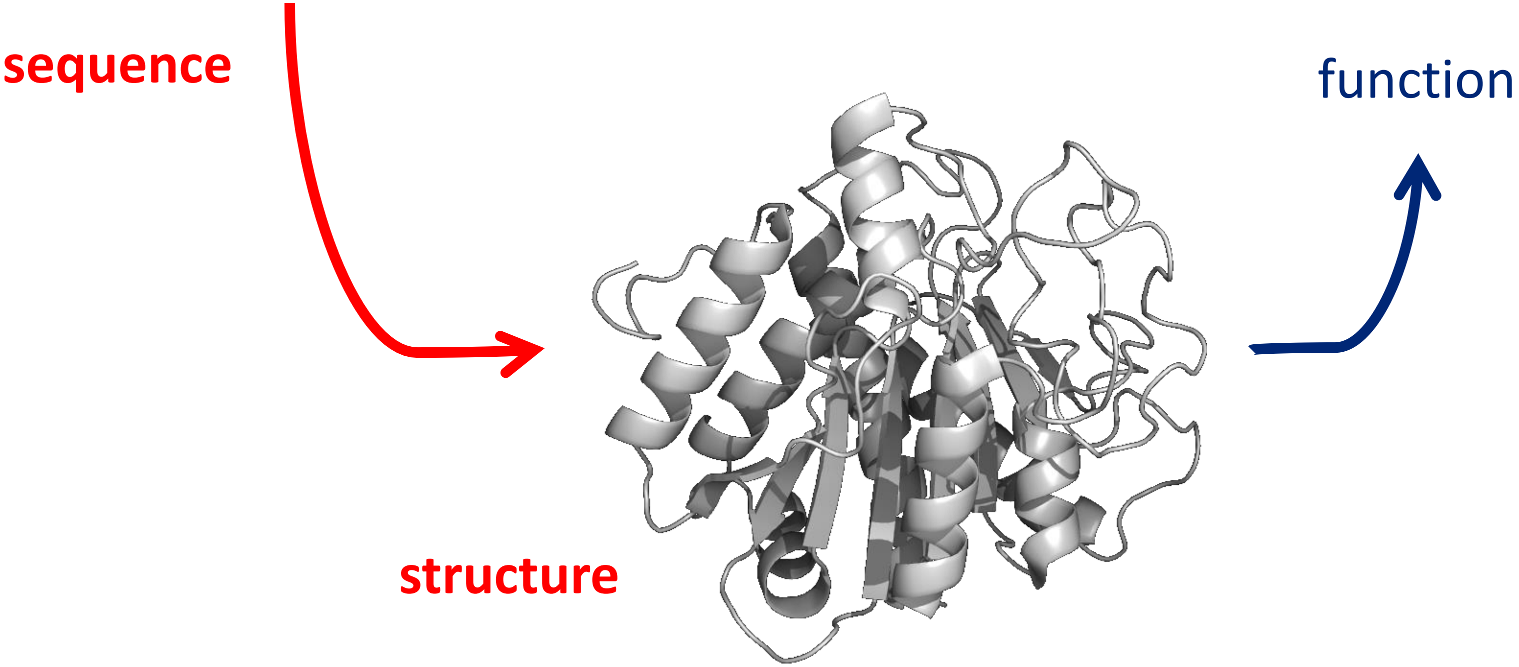




Sequence, Structure, Function Paradigm

- ❑ 3D structure is determined by the sequence
- ❑ Function is dictated by 3D structure

```
MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCA
GLGRLIACDLIGMGDSDKLDPSGPERYAYAEHRDYLDALWEALDLGDRVVLVV
HDWGSALGFDWARRHRERVQGIAYMEAIAMPIEWADFPEQDRDLFQAFRS
QAGEELVLQD
```



sequence

structure

function

A few words on **molecular** evolution

- Innovation happens at the sequence level
  - Mutations (***small changes***) introduced in DNA (***inheritable***)
  - Subsequently transcribed, processed, and translated into polypeptidic chains (proteins)
- ***Selective pressure*** operates at the function level
  - Proteins working ***better*** in their environments ***make individuals fitter***, adaptation occurred in human lineage

    Schaffner S. & Sabeti P (2008) Evolutionary adaptation in human lineage. Nature Education 1:14.

