

Functional annotation

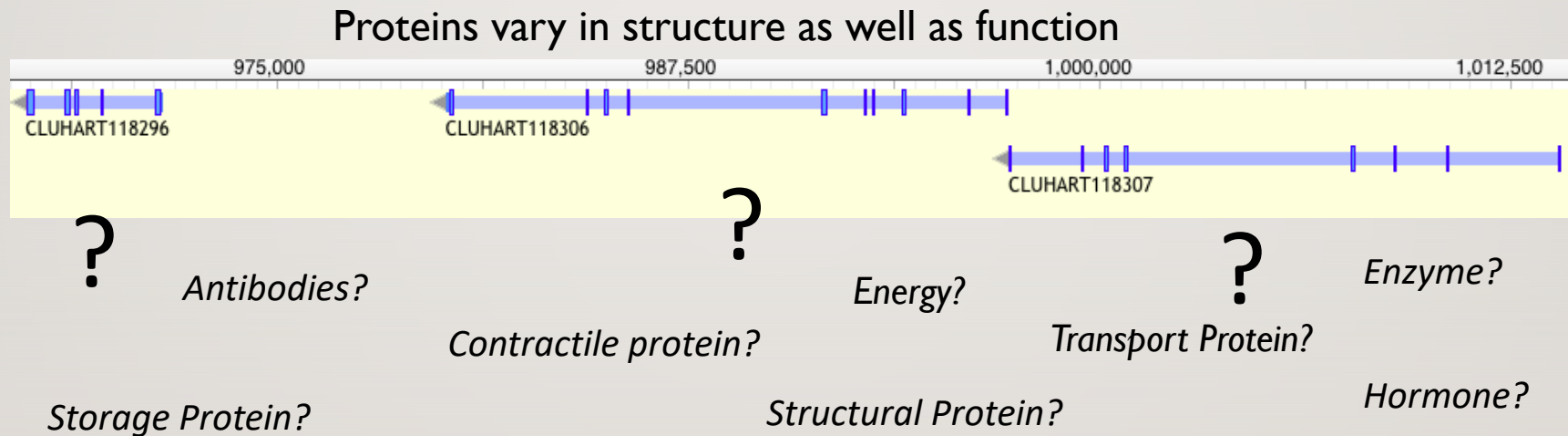


22/09/2025



Understanding the function of gene product is key to understanding how a limited number of interacting gene products can generate life, from simple unicellular organisms to the incredibly complex multi-cellular Homo sapiens.

Rison, S.C., Hodgman, T.C. and Thornton, J.M. (2000) Comparison of functional annotation schemes for genomes. *Funct. Integr. Genomics*, 1, 56–69.



Functional annotation – how?

- **Experimentally** (accurate)
 - Mutants, knockout, etc.



Mice homozygous for the diabetes 3J spontaneous mutation

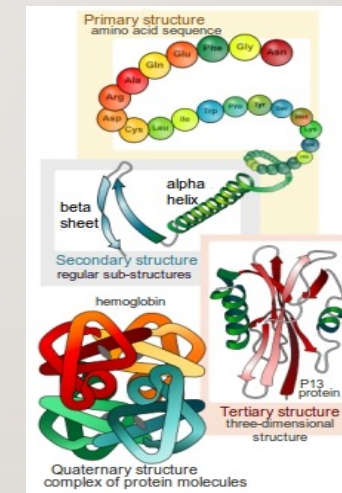
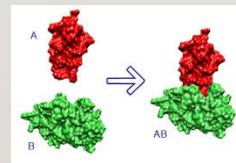
- **Computationally** (limited accuracy)

- Sequence-based



- Structure based

- Protein-protein interaction



- Based on similarity

=> Best blast hit

```
Q GLMDTAFEHIKATGGLTTESNYPYKGEDATCNS-KI
  GLM+ AFE+IK +GG+TTES YPY+ + TC++ +
S GLMENAFEYIKHSGGITTESAYPYRAANGTCDVR
```

