MCB 5472 Lecture #5:
Gene Prediction and Annotation
February 24, 2014

Note on the assignment

- Depending on your settings PSI-BLAST can take a while to run
  - Do not leave this until the last minute!
- Recall from Assignment Lecture #1: `nohup` can allow you to leave a job running on the cluster
  - E.g., `nohup [task] & > nohup.out`

Do you have a DNA sequence…

- Limited utility by itself
- Annotations describe what the DNA does
  - Structural: what features are present on the DNA?
  - Functional: what do those features do?

How to annotate: 2 methods

1. From first principles:
   - Experimental data in the literature
   - Algorithmic rules
2. From orthology / homology to previously annotated sequences

Annotation accuracy

- Manual annotation from experimental data in the literature is highly accurate
  - Although not all experiments are unequivocal
- Annotations using algorithms can be quite accurate
  - Depends on the complexity of the problem the algorithm is trying to solve
- Annotations based on orthology relies on the assumption that function is conserved
  - Depends on how rigorously orthologs are defined
  - Depends on functions not changing over time

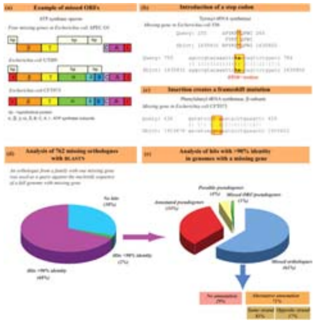
Gene annotation

- Gene and protein annotation is typically algorithmic
- Genes and proteins have specific features that algorithms use to define them
- Algorithms for bacteria and archaea work quite well, eukaryotes more difficult because of additional complexity, e.g., splicing

Prokaryote gene finding

- Glimmer, GeneMark: Markov Models
  - Genes modeled based on differences between coding and non-coding regions
    - E.g., typically start with ATG, end with stop codon
    - E.g., ORF overlap
    - E.g., ribosome binding regions
  - Often have difficulty to decide which strand is coding.
- Prodigal: summed likelihood of finding individual gene features
- Can be challenged by %GC bias
  - Better performance by training on known genome annotations

Fig. 1. Analysis of ORFs missing in one out of 30 completely annotated Escherichia genomes.

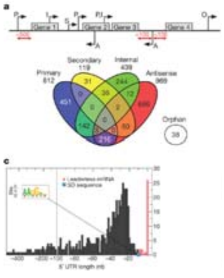


Poptsova M S , and Gogarten J P Microbiology 2010;156:1909-1917

SGM

Remember: genes are not transcripts!

- 5’ mRNA analysis in *Helicobacter pylori* shows much greater transcript diversity than evident from simple gene annotations
- Most NCBI annotations equate genes with transcripts



Sharma et al. 2010 Nature 464:250-255

Eukaryotic gene finding

- E.g., Augustus, GeneMark-ES
- *ab initio* methods work less well compared to prokaryotic genes
  - More complicated transcripts (e.g., splice variants)
  - Less information at promoter (e.g., Prodigal uses Shine-Delgarno sequences; -35 and -10 regions vs. single TATA box)
- NCBI annotations more clearly separate genes (includes pseudogenes), mRNA (typically spliced) & protein (spliced like mRNA)

Adding information to gene annotations

1. Combine multiple prediction methods
   - For prokaryotes, typically longest transcript chosen
   - For eukaryotes, typically all splice variants kept
2. Search for homologous genes in related taxa
   - True genes will be evolutionarily conserved
   - Annotation errors can be propagated
     - Annotations do not specify the evidence supporting them
3. Integrate RNAseq
   - Augustus can incorporate into its predictions directly
   - Rare for prokaryotes
   - Requires genes be expressed and detectable

Metagenomes and single-cell genomes

- Assemblies are typically much more fragmented than those of cultured microbes
- Requires dedicated gene prediction methods
  - Training information often missing/obscured
  - Gene fragments obscure genomic features used for gene prediction

Non-coding RNAs

- Some HMM-based software
  - RNAMMER (ribosomal RNAs)
  - tRNAscan-SE (tRNAs)
- Rfam: database of non-coding RNA families
  - Curated sequence alignments taking into account secondary structures
  - Infernal: software for searching DNA sequence databases using structured RNA molecule profiles
    - Takes RNA secondary structure into account via "covariance models"
  - Sister project to Pfam (see later)

Functional annotations

Manual annotation

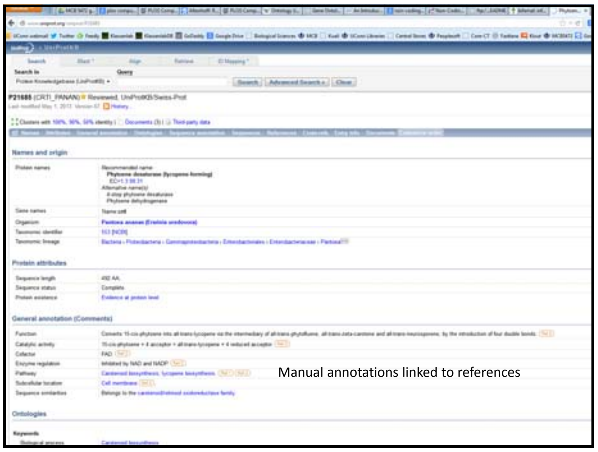
- Low-throughput
- High accuracy

SwissProt

- Started 1986 at the Swiss Institute for Bioinformatics, later developed at the European Bioinformatics Institute
- Goal: providing reliable protein sequences having a high level of annotation
  - Directly curated from literature information
  - Contrast to NCBI: a sequence repository with some automated annotation pipelines
- Current version (2014_02): 542,503 sequences annotated from 22,6190 references

UniProt

- Ultimately manual annotation couldn't keep up, parallel TrEMBL database created using automated annotation
- UniProtKB stores combined SwissProt/TrEMBL databases, incorporates Protein Information Resource (PIR), built on M. Dayhoff's atlas
- Syncs with EMBL/DDBJ/GenBank nucleotide databases
- Hosts several protein annotation schemes
- ExPASy – major proteomics analysis resource
- www.uniprot.org



P21685 (CRTI_PANAN) Reviewed, UniProtKB/Swiss-Prot

Names and origin

Protein names: Recommended name: Phytoene desaturase (lycopene-forming) EC=1.3.99.31; Alternative name(s): 4-step phytoene desaturase, Phytoene dehydrogenase

Gene names: Name: crtI

Organism: Pantoea ananas (Erwinia uredovora)

Taxonomic identifier: 553 [NCBI]

Taxonomic lineage: Bacteria › Proteobacteria › Gammaproteobacteria › Enterobacteriales › Enterobacteriaceae › Pantoea

Protein attributes

Sequence length: 492 AA.

