

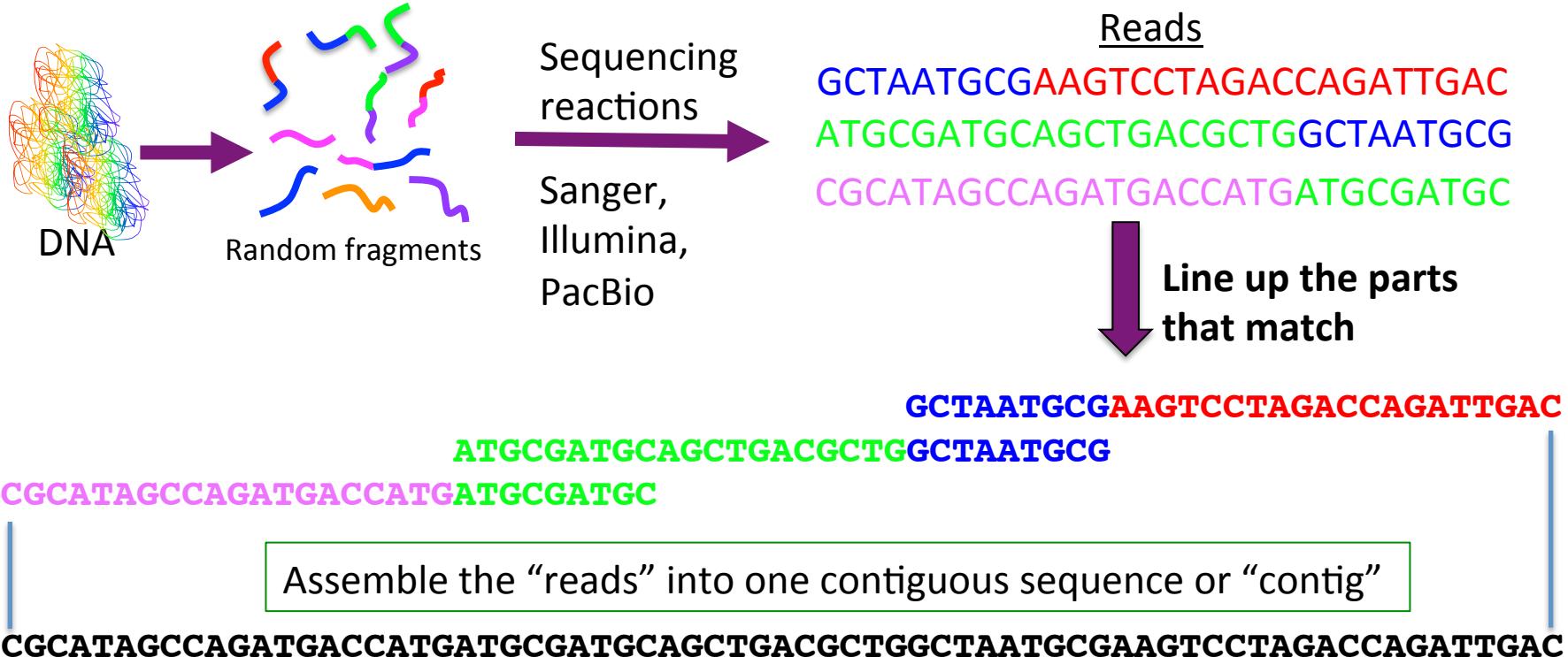
# The IGS Prokaryotic Annotation Pipeline

Michelle Gwinn Giglio

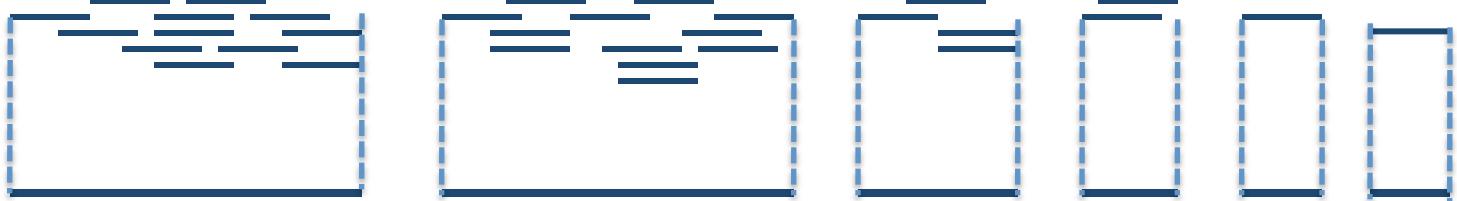
ASM Workshop 2016

## CloVR vs Analysis Engine

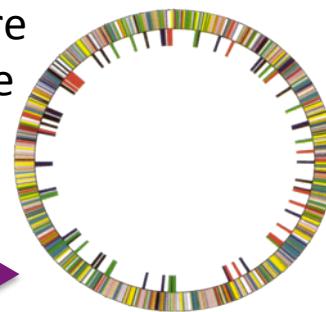
- Pipelines not identical
- Since CloVR pipeline can be distributed and run outside of IGS, there are licensing considerations
  - Center for Biological Analysis tools not included in CloVR
- Otherwise pipelines are the same



Of course real projects have millions of reads, and depending on the length and quality of reads this will result in many (or perhaps a few or one) contigs



Considerably more work and expense unless you are using PacBio

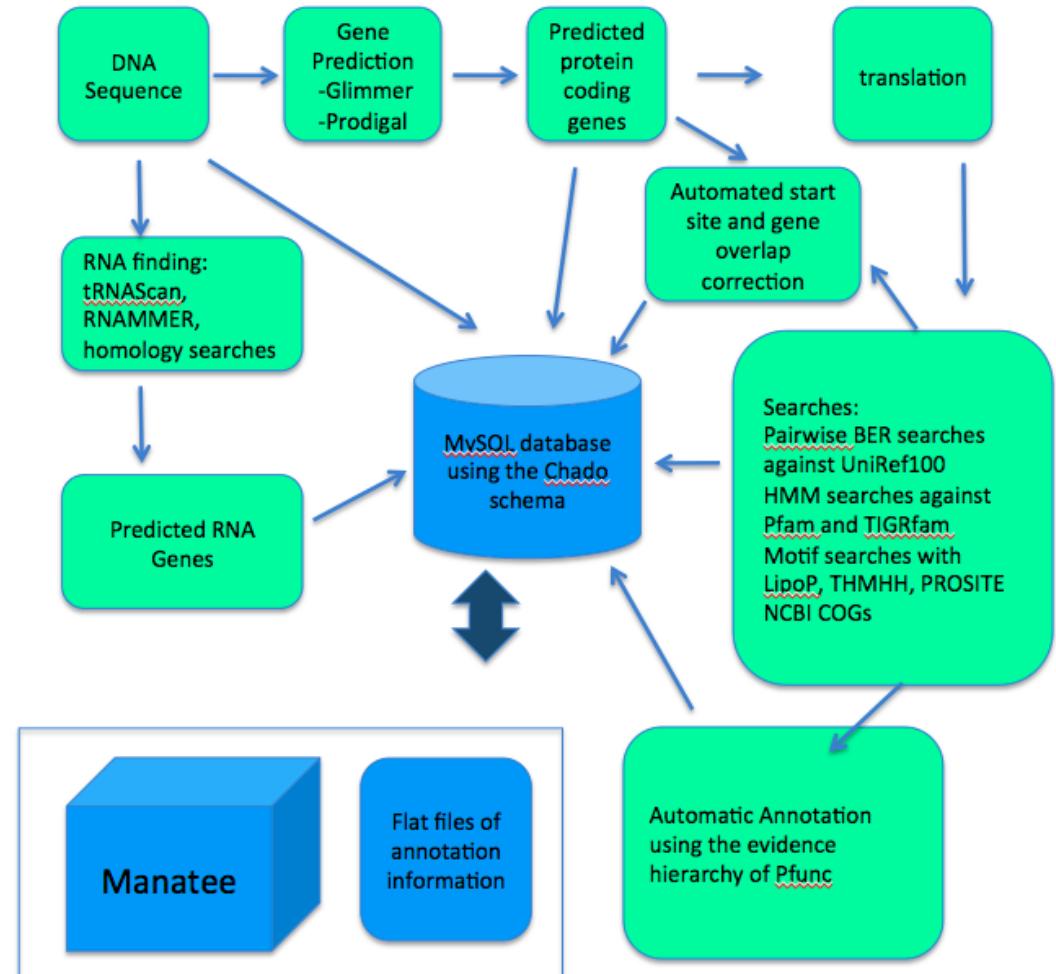


"Gold Standard" finished genome

# Steps in the pipeline

---

- Assembly
- Find genes (coding and non-coding)
- Gather evidence for functional annotation
- Make functional annotations
- View data

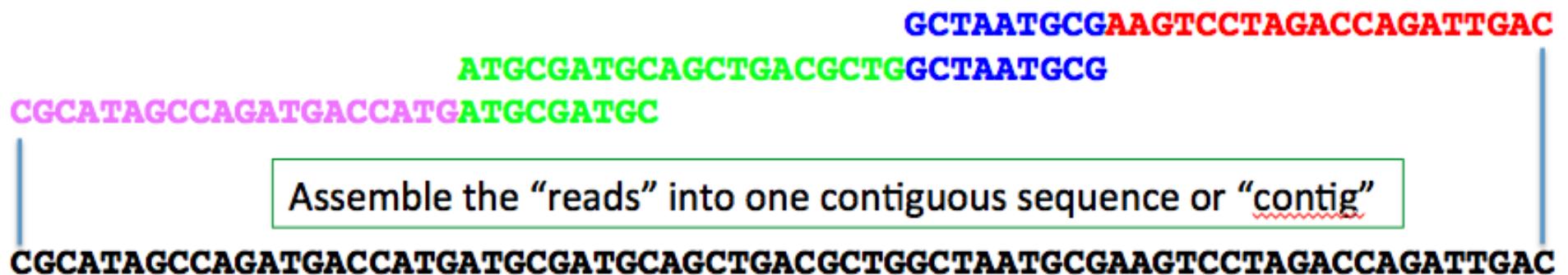


# Assembly

---

- Two methods available in CloVR, depending on the sequencing technology that was used
  - 454: Celera Assembler
  - Illumina: Spades
- Spades available as a service in Analysis Engine