- Based on Motif

=> Proscan, MEME, QuasiMotiFinder

D-X-[KR]-P-{WYF}-X5

Any amino acid either Any amino acid except

- Based on Profile

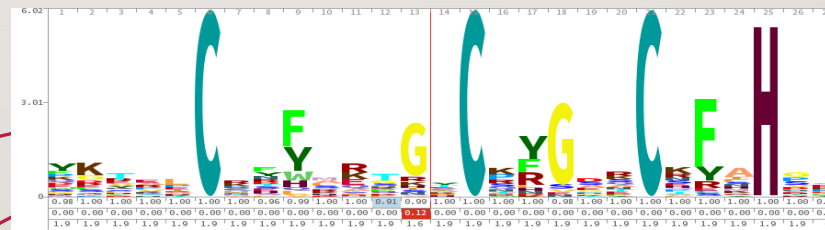
=> HMM or other statistical signature

Whole sequence
e.g. Psi-BLAST*, PIRSF

domain
e.g. PFAM

Localization (e.g. membrane, golgi, secreted)
e.g. SignalP, TMHMM

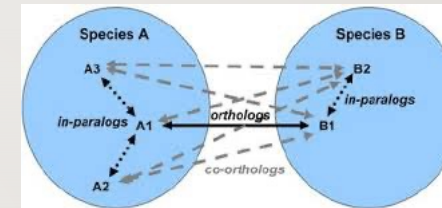
structural classification
e.g. SUPERFAMILY



Use of evolutionary relationship (Orthology)

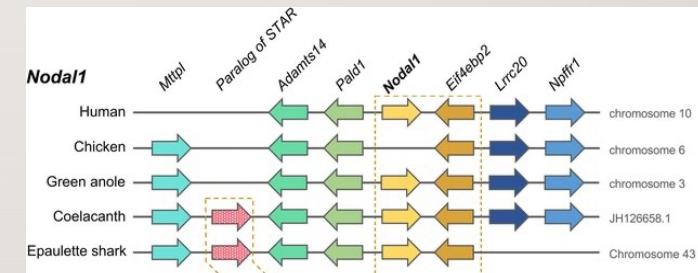
- Based on Clustering

=> KOG (Eukaryotic) / COG(Prokaryotic)



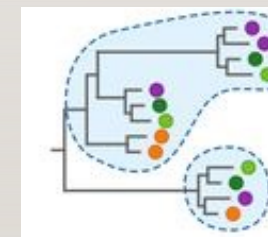
- Based on synteny

=> conservation of gene order/organization across species



- Based on phylogeny

=> Quite complicated at large scale



Focuses on the **3D conformation of a single protein** and similarity to known structures to infer function.

- **Global structure**

Classification of overall protein folds and domain (e.g. SCOP, CATH, FSSP)

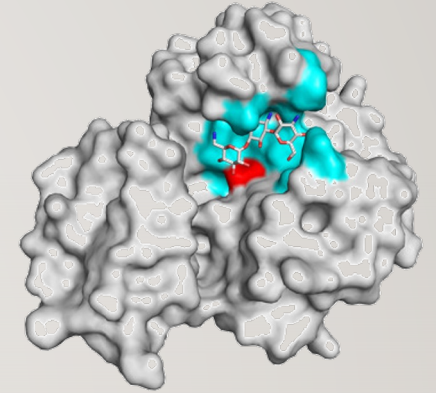
- **local structure**

Functionally relevant *parts of the structure* e.g. clefts, ligand-binding pockets, **interfaces**, surfaces.

It often suggests where ligands, cofactors, or partner proteins may interact (e.g. CASTp, fpocket, ProFunc).

- **active-site residues (atomic level)**

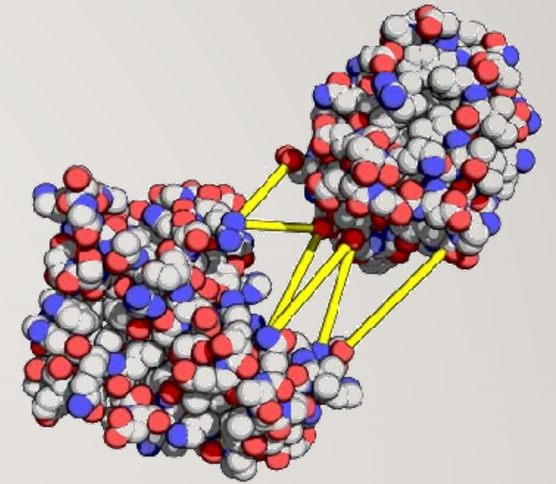
Specific catalytic or ligand-binding residues. Residues are often more conserved than the overall fold (e.g. PDBSiteScan, Catalytic Site Atlas)



- **Structure-guided interaction prediction (who interacts)**

Docking and modeling of protein complexes (e.g. HADDOCK, ClusPro)

Identifies potential interaction partners and functional interfaces.

