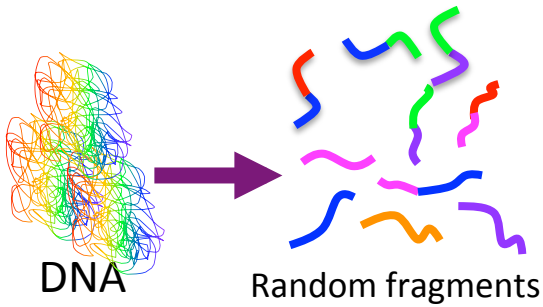
The IGS Prokaryotic Annotation Pipeline

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



Sequencing reactions

Sanger, Illumina, PacBio

Reads

GCTAATGCGAAGTCCTAGACCAGATTGAC
ATGCGATGCAGCTGACGCTGGCTAATGCG
CGCATAGCCAGATGACCATGATGCGATGC

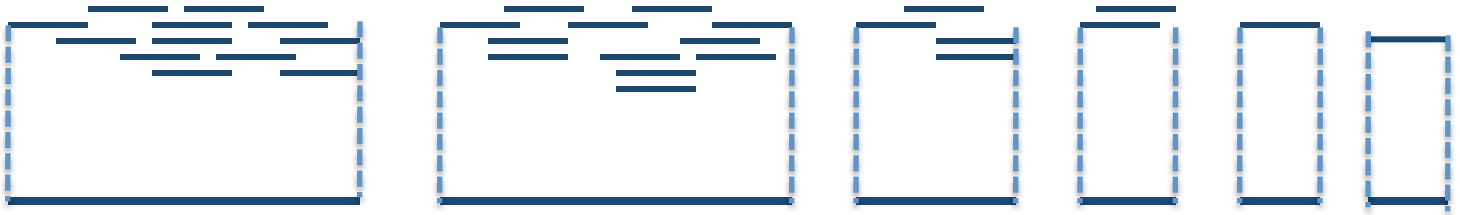
Line up the parts that match

GCTAATGCGAAGTCCTAGACCAGATTGAC
ATGCGATGCAGCTGACGCTGGCTAATGCG
CGCATAGCCAGATGACCATGATGCGATGC

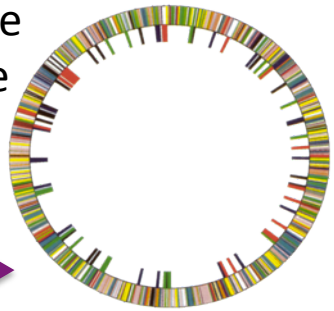
Assemble the "reads" into one contiguous sequence or "contig"