Sequence status: Complete.

Protein existence: Evidence at protein level

General annotation (Comments)

Function: Converts 15-cis-phytoene into all-trans-lycopene via the intermediary of all-trans-phytofluene, all-trans-zeta-carotene and all-trans-neurosporene, by the introduction of four double bonds.

Catalytic activity: 15-cis-phytoene + 4 acceptor = all-trans-lycopene + 4 reduced acceptor.

Cofactor: FAD.

Enzyme regulation: Inhibited by NAD and NADP.

Pathway: Carotenoid biosynthesis; lycopene biosynthesis.

Subcellular location: Cell membrane.

Sequence similarities: Belongs to the carotenoid/retinoid oxidoreductase family.

Ontologies

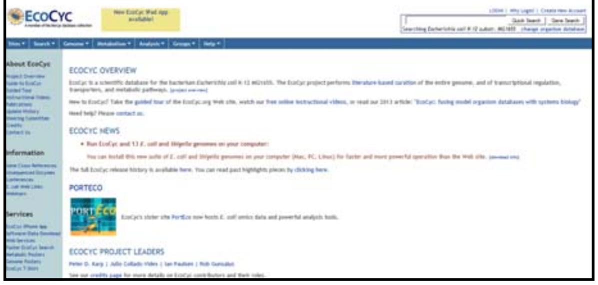
Keywords

Biological process: Carotenoid biosynthesis

Manual annotations linked to references

Ecocyc – an example manually edited model organism database

- http://ecocyc.org/



ECOCYC OVERVIEW

EcoCyc is a scientific database for the bacterium Escherichia coli K-12 MG1655. The EcoCyc project performs literature-based curation of the entire genome, and of transcriptional regulation, transporters, and metabolic pathways. [project overview]

New to EcoCyc? Take the guided tour of the EcoCyc.org Web site, watch our free online instructional videos, or read our 2013 article: "EcoCyc: fusing model organism databases with systems biology"

Need help? Please contact us.

ECOCYC NEWS

- Run EcoCyc and 13 E. coli and Shigella genomes on your computer:

  You can install this new suite of E. coli and Shigella genomes on your computer (Mac, PC, Linux) for faster and more powerful operation than the Web site. [detailed info]

The full EcoCyc release history is available here. You can read past highlights pieces by clicking here.

PORTECO

EcoCyc's sister site PortEco now hosts E. coli omics data and powerful analysis tools.

ECOCYC PROJECT LEADERS

Peter D. Karp | Julio Collado-Vides | Ian Paulsen | Rob Gunsalus

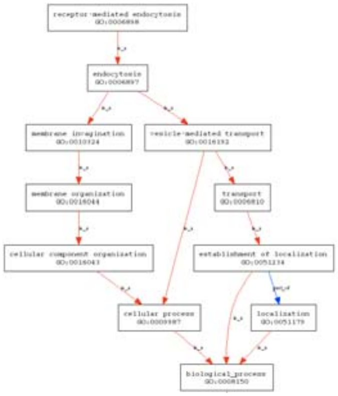
See our credits page for more details on EcoCyc contributors and their roles.

Ontologies

- Manual annotations originally used free-text labels, not standardized
  - Problem: free text is difficult for computers to make use of
- Ontologies: knowledge representation using standardized terms and interrelationships
  - Amenable to computation

E.g., GO

- Controlled vocabulary
- Defined relationships
- "Directed acyclic graph"
  - Links are directional
  - No individually circular paths



http://lopasgen677s09.weebly.com/gene-ontology.htm

Gene Ontology (GO)

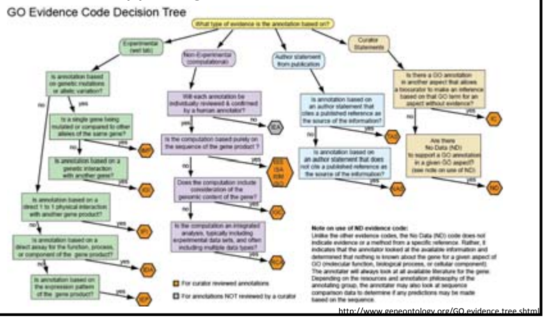
- http://www.geneontology.org/
- Consortium that defines standardized terms and relationships
- Centered on model organism databases
  - E.g., human, mouse, Drosophila, E.coli
  - Most curation derived from these sources, but do extend more broadly
- Linked and mapped to many other resources
- Used by many computational analysis tools

GO domains

- GO is divided into three domains, encompassing three separate functional properties
  - Biological process: what it does
  - Molecular function: how it does it
  - Cellular component: where it does it

GO evidence codes

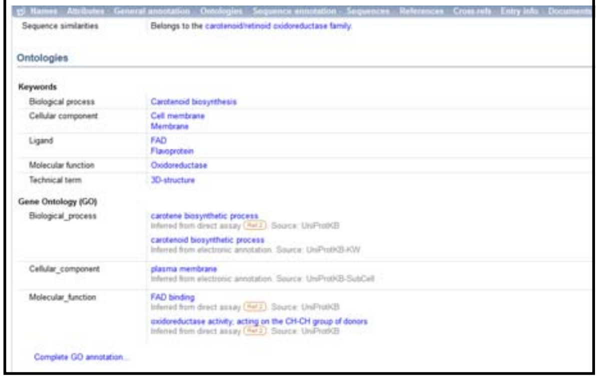
- GO uniquely has an ontology to describe the evidence supporting annotations