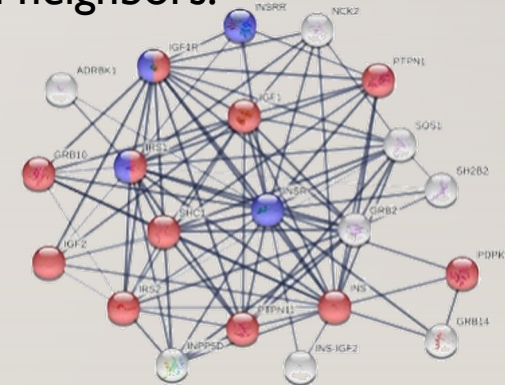


- **Pathway and network context (what role)**

Integration of predicted PPIs into biological networks (e.g. STRING).

Guilt-by-association: uncharacterized proteins may inherit function from well-annotated neighbors.

Supports placement of proteins into pathways and complexes.



- **Multidimensional problem:**

A protein can have a molecular function, a cellular role, and be part of a functional complex or pathway

- **Descriptive levels**

Molecular function can be illustrated at multiple levels e.g. '**enzyme**' category versus a more specific '**protease**' assignment

- **Similarities (structural or in sequence) function.**

- Similar sequences but different functions (new domain => new combination => different function)
- Different sequences may have same function (convergence) : Profiles helpful
- Two proteins may have a similar fold but different functions

- **Looks for conserved domains more reliable than whole sequence ?**

How to go from conserved domains to assigning a function to your protein?

=> Importance is to gather as much information as possible

Databases

Database	Information	Comment
CARD	Antibiotic Resistance	Comprehensive Antibiotic Resistance Database
Catalytic Site Atlas (CSA)	Sites	Database of catalytic residues and enzyme active sites
DIP	Interaction	Database of interacting proteins
eggNOG	Orthologous groups	Functional classification of proteins based on orthology (COG/KOG)
GO	Gene Ontology	Three structured, controlled vocabularies (ontologies) : biological processes, cellular components and molecular functions
KEGG	Pathway	Kyoto Encyclopedia of Genes and Genomes
PDB	Protein structures	Main repository of experimentally determined 3D protein structures
Pfam	Protein families	Multiple sequence alignments and hidden Markov models
Prosite	Protein domains	Protein domains and motifs Database of conserved protein motifs and functional sites
Reactome	Pathway	Curated and peer reviewed pathway database
Uniprot	Protein sequence and functional annotation database	Comprehensive resource for protein sequences, functional annotation, domains, post-translational modifications, subcellular localization, and cross-references to multiple biological databases
...	...	...


GENEONTOLOGY
Unifying Biology

[About](#)
[Ontology](#)
[Annotations](#)
[Downloads](#)
[Help](#)






Current release 2019-05-09: 45,006 GO terms | 6,307,350 annotations
1,164,920 gene products | 4,455 species

THE GENE ONTOLOGY RESOURCE

The mission of the GO Consortium is to develop a comprehensive, **computational model of biological systems**, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.

The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

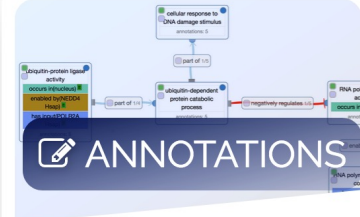


☒ Any
 ☐ Ontology
 ☐ Gene Product

GO Enrichment Analysis ?

Powered by PANTHER

Hint: can use UniProt ID/AC, Gene Name, Gene Symbols, MOD IDs


Gene Ontology (GO) is a major bioinformatics initiative that uses a **controlled vocabulary** (> 60 000 terms) to standardize the representation of gene and gene product attributes across all species.