CGCATAGCCAGATGACCATGATGCGATGCAGCTGACGCTGGCTAATGCGAAGTCCTAGACCAGATTGAC

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



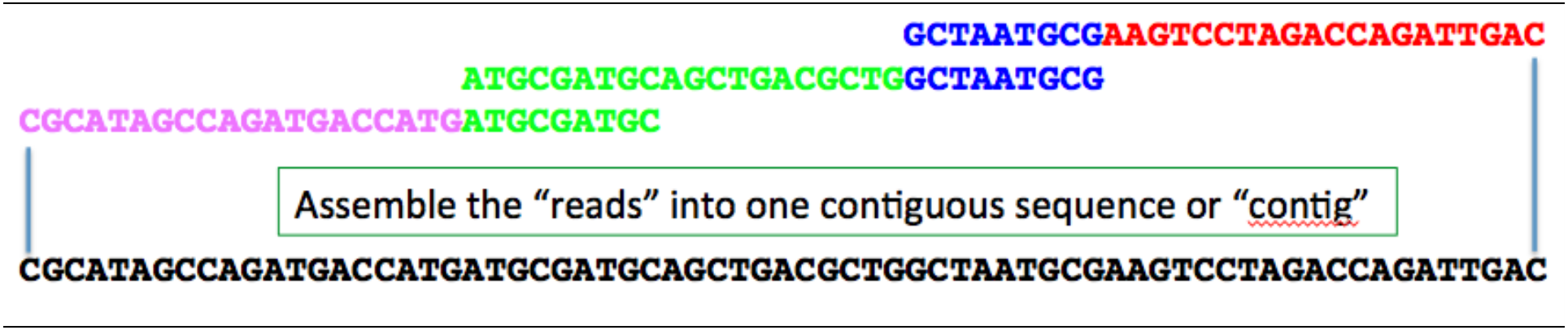
"Gold Standard" finished genome

Steps in the pipeline

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

Assembly

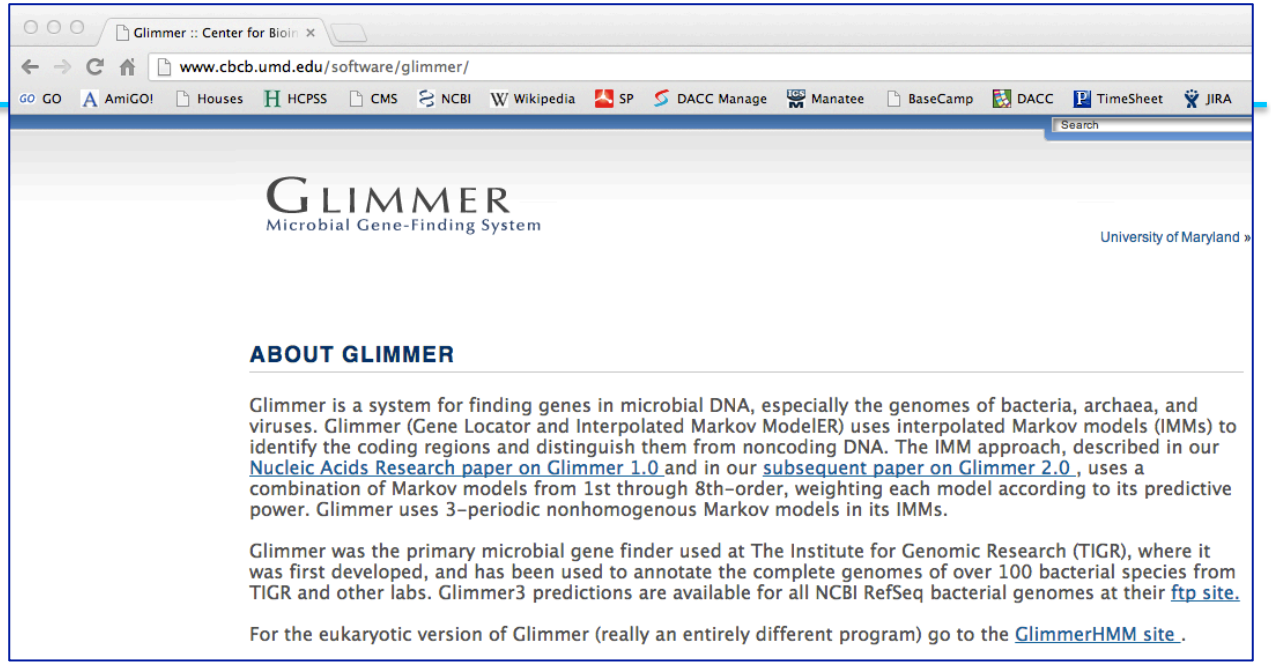
- Two methods available in CloVR, depending on the sequencing technology that was used
  - 454: Celera Assembler
  - Illumina: Spades
- Spades available as a service in Analysis Engine