---

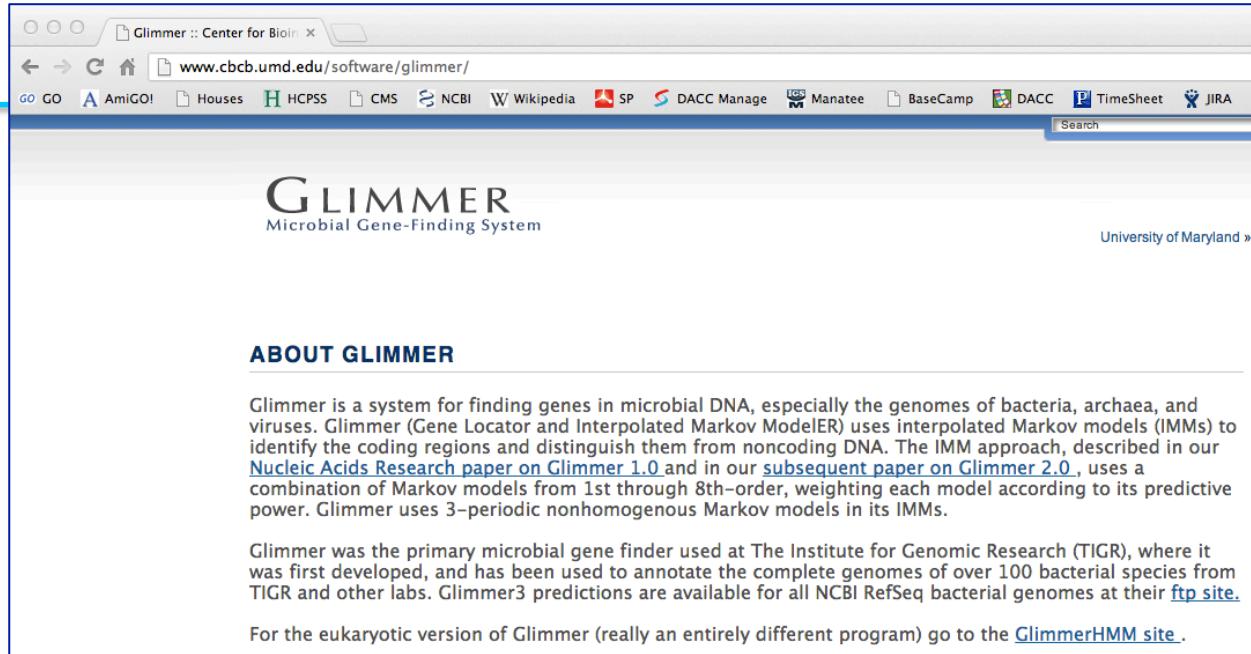


GCTAATGCGAAGTCCTAGACCAGATTGAC  
ATGCGATGCAGCTGACGCTGGCTAATGCG  
CGCATAGCCAGATGACCATGATGCGATGC  
CGCATAGCCAGATGACCATGATGCGATGCAGCTGACGCTGGCTAATGCGAAGTCCTAGACCAGATTGAC

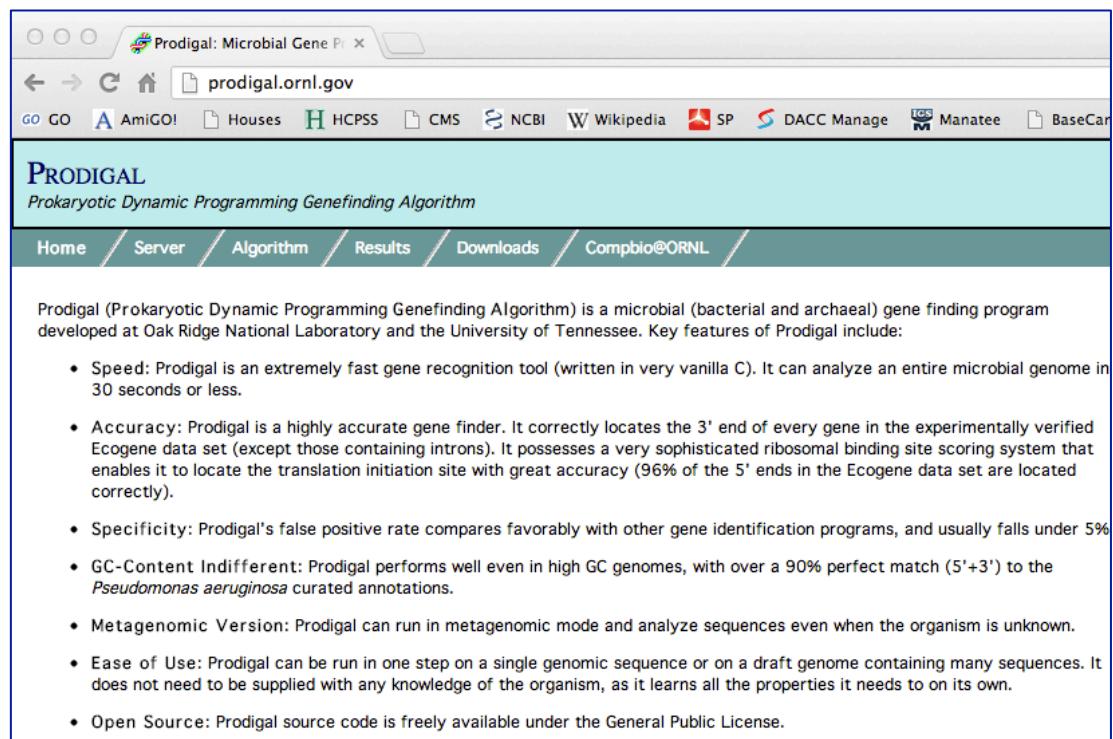
Assemble the “reads” into one contiguous sequence or “contig”

# Gene Finding

- Analysis Engine  
(choose from)
  - Glimmer3
  - Prodigal
- CloVR
  - Glimmer3



The screenshot shows a web browser window for the Glimmer Microbial Gene-Finding System. The URL is [www.cbcn.umd.edu/software/glimmer/](http://www.cbcn.umd.edu/software/glimmer/). The page title is "GLIMMER Microbial Gene-Finding System". A sidebar on the right includes a link to "University of Maryland >". The main content area is titled "ABOUT GLIMMER" and contains text about the system's history and methodology. It mentions that Glimmer uses interpolated Markov models (IMMs) to identify coding regions and distinguish them from noncoding DNA. The IMM approach, described in a [Nucleic Acids Research paper on Glimmer 1.0](#) and a [subsequent paper on Glimmer 2.0](#), uses a combination of Markov models from 1st through 8th-order, weighting each model according to its predictive power. Glimmer uses 3-periodic nonhomogenous Markov models in its IMMs. Below this, it states that Glimmer was the primary microbial gene finder used at The Institute for Genomic Research (TIGR), where it was first developed, and has been used to annotate the complete genomes of over 100 bacterial species from TIGR and other labs. Glimmer3 predictions are available for all NCBI RefSeq bacterial genomes at their [ftp site](#). A note at the bottom indicates that the eukaryotic version of Glimmer (really an entirely different program) can be found at the [GlimmerHMM site](#).



The screenshot shows a web browser window for the Prodigal Prokaryotic Dynamic Programming Genefinding Algorithm. The URL is [prodigal.orgnl.gov](http://prodigal.orgnl.gov). The page title is "PRODIGAL Prokaryotic Dynamic Programming Genefinding Algorithm". The navigation menu includes "Home", "Server", "Algorithm", "Results", "Downloads", and "Compbio@ORNL". The main content area discusses the features of Prodigal, stating it is a microbial (bacterial and archaeal) gene finding program developed at Oak Ridge National Laboratory and the University of Tennessee. Key features listed include:

- Speed: Prodigal is an extremely fast gene recognition tool (written in very vanilla C). It can analyze an entire microbial genome in 30 seconds or less.
- Accuracy: Prodigal is a highly accurate gene finder. It correctly locates the 3' end of every gene in the experimentally verified Ecogene data set (except those containing introns). It possesses a very sophisticated ribosomal binding site scoring system that enables it to locate the translation initiation site with great accuracy (96% of the 5' ends in the Ecogene data set are located correctly).
- Specificity: Prodigal's false positive rate compares favorably with other gene identification programs, and usually falls under 5%.
- GC-Content Indifferent: Prodigal performs well even in high GC genomes, with over a 90% perfect match (5'+3') to the *Pseudomonas aeruginosa* curated annotations.
- Metagenomic Version: Prodigal can run in metagenomic mode and analyze sequences even when the organism is unknown.
- Ease of Use: Prodigal can be run in one step on a single genomic sequence or on a draft genome containing many sequences. It does not need to be supplied with any knowledge of the organism, as it learns all the properties it needs to on its own.
- Open Source: Prodigal source code is freely available under the General Public License.

# tRNAscan: <http://selab.janelia.org/tRNAscan-SE/>

## Brief Description

tRNAscan-SE identifies transfer RNA genes in genomic DNA or RNA sequences. It combines the specificity of the Cove probabilistic RNA prediction package (Eddy & Durbin, 1994) with the speed and sensitivity of tRNAscan 1.3 (Fichant & Burks, 1991) plus an implementation of an algorithm described by Pavesi and colleagues (1994) which searches for eukaryotic pol III tRNA promoters (our implementation referred to as EufindtRNA). tRNAscan and EufindtRNA are used as first-pass prefilters to identify "candidate" tRNA regions of the sequence. These subsequences are then passed to Cove for further analysis, and output if Cove confirms the initial tRNA prediction. In this way, tRNAscan-SE attains the best of both worlds:

- a false positive rate of less than one per 15 billion nucleotides of random sequence
- the combined sensitivities of tRNAscan and EufindtRNA (detection of 99% of true tRNAs)
- search speed 1,000 to 3,000 times faster than Cove analysis and 30 to 90 times faster than the original tRNAscan 1.3 (tRNAscan-SE uses both a code-optimized version of tRNAscan 1.3 which gives a 650-fold increase in speed, and a fast C implementation of the Pavesi *et al.* algorithm).

This program and results of its analysis of a number of genomes have been published in Lowe & Eddy, *Nucleic Acids Research* **25**: 955-964 (1997).