GO Evidence Code Decision Tree
What type of evidence is the annotation based on?
Experimental (wet lab)
Non-Experimental (computational)
Author statement from publication
Curator Statements
Is annotation based on genetic mutations or allelic variation?
Is a single gene being mutated or compared to other alleles of the same gene?
IMP
Is annotation based on a genetic interaction with another gene?
IGI
Is annotation based on a direct 1 to 1 physical interaction with another gene product?
IPI
Is annotation based on a direct assay for the function, process, or component of the gene product?
IDA
Is annotation based on the expression pattern of the gene product?
IEP
Will each annotation be individually reviewed & confirmed by a human annotator?
IEA
Is the computation based purely on the sequence of the gene product?
ISS
ISA
ISM
ISO
Does the computation include consideration of the genomic context of the gene?
IGC
Is the computation an integrated analysis, typically including experimental data sets, and often including multiple data types?
RCA
Is annotation based on an author statement that cites a published reference as the source of the information?
TAS
Is annotation based on an author statement that does not cite a published reference as the source of the information?
NAS
Is there a GO annotation in another aspect that allows a biocurator to make an inference based on that GO term for an aspect without evidence?
IC
Are there No Data (ND) to support a GO annotation in a given GO aspect? (see note on use of ND)
ND
yes
no
For curator reviewed annotations
For annotations NOT reviewed by a curator

Note on use of ND evidence code:
Unlike the other evidence codes, the No Data (ND) code does not indicate evidence or a method from a specific reference. Rather, it indicates that the annotator looked at the available information and determined that nothing is known about the gene for a given aspect of GO (molecular function, biological process, or cellular component). The annotator will always look at all available literature for the gene. Depending on the resources and annotation philosophy of the annotating group, the annotator may also look at sequence comparison data to determine if any predictions may be made based on the sequence.

http://www.geneontology.org/GO.evidence.tree.shtml

GO on UniProt


Names · Attributes · General annotation · Ontologies · Sequence annotation · Sequences · References · Cross-refs · Entry info · Documents

Sequence similarities — Belongs to the carotenoid/retinoid oxidoreductase family.

Ontologies

Keywords

Biological process	Carotenoid biosynthesis
Cellular component	Cell membrane Membrane
Ligand	FAD Flavoprotein
Molecular function	Oxidoreductase
Technical term	3D-structure

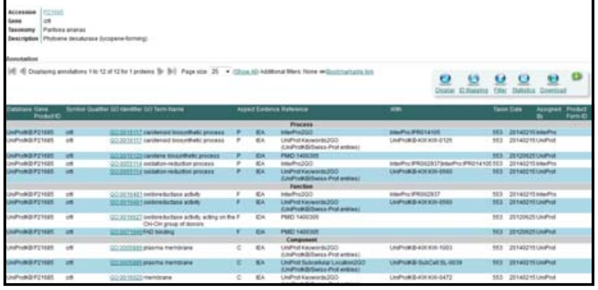
Gene Ontology (GO)

Biological_process	carotene biosynthetic process Inferred from direct assay (Ref.2). Source: UniProtKB carotenoid biosynthetic process Inferred from electronic annotation. Source: UniProtKB-KW
Cellular_component	plasma membrane Inferred from electronic annotation. Source: UniProtKB-SubCell
Molecular_function	FAD binding Inferred from direct assay (Ref.2). Source: UniProtKB oxidoreductase activity, acting on the CH-CH group of donors Inferred from direct assay (Ref.2). Source: UniProtKB

Complete GO annotation...

GO on UniProt

- Some GO annotations added manually
- Some mapped to other term databases



Accession	P21685
Gene	crtI
Taxonomy	Pantoea ananas
Description	Phytoene desaturase (lycopene-forming)

Annotation

Displaying annotations 1 to 12 of 12 for 1 proteins Page size: 25 (Show All) Additional filters: None Bookmarkable link

Database	Gene Product ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference	With	Taxon	Date	Assigned By	Product Form ID
					Process								
UniProtKB	P21685	crtI		GO:0016117	carotenoid biosynthetic process	P	IEA	InterPro2GO	InterPro:IPR014105	553	20140215	InterPro	
UniProtKB	P21685	crtI		GO:0016117	carotenoid biosynthetic process	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0125	553	20140215	UniProt	
UniProtKB	P21685	crtI		GO:0016120	carotene biosynthetic process	P	IDA	PMID 1400305		553	20120625	UniProt	
UniProtKB	P21685	crtI		GO:0055114	oxidation-reduction process	P	IEA	InterPro2GO	InterPro:IPR002937|InterPro:IPR014105	553	20140215	InterPro	
UniProtKB	P21685	crtI		GO:0055114	oxidation-reduction process	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0560	553	20140215	UniProt	
					Function								
UniProtKB	P21685	crtI		GO:0016491	oxidoreductase activity	F	IEA	InterPro2GO	InterPro:IPR002937	553	20140215	InterPro	
UniProtKB	P21685	crtI		GO:0016491	oxidoreductase activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0560	553	20140215	UniProt	
UniProtKB	P21685	crtI		GO:0016627	oxidoreductase activity, acting on the CH-CH group of donors	F	IDA	PMID 1400305		553	20120625	UniProt	
UniProtKB	P21685	crtI		GO:0071949	FAD binding	F	IDA	PMID 1400305		553	20120625	UniProt	
					Component								
UniProtKB	P21685	crtI		GO:0005886	plasma membrane	C	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-1003	553	20140215	UniProt	
UniProtKB	P21685	crtI		GO:0005886	plasma membrane	C	IEA	UniProt Subcellular Location2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-SubCell:SL-0039	553	20140215	UniProt	
UniProtKB	P21685	crtI		GO:0016020	membrane	C	IEA	UniProt Keywords2GO	UniProtKB-KW:KW-0472	553	20140215	UniProt	

Annotation families

- There are many different types of protein annotations, often with different foci and methods
- Hand vs. automatically generated
- Entire vs partial proteins