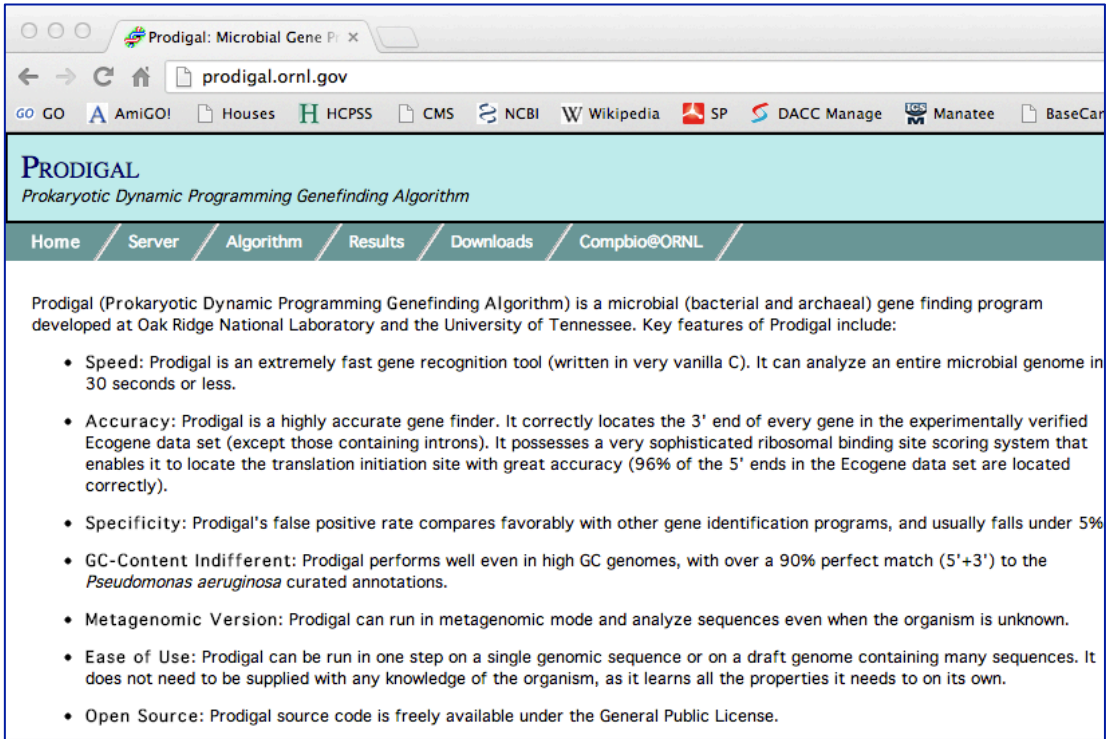


Gene Finding

- Analysis Engine
  (choose from)
  - Glimmer3
  - Prodigal

- CloVR
  - Glimmer3





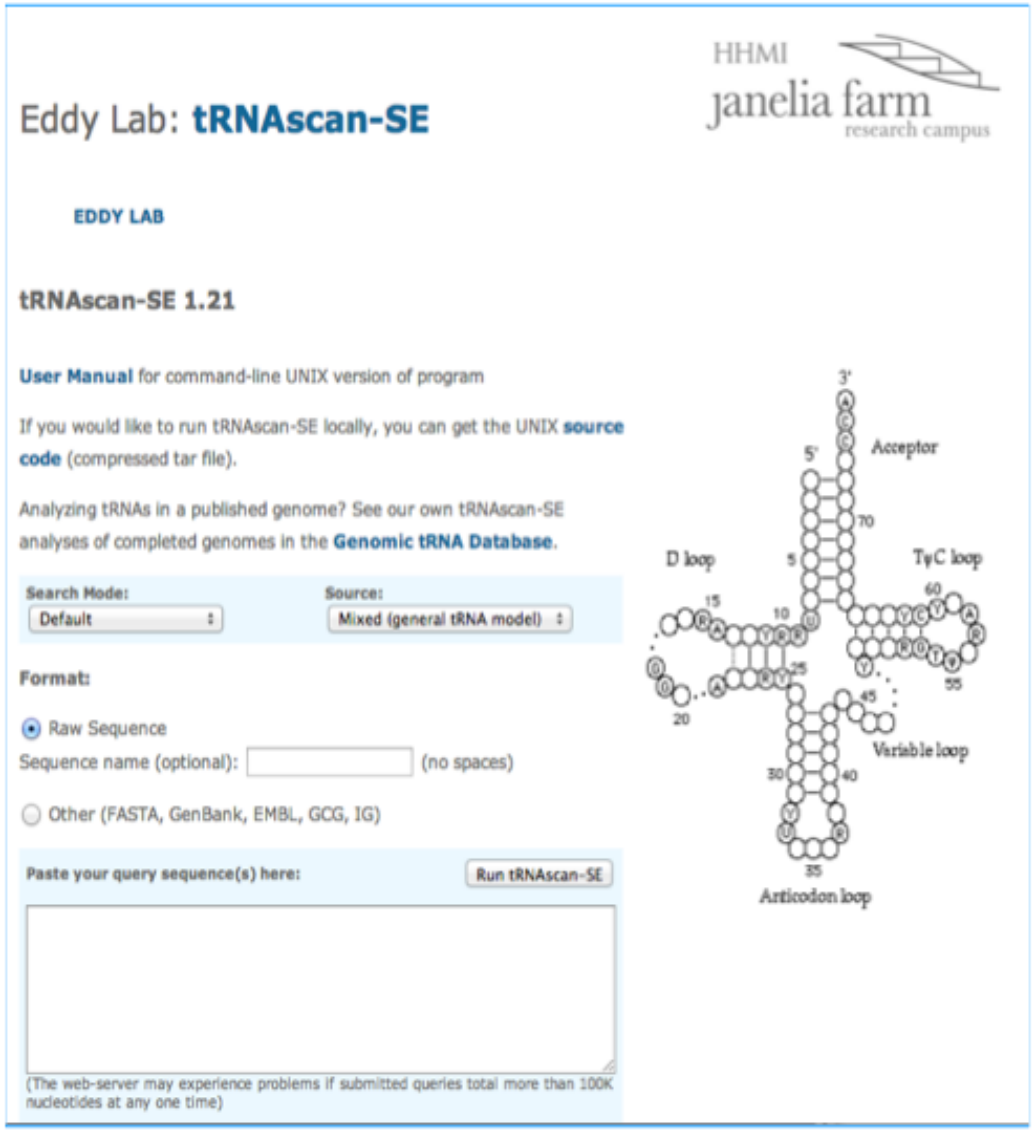
tRNAscan: http://selab.janelia.org/tRNAscan-SE/