HHMI  
janelia farm  
research campus

Eddy Lab: **tRNAscan-SE**

EDDY LAB

tRNAscan-SE 1.21

User Manual for command-line UNIX version of program

If you would like to run tRNAscan-SE locally, you can get the UNIX [source code](#) (compressed tar file).

Analyzing tRNAs in a published genome? See our own tRNAscan-SE analyses of completed genomes in the [Genomic tRNA Database](#).

Search Mode: Default      Source: Mixed (general tRNA model)

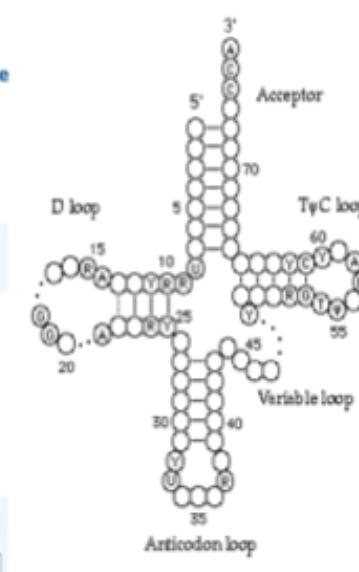
Format:

Raw Sequence  
Sequence name (optional):  (no spaces)

Other (FASTA, GenBank, EMBL, GCG, IG)

Paste your query sequence(s) here:  Run tRNAscan-SE

(The web-server may experience problems if submitted queries total more than 100K nucleotides at any one time)



# Other ncRNAs

The screenshot shows a web browser window for the Infernal website at [infernal.janelia.org](http://infernal.janelia.org). The title bar says "Infernal: inference of RNA alignments". The page features a large green logo on the left with the word "infernal" written vertically. Below the logo, the text "Infernal: inference of RNA alignments" is centered. At the bottom, there are links to "infernal home", "rfam database", "eddy lab", and "janella farm".

The screenshot shows a web browser window for the Rfam website at [rfam.sanger.ac.uk](http://rfam.sanger.ac.uk). The title bar says "Rfam: Home page". The header includes the Wellcome Trust Sanger Institute logo and links for HOME, SEARCH, BROWSE, BIOMART, BLOG, and HELP. The main content area displays "Rfam 11.0 (August 2012, 2208 families)". It describes the Rfam database as a collection of RNA families, each represented by multiple sequence alignments, consensus secondary structures, and covariance models (CMs). It also features a "QUICK LINKS" section with links to Sequence Search, View an RFAM Family, View an RFAM CLAN, Keyword Search, and Taxonomy Search. A "JUMP TO" search bar is present, along with a link to the help pages.

# Functional Annotation

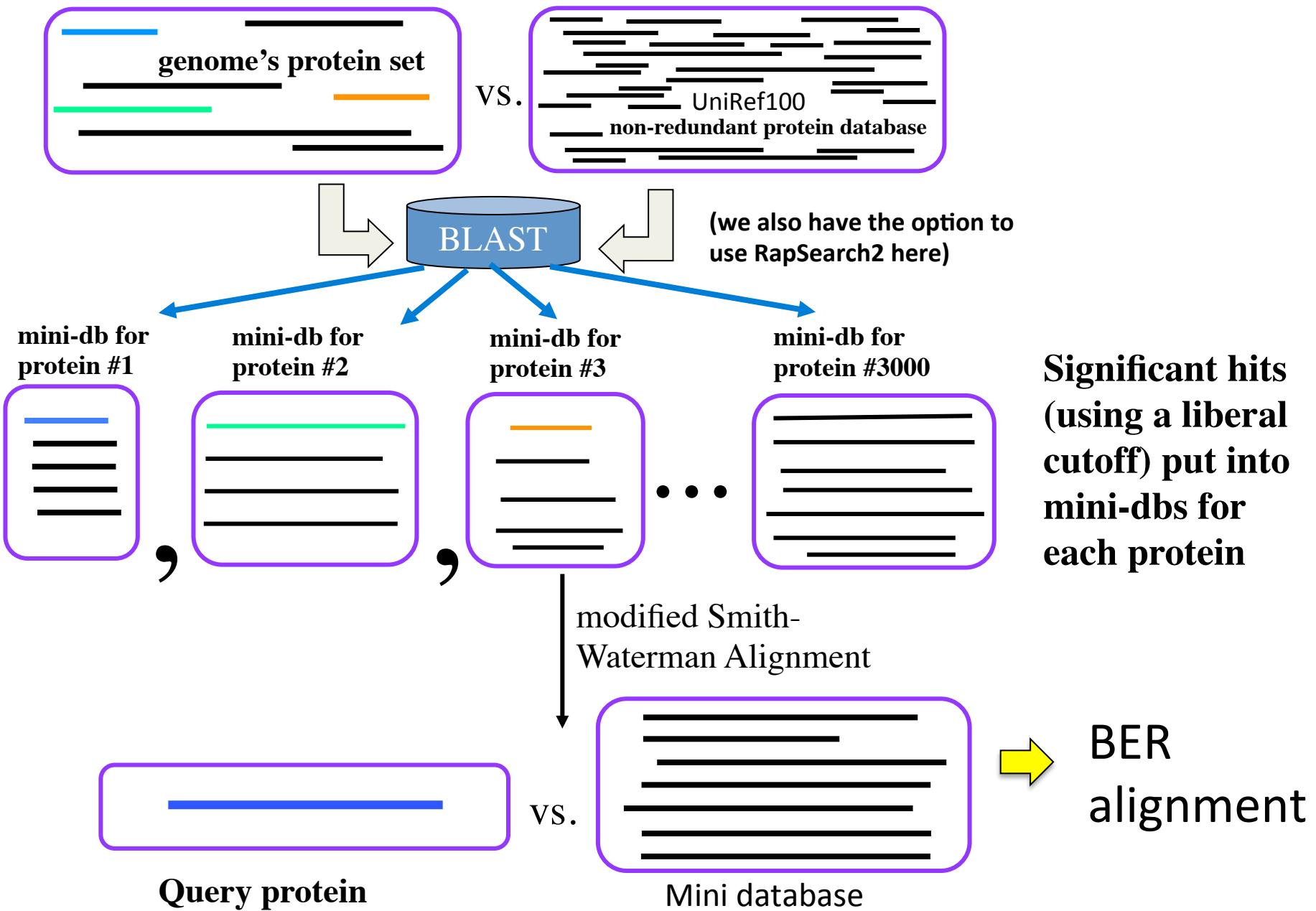
- Sequence similarity is the basis of functional assignments in our pipeline
  - Protein (not nucleotide) alignments for functional prediction
- Remember: one aa difference can change the function of a protein
  - All sequence-based annotations are putative until experimentally confirmed.

		nonpolar polar basic acidic (stop codon)			
		2nd base			
		U	C	A	G
U	UUU	(Phe/F) Phenylalanine	UCU (Ser/S) Serine	UAU (Tyr/Y) Tyrosine	UGU (Cys/C) Cysteine
	UUC	(Phe/F) Phenylalanine	UCC (Ser/S) Serine	UAC (Tyr/Y) Tyrosine	UGC (Cys/C) Cysteine
	UUA	(Leu/L) Leucine	UCA (Ser/S) Serine	UAA Stop (Ochre)	UGA Stop (Opal)
	UUG	(Leu/L) Leucine	UCG (Ser/S) Serine	UAG Stop (Amber)	UGG (Trp/W) Tryptophan
C	CUU	(Leu/L) Leucine	CCU (Pro/P) Proline	CAU (His/H) Histidine	CGU (Arg/R) Arginine
	CUC	(Leu/L) Leucine	CCC (Pro/P) Proline	CAC (His/H) Histidine	CGC (Arg/R) Arginine
	CUA	(Leu/L) Leucine	CCA (Pro/P) Proline	CAA (Gln/Q) Glutamine	CGA (Arg/R) Arginine
	CUG	(Leu/L) Leucine	CCG (Pro/P) Proline	CAG (Gln/Q) Glutamine	CGG (Arg/R) Arginine
A	AUU	(Ile/I) Isoleucine	ACU (Thr/T) Threonine	AAU (Asn/N) Asparagine	AGU (Ser/S) Serine
	AUC	(Ile/I) Isoleucine	ACC (Thr/T) Threonine	AAC (Asn/N) Asparagine	AGC (Ser/S) Serine
	AUA	(Ile/I) Isoleucine	ACA (Thr/T) Threonine	AAA (Lys/K) Lysine	AGA (Arg/R) Arginine
	AUG <sup>[A]</sup>	(Met/M) Methionine	ACG (Thr/T) Threonine	AAG (Lys/K) Lysine	AGG (Arg/R) Arginine
G	GUU	(Val/V) Valine	GCU (Ala/A) Alanine	GAU (Asp/D) Aspartic acid	GGU (Gly/G) Glycine
	GUC	(Val/V) Valine	GCC (Ala/A) Alanine	GAC (Asp/D) Aspartic acid	GGC (Gly/G) Glycine
	GUA	(Val/V) Valine	GCA (Ala/A) Alanine	GAA (Glu/E) Glutamic acid	GGG (Gly/G) Glycine
	GUG	(Val/V) Valine	GCG (Ala/A) Alanine	GAG (Glu/E) Glutamic acid	GGG (Gly/G) Glycine

# Gather Evidence

---

# Pairwise Alignments – IGS uses BER (blast\_extend\_repraze)



# BER Alignment

ORF04813( 7 - 350 of 351 aa)

SP|P12996|BIOB\_ECOLI(4 - 346 of 346) Biotin synthase (EC 2.8.1.6) (Biotin synthetase).