<http://www.geneontology.org/>

Gene Ontology (GO) organizes function into three complementary aspects:

GO term prediction

Biological Process

- [GO:0006631](#) fatty acid metabolic process
- [GO:0006635](#) fatty acid beta-oxidation
- [GO:0008152](#) metabolic process
- [GO:0055114](#) oxidation-reduction process

Molecular Function

- [GO:0003824](#) catalytic activity
- [GO:0003857](#) 3-hydroxyacyl-CoA dehydrogenase activity
- [GO:0004300](#) enoyl-CoA hydratase activity
- [GO:0016491](#) oxidoreductase activity
- [GO:0016616](#) oxidoreductase activity, acting on the CH-OH group of donors
- [GO:0050662](#) coenzyme binding

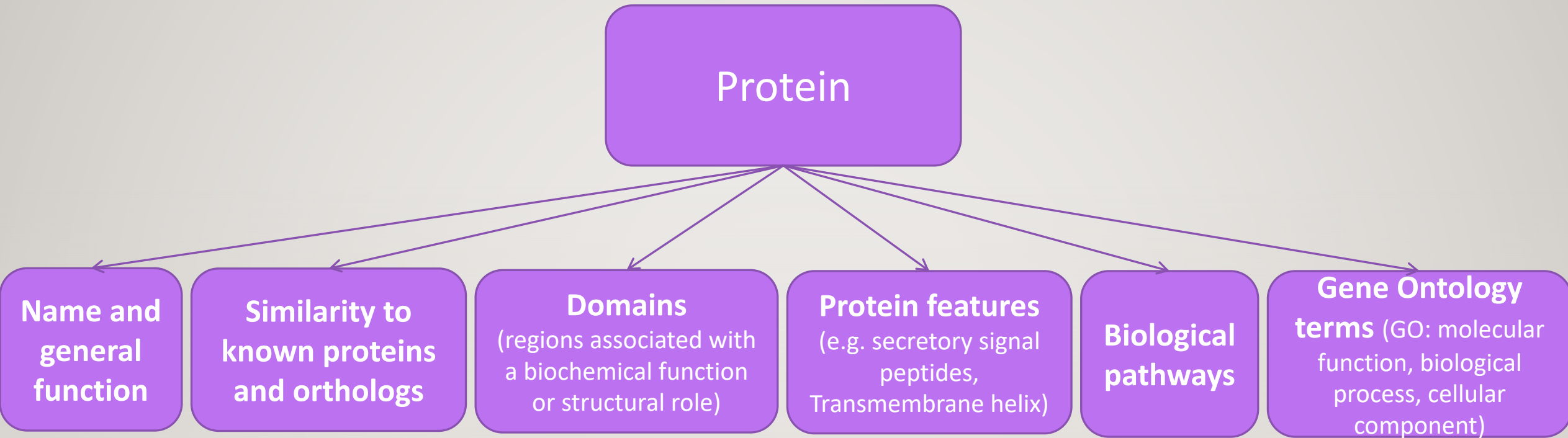
Cellular Component

- [GO:0005739](#) mitochondrion
- [GO:0016507](#) mitochondrial fatty acid beta-oxidation multienzyme complex

pathways and larger processes
made up of the activities
of multiple gene products.

molecular activities
of gene products

where gene products are active



These annotations exist at different biological scales:

- Molecular level → domains, catalytic activity
- Cellular level → localization, pathways
- Systems level → biological processes and networks

Protein

Name and
general function

Similarity to
known proteins
and orthologs

Tools: BlastP , MMseqs2, Diamond, ...
Database: Uniprot, eggno, ...