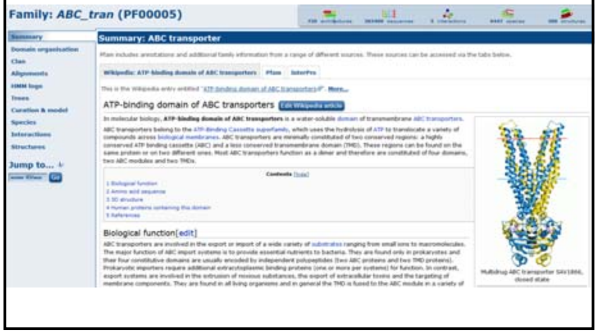
Pfam

- http://pfam.sanger.ac.uk/
- Originally constructed in the late 1990's for annotation of the *C. elegans* genome
- Developed & maintained by the Sanger Institute and S. Eddy (now Howard Hughes)
- Purpose: to overcome the % alignment problem inherit to BLAST
  - i.e., BLAST hits may not reflect homology over the entire query and/or reference sequence
- Currently (v27.0) 14,831 manually curated protein domain families

Pfam

- Pfam-A: manually selected and aligned alignments and HMMs of protein domains
  - v27.0: 14,831 families
  - At least 1 domain in 80% of proteins in UniProt
    - Figure is still scaling with database sizes
    - Represents 58% of total sequence in UniProt
- Pfam-B: automatically-generated families for domains not in Pfam-A
  - Mostly families with only a few members

Pfam example



Family: *ABC_tran* (PF00005)

Summary: ABC transporter

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

This is the Wikipedia entry entitled "ATP-binding domain of ABC transporters". More...

ATP-binding domain of ABC transporters

In molecular biology, **ATP-binding domain of ABC transporters** is a water-soluble domain of transmembrane ABC transporters.

ABC transporters belong to the ATP-Binding Cassette superfamily, which uses the hydrolysis of ATP to translocate a variety of compounds across biological membranes. ABC transporters are minimally constituted of two conserved regions: a highly conserved ATP binding cassette (ABC) and a less conserved transmembrane domain (TMD). These regions can be found on the same protein or on two different ones. Most ABC transporters function as a dimer and therefore are constituted of four domains, two ABC modules and two TMDs.

Contents [hide]

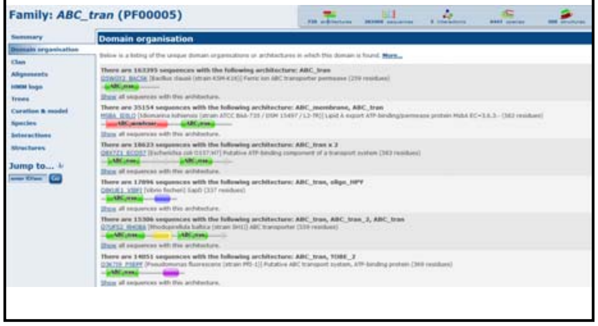
1 Biological function
2 Amino acid sequence
3 3D structure
4 Human proteins containing this domain
5 References

Biological function[edit]

ABC transporters are involved in the export or import of a wide variety of substrates ranging from small ions to macromolecules. The major function of ABC import systems is to provide essential nutrients to bacteria. They are found only in prokaryotes and their four constitutive domains are usually encoded by independent polypeptides (two ABC proteins and two TMD proteins). Prokaryotic importers require additional extracytoplasmic binding proteins (one or more per systems) for function. In contrast, export systems are involved in the extrusion of noxious substances, the export of extracellular toxins and the targeting of membrane components. They are found in all living organisms and in general the TMD is fused to the ABC module in a variety of

Multidrug ABC transporter SAV1866, closed state

Pfam example



Family: *ABC_tran* (PF00005)

Domain organisation

Below is a listing of the unique domain organisations or architectures in which this domain is found. More...

There are 163395 sequences with the following architecture: ABC_tran
Q5WDI2_BACSK [Bacillus clausii (strain KSM-K16)] Ferric ion ABC transporter permease (259 residues)
Show all sequences with this architecture.

There are 35154 sequences with the following architecture: ABC_membrane, ABC_tran
MSBA_IDILO [Idiomarina loihiensis (strain ATCC BAA-735 / DSM 15497 / L2-TR)] Lipid A export ATP-binding/permease protein MsbA EC=3.6.3.- (582 residues)
Show all sequences with this architecture.

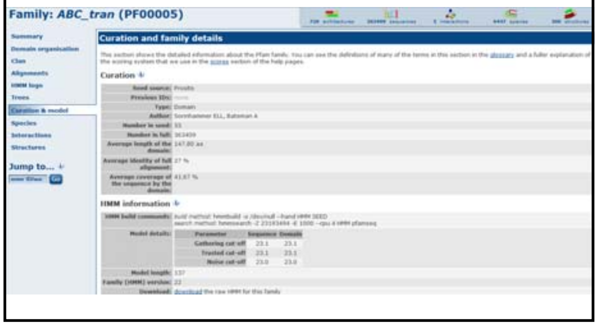
There are 18623 sequences with the following architecture: ABC_tran x 2
Q8X721_ECO57 [Escherichia coli O157:H7] Putative ATP-binding component of a transport system (503 residues)
Show all sequences with this architecture.

There are 17896 sequences with the following architecture: ABC_tran, oligo_HPY
Q8KUE1_VIBFI [Vibrio fischeri] SapD (337 residues)
Show all sequences with this architecture.

There are 15306 sequences with the following architecture: ABC_tran, ABC_tran_2, ABC_tran
Q7UFS2_RHOBA [Rhodopirellula baltica (strain SH1)] ABC transporter (559 residues)
Show all sequences with this architecture.

There are 14051 sequences with the following architecture: ABC_tran, TOBE_2
Q3K7I9_PSEPF [Pseudomonas fluorescens (strain Pf0-1)] Putative ABC transport system, ATP-binding protein (369 residues)
Show all sequences with this architecture.

Pfam example



Family: *ABC_tran* (PF00005)

Curation and family details

This section shows the detailed information about the Pfam family. You can see the definitions of many of the terms in this section in the glossary and a fuller explanation of the scoring system that we use in the scores section of the help pages.