Brief Description

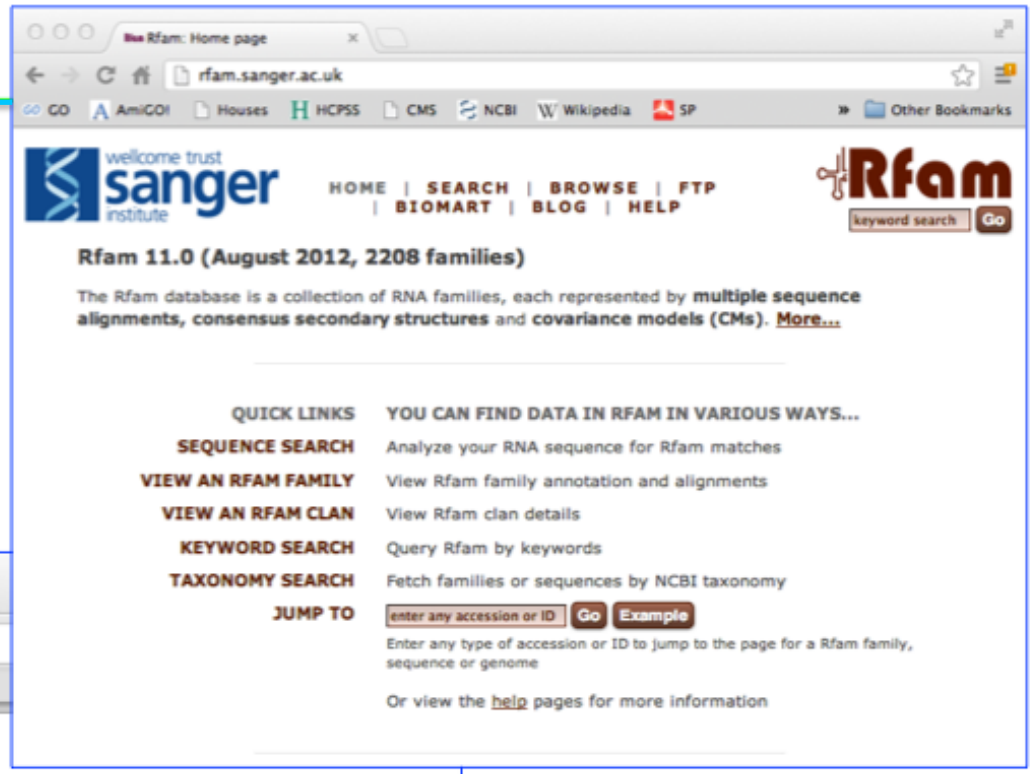
tRNAscan-SE identifies transfer RNA genes in genomic DNA or RNA sequences. It combines the specificity of the Cove probabilistic RNA prediction package (Eddy & Durbin, 1994) with the speed and sensitivity of tRNAscan 1.3 (Fichant & Burks, 1991) plus an implementation of an algorithm described by Pavesi 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:

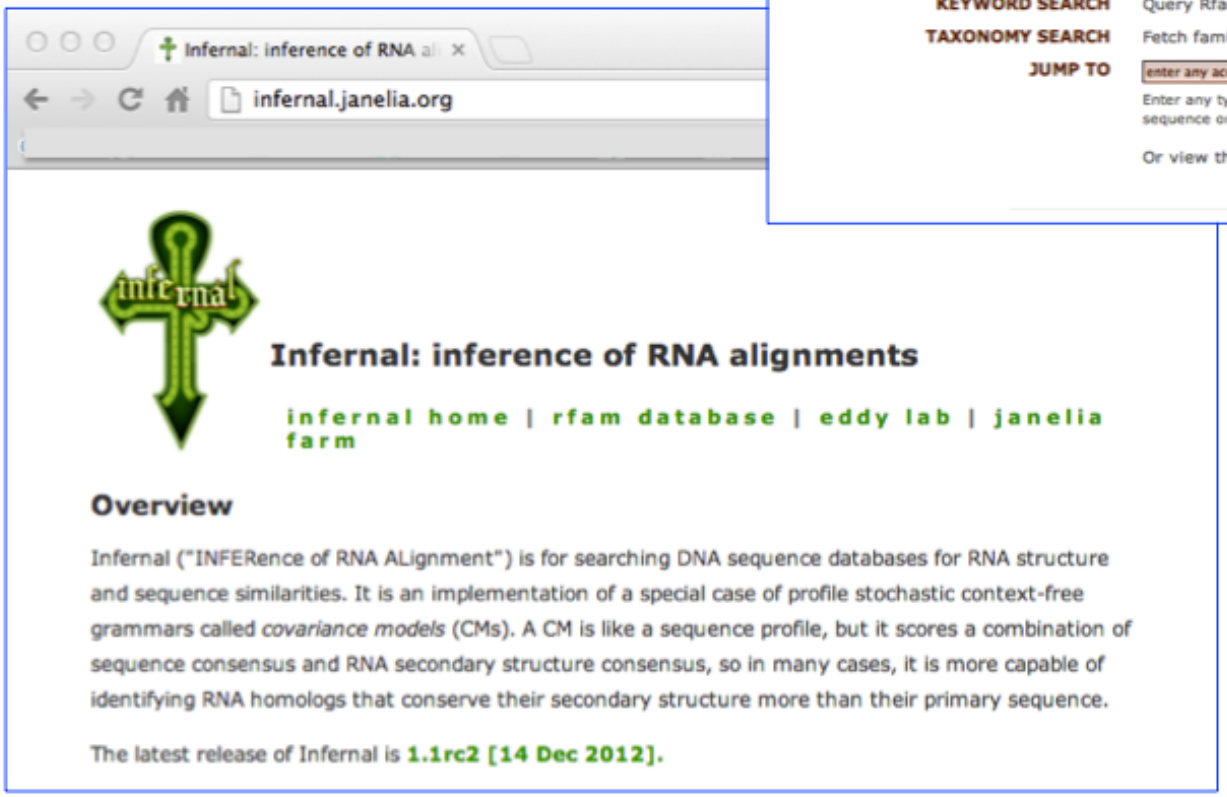
- a false positive rate of less than one per 15 billion nucleotides of random sequence
- the combined sensitivities of tRNAscan and EufindtRNA (detection of 99% of true tRNAs)
- search speed 1,000 to 3,000 times faster than Cove analysis and 30 to 90 times faster than the original tRNAscan 1.3 (tRNAscan-SE 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).