%Match = 42.2

%Identity = 66.0 %Similarity = 79.7

Matches = 227 Mismatches = 69 Conservative Sub.s = 47

Gaps = 1 InDels = 3 Frame Shifts = 0

Primary Frame = 1 [343, 0, 0]

tcctgtcccacgcacgctgccacggcgttataaggatgctcacacgtacagtagattggactttttagtgcatttc  
gaaatagagctccggtgagtcgaaataacggaaagcaaggaagaagttagcaatccttaaaacttataaagtctctagtgac  
ctaattctaatatcctgctacaagagcggagggtgtcatgctatataatcaagcagggtataaagacgtcgaaaataggacga  
-81 -71 -61 -51 -41 -31 -21 -11

CHQ\*VYGHPIPARIPLGHCVPRKQE**VHESWRYKGAKYGRYQSQNVRNCA**\*KLTALEN\***SVVNLHYWLALT**V\*GLRLS\*P

ataaaaaagttatctcgccgtacggagggttgcacagtttagcaaccgggtgcaggcaactttaaagggtcggtattccagctg  
gaacataggctcatatgaagagaataacttcttaatttacagtagaaaaacaatatggttctacgcgcagaacgcaggcga  
gataaaaagtagggattttggaaacacatggggtcaatacctcctagttcaggcccaggcatgtttttattgggtcc

-1 10 20 30 40 50 60 70

R\*NTKIKGCS**MSQLQVRHDWKREEIEALFALPMNDLLFKAHSIHRE**EYDPNE**VQISRLLSIKTGAC**PEDCKYCPQSARYD

: | | | :: || | : || || : | : || : || : || || || || || || || : ||

MAHRPRWTLSQVTELFEKPLLDLLFEAQ**QVHRQHF**DPRQV**QV**STLLSIKTGACPEDCKYCPQSRYK

10 20 30 40 50 60

agcgagcctgagagcaggcagagggtcttagggtcacagagactcacagcgagcgagataatgataggcgagtgggc  
cgtaaagtctacttacggcaccggcggtgtccggacaaaatcataattataactgtacgtctgtcaacaatcacgt  
tctaaagtcaaggacgcccagtcgaggctgtctggctgtcgatatgaccgaggaggaccggactgcaggatcgactggcaact

80 90 100 110 120 130 140 150

TGLEKERLLAM**ETVLTE**ARSAKAAGASRFC**MGA**AWRNPKDKD**MPYLKQM**VQE**VKALG**MET**CMLGML**SAEQANE**LAEAGL**

||||| ||| : | || |||| : || || ||| : || : || ||| : || | || | || | || | || | || |

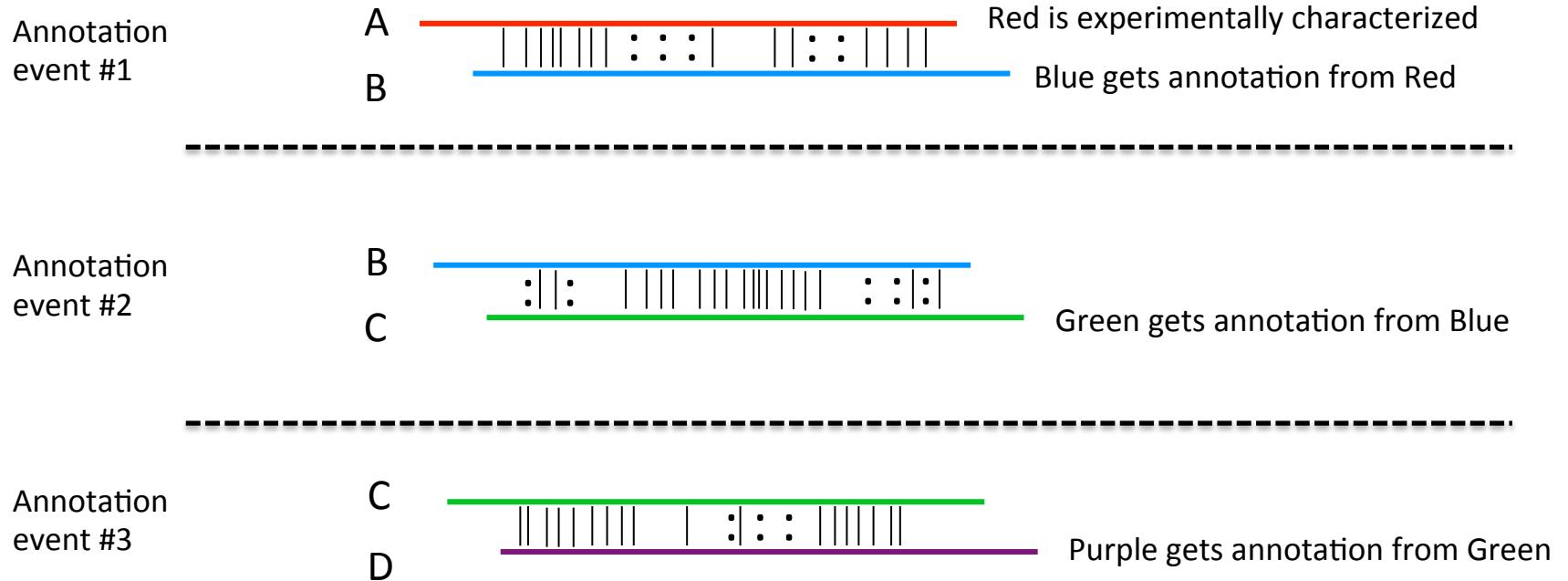
TGLEAERLMEVE**QVLES**ARKAKAAGSTRFC**MGA**AWKNPHERD**MPYL**EQMV**QGV**KAMGLEAC**MTLGT**LSes**QAQRL**ANAGL

80 90 100 110 120 130 140

# Transitive Annotation

---

Transitive Annotation is the process of passing annotation from one protein (or gene) to another based on sequence similarity:



# The Pitfalls of Transitive Annotation

---

If we compared A and D directly would we assert they likely had the same function?



NO, not in this case!  
Thus a transitive annotation error has occurred.

- Current public datasets contain significant numbers of such errors
- Always remember, unless it is experimentally verified, any function you see assigned to a protein is putative.
- Criteria we use
  - 40-50% identity over full length of proteins
  - Match must be to trusted protein that has evidence of experimental characterization
    - Trusted proteins come from Gene Ontology and UniRef evidence annotations
    - If not, we qualify the annotation in some way, e.g. use of “putative”

# UniProt

---

## WELCOME

The mission of [UniProt](#) is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

### What we provide

UniProtKB	Protein knowledgebase, consists of two sections: <ul style="list-style-type: none"><li>★ Swiss-Prot, which is manually annotated and reviewed.</li><li>★ TrEMBL, which is automatically annotated and is <b>not</b> reviewed.</li></ul> Includes <a href="#">complete and reference proteome sets</a> .
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	<a href="#">Literature citations, taxonomy, keywords, subcellular locations, cross-referenced databases and more</a> .

### Getting started

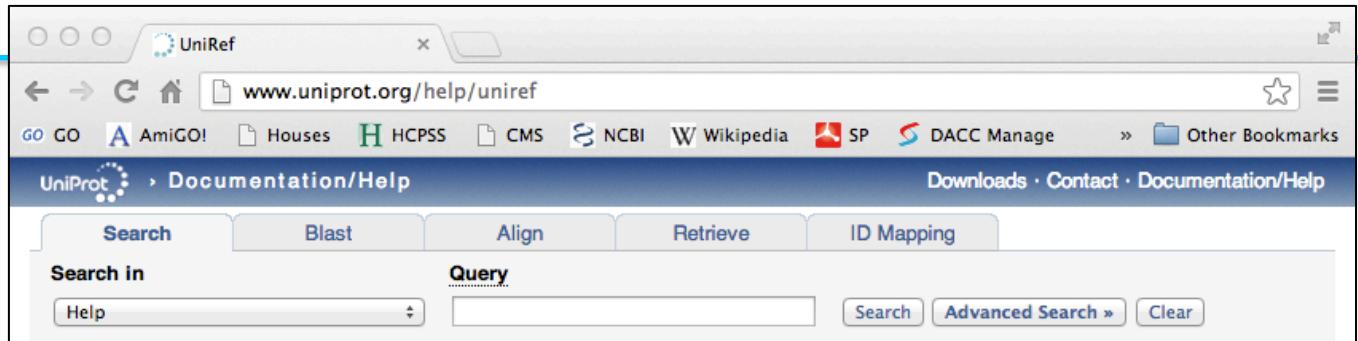
- [Text search](#)
- [Sequence similarity searches \(BLAST\)](#)
- [Sequence alignments](#)
- [Batch retrieval](#)
- [Database identifier mapping \(ID Mapping\)](#)



- Rich source of annotation information
- Includes info on function, subunit structures, protein sequence/structural features, references
- Search features like those found at GenBank
- Reference Proteomes

# UniRef

- UniRef 100
- UniRef 90
- UniRef 50



The screenshot shows a web browser window with the address bar displaying "www.uniprot.org/help/uniref". The page header includes the UniProt logo and navigation links for "Documentation/Help", "Downloads", "Contact", and "Documentation/Help". Below the header is a menu bar with tabs: "Search", "Blast", "Align", "Retrieve", and "ID Mapping". Under "Search", there are dropdown menus for "Search in" (set to "Help") and "Query", along with "Search", "Advanced Search", and "Clear" buttons.

## UniRef

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the [UniProt Knowledgebase](#) (including isoforms) and selected [UniParc](#) records in order to obtain complete coverage of the sequence space at several resolutions while hiding redundant sequences (but not their descriptions) from view. Unlike in UniParc, sequence fragments are merged in UniRef: The UniRef100 database combines identical sequences and sub-fragments with 11 or more residues from any organism into a single UniRef entry, displaying the sequence of a representative protein, the accession numbers of all the merged entries and links to the corresponding UniProtKB and UniParc records. UniRef90 is built by clustering UniRef100 sequences with 11 or more residues using the [CD-HIT algorithm](#) (Li W. and Godzik A., *Bioinformatics*, 22: 1658-1659, 2006) such that each cluster is composed of sequences that have at least 90% sequence identity to and 80% overlap with the longest sequence (a.k.a. seed sequence) of the cluster. Similarly, UniRef50 is built by clustering UniRef90 seed sequences that have at least 50% sequence identity to and 80% overlap with the longest sequence in the cluster. Prior to 2013 there was no overlap threshold, so clusters were more heterogeneous in length. UniRef90 and UniRef50 yield a database size reduction of approximately 58% and 79%, respectively, providing for significantly faster sequence similarity searches. The seed sequences are the longest members of the cluster. However, the longest sequence is not always the most informative. There is often more biologically relevant information (name, function, cross-references) available on other cluster members. All the proteins in a cluster are therefore ranked as follows to facilitate the selection of a biologically relevant representative for the cluster:

1. quality of the entry: manually reviewed entries (from the UniProtKB/Swiss-Prot section) are preferred
2. meaningful name: entries with names that do not contain words such as hypothetical, probable, etc. are preferred
3. organism: entries from model organisms are preferred
4. length of the sequence: longest sequence is preferred

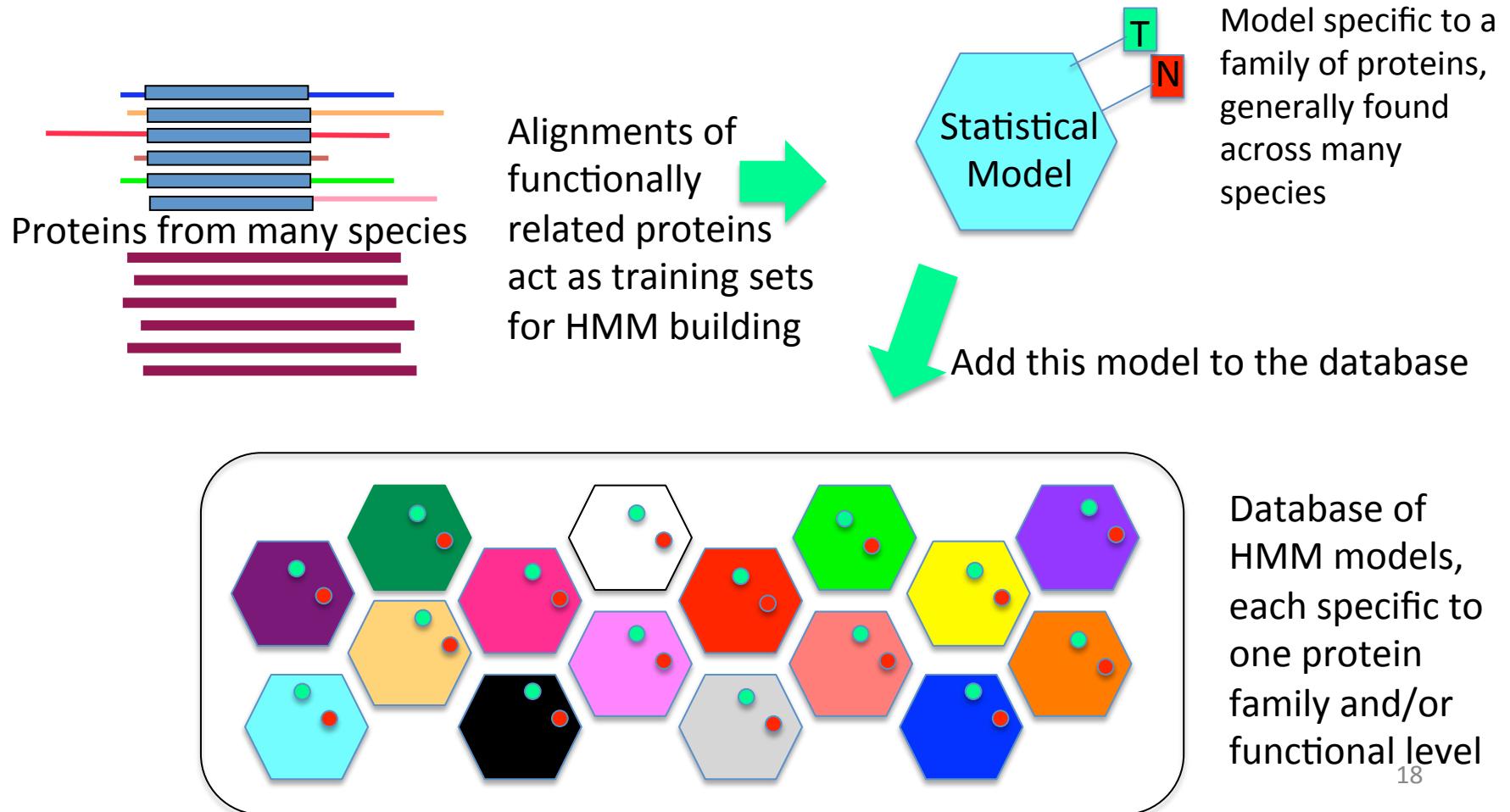
## UniRef100

UniRef100 contains all UniProt Knowledgebase records plus selected UniParc records (see below). In UniRef100, all identical sequences and subfragments with 11 or more residues are placed into a single record. UniRef50 and UniRef90 are built based on UniRef100.

# HMM databases

Statistical models of patterns of amino acids within an alignment of proteins – can be used to predict membership of a protein in a functional family. Some families are functionally specific, some are more general.

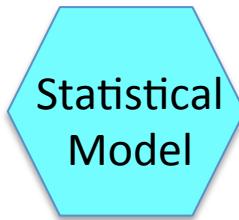
## Pfam and TIGRFAM



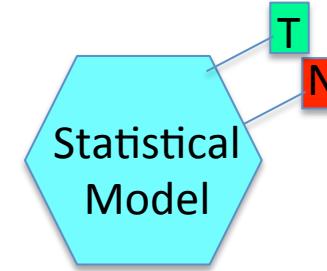
# HMM Scores

---

- When a protein is searched against an HMM it receives a BITS score and an e-value indicating the significance of the match



The person building the HMM will  
search the new HMM against a →  
protein database and decide on the  
trusted and noise cutoff scores



- The search protein's score is compared with the trusted and noise cutoff scores attached to the HMM
  - proteins scoring above the trusted cutoff can be assumed to be members of the family
  - proteins scoring below the noise cutoff can be assumed NOT to be members of the family
  - when proteins score in-between the trusted and noise cutoffs, the protein may be a member of the family and may not.

# Annotation attached to HMMs

---

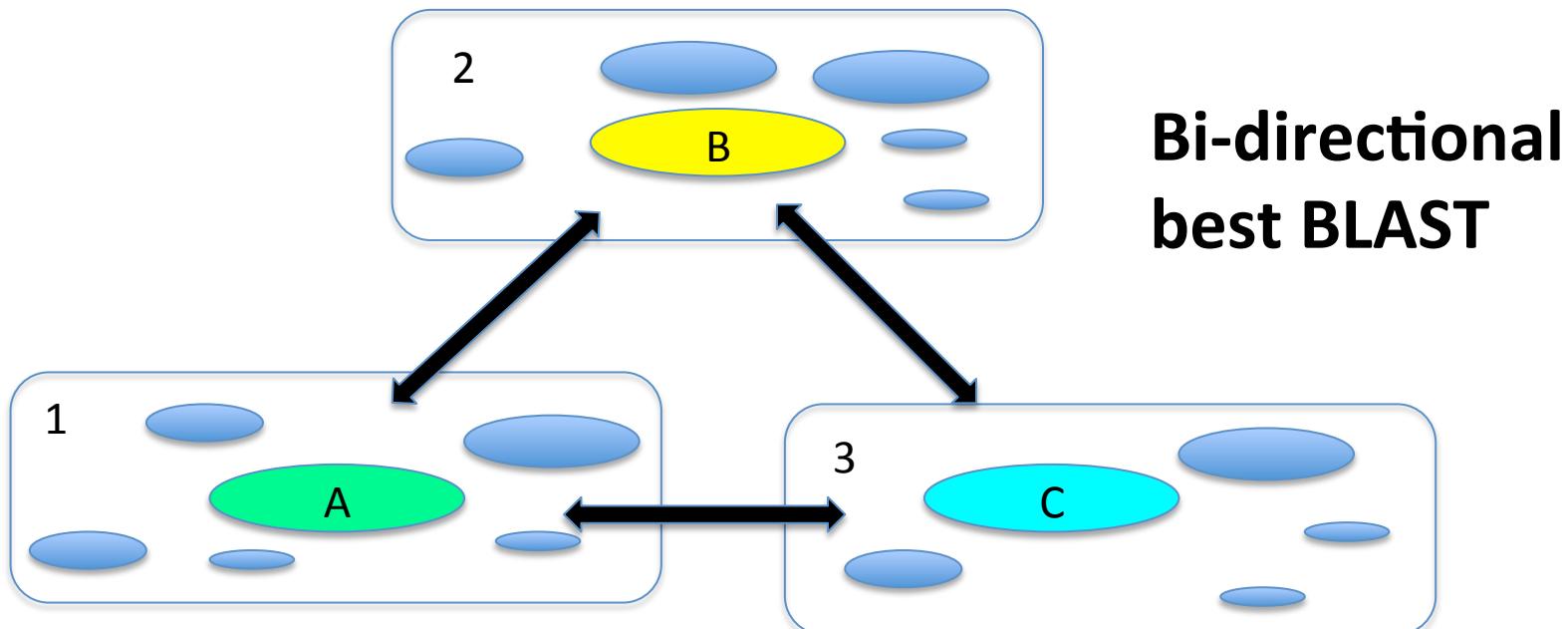
- Functionally specific HMMs have specific annotations  
**TIGR00433** (accession number for the model)
  - name: biotin synthase
  - category: equivalog
  - EC: 2.8.1.6
  - gene symbol: bioB
  - Roles:
    - biotin biosynthesis (TIGR 77/GO:0009102)
    - biotin synthase activity (GO:0004076)
- Functionally general HMMs have general annotations  
**PF04055**
  - name: radical SAM domain protein
  - category: domain
  - EC: not applicable
  - gene symbol: not applicable
  - Roles:
    - enzymes of unknown specificity (TIGR role 703)
    - catalytic activity (GO:0003824)
    - metabolism (GO:0008152)



# Orthologous Groups

---

- COGs
  - “clusters of orthologous genes”
  - a product of NCBI
- eggNOG
  - “evolutionary genealogy of genes: Non-supervised Orthologous Groups”
  - A product of EMBL



# Motif Searches

---

- PROSITE - <http://www.expasy.org/prosite/>
  - “consists of documentation entries describing protein domains, families and functional sites as well as associated patterns to identify them.”
- Center for Biological Sequence Analysis -  
<http://www.cbs.dtu.dk/>
  - Protein Sorting (7 tools)
    - **Signal P** finds potential secreted proteins
    - **LipoP** finds potential lipoproteins
    - TargetP predicts subcellular location of proteins
  - Protein function and structure (9 tools)
    - **TmHMM** finds potential membrane spans
  - Post-translational modifications (14 tools)
  - Immunological features (9 tools)
  - Gene finding and splice sites (9 tools)
  - DNA microarray analysis (2 tools)
  - Small molecules (2 tools)