Curation

Seed source:	Prosite
Previous IDs:	none
Type:	Domain
Author:	Sonnhammer ELL, Bateman A
Number in seed:	55
Number in full:	363409
Average length of the domain:	147.80 aa
Average identity of full alignment:	27 %
Average coverage of the sequence by the domain:	41.67 %

HMM information

HMM build commands: build method: hmmbuild -o /dev/null --hand HMM SEED
search method: hmmsearch -Z 23193494 -E 1000 --cpu 4 HMM pfamseq

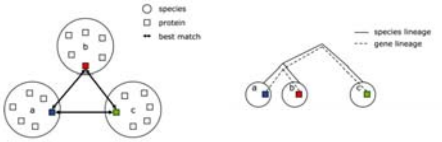
Model details:

Parameter	Sequence	Domain
Gathering cut-off	23.1	23.1
Trusted cut-off	23.1	23.1
Noise cut-off	23.0	23.0

Model length: 137
Family (HMM) version: 22
Download: download the raw HMM for this family

Clusters of Orthologus Groups (COG)

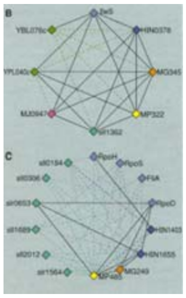
- One of the earliest attempts to define protein families by orthology (Tatusov et al 1997 Science)
- Used BLAST between proteins from multiple genomes to define triangles, i.e., triplets where each is a best match to the others



Kristensen et al. 2010 Bioinformatics 26:1481-1487

COG triangles

- Allows single-direction best hits
- Start with central triangle and add edges whenever possible
- Causes paralogs to be linked
- Allows distant & fast evolving homologs to be linked through intermediates



Sold lines: RBHs
Dotted lines: single direction

Tatusov et al 1997 Science 278: 631-637

COGs

- Bacterial COGs not updated often (last 2003)
- COGs more recently defined for other groups:
  - KOGs (eukaryotes)
  - arCOGs (archaea)
  - POGs (phages)
- Each COG family has a free-text annotation
  - 4873 families total
- Grouped into 24 superfamilies
  - COGs can belong to >1 superfamilies

eggNOG

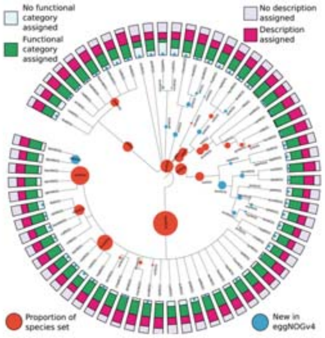
- ‘evolutionary genealogy of genes: Non-supervised Orthologous Groups’
- Constructed & maintained by EMBL (Peer Bork)
- Attempt to extend and update COG/KOG database annotations without requiring manual annotations (which do not scale)

eggNOG: method

- Use BLAST/fasta/Smith-Watterman alignments to find best matches
- Represent in-paralogs by single sequences
- Map sequences to COG/KOGs
- Triangle cluster non-matching sequences
- Add single RBH hits to clusters
- Automatically split multi-domain proteins
- Derive annotations by consensus within groups derived from multiple annotation sources

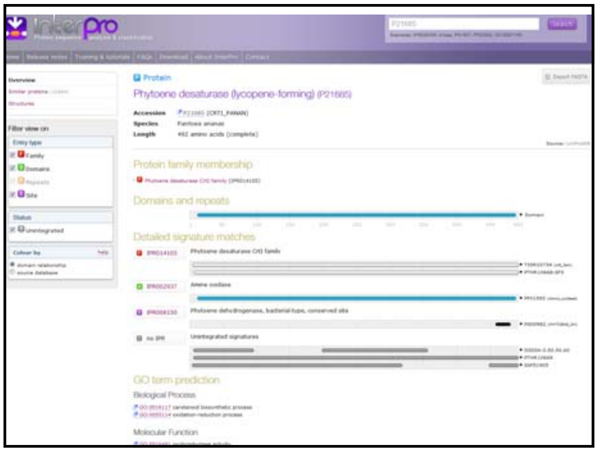
eggNOG:

- http://eggnog.embl.de/version_4.0.beta/
- 107 different annotation levels
- 1.7 million ortholog groups
- 7.7 million proteins
- Probably the currently most comprehensive ortholog database
- Can use to construct PSSMs/HMMs


No functional category assigned
Functional category assigned
No description assigned
Description assigned
Proportion of species set
New in eggNOGv4

Interpro

- Classifies proteins according to a combination of multiple protein motifs
- Multiple sources synthesized into single Interpro classification system
  - Four broad annotation types: Family, Domains, Repeats, Sites
- Interpro terms mapped to GO
- InterProScan – resource to annotate proteins using all member databases
  - HMM and regular expression-based classifications



Protein

Phytoene desaturase (lycopene-forming) (P21685)

Accession P21685 (CRTI_PANAN)

Species Pantoea ananas

Length 492 amino acids (complete)

Protein family membership

Phytoene desaturase CrtI family (IPR014105)

Domains and repeats

Detailed signature matches

IPR014105 Phytoene desaturase CrtI family

IPR002937 Amine oxidase

IPR008150 Phytoene dehydrogenase, bacterial-type, conserved site

no IPR Unintegrated signatures

GO term prediction

Biological Process

GO:0016117 carotenoid biosynthetic process

GO:0055114 oxidation-reduction process

Molecular Function

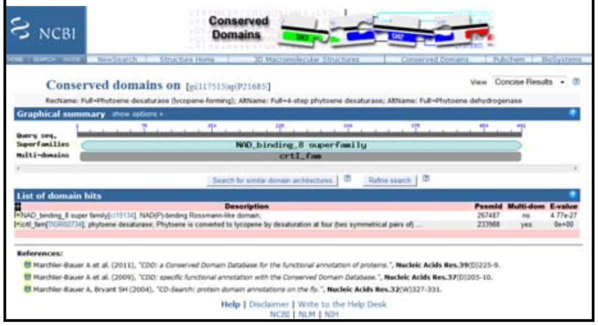
Interpro: member databases

- Pfam (domains, curated; Sanger)
- PROSITE (diagnostic motifs; SIB)
- HAMAP (homologs, curated; SIB)
- PRINTS (conserved motifs; U. Manchester)
- ProDom (domains, automatic via PSI-BLAST; PRABI Villerubanne)
- SMART (domains and architectures esp. signaling, curated; EMBL)
- TIGRFAMs (homologs, curated; JCVI)
- PIRSF (homologs & domains, ; Georgetown)
- SUPERFAMILY (structures, curated, U Bristol)
- CATH-Gene3D (homologs, mapped to structures, automatic via Markov clustering; University College London)
- PANTHER (functional homologs, curated, USC)

Conserved Domains Database (CDD)