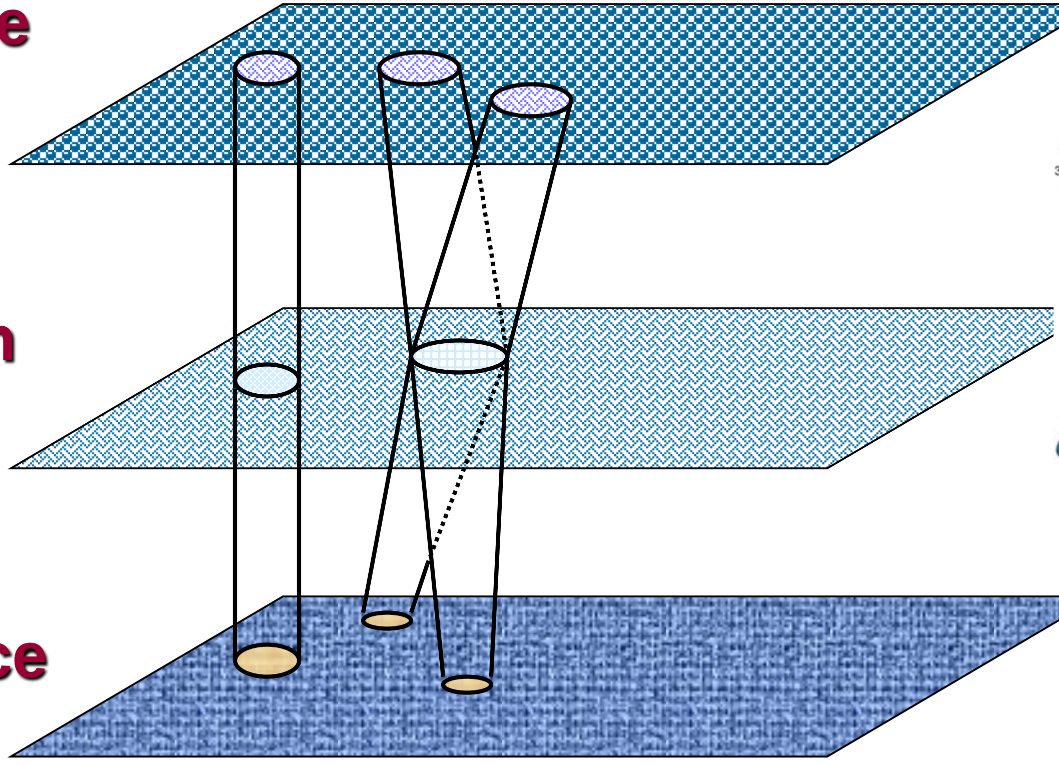
A few words on **molecular** evolution

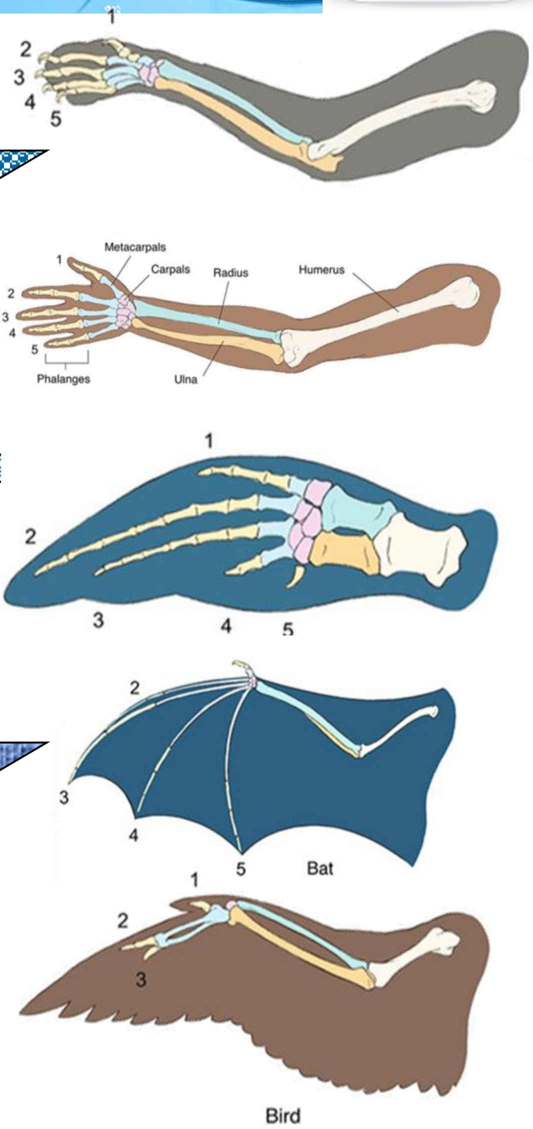
Diversity

Structure

Function

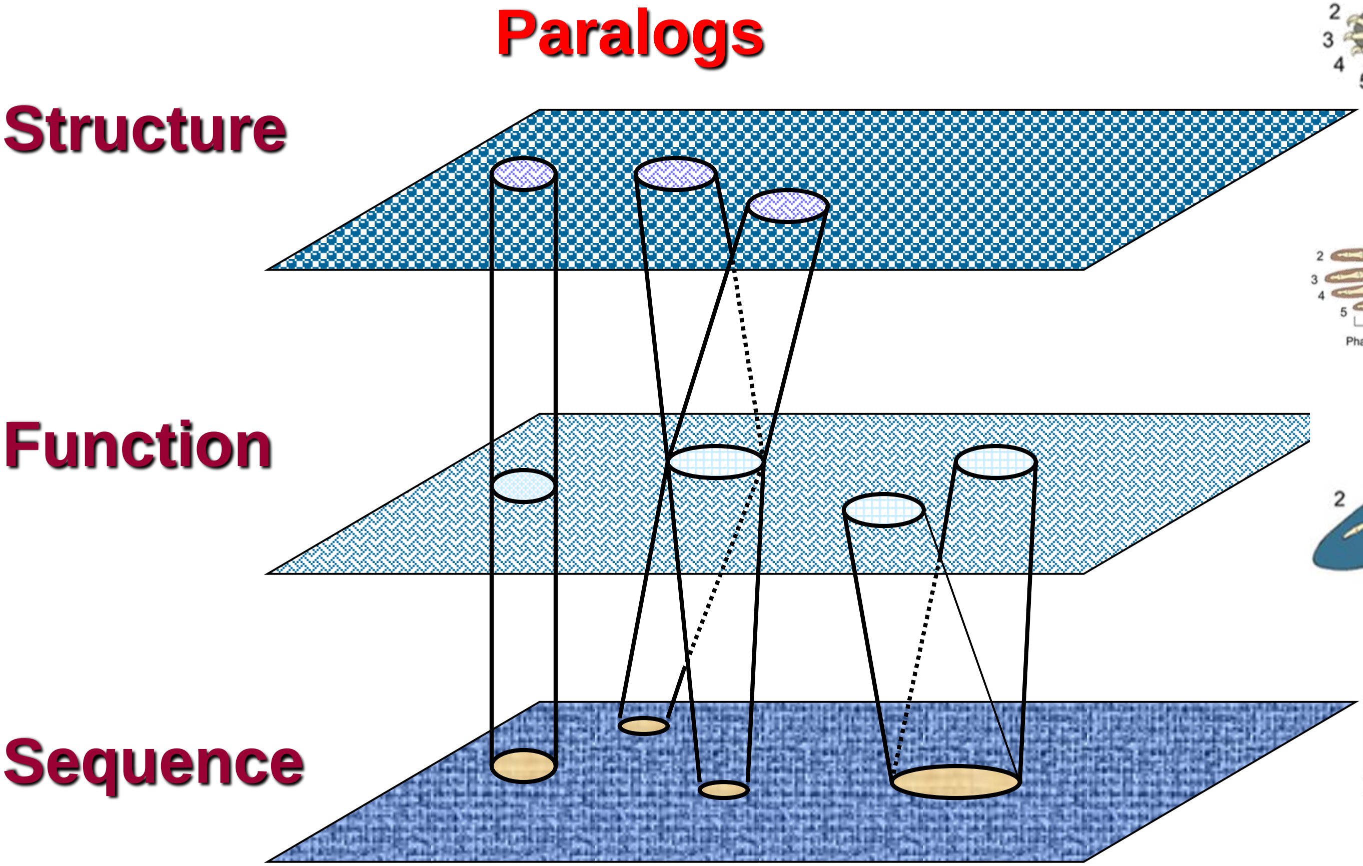
Sequence





Homology: two proteins are homologous if they are the products of genes that evolved from the same ancestor

A few words on **molecular** evolution

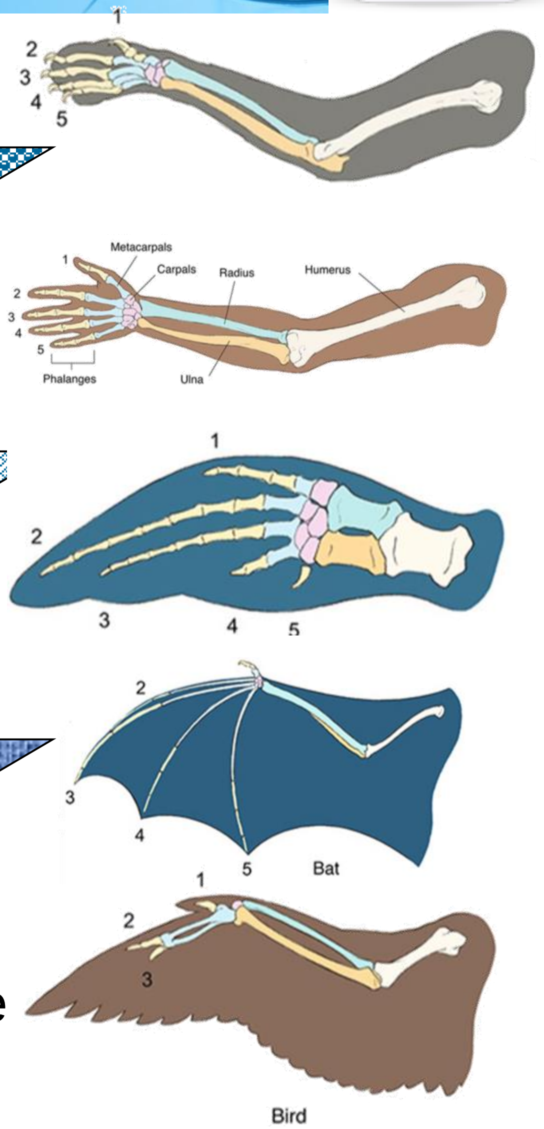


Paralogs

Structure

Function

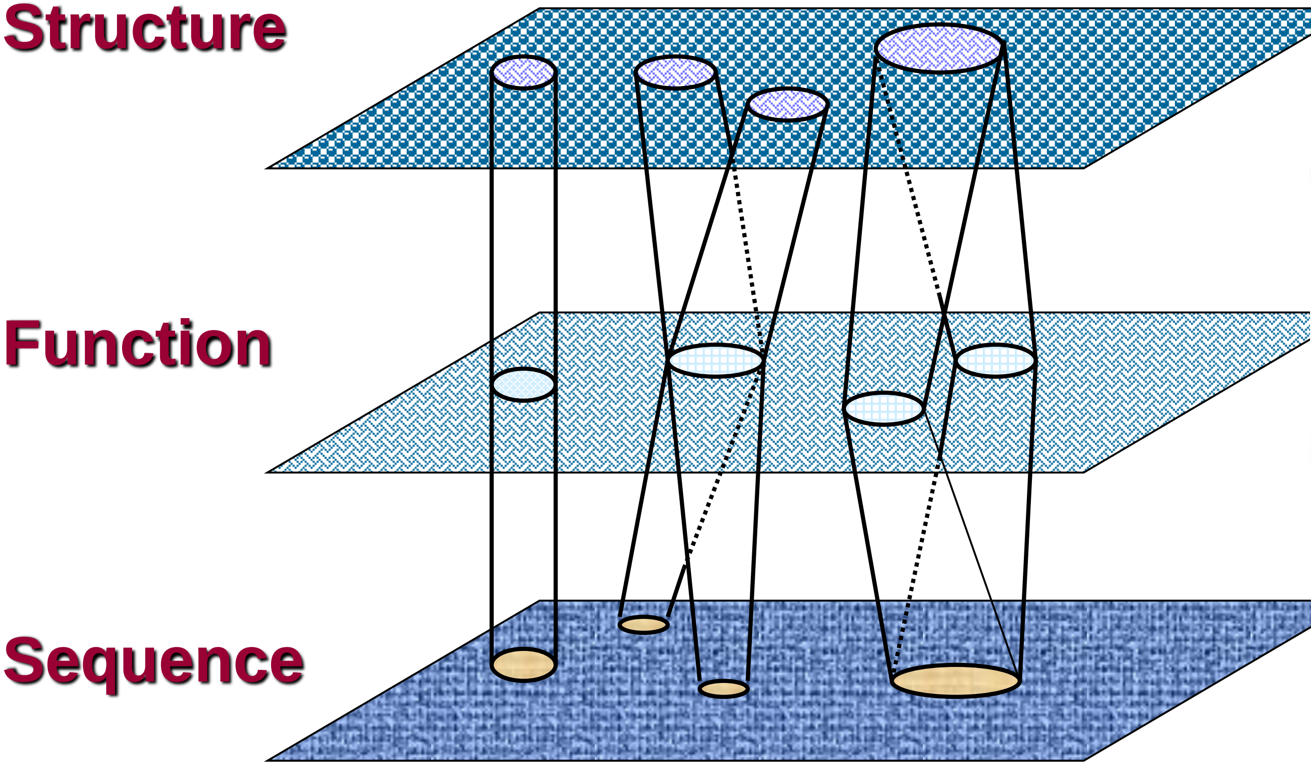
Sequence



Homology: two proteins are homologous if they are the products of genes that evolved from the same ancestor

A few words on **molecular** evolution

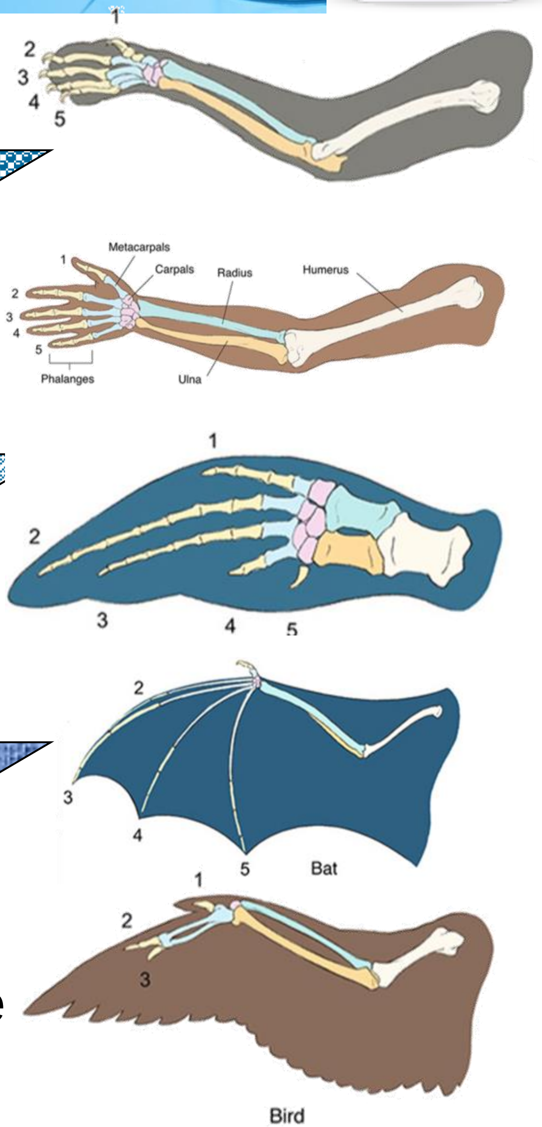
Annotation problem



Structure

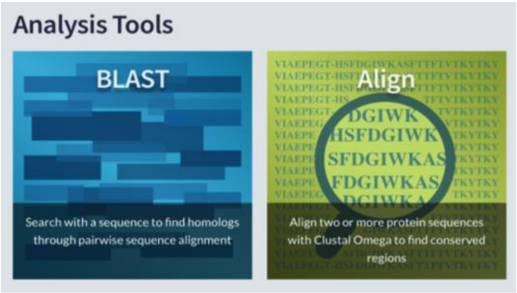
Function

Sequence



Homology: two proteins are homologous if they are the products of genes that evolved from the same ancestor

Sequence alignments



Analysis Tools

BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

Align two or more protein sequences with Clustal Omega to find conserved regions

Alignments are models that aim to pair the most similar parts among different proteins.

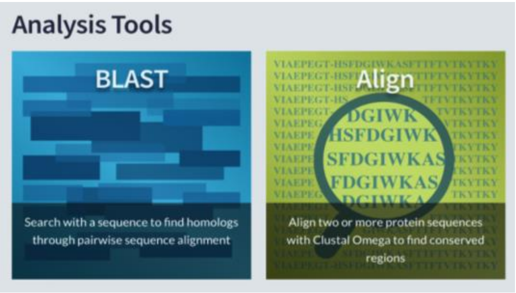
Global alignments: consider similarity across the entire sequence

Local alignments: consider similarity across sequence fragments

Pairwise alignments: two sequences compared

Multiple sequence alignments: multiple

Sequence alignments



Analysis Tools

BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

Align two or more protein sequences with Clustal Omega to find conserved regions

Alignments are models that aim to pair the most similar parts among different proteins.