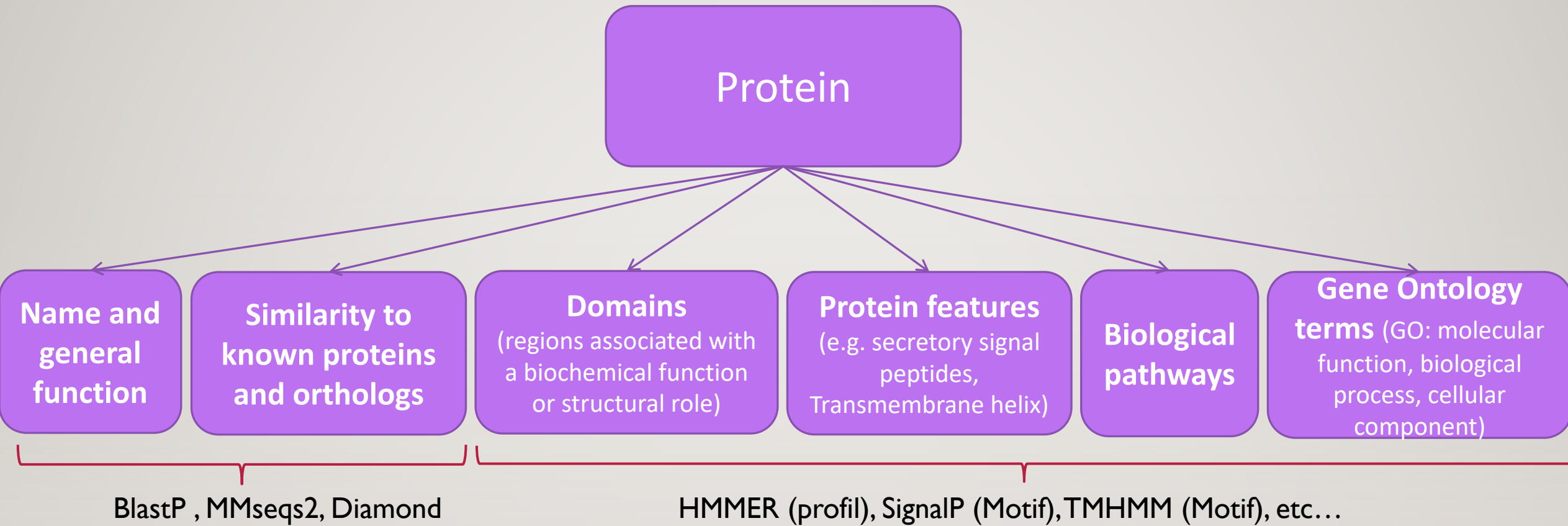
Sequence Similarity :

- The most used (popular)
- Quick
- Easy to use
- Accurate (>70%)
- Many resources: even structural domains information
- Less computationally demanding

Watson JD *et al.* : Towards fully automated structure-based function prediction in structural genomics: a case study. J Mol Biol. 2007, 367: 1511-1522. 10.1016/j.jmb.2007.01.063.

Limits

- Reliability of the database used e.g. Uniprot (exhaustive) vs Swissprot (reliable)
- Orthology not certain - best blast-hit does not equal orthologous!
- Minimum e-value (arbitrary)
- Bias due to well conserved domains
- Best Hit (used as template) is not necessary the best annotated sequence to use => Could apply a prioritization rule (Species X first, then Y, etc).



Tool	Approach	Comment
Trinotate	Best blast hit + protein domain identification (HMMER/PFAM) + protein signal peptide and transmembrane domain prediction (signalP/tmHMM), and leveraging various annotation databases (eggNOG/GO/Kegg databases).	Partially automated
Annocript	Best blast hit	Collects the best-hit and related annotations (proteins, domains, GO terms, Enzymes, pathways, short)
Annot8r	Best blast hits	A tool for Gene Ontology, KEGG biochemical pathways and Enzyme Commission EC number annotation of nucleotide and peptide sequences.
Sma3s	Best blast hit + Best reciprocal blast hit + clusterisation	3 annotation levels
afterParty	BLAST, InterProScan	web application
Interproscan	Run separate search applications HMMs, fingerprints, patterns => InterPro	Created to unite secondary databases
Blast2Go	Best* blast hits (picks out GO terms based on blast hits and uniprot with statistical significance test)	Commercial ! Contain nice downstream analysis/visualization components Plugin for Geneious or CLC bio Workbench

- **InterPro**

is an integrative database that combines the main **protein signature databases** (Pfam, PROSITE, SMART, CATH, SUPERFAMILY, etc.).

- **A signature**

is a motif, a profile, or an HMM (Hidden Markov Model) provided by a member database that describes a protein family, a domain, an active site, or another functional region.

- **Integrated signature**

is a signatures linked to an InterPro entry

- **InterPro entries**

is a curated biological interpretations combining multiple integrated signatures.

- **InterProScan**

is the software that compares a given protein sequence against all these signature databases to assign potential functions.

InterPro 106.0 • 19th June 2025

Features include:

- The addition of 683 InterPro entries.
- An update to Pfam [37.4].
- Integration of 727 new methods from the SMART (1), Pfam (613), NCBIFAM (78), SUPERFAMILY (1), PANTHER (4), CDD (30) databases.