- Protein classification database maintained by NCBI
- CDD database based on domains curated by NCBI using structural alignments
- Also includes external resources: Pfam, SMART, COG, PRK, TIGRFAM
- Downloadable PSSMs for each CDD family for querying via RPS-BLAST

CD-Search



Conserved domains on [gi|117515|sp|P21685]

View Concise Results

RecName: Full=Phytoene desaturase (lycopene-forming); AltName: Full=4-step phytoene desaturase; AltName: Full=Phytoene dehydrogenase

Graphical summary show options »

Query seq.
Superfamilies NAD_binding_8 superfamily
Multi-domains crtI_fam

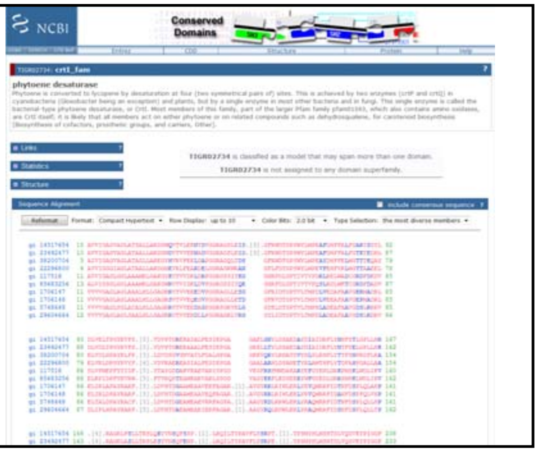
Search for similar domain architectures | Refine search

List of domain hits

Description	Pssmld	Multi-dom	E-value
NAD_binding_8 super family[cl19134], NAD(P)-binding Rossmann-like domain;	267487	no	4.77e-27
crtI_fam[TIGR02734], phytoene desaturase; Phytoene is converted to lycopene by desaturation at four (two symmetrical pairs of) ...	233988	yes	0e+00

References:
- Marchler-Bauer A et al. (2011), "CDD: a Conserved Domain Database for the functional annotation of proteins.", Nucleic Acids Res.39(D)225-9.
- Marchler-Bauer A et al. (2009), "CDD: specific functional annotation with the Conserved Domain Database.", Nucleic Acids Res.37(D)205-10.
- Marchler-Bauer A, Bryant SH (2004), "CD-Search: protein domain annotations on the fly.", Nucleic Acids Res.32(W)327-331.

Help | Disclaimer | Write to the Help Desk
NCBI | NLM | NIH



TIGR02734: crtI_fam

phytoene desaturase

Phytoene is converted to lycopene by desaturation at four (two symmetrical pairs of) sites. This is achieved by two enzymes (crtP and crtQ) in cyanobacteria (Gloeobacter being an exception) and plants, but by a single enzyme in most other bacteria and in fungi. This single enzyme is called the bacterial-type phytoene desaturase, or CrtI. Most members of this family, part of the larger Pfam family pfam01593, which also contains amino oxidases, are CrtI itself; it is likely that all members act on either phytoene or on related compounds such as dehydrosqualene, for carotenoid biosynthesis [Biosynthesis of cofactors, prosthetic groups, and carriers, Other].

Links
Statistics
Structure

TIGR02734 is classified as a model that may span more than one domain.
TIGR02734 is not assigned to any domain superfamily.

Sequence Alignment ☐ include consensus sequence

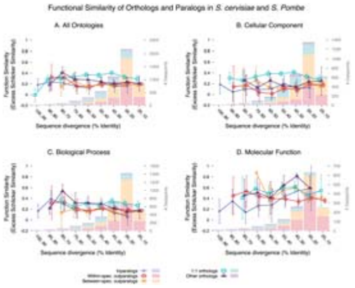
Reformat Format: Compact Hypertext Row Display: up to 10 Color Bits: 2.0 bit Type Selection: the most diverse members

Are functions actually conserved?

- All of the protein annotation methods that we have discussed assume the hypothesis that function is evolutionarily conserved
- But we know that this can be confounded by duplication/loss and xenology
  - Can be addressed by better methods of determining orthology
  - Not typically accommodated by annotation databases
- Even orthologous functions can drift and/or be promiscuous

Are functions actually conserved?

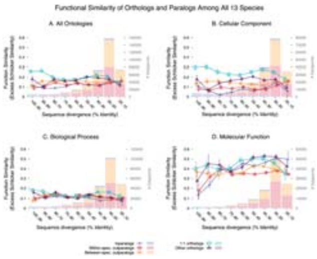
- Compare curated GO annotations of orthologs and paralogs
  - Corrected for annotation biases
- Functions of orthologs more similar than paralogs, but not perfectly



Altenhoff et al. (2012) PLoS Comput. Biol. 8:e1002514

Are functions actually conserved?

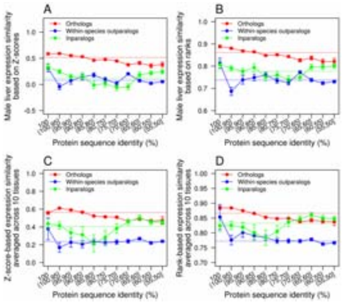
- Same for 13 species instead of just 2
  - Paralogy potentially a greater confounder
  - Differences in GO annotation completeness



Altenhoff et al. (2012) PLoS Comput. Biol. 8:e1002514

Are functions actually conserved?

- Define function as expression similarity between same human and mice tissues
- Same trend: ortholog function more conserved than paralogs, not absolute



Chen & Zhange (2012) PLoS Comput. Biol. 8:e1002784

Are functions actually conserved?