Other ncRNAs





Functional Annotation

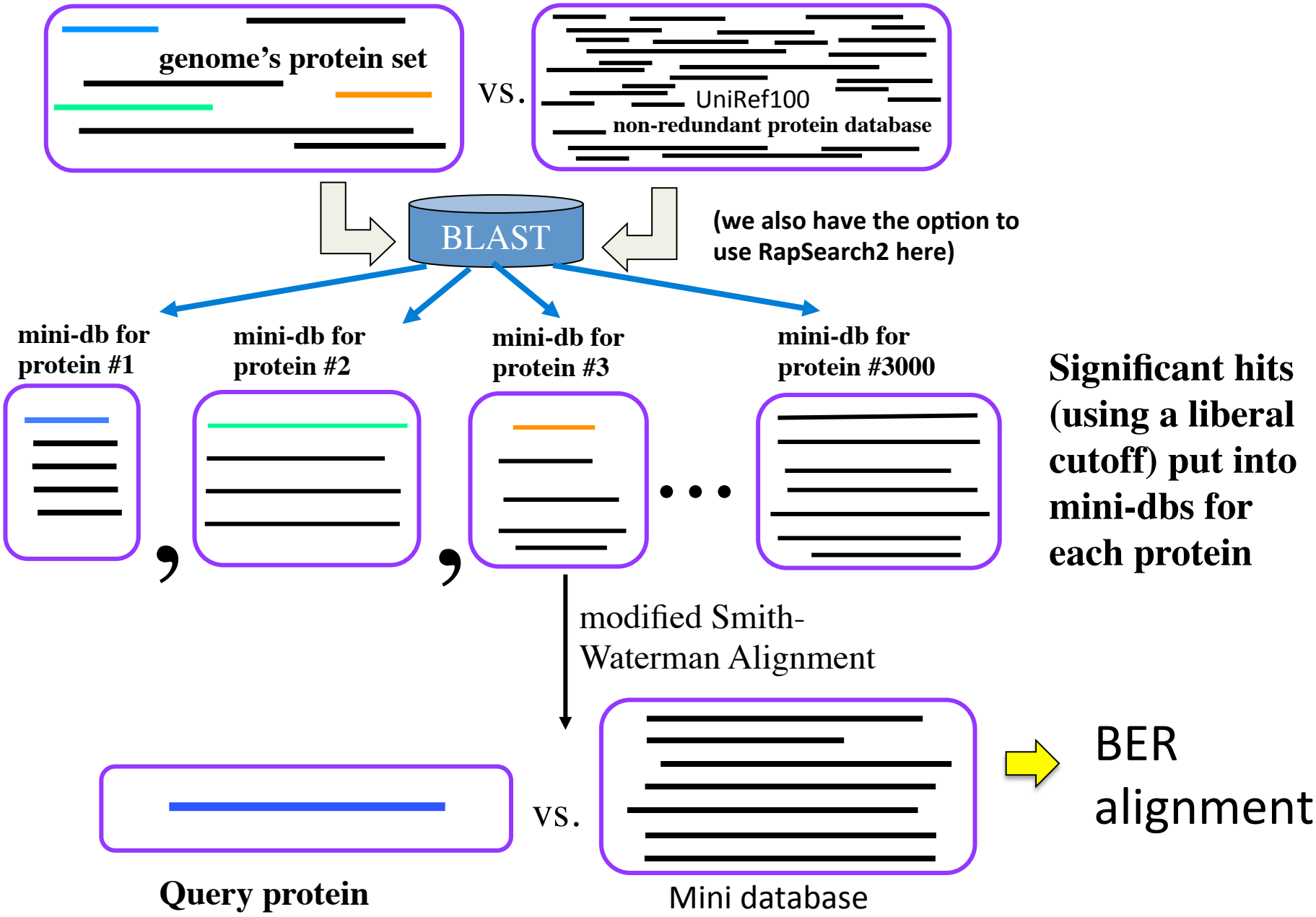
- Sequence similarity is the basis of functional assignments in our pipeline
  - Protein (not nucleotide) alignments for functional prediction
- Remember: one aa difference can change the function of a protein
  - All sequence-based annotations are putative until experimentally confirmed.

nonpolar polar basic acidic (stop codon)