Contents and coverage

InterPro protein matches are now calculated for all UniProtKB and UniParc proteins. InterPro release 106.0 contains 48 679 entries (last entry: IPR058129), representing:

H	Homologous Superfamily	3 512
F	Family	26 622
D	Domain	17 171
R	Repeat	385
S	Site	
	L. Active Site	133
	L. Binding Site	82
	L. Conserved Site	757
	L. PTM	17

InterPro2GO

We have a total number of 30 207 GO terms mapped to InterPro entries.

List of InterPro2GO mappings. These are also available through the EBI GO browser QuickGO

pathway :

- MetaCyc
- Reactome

other licensed analysis available:

- Phobius (licensed software)
- SignalP
- SMART (licensed components)
- TMHMM

Member database information

	Signature database	Version	Signatures ⓘ	Integrated signatures ⓘ
	CATH-Gene3D	4.3.0	6 631	2 823 (42.6%)
	CDD	3.21	19 902	4 948 (24.9%)
	HAMAP	2025_01	2 391	2 386 (99.8%)
	NCBIFAM	17.0	31 915	6 007 (18.8%)
	PANTHER	19.0	15 683	10 528 (67.1%)
	Pfam	37.4	24 736	23 862 (96.5%)
	PIRSF	3.10	3 285	3 230 (98.3%)
	PRINTS	42.0	2 106	1 939 (92.1%)
	PROSITE patterns	2025_01	1 311	1 282 (97.8%)
	PROSITE profiles	2025_01	1 399	1 303 (93.1%)
	SFLD	4	303	159 (52.5%)
	SMART	9.0	1 322	1 274 (96.4%)
	SUPERFAMILY	1.75	2 019	1 645 (81.5%)

Total 113 003 61386

Search against signature databases

InterProScan runs searches in parallel (Each database uses its own search engine:

- Pfam → HMMER (HMM profiles).
- PROSITE → regular expressions or profiles.
- SUPERFAMILY → structural HMMs.
- COILS → prediction of coiled-coil regions.

Result aggregation

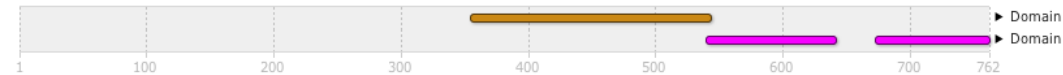
All matches are collected and for each match, InterProScan adds:

- Associated **InterPro ID** (if --iprlookup is used e.g. IPR002568) <= It unifies results coming from different databases.
- Associated **GO terms** (if --goterms is used).
- Associated **pathways** (if --pa is used).