# Types of Annotations

---

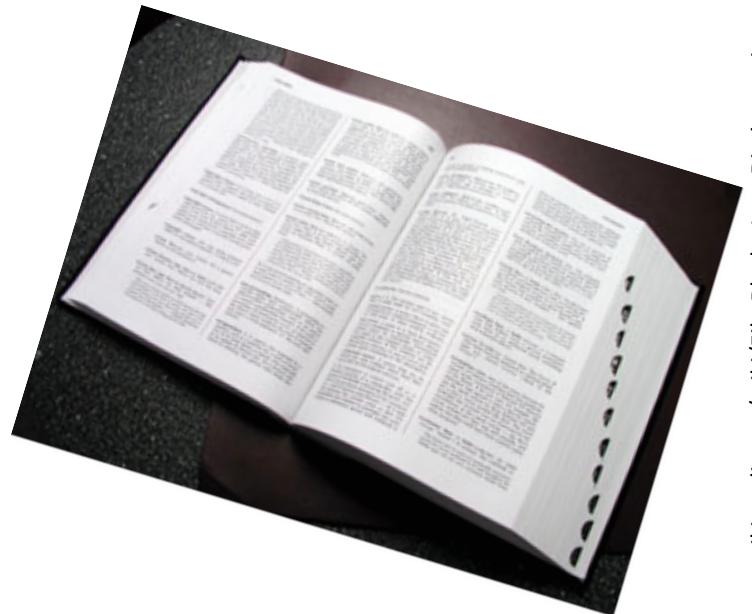
# Types of Annotation

- Gene product names
- Gene symbols
- Locus/allele identifiers
- EC numbers
- Gene Ontology terms
- Phenotype terms

# Protein names can be problematic....

---

- ....because humans do not always use precise and consistent terminology
- Our language is riddled with
  - Synonyms – different names for the same thing
    - Alternate names for enzymatic reactions
  - Homonyms – different things with the same name
    - Reproductive vs. survival sporulation
- This makes data mining/query difficult
  - What name should you assign?
  - What name should you use when you search a database?
- How do we fix this?
  - **Controlled vocabularies**
  - **Ontologies!**



# Enzyme Commission

Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes by the Reactions they Catalyze



- not sequence based
- categorized collection of enzymatic reactions
- reactions have accession numbers indicating the type of reaction, for example EC 1.2.1.5
- <http://www.chem.qmul.ac.uk/iubmb/enzyme/>
- <http://www.expasy.ch/enzyme/>

Top-level EC numbers <sup>[5]</sup>			
Group	Reaction catalyzed	Typical reaction	Enzyme example(s) with trivial name
<b>EC 1 <i>Oxidoreductases</i></b>	To catalyze oxidation/reduction reactions; transfer of H and O atoms or electrons from one substance to another	$AH + B \rightarrow A + BH$ (reduced) $A + O \rightarrow AO$ (oxidized)	Dehydrogenase, oxidase
<b>EC 2 <i>Transferases</i></b>	Transfer of a functional group from one substance to another. The group may be methyl-, acyl-, amino- or phosphate group	$AB + C \rightarrow A + BC$	Transaminase, kinase
<b>EC 3 <i>Hydrolases</i></b>	Formation of two products from a substrate by hydrolysis	$AB + H_2O \rightarrow AOH + BH$	Lipase, amylase, peptidase
<b>EC 4 <i>Lyases</i></b>	Non-hydrolytic addition or removal of groups from substrates. C-C, C-N, C-O or C-S bonds may be cleaved	$RCOCOOH \rightarrow RCOH + CO_2$ or $[X-A-B-Y] \rightarrow [A=B + X-Y]$	Decarboxylase
<b>EC 5 <i>Isomerases</i></b>	Intramolecule rearrangement, i.e. isomerization changes within a single molecule	$AB \rightarrow BA$	Isomerase, mutase
<b>EC 6 <i>Ligases</i></b>	Join together two molecules by synthesis of new C-O, C-S, C-N or C-C bonds with simultaneous breakdown of ATP	$X + Y + ATP \rightarrow XY + ADP + Pi$	Synthetase

# EC Hierarchy

All ECs starting with #1 are some kind of oxidoreductase

Further numbers narrow specificity of the type of enzyme

A four-position EC number describes one particular reaction

## EC 1 Oxidoreductases

Number	Name	Enzyme file type
<b>EC 1.1</b>	<b>Acting on the CH-OH group of donors</b>	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.1.1	With NAD <sup>+</sup> or NADP <sup>+</sup> as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.1.2	With a cytochrome as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.1.3	With oxygen as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.1.4	With a disulfide as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.1.5	With a quinone or similar compound as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.1.99	With other acceptors	<a href="#">separate</a> <a href="#">up to 50</a>
<b>EC 1.2</b>	<b>Acting on the aldehyde or oxo group of donors</b>	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.2.1	With NAD <sup>+</sup> or NADP <sup>+</sup> as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.2.2	With a cytochrome as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.2.3	With oxygen as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.2.4	With a disulfide as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.2.7	With an iron-sulfur protein acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.2.99	With other acceptors	<a href="#">separate</a> <a href="#">up to 50</a>
<b>EC 1.3</b>	<b>Acting on the CH-CH group of donors</b>	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.3.1	With NAD <sup>+</sup> or NADP <sup>+</sup> as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.3.2	With a cytochrome as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.3.3	With oxygen as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.3.5	With a quinone or related compound as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.3.7	With an iron-sulfur protein as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.3.99	With other acceptors	<a href="#">separate</a> <a href="#">up to 50</a>
<b>EC 1.4</b>	<b>Acting on the CH-NH<sub>2</sub> group of donors</b>	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.4.1	With NAD <sup>+</sup> or NADP <sup>+</sup> as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.4.2	With a cytochrome as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.4.3	With oxygen as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.4.4	With a disulfide as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.4.7	With an iron-sulfur protein as acceptor	<a href="#">separate</a> <a href="#">up to 50</a>
EC 1.4.99	With other acceptors	<a href="#">separate</a> <a href="#">up to 50</a>

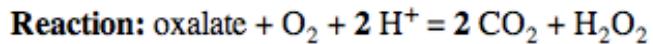
# Specific Example

---

IUBMB Enzyme Nomenclature

## EC 1.2.3.4

**Accepted name:** oxalate oxidase



**Other name(s):** aero-oxalo dehydrogenase; oxalic acid oxidase

**Systematic name:** oxalate:oxygen oxidoreductase

**Comments:** Contains Mn<sup>2+</sup> as a cofactor. The enzyme is not a flavoprotein as had been thought [3].

**Links to other databases:** [BRENDA](#), [EXPASY](#), [KEGG](#), [ERGO](#), [PDB](#), CAS registry number: 9031-79-2

### References:

1. Datta, P.K., Meeuse, B.J.D., Engstrom-Heg, V. and Hilal, S.H. Moss oxalic acid oxidase - a flavoprotein. *Biochim. Biophys. Acta* 17 (1955) 602-603. [PMID: [13250021](#)]
2. Kotsira, V.P. and Clonis, Y.D. Oxalate oxidase from barley roots: purification to homogeneity and study of some molecular, catalytic, and binding properties. *Arch. Biochem. Biophys.* 340 (1997) 239-249. [PMID: [9143327](#)]
3. Requena, L., and Bornemann, S. Barley (*Hordeum vulgare*) oxalate oxidase is a manganese-containing enzyme. *Biochem. J.* 343 (1999) 185-190. [PMID: [10493928](#)]

[EC 1.2.3.4 created 1961]

# Gene Ontology

---

- [www.geneontology.org](http://www.geneontology.org)
- Started as a collaboration between the model organism dbs for mouse, fly, and yeast
- It has now become one of the key standards for functional annotation





## Three controlled vocabularies

- Molecular Function
  - What the gene product is doing
- Biological Process
  - Why the gene product is doing what it does
- Cellular component
  - Where a gene product is doing what it does
- Make an annotation by linking a gene product to a GO term.

### Annotation production status<sup>a</sup>

Total number of GO terms	41 775
Biological process terms	27 284
Molecular function terms	10 733
Cellular component terms	3758
Species with annotations	461 573
Total annotated gene products <sup>b</sup>	53 042 843
Manually annotated (experimental) gene products	311 335
Manually annotated (phylogenetic) gene products	79 839
Total annotations	4 185 487

<sup>a</sup>As of September 2014.

<sup>b</sup>Includes isoforms.

# Effective capture of information

---

Annotation becomes a series of cv term ids linked to proteins/genes/features

A protein is integral to the plasma membrane and is part of an ATP-binding cassette (ABC) transporter complex. It functions as part of a transporter to accomplish the transport of sulfate across the plasma membrane using ATP hydrolysis as an energy source.