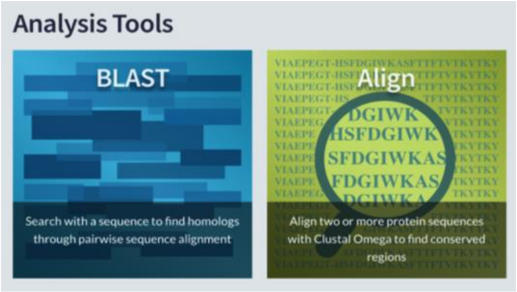
Pairwise alignment techniques

- DotPlot methods
- Dynamic programming algorithm
  - Needelman & Wunsch (Global)
  - Smith & Waterman (Local)
- Word methods

Multiple sequence alignment techniques:

- Dynamic programming
- Progressive methods
- Iterative methods

Sequence alignments



Analysis Tools

BLAST

Search with a sequence to find homologs through pairwise sequence alignment

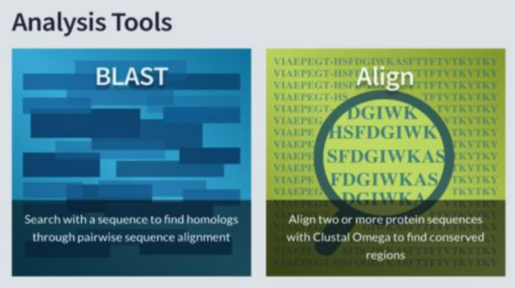
Align

Align two or more protein sequences with Clustal Omega to find conserved regions

Alignments are models that aim to pair the most similar parts among different proteins.

How can similarity among different parts of proteins be measured?

Sequence alignments



Analysis Tools

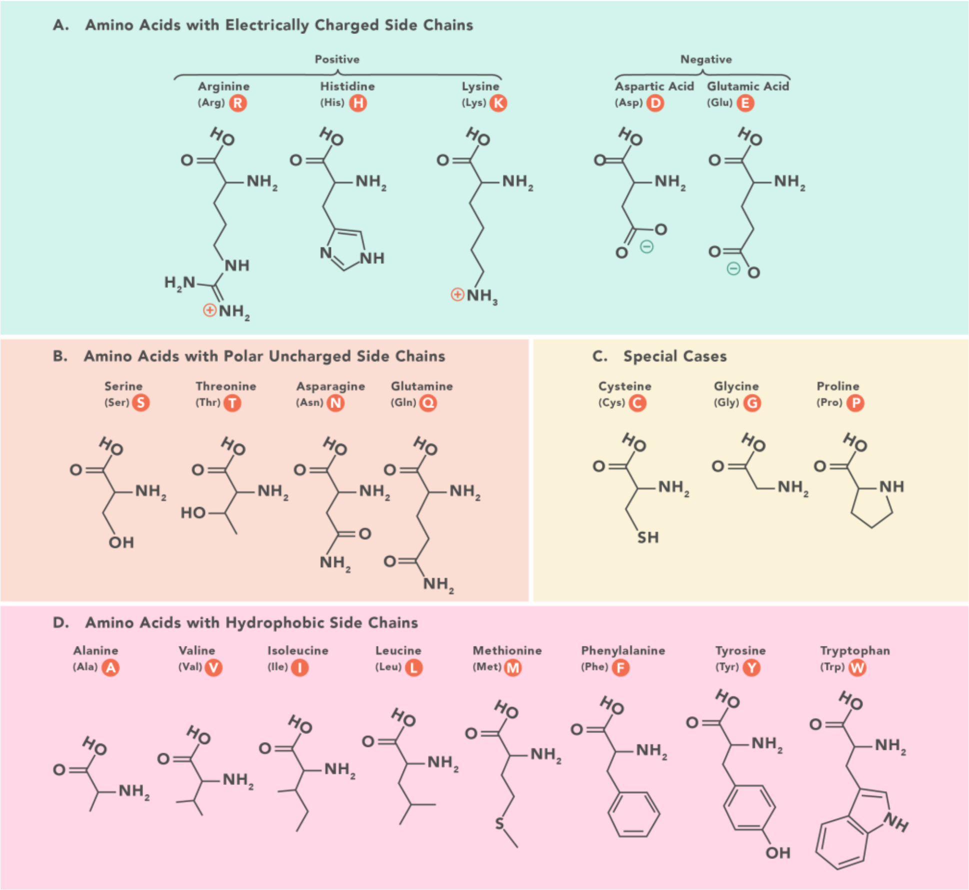
BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

Align two or more protein sequences with Clustal Omega to find conserved regions

Similitude in between amino-acids:

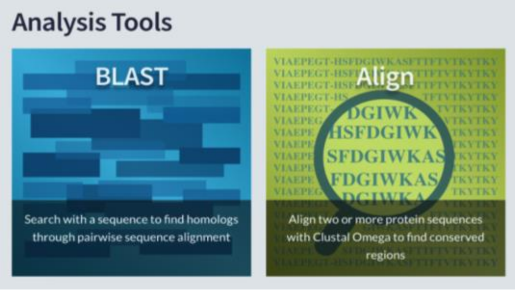


Sequence alignments

How can similarity among different parts of proteins be measured?

Assessing similarity in pairs of Amino-acids:

- Each possible pair of amino-acids is given a substitution score (substitution matrix)
- Amino-acids from the (two) sequences should be paired such as the total alignment score is optimized.
- Sometimes no good pairing can be found and a ***gap*** needs to be introduced.
- Gaps require a special penalty (negative score) in order to force longer and biologically meaningful alignments.



Analysis Tools

BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

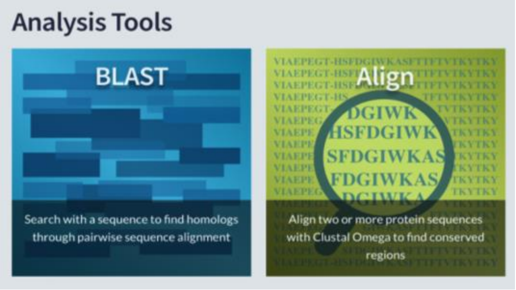
Align two or more protein sequences with Clustal Omega to find conserved regions

Sequence alignments

How can similarity among different parts of proteins be measured?

- Identity matrix (**Dot-matrix plots**):
  - 1 if same amino-acid
  - 0 otherwise

  → Limited model: forces the introduction of too many gaps.



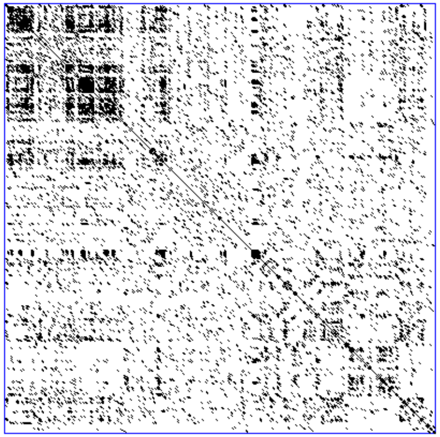
Analysis Tools

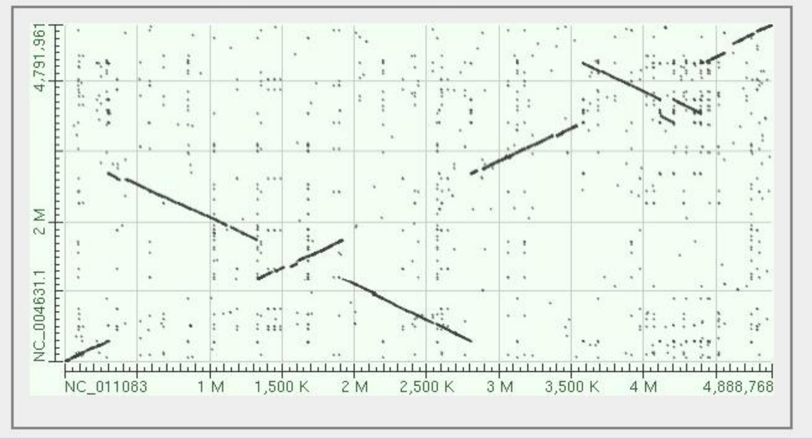
BLAST

Search with a sequence to find homologs through pairwise sequence alignment

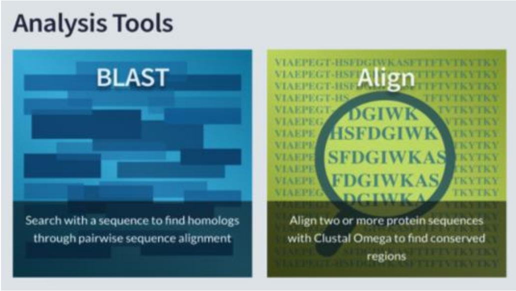
Align

Align two or more protein sequences with Clustal Omega to find conserved regions





Sequence alignments



Analysis Tools

BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

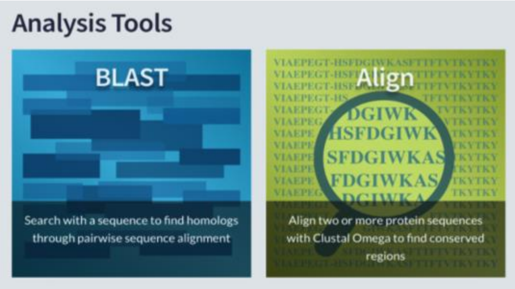
Align two or more protein sequences with Clustal Omega to find conserved regions

How can similarity among different parts of proteins be measured?