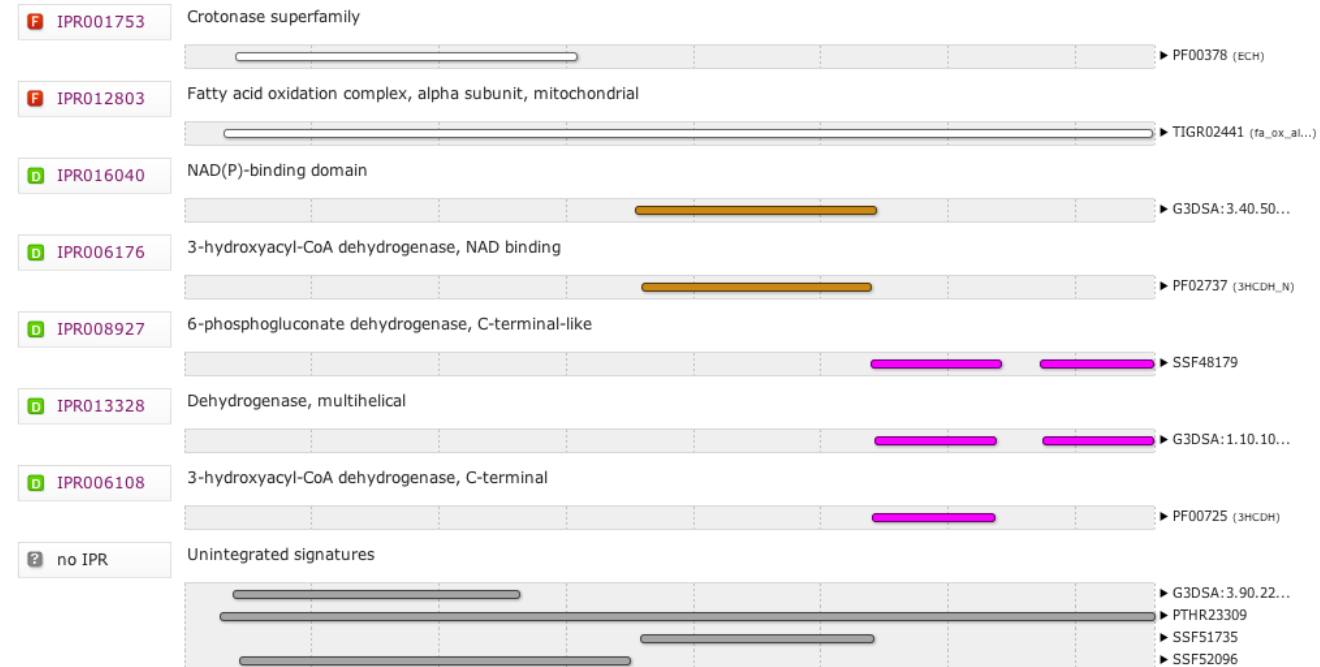
Protein family membership

- F Crotonase superfamily (IPR001753)
 - F Fatty acid oxidation complex, alpha subunit, mitochondrial (IPR012803)

Domains and repeats



Detailed signature matches



Categorizations of gene function (e.g GO) in a hierarchy of categories is helpful

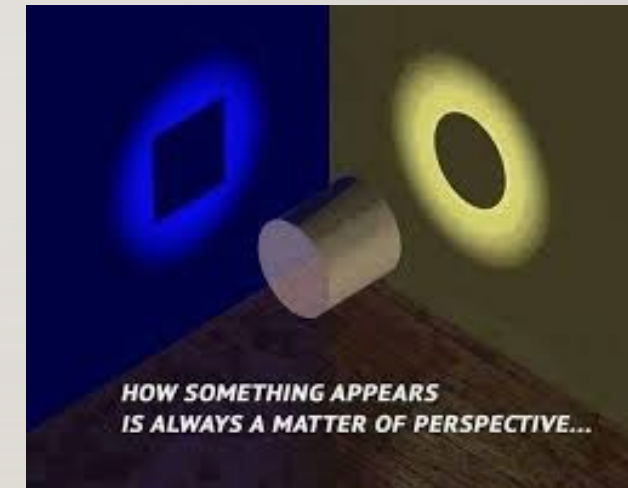
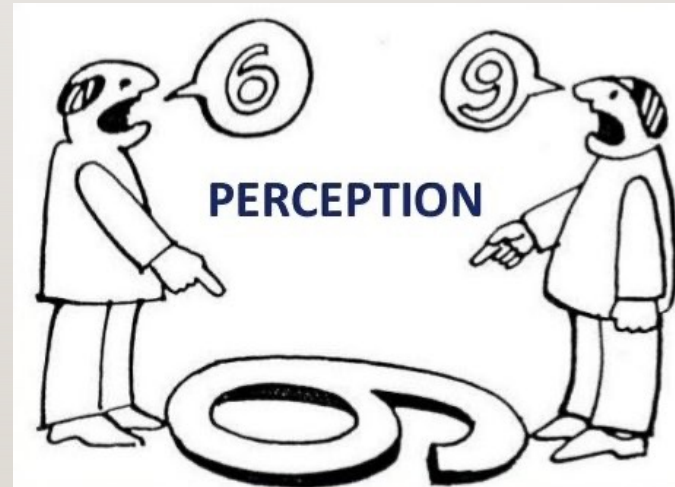
BUT