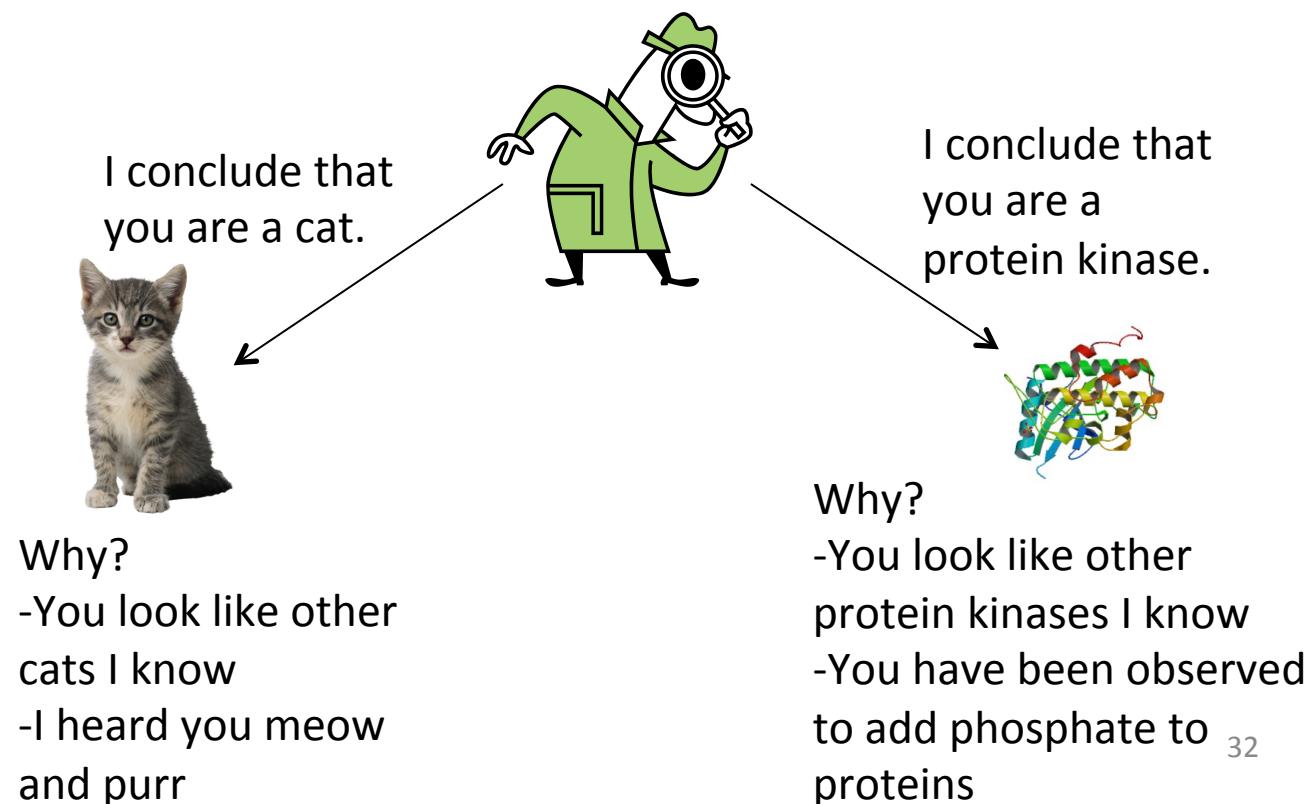


- GO:0008272
  - sulfate transport
- GO:0015419
  - sulfate transmembrane-transporting ATPase activity
- GO:0043190
  - ATP-binding cassette (ABC) transporter complex
- GO:0005887
  - integral component of plasma membrane

# The Importance of Evidence Tracking

---

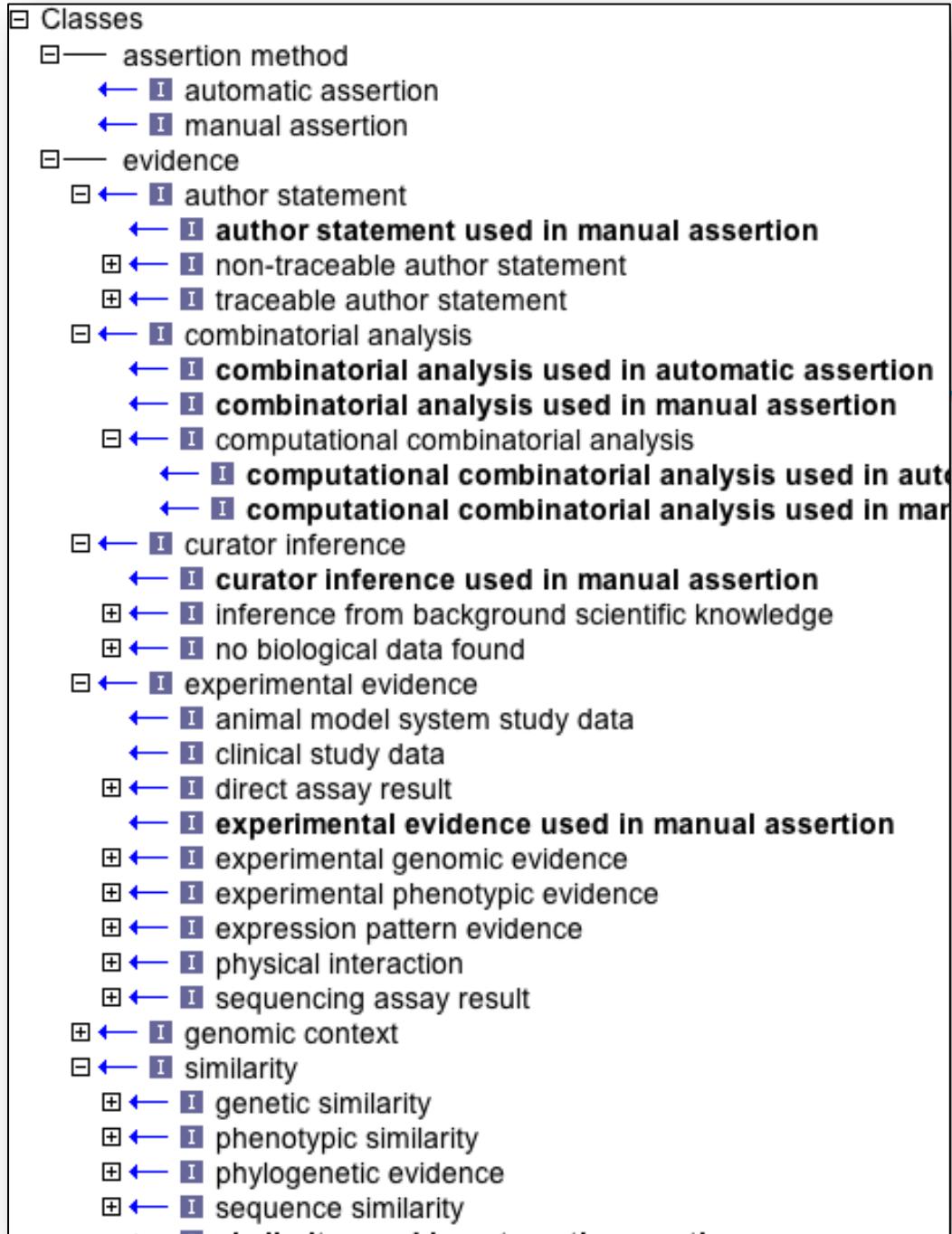
- The process of functional annotation involves assessing available evidence and reaching a conclusion about what you think the protein is doing in the cell and why.
- Ideally, all evidence that led to the annotation conclusions that were made must be stored in a way that allows users of the data to assess the value of the annotations.
- In addition, detailed documentation of methodologies and general rules or guidelines used in any annotation process should be provided.



# Types of Evidence

- Experiments (the only “truth”)
  - Pairwise/multiple alignments
  - HMM/domain matches scoring above trusted cutoff
  - Metabolic Pathway analysis
  - Match to an ortholog group (COG,eggNOG)
  - Motifs
- 
- **Who tracks evidence for annotations?**
    - UniProt
    - Gene Ontology





# The Evidence Ontology



Ids for the type of evidence can be linked to annotations attached to proteins (or other objects)

Modified from slide from Marcus Chibucos



# The Open Biological and Biomedical Ontologies

[Ontologies](#)[Resources](#)[Participate](#)[About](#)

The OBO Foundry is a collaborative experiment involving developers of science-based ontologies who are establishing a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain. The groups developing ontologies who have expressed an interest in this goal are listed below, followed by other relevant efforts in this domain.

In addition to a listing of OBO ontologies, this site also provides a statement of the OBO Foundry principles, discussion fora, technical infrastructure, and other services to facilitate ontology development. We welcome feedback and encourage participation.

Click any column header to sort the table by that column. The s link to the term request trackers for the listed ontologies.

More than 100  
ontologies

OBO Foundry ontologies				
Title	Domain	Prefix	File	
<a href="#">Biological process</a>	biological process	GO	<a href="#">go.obo</a> 	
<a href="#">Cellular component</a>	anatomy	GO	<a href="#">go.obo</a> 	
<a href="#">Chemical entities of biological interest</a>	biochemistry	CHEBI	<a href="#">chebi.obo</a> 	
<a href="#">Molecular function</a>	biological function	GO	<a href="#">go.obo</a> 	
<a href="#">Ontology for biomedical investigations</a>	experiments	OBI	<a href="#">obi.owl</a> 	
<a href="#">Phenotypic quality</a>	phenotype	PATO	<a href="#">quality.obo</a> 	
<a href="#">Plant Ontology</a>	anatomy and development	PO	<a href="#">plant_ontology.obo?view=co</a> 	
<a href="#">PRotein Ontology (PRO)</a>	proteins	PR	<a href="#">pro.obo</a> 	
<a href="#">Xenopus anatomy and development</a>	anatomy	XAO	<a href="#">xenopus_anatomy.obo</a> 	
<a href="#">Zebrafish anatomy and development</a>	anatomy	ZFA	<a href="#">zfa.obo</a> 	

OBO Foundry candidate ontologies and other ontologies of interest				
Title	Domain	Prefix	File	Last changed
<a href="#">Adverse Event Reporting Ontology</a>	health	AERO	<a href="#">aero.owl</a>	
<a href="#">Anatomical Entity Ontology</a>	anatomy	AEON	<a href="#">aeo.obo</a>	2012/06/01
<a href="#">Ascomycete phenotype ontology</a>	phenotype	APO	<a href="#">ascomycete_phenotype.obo</a>	2014/06/30
<a href="#">Basic Formal Ontology</a>	upper	BFO	<a href="#">1.1</a>	
<a href="#">Beta Cell Genomics Ontology</a>		BCGO	<a href="#">bcgo.owl</a> 	
<a href="#">Biological Collections Ontology</a>		BCO	<a href="#">bco.owl</a> 	
<a href="#">Biological imaging methods</a>	experiments	FBbi	<a href="#">image.owl</a>	2013/11/05
<a href="#">Biological Spatial Ontology</a>	anatomy	BSPO	<a href="#">bspo.obo</a> 	
<a href="#">BRENDA tissue / enzyme source</a>	anatomy	BTO	<a href="#">BrendaTissueOBO</a>	
<a href="#">C. elegans development</a>	anatomy	WBIs	<a href="#">worm_development.obo</a>	
<a href="#">C. elegans gross anatomy</a>	anatomy	WBbt	<a href="#">wbbt.owl</a> 	
<a href="#">C. elegans phenotype</a>	phenotype	WBPhenotype	<a href="#">wbphenotype.owl</a>	
<a href="#">Cardiovascular Disease Ontology</a>	health	CVDO	<a href="#">cvdo.owl</a> 	
<a href="#">Cell Line Ontology</a>		CLO	<a href="#">clo.owl</a>	
<a href="#">Cell type</a>	anatomy	CL	<a href="#">cl.owl</a> 	
<a href="#">Chemical Information Ontology</a>	biochemistry	CHEMINF	<a href="#">cheminf.owl</a>	
<a href="#">Chemical Methods Ontology</a>	health	CHMO	<a href="#">chmo.owl</a> 	
<a href="#">Common Anatomy Reference Ontology</a>	anatomy	CARO	<a href="#">caro.owl</a> 	2011/12/14