- Identity matrix (Dot-matrix plots):
  - 1 if same amino-acid
  - 0 otherwise

  → Limited model: forces the introduction of too many gaps.
- Substitution models:
  - Score depending on the probability of observing a substitution (mutation) of one particular Aa for another (i.e. Arg → Lys should score better than Arg → Glu)

Sequence alignments



Analysis Tools

BLAST

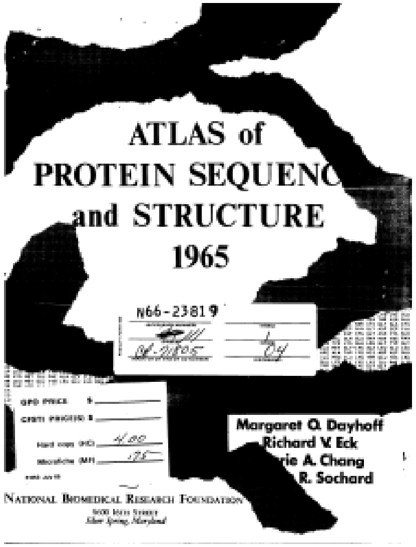
Search with a sequence to find homologs through pairwise sequence alignment

Align

Align two or more protein sequences with Clustal Omega to find conserved regions

Substitution models include evolutionary information

Margaret Dayhoff
Atlas of protein sequence and structure



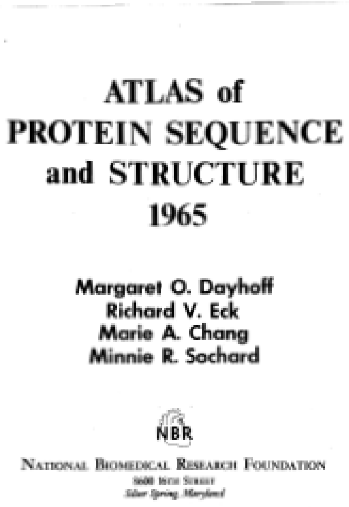
ATLAS of
PROTEIN SEQUENC
and STRUCTURE
1965

N66-23819

GPO PRICE $
CFSTI PRICE(S) $
Hard copy (HC) 4.00
Microfiche (MF) .75

Margaret O. Dayhoff
Richard V. Eck
rie A. Chang
R. Sochard

NATIONAL BIOMEDICAL RESEARCH FOUNDATION
8600 16TH STREET
Silver Spring, Maryland

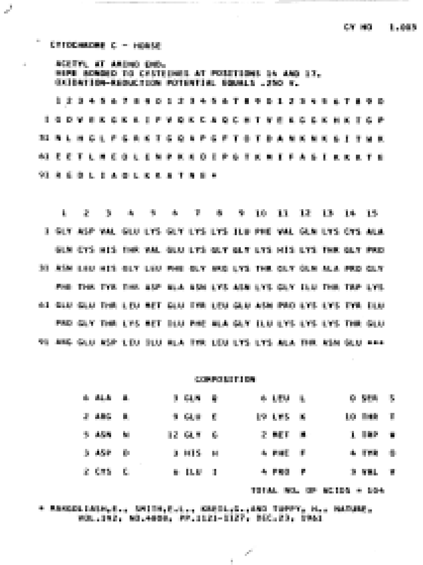


ATLAS of
PROTEIN SEQUENCE
and STRUCTURE
1965

Margaret O. Dayhoff
Richard V. Eck
Marie A. Chang
Minnie R. Sochard

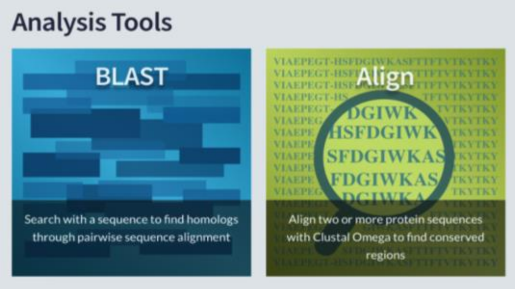
NBR

NATIONAL BIOMEDICAL RESEARCH FOUNDATION
8600 16TH STREET
Silver Spring, Maryland



```
                                                    CY HO   1.085
 CYTOCHROME C - HORSE

  ACETYL AT AMINO END.
  HEME BONDED TO CYSTEINES AT POSITIONS 14 AND 17.
  OXIDATION-REDUCTION POTENTIAL EQUALS .250 V.

   1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0

 1 G D V E K G K K I F V Q K C A Q C H T V E K G G K H K T G P
31 N L H G L F G R K T G Q A P G F T Y T D A N K N K G I T W K
61 E E T L M E Y L E N P K K Y I P G T K M I F A G I K K K T E
91 R E D L I A Y L K K A T N E *

    1   2   3   4   5   6   7   8   9   10  11  12  13  14  15
 1 GLY ASP VAL GLU LYS GLY LYS LYS ILU PHE VAL GLN LYS CYS ALA
   GLN CYS HIS THR VAL GLU LYS GLY GLY LYS HIS LYS THR GLY PRO
31 ASN LEU HIS GLY LEU PHE GLY ARG LYS THR GLY GLN ALA PRO GLY
   PHE THR TYR THR ASP ALA ASN LYS ASN LYS GLY ILU THR TRP LYS
61 GLU GLU THR LEU MET GLU TYR LEU GLU ASN PRO LYS LYS TYR ILU
   PRO GLY THR LYS MET ILU PHE ALA GLY ILU LYS LYS LYS THR GLU
91 ARG GLU ASP LEU ILU ALA TYR LEU LYS LYS ALA THR ASN GLU ***

                          COMPOSITION

        6 ALA  A       3 GLN  Q       6 LEU  L       0 SER  S
        2 ARG  R       9 GLU  E      19 LYS  K      10 THR  T
        5 ASN  N      12 GLY  G       2 MET  M       1 TRP  W
        3 ASP  D       3 HIS  H       4 PHE  F       4 TYR  Y
        2 CYS  C       6 ILU  I       4 PRO  P       3 VAL  V
                                        TOTAL NO. OF ACIDS = 104

 * MARGOLIASH,E., SMITH,E.L., KREIL,G., AND TUPPY, H., NATURE,
      VOL.192, NO.4808, PP.1121-1127, DEC.23, 1961
```

Sequence alignments



Analysis Tools

BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

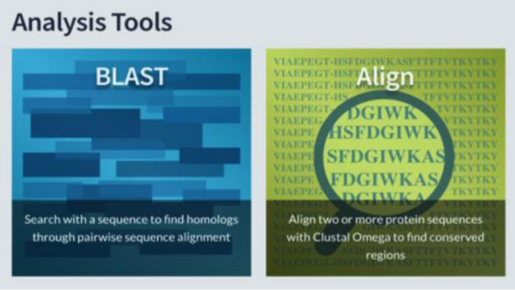
Align two or more protein sequences with Clustal Omega to find conserved regions

Substitution models include evolutionary information

Dayhoff Mutation Data Matrix

- Score is based on the concept of **Point Accepted Mutation (PAM)**
- Evolutionary distance 1 PAM = time in which 1/100 amino acids are expected to mutate.
- Higher evolutionary times inferred from a Markov chain model: PAM matrix product.
- 250 PAM matrix – targets the limit where is safe to infer homology in proteins (*twilight*).
- Limitation: derived from 1572 observed mutations in (manual) alignment of sequences >85% identical

Sequence alignments



Analysis Tools

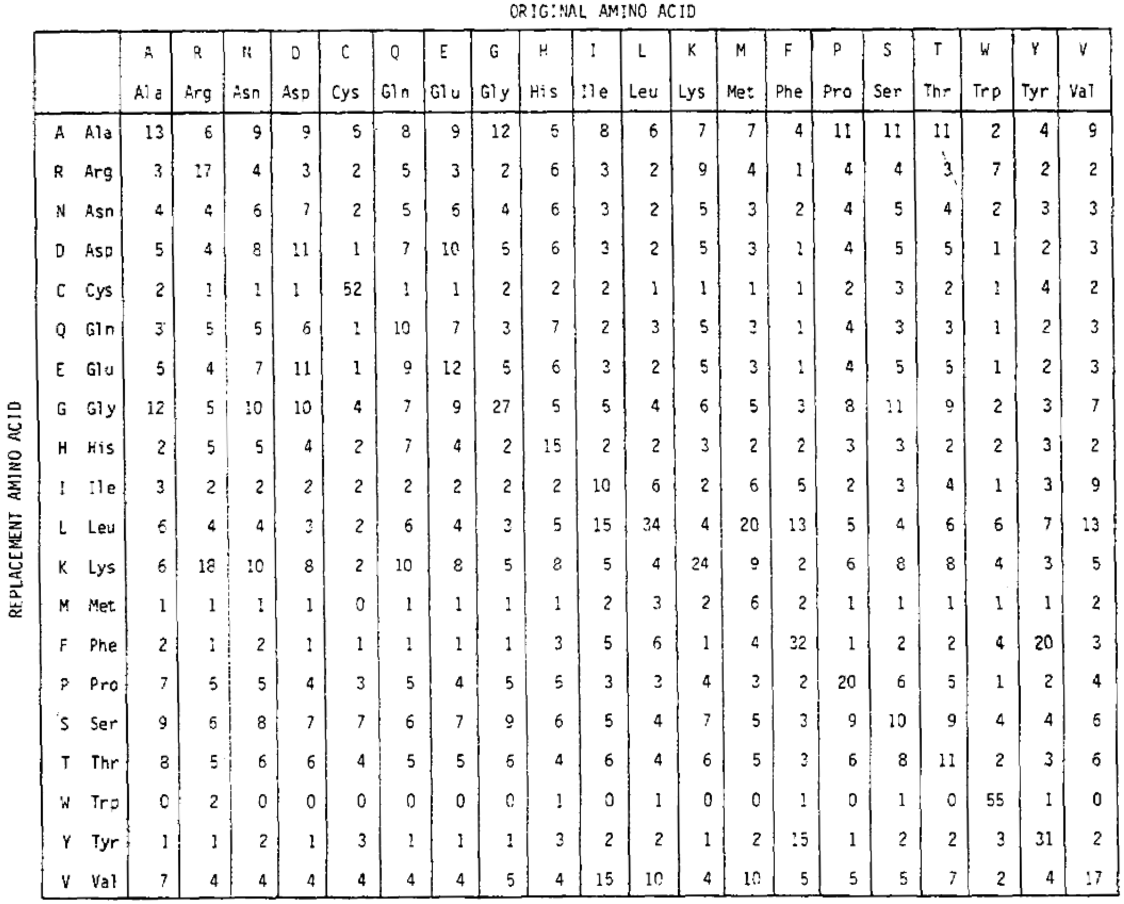
BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

Align two or more protein sequences with Clustal Omega to find conserved regions

Substitution models include evolutionary information PAM250



ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID

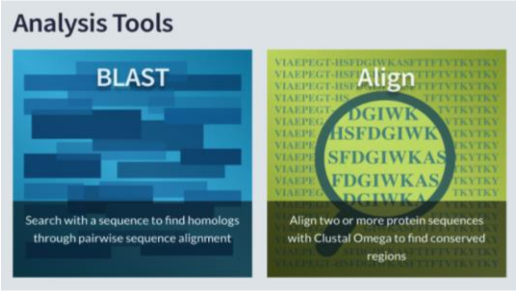
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A	Ala	13	6	9	9	5	8	9	12	5	8	6	7	7	4	11	11	11	2	4	9
R	Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	Gln	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	Pro	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Sequence alignments

Substitution models include evolutionary information

BLOSSUM matrices

- BLOcks SUbstitution Matrix
- Derived from blocks of aligned sequences in BLOCKS database – implicitly represents distant relationships.
- bias from identical sequences is removed by clustering at a sequence identity threshold
- BLOSUM62 = matrix derived from sequences clustered at 62% or greater identity



Analysis Tools

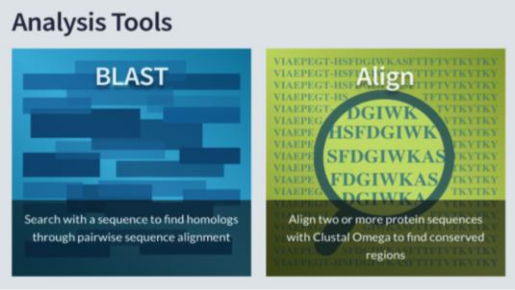
BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