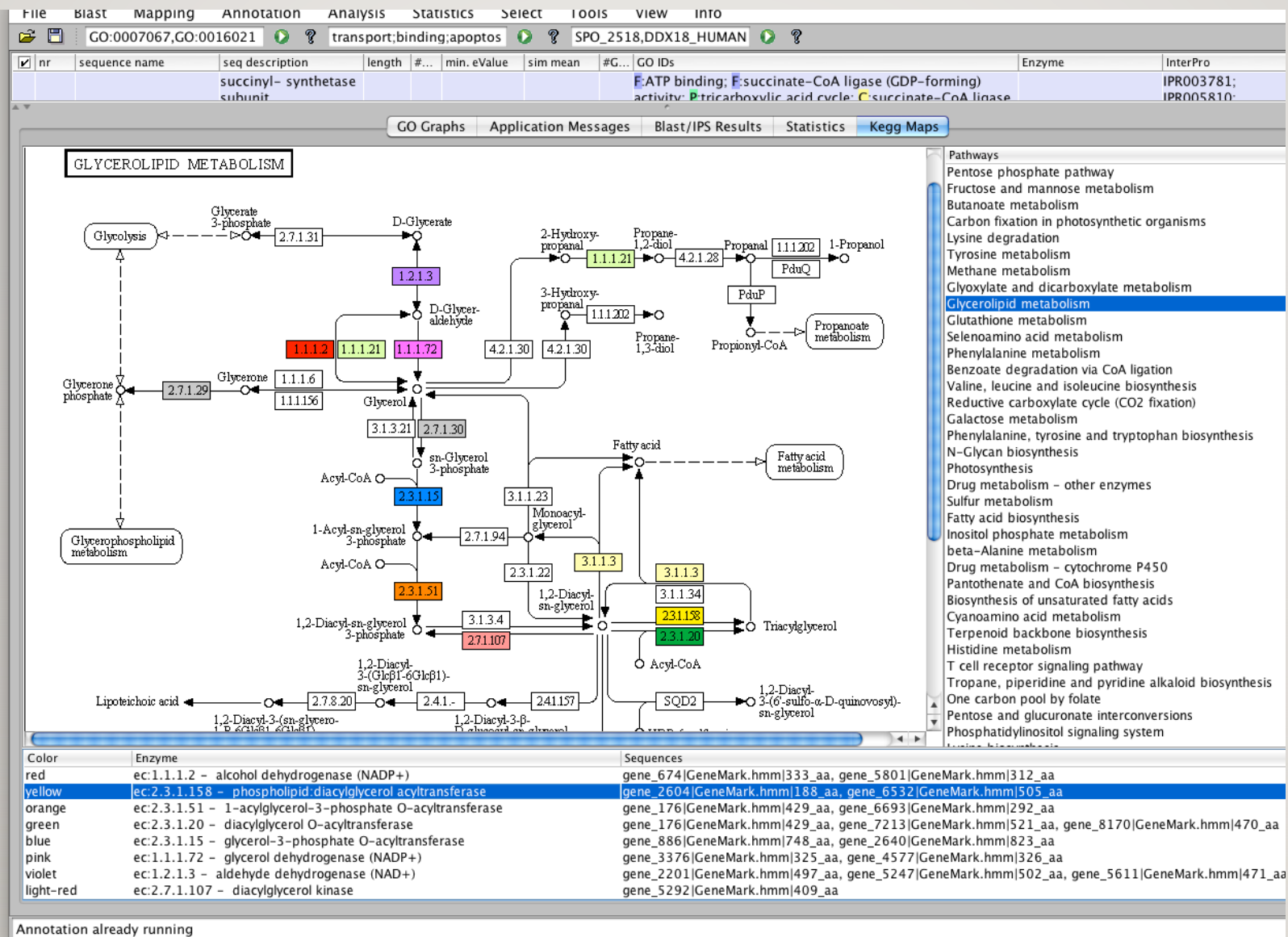
gene has no function alone

⇒ Pathways / regulatory networks explain how genes interact so what they are doing!

E.g. databases for pathway :

- KEGG
- MetaCyc
- Reactome
- UniPathway







- Structural annotation matters

Functional annotation depends strongly on the quality of structural annotation

Common issues: gene fusions, gene splits, partial genes, over-predictions, wrong ORFs



- Limits of functional annotation

Errors can propagate from databases => experimental validation is essential

Annotation quality depends on the database used

Many proteins remain *hypothetical* or *uncharacterized* (Depends largely on conventional experiments)

- Beyond sequence similarity

High sequence similarity $\neq$ identical function

Functional understanding requires additional context:

- Pathway integration
- Spatial/temporal regulation
- Biological networks

The End

Thank you for your attention