# Making the Assignments

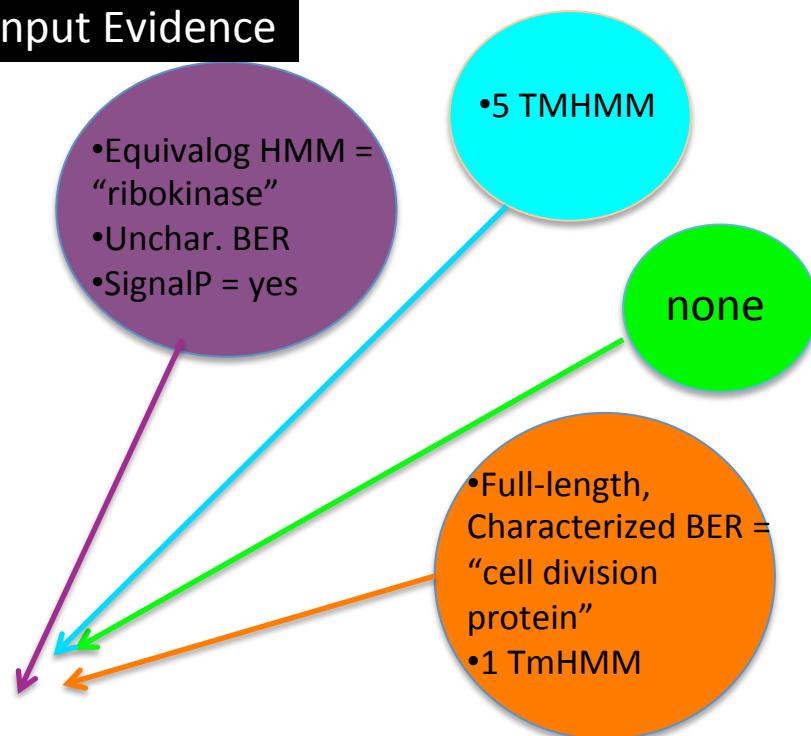
---

# Automated Functional Annotation

---

- Manual annotation is the best
  - Because humans are good at drawing conclusions from multiple pieces of information
- But it is not feasible in most cases
- So we need to try to mimic what a human would do using automated systems.

## Input Evidence



1. HMM modeling a specific function with a score above the trusted cutoff
2. Full length match that is experimentally characterized
3. HMM modeling a general function with a score above the trusted cutoff
4. > or = 5 TMHMM regions
5. Default – “hypothetical protein”

## Evidence Hierarchy

Automatic annotation is based on an evidence hierarchy. Proteins receive annotation at the highest level ranking for which they have the required evidence.

## Output annotations

Name: ribokinase  
EC:2.7.1.15  
GO:0004747  
GO:0006014  
GO:0005737

Name: cell division protein  
EC: none  
GO:0051301  
GO:0003674  
GO:0005575

Name: integral membrane protein  
EC: none  
GO:0008150  
GO:0003674  
GO:0005887

Name: hypothetical protein  
EC: none  
GO:0008150  
GO:0003674  
GO:0005575

# pFunc: the actual hierarchy in the pipeline

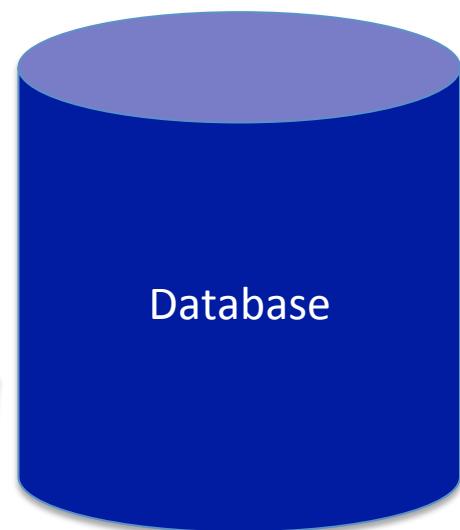
---

Evidence	Criteria	Query	Match	Rank
HMM	Equivalog	N/A	N/A	1
BER	Trusted	Full	Full	2
HMM	Equivalog Domain	Full	Full	3
BER	Trusted	Partial	Full	4
HMM	Subfamily	N/A	N/A	5
HMM	Superfamily	N/A	N/A	6
HMM	Subfamily Domain	N/A	N/A	7
HMM	Domain	Partial	Full	8
HMM	Pfam	Full	Full	9
BER	Trusted	Full	Partial	10
TMHMM	> 5 membrane spans	N/A	N/A	11
LipoP	Presence of prediction	N/A	N/A	12
HMM	Hypothetical Equivalog	N/A	N/A	13
BER	Not trusted	Full	Full	14
BER	Not trusted	Partial	Full	15
BER	Not trusted	Full	Partial	16
BER	With ambiguous term	Full/Partial	Full/Partial	17

Current evidence hierarchy in the IGS prokaryotic pipeline.

Evidence	Criteria	Query	Match	Rank
HMM	Equivalog	N/A	N/A	1
BER	Trusted	Full	Full	2
HMM	Equivalog Domain	Full	Full	3
BER	Trusted	Partial	Full	4
HMM	Subfamily	N/A	N/A	5
HMM	Superfamily	N/A	N/A	6
HMM	Subfamily Domain	N/A	N/A	7
HMM	Domain	Partial	Full	8
HMM	Pfam	Full	Full	9
BER	Trusted	Full	Partial	10
TMHMM	> 5 membrane spans	N/A	N/A	11
LipoP	Presence of prediction	N/A	N/A	12
HMM	Hypothetical Equivalog	N/A	N/A	13
BER	Not trusted	Full	Full	14
BER	Not trusted	Partial	Full	15
BER	Not trusted	Full	Partial	16
BER	With ambiguous term	Full/Partial	Full/Partial	17

Flat files



Database



Flat files

# Visualization of the Data

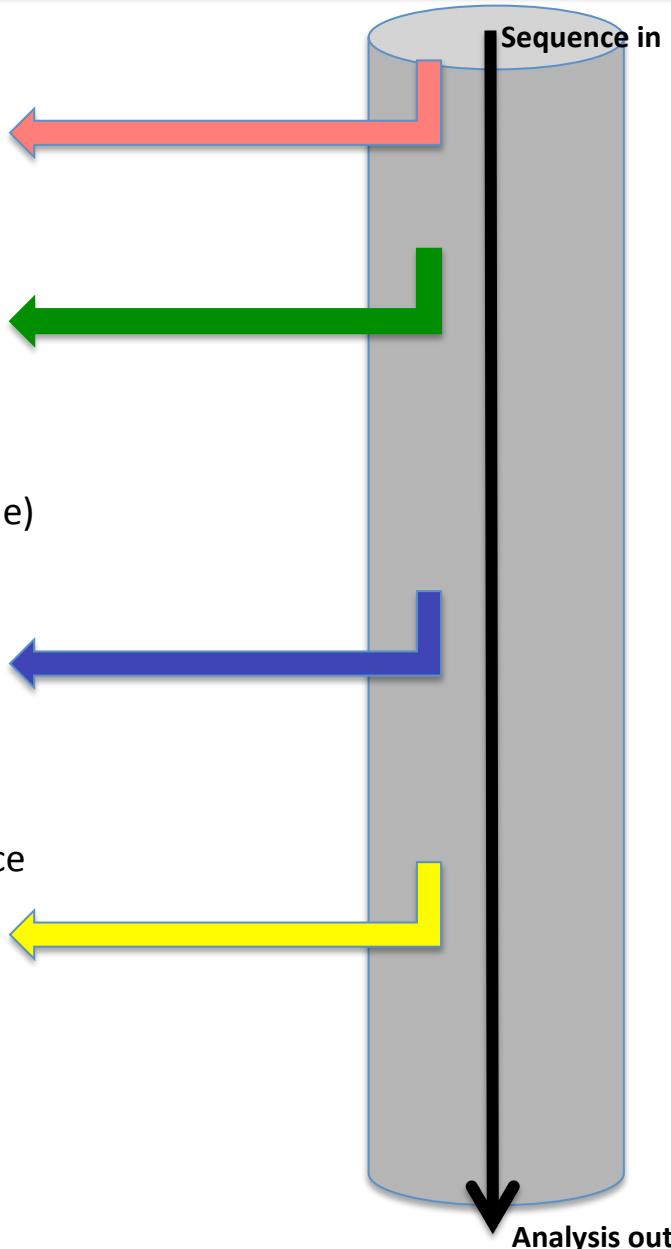
---



- Manatee is a web-based manual annotation tool for accessing and editing annotation data
- Manatee draws information from an underlying database for its displays
- Manatee sends information entered by annotators to the underlying database for storage
- Multiple users can access the same database from different computers when Manatee is run on a server
- MORE ON THIS LATER

# Summary of Annotation Pipeline outputs

- **Structural annotations**
  - Coding genes
  - tRNAs
  - ncRNAs
- **Evidence for functional annotations**
  - Pairwise alignments
  - HMM search results
  - Prosite motifs
  - SignalP, LipoP, TMHMM (not in CloVR pipeline)
  - COGs
  - eggNOG
- **Functional annotations**
  - Protein name
  - Gene symbol
  - EC number
  - Gene Ontology terms and associated evidence
- **Multiple file outputs**
  - Sequences of predicted genes/proteins
  - Gene coordinates
  - GO gene association
  - GenBank
  - Gff3
  - Tab-delimited functional annotation
  - Many other options by request



## Acknowledgments

---

- Shaun Adkins
  - James Matsumura
  - Sean Daugherty
  - Heather Creasy
  - Suvarna Nadendla
  - Anup Mahurkar
  - Owen White
- 
- The many past IGS people who have contributed to this pipeline.