Align two or more protein sequences with Clustal Omega to find conserved regions

Sequence alignments



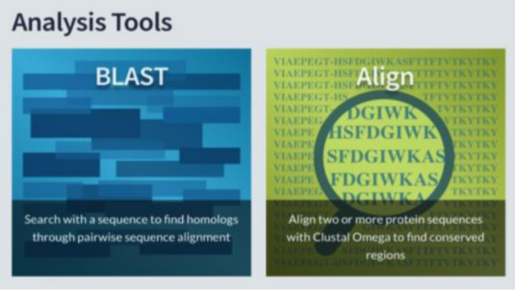
Analysis Tools

Search with a sequence to find homologs through pairwise sequence alignment

Align two or more protein sequences with Clustal Omega to find conserved regions

PAM	BLOSUM
Similar proteins compared as whole	Conserved BLOKS (fragments) compared
PAM1 corresponds to 1 ≠ residue in 100 → 99% ID	BLOSUM1 corresponds to 1% ID
Other PAM matrices extrapolated from PAM1	Each matrix based on observed alignments
Higher numbers, more evolutionary distance	Higher numbers, more similarity (less evolutionary distance)
100	90
120	80
160	62
200	50
250	45

Sequence alignments



Analysis Tools

BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

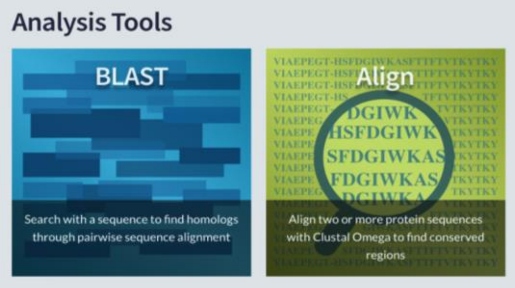
Align two or more protein sequences with Clustal Omega to find conserved regions

Dynamic Programing Algorithm

Matrix:

- Each dimension corresponds to one of the proteins to be aligned.
- Each cell contains the score value from the substitution model corresponding to the residue pair.
- Diagonal transitions represent aligned positions
- Vertical and horizontal transitions represent gaps and are penalized.
- The final alignment corresponds to the path in the matrix that maximizes the score.

Sequence alignments



Analysis Tools

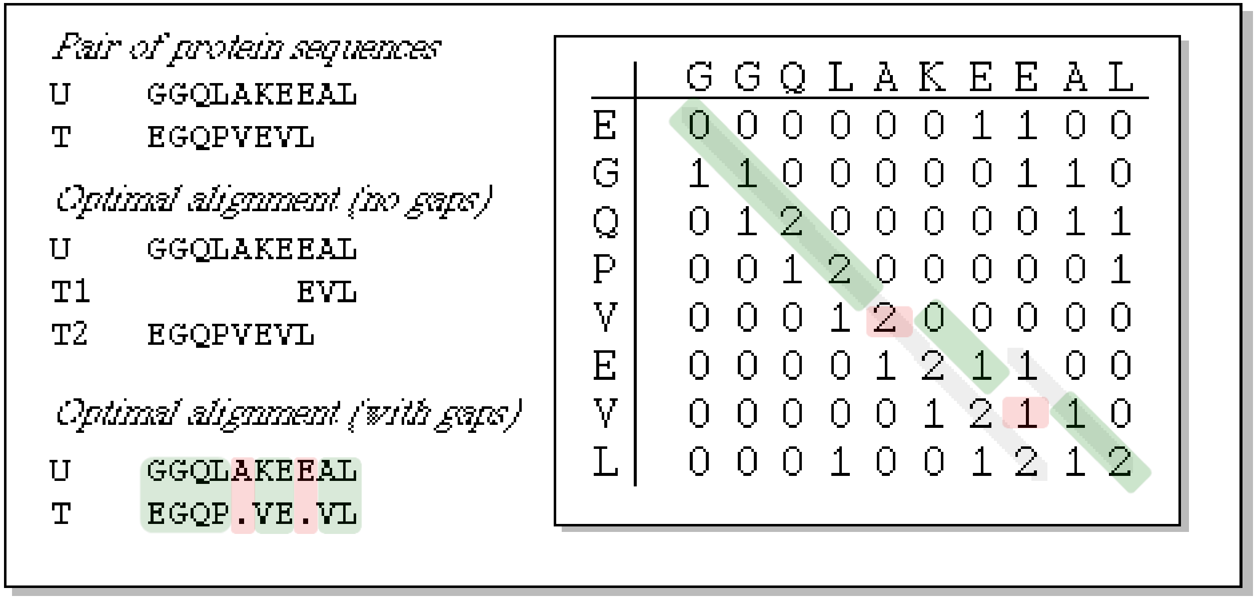
BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

Align two or more protein sequences with Clustal Omega to find conserved regions

Dynamic Programing Algorithm



Pair of protein sequences

```
U     GGQLAKEEAL
T     EGQPVEVL
```

Optimal alignment (no gaps)

```
U     GGQLAKEEAL
T1           EVL
T2    EGQPVEVL
```

Optimal alignment (with gaps)

```
U     GGQLAKEEAL
T     EGQP.VE.VL
```

	G	G	Q	L	A	K	E	E	A	L
E	0	0	0	0	0	0	1	1	0	0
G	1	1	0	0	0	0	0	1	1	0
Q	0	1	2	0	0	0	0	0	1	1
P	0	0	1	2	0	0	0	0	0	1
V	0	0	0	1	2	0	0	0	0	0
E	0	0	0	0	1	2	1	1	0	0
V	0	0	0	0	0	1	2	1	1	0
L	0	0	0	1	0	0	1	2	1	2

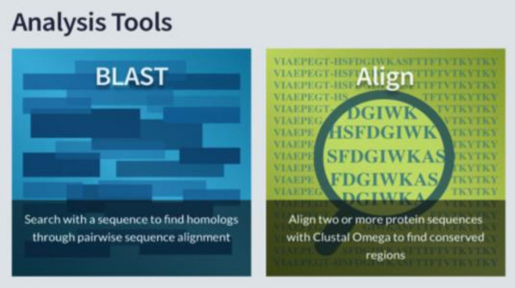
Back-trace from bottom-right

Global: Needelman & Wunsch. From the corner

Local: Smith & Waterman. From any position.

☺ **DETERMINISTIC** ☹ **Comp. expensive**

Sequence alignments



Analysis Tools

BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

Align two or more protein sequences with Clustal Omega to find conserved regions

Word methods

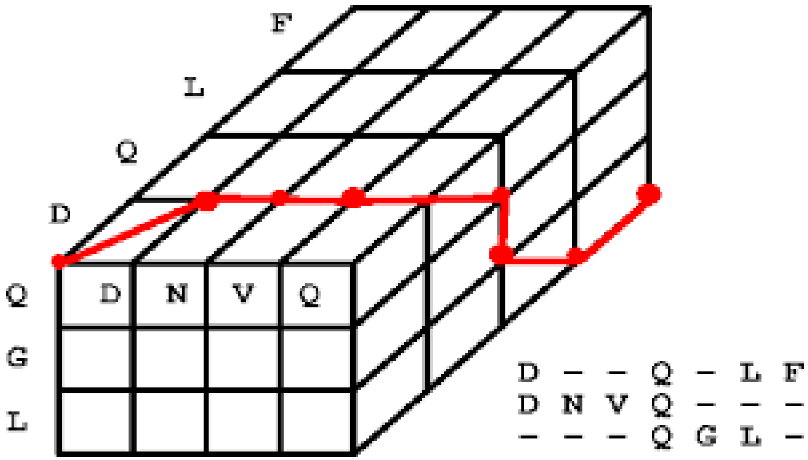
- Short non-overlapping sequence stretches (k-tuples or ***words***) are identified in the ***query*** sequence and matched in ***target*** sequence(s).
- Relative positions of the matching region define an ***offset*** (subtraction)
- Multiple words matching with similar offset define a region prone to alignment.
- Alignments are subsequently extended in alingment-prone regions.

- ☹ **HEURISTIC**, optimal align not guaranteed.
- ☺ Efficient for database searches.
- BLAST, FASTA.

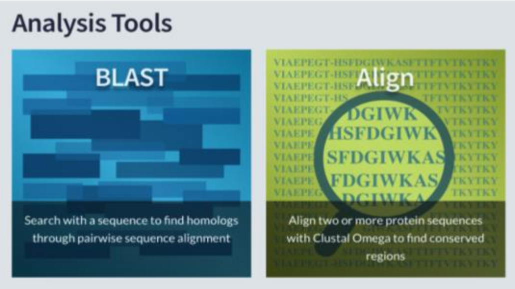
Sequence alignments

Multiple sequence alignments

- Dynamic programming algorithm (N-dimensional matrix)



```
D – – Q – L F
D N V Q – – –
– – – Q G L –
```



Analysis Tools

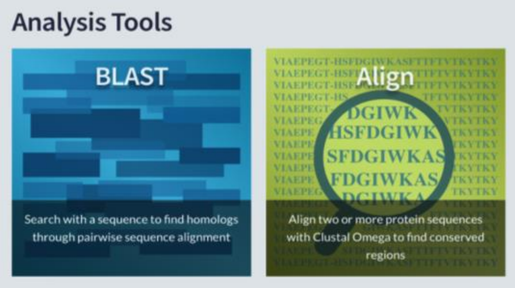
BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

Align two or more protein sequences with Clustal Omega to find conserved regions

Sequence alignments



Analysis Tools

BLAST

Search with a sequence to find homologs through pairwise sequence alignment

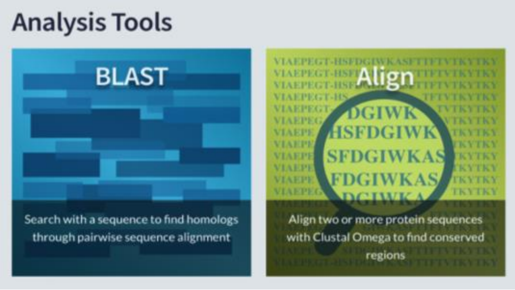
Align

Align two or more protein sequences with Clustal Omega to find conserved regions

Multiple sequence alignments

- **Dynamic programming algorithm**
- **Progressive methods**
  - First align the most similar pair
  - Subsequently add less similar sequences
  - Sensitive to similarity inaccuracy (i.e. due to differences in sequence length)
  - CLUSTAL
  - Additional info considered: T-Coffee (slow)
- **Iterative methods**

Sequence alignments



Analysis Tools

BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

Align two or more protein sequences with Clustal Omega to find conserved regions

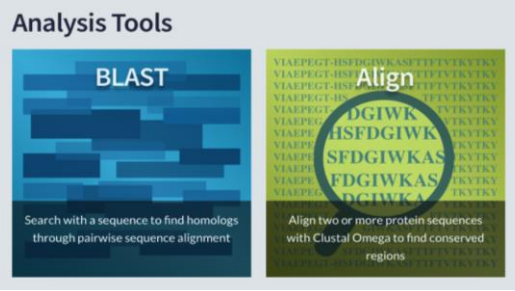
Multiple sequence alignments

- **Dynamic programming algorithm**
- **Progressive methods**
- **Iterative methods**
  - Initial global alignment
  - Objective function (based on score) to optimise similarity assessment. Chose best.
  - All possible remaining sequence subsets re-aligned and re-scored
  - Best subset included in the alignment/iter.
  - Typically slower, more accurate
  - MUSCLE, MAFT.