- Yes, but not perfectly even for highly conserved sequences
- Likely depends on definition of “function”
- Annotated functions are likely quite broad in most cases

Protein database vs. pathways and reactions

- Protein databases are based on homology
  - Hypothesis that function is conserved
- Reaction databases classify function without reference to homology
  - Function can be due to evolutionary convergence
  - GO is an example of this we have already seen
- Reaction and pathway annotations are therefore closer to function but further from underlying evolutionary mechanism

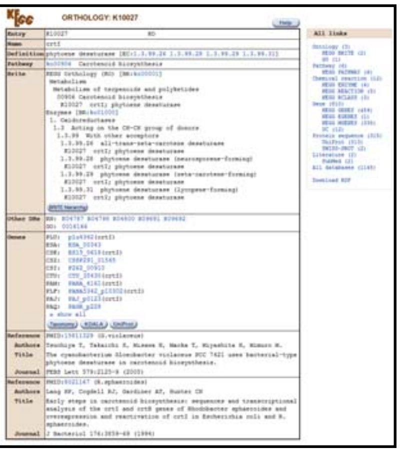
Enyzme Commission

- One of the oldest functional annotation schemes, arising out of biochemistry
- Four part numerical nomenclature having increasing specificity
  - EC 3: hydrolases
  - EC 3.4: hydrolases acting on peptide bonds
  - EC 3.4.11: hydrolases cleaving amino-terminal amino acids from a peptide
  - EC 3.4.11.4: hydrolases cleaving amino-terminal amino acids from a tripeptide
- Database updates are infrequent

Kyoto Encyclopedia of Genes and Genomes (KEGG)

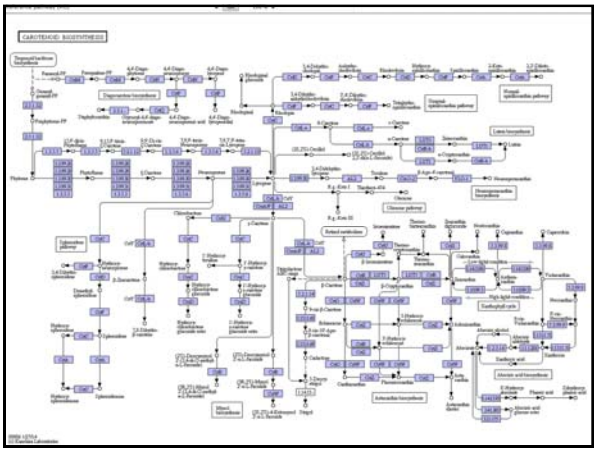
- Manually edited pathway database
- Orthologs defined in other genomes
- Reactions combined into metabolic maps
  - Pathways are typically quite general
- Individual proteins can be freely queried via web
- Individual genomes can be annotated via KAAS server
- Underlying database NO LONGER FREE

KEGG example

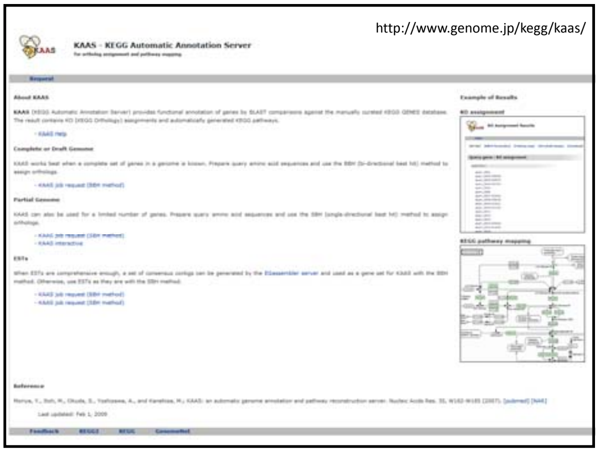

ORTHOLOGY: K10027

Entry	K10027 KO
Name	crtI
Definition	phytoene desaturase [EC:1.3.99.26 1.3.99.28 1.3.99.29 1.3.99.31]
Pathway	ko00906 Carotenoid biosynthesis
Brite	KEGG Orthology (KO) [BR:ko00001] Metabolism Metabolism of terpenoids and polyketides 00906 Carotenoid biosynthesis K10027 crtI; phytoene desaturase Enzymes [BR:ko01000] 1. Oxidoreductases 1.3 Acting on the CH-CH group of donors 1.3.99 With other acceptors 1.3.99.26 all-trans-zeta-carotene desaturase K10027 crtI; phytoene desaturase 1.3.99.28 phytoene desaturase (neurosporene-forming) K10027 crtI; phytoene desaturase 1.3.99.29 phytoene desaturase (zeta-carotene-forming) K10027 crtI; phytoene desaturase 1.3.99.31 phytoene desaturase (lycopene-forming) K10027 crtI; phytoene desaturase BRITE hierarchy
Other DBs	RN: R04787 R04798 R04800 R09691 R09692 GO: 0016166
Genes	PLU: plu4342(crtI) ESA: ESA_00343 CSK: ES15_0418(crtI) CSZ: CSSP291_01565 CSI: P262_00913 CTU: CTU_35430(crtI) PAM: PANA_4142(crtI) PLF: PANA5342_p10302(crtI) PAJ: PAJ_p0123(crtI) PAQ: PAGR_p228 » show all Taxonomy KOALA UniProt
Reference	PMID:15811329 (G.violaceus)
Authors	Tsuchiya T, Takaichi S, Misawa N, Maoka T, Miyashita H, Mimuro M.
Title	The cyanobacterium Gloeobacter violaceus PCC 7421 uses bacterial-type phytoene desaturase in carotenoid biosynthesis.
Journal	FEBS Lett 579:2125-9 (2005)
Reference	PMID:8021147 (R.sphaeroides)
Authors	Lang HP, Cogdell RJ, Gardiner AT, Hunter CN
Title	Early steps in carotenoid biosynthesis: sequences and transcriptional analysis of the crtI and crtB genes of Rhodobacter sphaeroides and overexpression and reactivation of crtI in Escherichia coli and R. sphaeroides.
Journal	J Bacteriol 176:3859-69 (1994)

All links
Ontology (3)
KEGG BRITE (2)
GO (1)
Pathway (4)
KEGG PATHWAY (4)
Chemical reaction (12)
KEGG ENZYME (4)
KEGG REACTION (5)
KEGG RCLASS (3)
Gene (515)
KEGG GENES (494)
KEGG EGENES (1)
KEGG MGENES (10)
DC (12)
Protein sequence (515)
UniProt (513)
SWISS-PROT (2)
Literature (2)
PubMed (2)
All databases (1145)
Download RDF



http://www.genome.jp/kegg/kaas/



KAAS - KEGG Automatic Annotation Server

For ortholog assignment and pathway mapping

Request

About KAAS

KAAS (KEGG Automatic Annotation Server) provides functional annotation of genes by BLAST comparisons against the manually curated KEGG GENES database. The result contains KO (KEGG Orthology) assignments and automatically generated KEGG pathways.