Sequence alignments

Beyond pure sequences: patterns and models

- Aligned sequences can be used to define patterns, that can then be used to perform searches in databases.
- Position Specific Scoring Matrices
- Hidden Markov Models



Analysis Tools

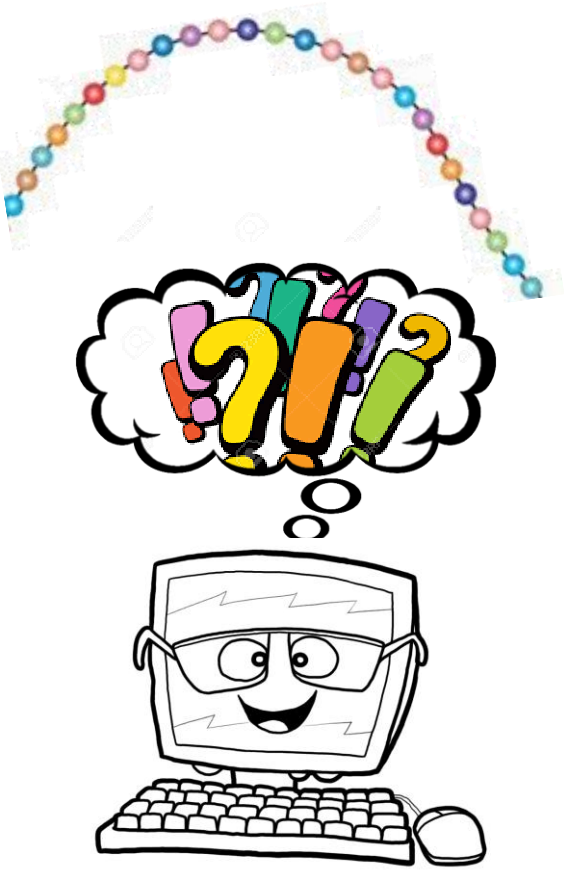
BLAST

Search with a sequence to find homologs through pairwise sequence alignment

Align

Align two or more protein sequences with Clustal Omega to find conserved regions

Summary of 1D predictions



Different protein properties or characteristics can be predicted from its primary sequence:

- Secondary structure
- Solvent accessibility
- Solubility/expressability
- Transmembrane regions

The methods that do such predictions improve if they consider ***evolutionary information***

Secondary structure prediction

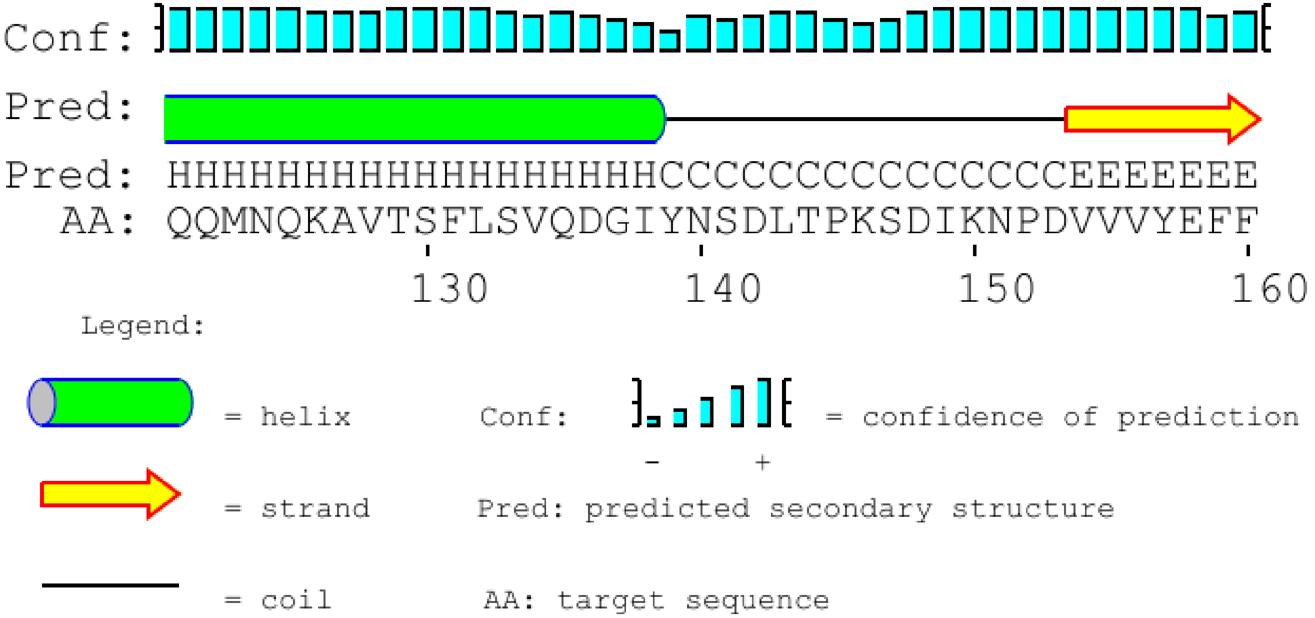
- prediction of the conformational state of each amino acid (AA) residue of a protein sequence as one of the possible states:
  - helix (H)
  - strand (S)
  - coil (C)

Secondary structure prediction

- amino acid propensities derived from known 3D structures
  - probability of a particular AA for a particular secondary structure state
  - first-generation methods – low accuracy
- propensities of segments of adjacent residues
  - local environment of residues considered (3-51 consecutive residues)
  - second-generation methods – accuracy ~ 60 % - 65 %
- evolutionary information combined with machine learning
  - training set – sequence profiles associated with a particular secondary structure arrangement (based on known 3D structures)
  - sequence profiles derived from family sequence alignments
  - state-of-the-art methods – accuracy ~ 70 % - 80 %

Secondary structure prediction programs

- PSI-PRED
  - http://bioinf.cs.ucl.ac.uk/psipred/
  - combination of PSI-BLAST profiles and neural networks
  - careful selection of sequences used for profile construction

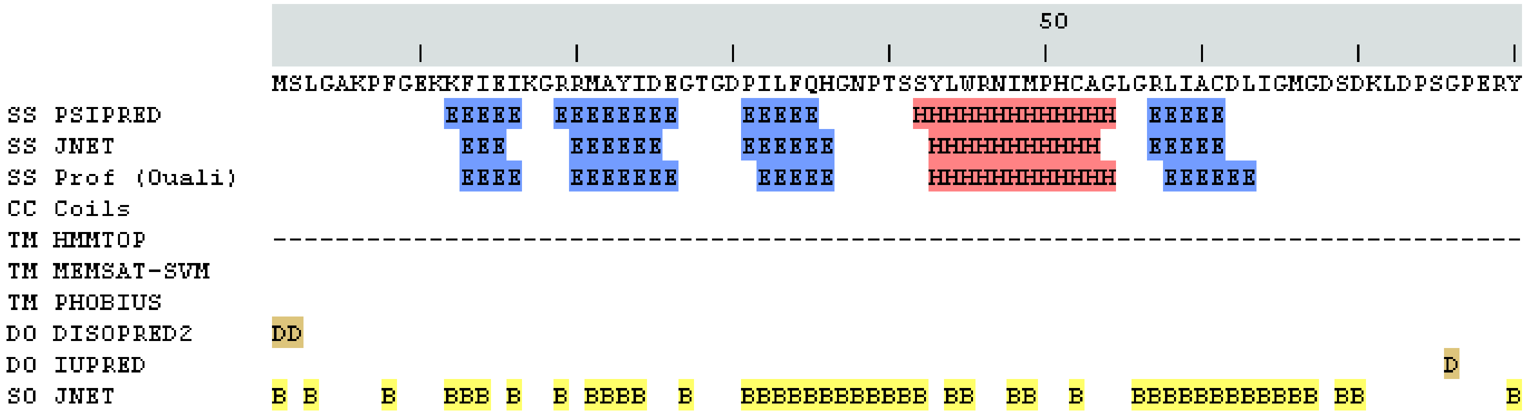


```
Conf:
Pred:
Pred: HHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCEEEEEEE
  AA: QQMNQKAVTSFLSVQDGIYNSDLTPKSDIKNPDVVVYEFF
                 130         140         150         160
```

Legend:
= helix
= strand
= coil
Conf:
-
+
= confidence of prediction
Pred: predicted secondary structure
AA: target sequence

Secondary structure prediction programs

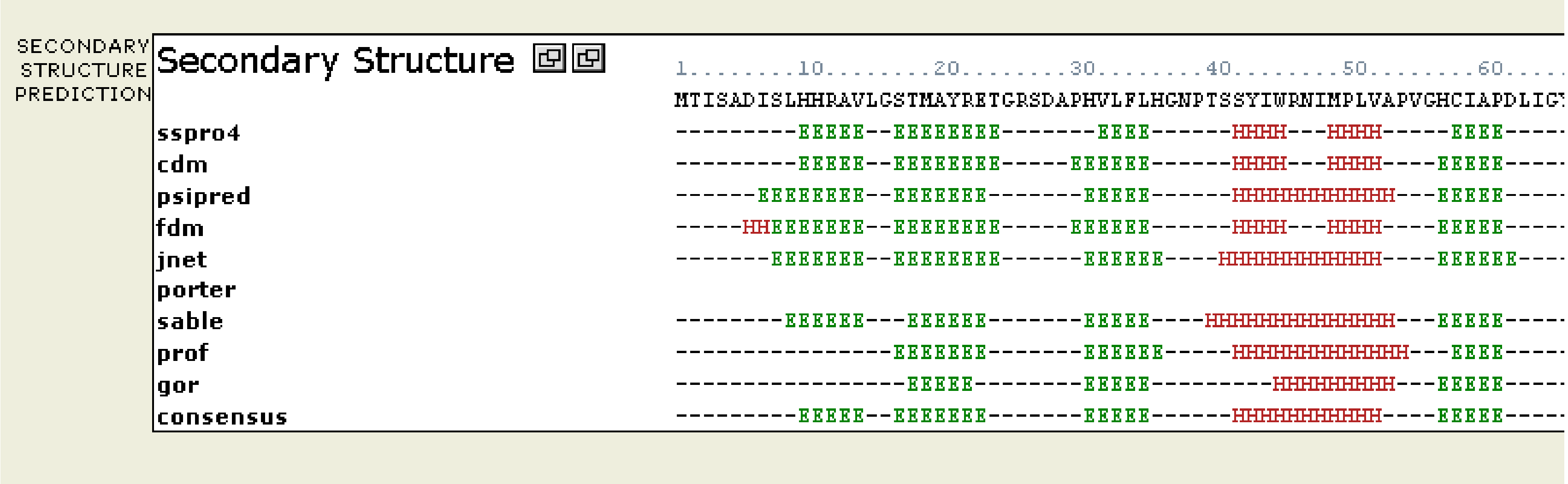
- Quick2D (MPI toolkit)
  - https://toolkit.tuebingen.mpg.de/tools/quick2d
  - overview of secondary structure features (α-helices, extended β-strands, coiled coils, transmembrane helices, disorder regions)
  - predictions by PSI-PRED, JNET, Prof, Coils, MEMSAT2, HMMTOP,...



```
                                                                    50
                             |         |         |         |         |         |         |         |
                    MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIACDLIGMGDSDKLDPSGPERY
SS PSIPRED                     EEEEE  EEEEEEEE    EEEEE      HHHHHHHHHHHHH  EEEEE
SS JNET                         EEE    EEEEEE     EEEEEE      HHHHHHHHHHH   EEEEE
SS Prof (Ouali)                 EEEE   EEEEEEE     EEEEE      HHHHHHHHHHHH   EEEEEE
CC Coils
TM HMMTOP           --------------------------------------------------------------------------------
TM MEMSAT-SVM
TM PHOBIUS
DO DISOPRED2        DD
DO IUPRED                                                                              D
SO JNET             B B    B   BBB B  B BBBB  B   BBBBBBBBBBBB BB  BB  B   BBBBBBBBBBBB BB         B
```

Secondary structure prediction programs

- GeneSilico metaserver
  - https://genesilico.pl/meta2/
  - meta-server for protein structure prediction, including secondary structure prediction



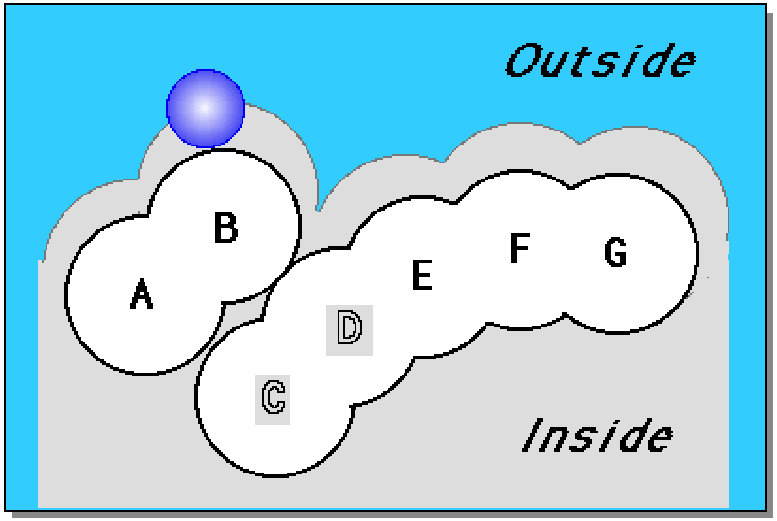
SECONDARY STRUCTURE PREDICTION

Secondary Structure

```
                1........10........20........30........40........50........60....
                MTISADISLHHRAVLGSTMAYRETGRSDAPHVLFLHGNPTSSYIWRNIMPLVAPVGHCIAPDLIG
sspro4          ---------EEEEE--EEEEEEEE-------EEEE------HHHH---HHHH-----EEEE-----
cdm             ---------EEEEE--EEEEEEEE-----EEEEEE------HHHH---HHHH----EEEEE-----
psipred         ------EEEEEEEE--EEEEEEE-------EEEEE------HHHHHHHHHHHH---EEEEE-----
fdm             -----HHEEEEEEE--EEEEEEEE-----EEEEEE------HHHH---HHHH----EEEEE-----
jnet            -------EEEEEEE--EEEEEEEE------EEEEEE----HHHHHHHHHHHH----EEEEEE----
porter
sable           --------EEEEEE---EEEEEE-------EEEEE----HHHHHHHHHHHHHH---EEEEE-----
prof            ----------------EEEEEEE-------EEEEEE-----HHHHHHHHHHHHH---EEEE-----
gor             -----------------EEEEE--------EEEEE---------HHHHHHHHH---EEEEE-----
consensus       ---------EEEEE--EEEEEEE-------EEEEE------HHHHHHHHHHH----EEEEE-----
```

Solvent accessibility prediction

- prediction of the extent to which a residue embedded in a protein structure is accessible to solvent
  - comparison of accessibility of different amino acids – relative values (actual area as percentage of maximally accessible area)
  - simplified two state description – buried vs. exposed residues



Solvent accessibility prediction

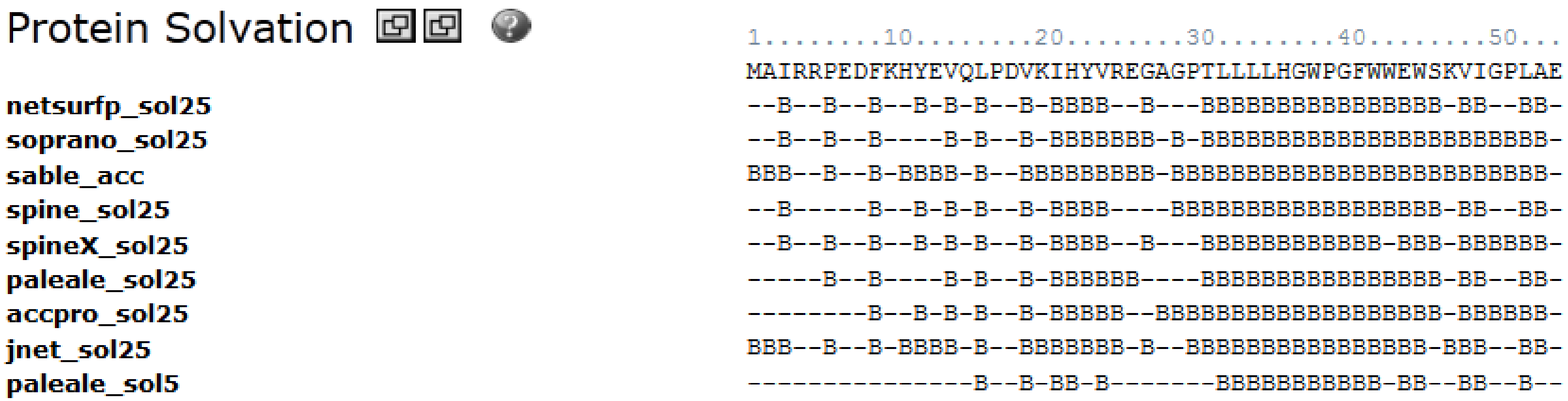
- residue hydrophobicity
  - very hydrophobic stretches are predicted as buried
- propensities of single residues or segments of residues to be solvent accessible
  - superior to simple hydrophobicity analyses
- evolutionary information
  - solvent accessibility at each position of protein structure is evolutionary conserved within sequence families → methods using multiple sequence alignment information
  - prediction accuracy above 75%

Solvent accessibility prediction programs

- PHD
  - http://www.predictprotein.org/
  - combination of evolutionary information with neural network
- PROFphd
  - http://www.predictprotein.org/
  - improved version of PHD
  - combination of evolutionary information and secondary structure prediction with neural network
  - trained only on high resolution structures

Solvent accessibility prediction programs

- SABLE2
  - http://sable.cchmc.org/
  - combines solvent accessibility and secondary structure predictions
- GeneSilico metaserver
  - https://genesilico.pl/meta2/
  - meta-server for structure prediction, including solvent accessibility



Protein Solvation

```
                1........10........20........30........40........50...
                MAIRRPEDFKHYEVQLPDVKIHYVREGAGPTLLLLHGWPGFWWEWSKVIGPLAE
netsurfp_sol25  --B--B--B--B-B-B--B-BBBB--B---BBBBBBBBBBBBBBBB-BB--BB-
soprano_sol25   --B--B--B----B-B--B-BBBBBBB-B-BBBBBBBBBBBBBBBBBBBBBBB-
sable_acc       BBB--B--B-BBBB-B--BBBBBBBBB-BBBBBBBBBBBBBBBBBBBBBBBBB-
spine_sol25     --B-----B--B-B-B--B-BBBB----BBBBBBBBBBBBBBBBBB-BB--BB-
spineX_sol25    --B--B--B--B-B-B--B-BBBB--B---BBBBBBBBBBBB-BBB-BBBBBB-
paleale_sol25   -----B--B----B-B--B-BBBBBB----BBBBBBBBBBBBBBBB-BB--BB-
accpro_sol25    --------B--B-B-B--B-BBBBB--BBBBBBBBBBBBBBBBBBB-BBBBBB-
jnet_sol25      BBB--B--B-BBBB-B--BBBBBBB-B--BBBBBBBBBBBBBBBB-BBB--BB-
paleale_sol5    ---------------B--B-BB-B-------BBBBBBBBBBB-BB--BB--B--
```

Solubility and expressability prediction

- Complicated definition of the property
- Prediction of the extent to which a given sequence will produce a soluble protein in a given expression system or
- Prediction of aggregation propensity
- Methods heavily rely on machine learning.