- KAAS help

Complete or Draft Genome

KAAS works best when a complete set of genes in a genome is known. Prepare query amino acid sequences and use the BBH (bi-directional best hit) method to assign orthologs.

- KAAS job request (BBH method)

Partial Genome

KAAS can also be used for a limited number of genes. Prepare query amino acid sequences and use the SBH (single-directional best hit) method to assign orthologs.

- KAAS job request (SBH method)
- KAAS interactive

ESTs

When ESTs are comprehensive enough, a set of consensus contigs can be generated by the EGassembler server and used as a gene set for KAAS with the BBH method. Otherwise, use ESTs as they are with the SBH method.

- KAAS job request (BBH method)
- KAAS job request (SBH method)

Reference

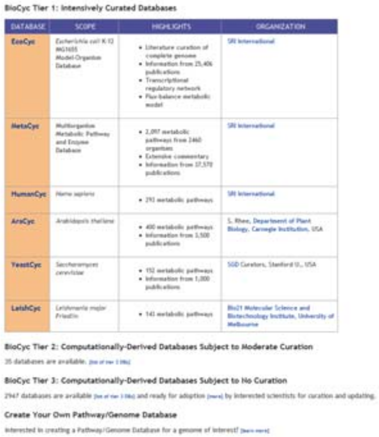
Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A., and Kanehisa, M.; KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35, W182-W185 (2007). [pubmed] [NAR]

Last updated: Feb 1, 2009

Feedback KEGG2 KEGG GenomeNet

Biocyc

- Collection of curated metabolic pathways
- Typically smaller modules compared to KEGG
- www.biocyc.org



BioCyc Tier 1: Intensively Curated Databases

DATABASE	SCOPE	HIGHLIGHTS	ORGANIZATION
EcoCyc	*Escherichia coli* K-12 MG1655 Model-Organism Database	• Literature curation of complete genome • Information from 25,406 publications • Transcriptional regulatory network • Flux-balance metabolic model	SRI International
MetaCyc	Multiorganism Metabolic Pathway and Enzyme Database	• 2,097 metabolic pathways from 2460 organisms • Extensive commentary • Information from 37,570 publications	SRI International
HumanCyc	*Homo sapiens*	• 293 metabolic pathways	SRI International
AraCyc	*Arabidopsis thaliana*	• 400 metabolic pathways • Information from 3,500 publications	S. Rhee, Department of Plant Biology, Carnegie Institution, USA
YeastCyc	*Saccharomyces cerevisiae*	• 152 metabolic pathways • Information from 1,000 publications	SGD Curators, Stanford U., USA
LeishCyc	*Leishmania major* Friedlin	• 141 metabolic pathways	Bio21 Molecular Science and Biotechnology Institute, University of Melbourne

BioCyc Tier 2: Computationally-Derived Databases Subject to Moderate Curation

35 databases are available. [list of tier 2 DBs]

BioCyc Tier 3: Computationally-Derived Databases Subject to No Curation

2947 databases are available [list of tier 3 DBs] and ready for adoption [more] by interested scientists for curation and updating.

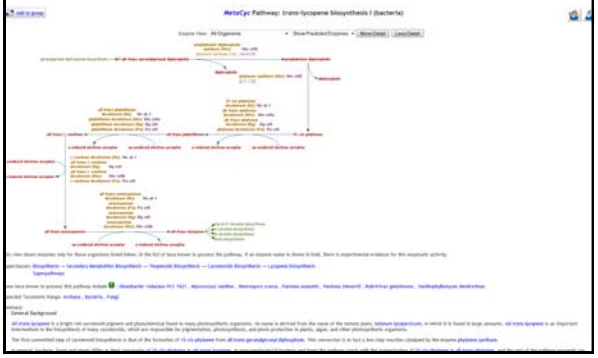
Create Your Own Pathway/Genome Database

Interested in creating a Pathway/Genome Database for a genome of interest? [learn more]

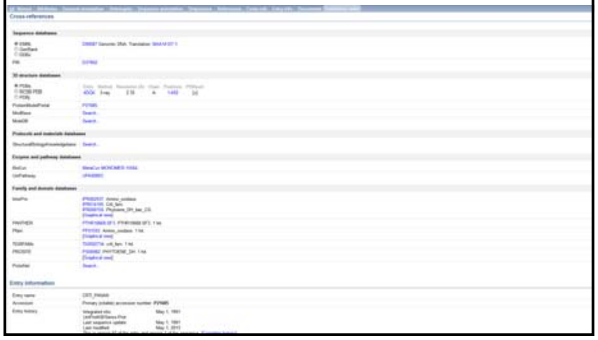
Metacyc example

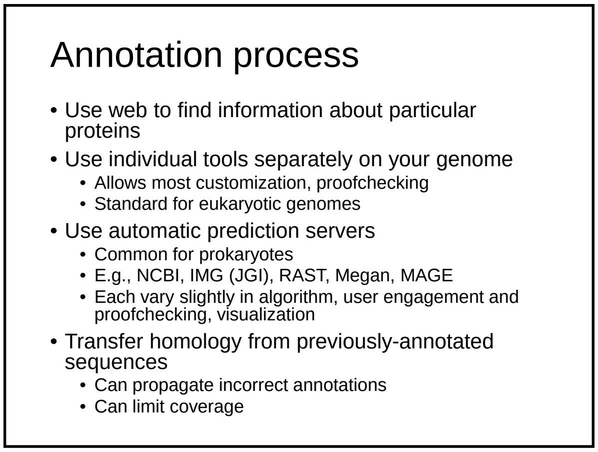
Metacyc example

Metacyc example



Uniprot cross-references Interpro, metacyc





Annotation process

- Use web to find information about particular proteins
- Use individual tools separately on your genome
  - Allows most customization, proofchecking
  - Standard for eukaryotic genomes
- Use automatic prediction servers
  - Common for prokaryotes
  - E.g., NCBI, IMG (JGI), RAST, Megan, MAGE
  - Each vary slightly in algorithm, user engagement and proofchecking, visualization
- Transfer homology from previously-annotated sequences
  - Can propagate incorrect annotations
  - Can limit coverage