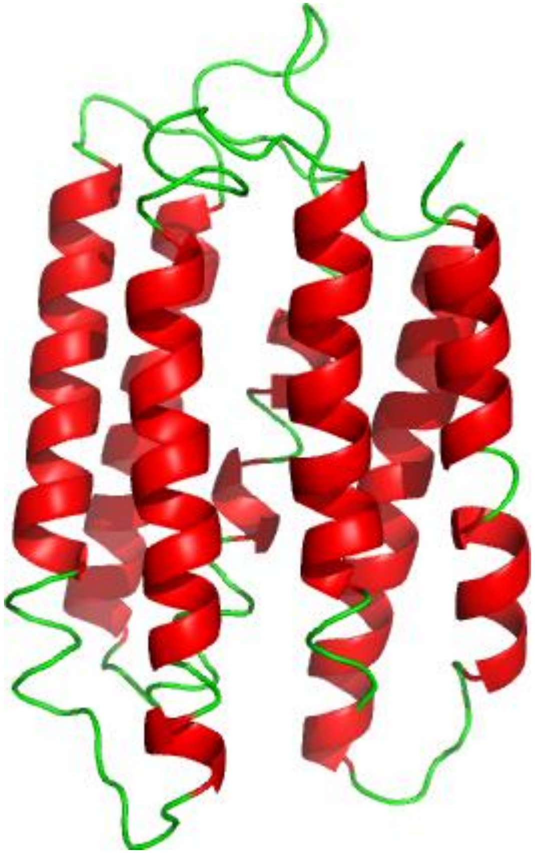
Solubility and expressability prediction

- Methods based on:
  - Plain protein sequences
    - Evolutionary information implicit in the learning data
    - SOLpro http://scratch.proteomics.ics.uci.edu
    - ESPRESSO http://mbs.cbrc.jp/ESPRESSO
    - SoluProt https://loschmidt.chemi.muni.cz/soluprot/
  - Sequence profiles
    - Evolutionary Information implicit in the profile
    - AGGRESCAN http://bioinf.uab.es/aggrescan/
    - TANGO http://tango.crg.es
    - PASTA http://protein.cribi.unipd.it/pasta/

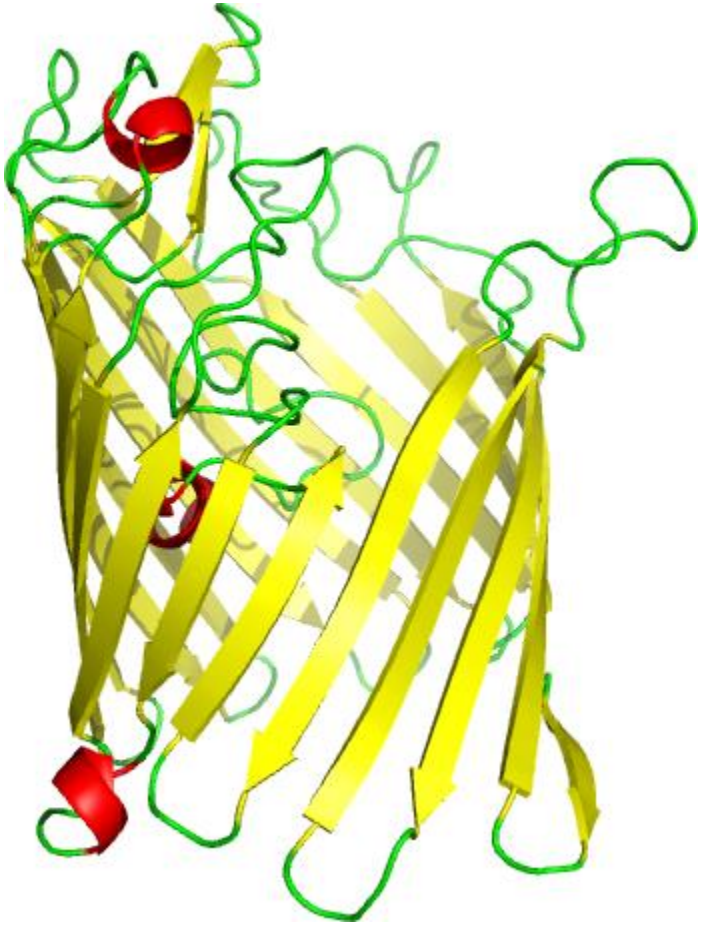
Transmembrane region prediction

- transmembrane (TM) proteins – challenge for experimental determination of 3D structure → structure prediction needed even more than for globular water-soluble proteins
- two major classes of integral membrane proteins
  - transmembrane helices (TMH)
  - transmembrane beta-strand barrels (TMB)

Transmembrane region prediction



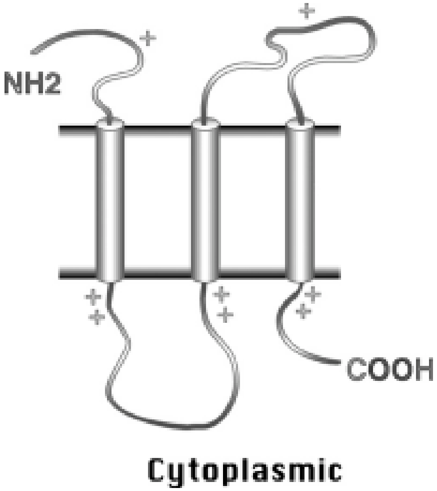
TMH: bacteriorhodopsin (PDB-ID 1ap9)



TMB: matrix porin (PDB-ID 2omf)

Transmembrane region prediction

- prediction of TMH simplified by strong environmental constraints – lipid bilayer of the membrane
  - TMHs are predominantly apolar and 12-35 residues long (hydrophobicity)
  - specific distribution of Arg and Lys (positively charged) → connecting loop regions at the inside of the membrane have more positive charges than loop regions at the outside = positive-inside rule



Transmembrane region prediction

- prediction of TMB
  - transmembrane beta-strands contain 10 - 25 residues
  - only every second residue faces the lipid bilayers and is hydrophobic, other residues face the pore of the β-barrel and are more hydrophilic
    → analysis of hydrophobicity NOT useful for TMB prediction

Transmembrane region prediction

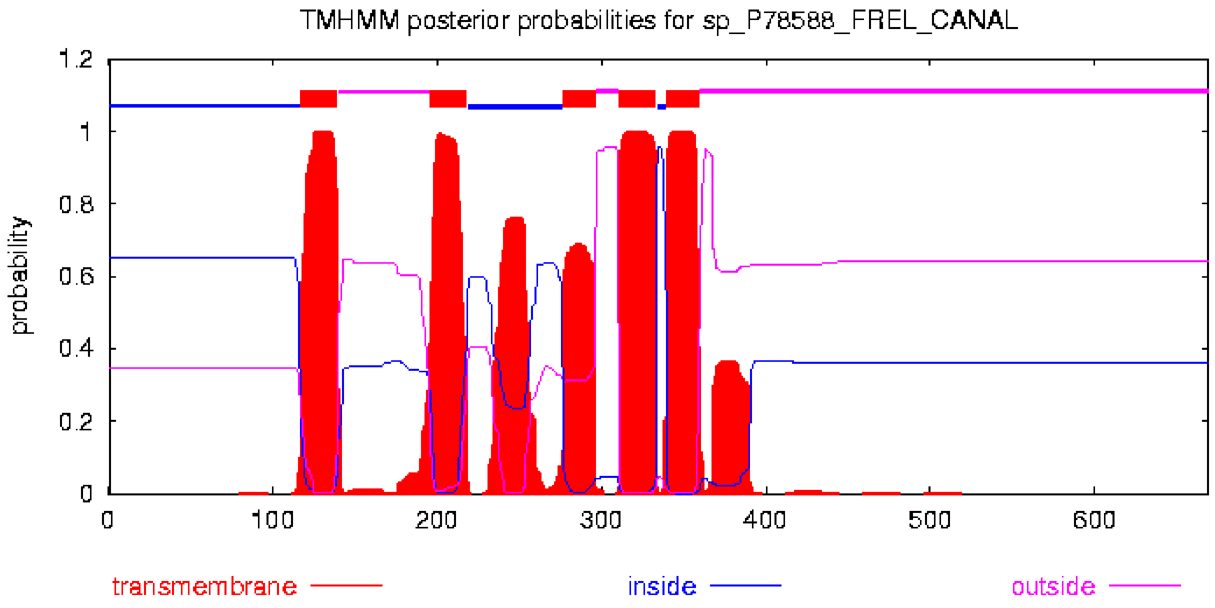
- hydrophobicity-based methods (for TMH)
  - hydrophobicity along the sequence, hydrophobic moment or other membrane-specific amino acid preferences
  - averaging hydrophobicity values over windows of adjacent residues
  - prediction of orientation of TMH using positive-inside rule
- evolutionary information combined with machine learning or hidden Markov models (for TMH)
  - superior to methods based solely on hydrophobicity
- evolutionary information combined with machine learning or hidden Markov models (for TMB)

Transmembrane region prediction programs

- no appropriate estimate of performance available
  - insufficient number of high-resolution structures (needed for a statistically significant analysis)
  - in the papers, accuracy of methods usually largely overestimated – methods perform much better on proteins for which they were developed than on new proteins
  - the best methods for TMH estimated to have ~70% accuracy

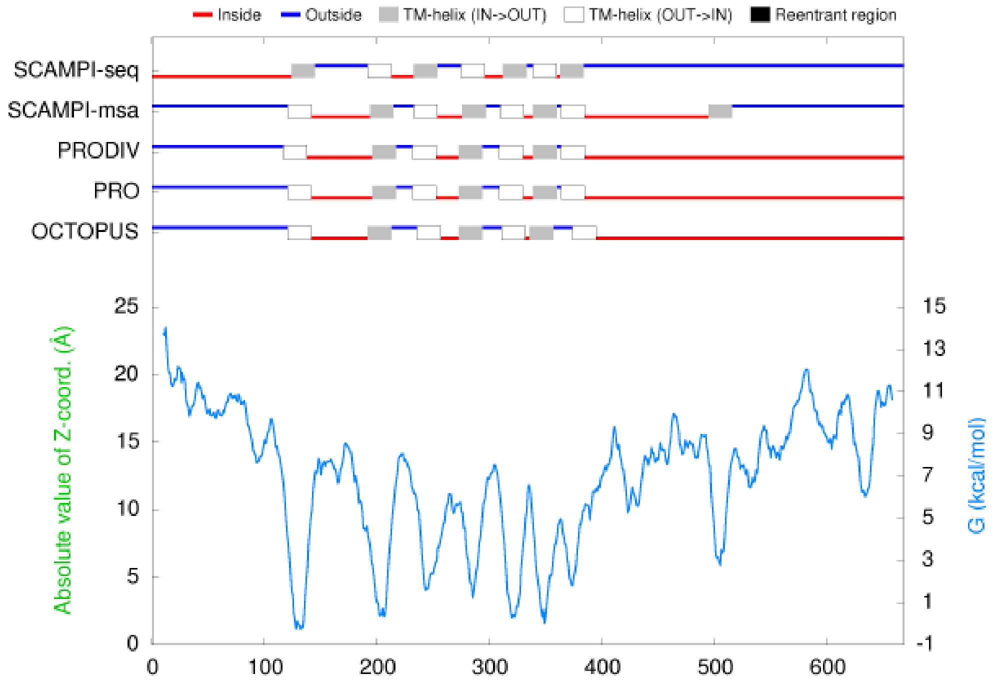
Transmembrane region prediction programs

- ❑ TMHMM 2.0
  - http://www.cbs.dtu.dk/services/TMHMM/
  - a number of statistical preferences and rules embedded in hidden Markov model → localization and orientation of TMH



Transmembrane region prediction programs

- TOPCONS
  - http://topcons.cbr.su.se/
  - consensus prediction of TMHs



Transmembrane region prediction programs

- TBBpred
  - http://www.imtech.res.in/raghava/tbbpred/
  - prediction of TMB using machine learning

- PROFtmb
  - http://www.predictprotein.org/
  - profile-based hidden Markov model
  - prediction of bacterial TMB

- …

References

- Gu, J. & Bourne, P. E. (2009). **Structural Bioinformatics, 2nd Edition**, Wiley-Blackwell, Hoboken, p. 1067.
- Xiong, J. (2006). **Essential Bioinformatics**. Cambridge University Press, New York, p. 352.
- Schwede, T. & Peitsch, M. C. (2008). **Computational Structural Biology: Methods and Applications**, World Scientific Publishing Company, Singapore, p. 700.