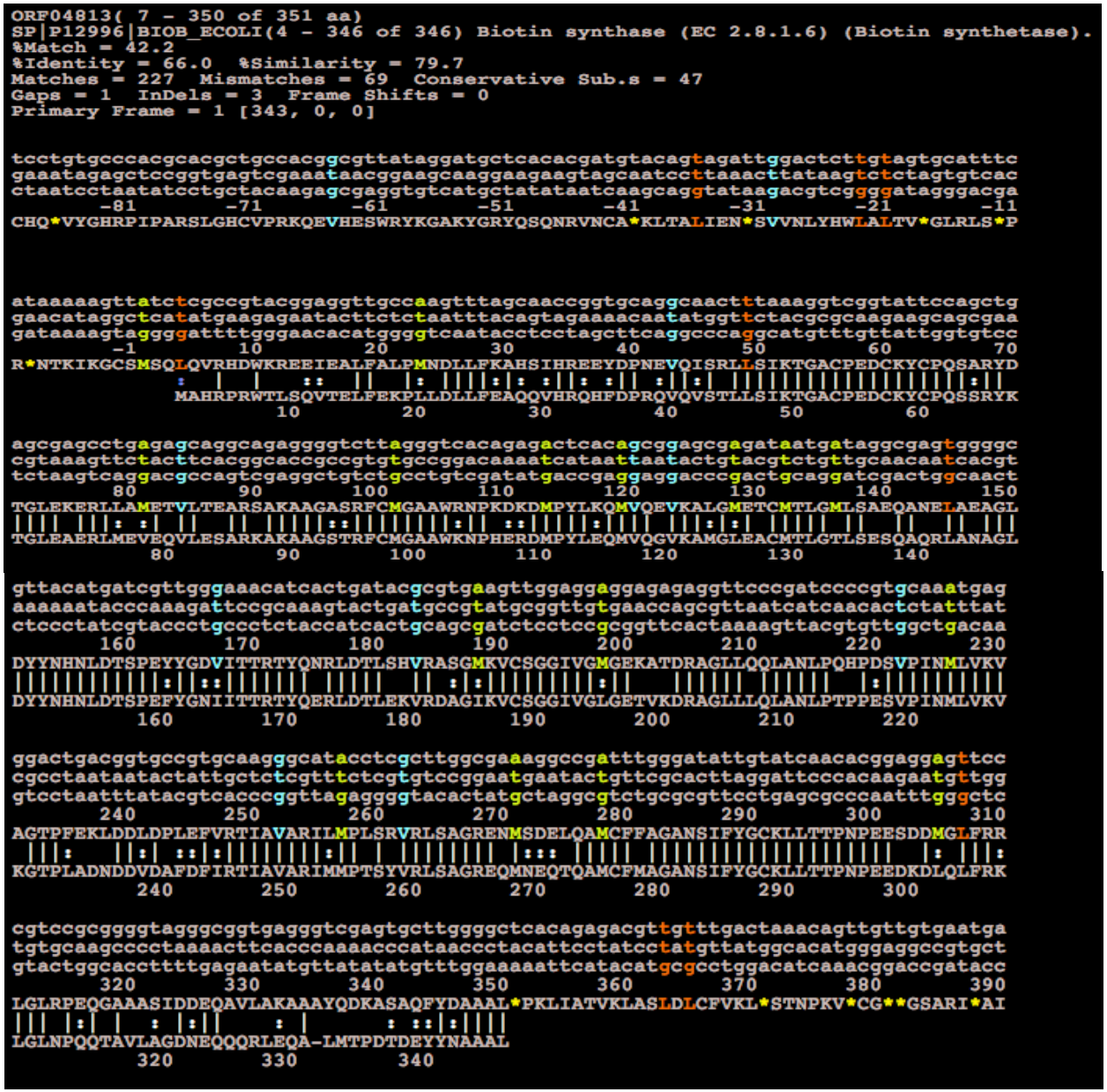
		2nd base							
		U		C		A		G	
1st base	U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine
		UUC	(Phe/F) Phenylalanine	UCC	(Ser/S) Serine	UAC	(Tyr/Y) Tyrosine	UGC	(Cys/C) Cysteine
		UUA	(Leu/L) Leucine	UCA	(Ser/S) Serine	UAA	Stop (*Ochre*)	UGA	Stop (*Opal*)
		UUG	(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[A]	(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	GGA	(Gly/G) Glycine
		GUG	(Val/V) Valine	GCG	(Ala/A) Alanine	GAG	(Glu/E) Glutamic acid	GGG	(Gly/G) Glycine

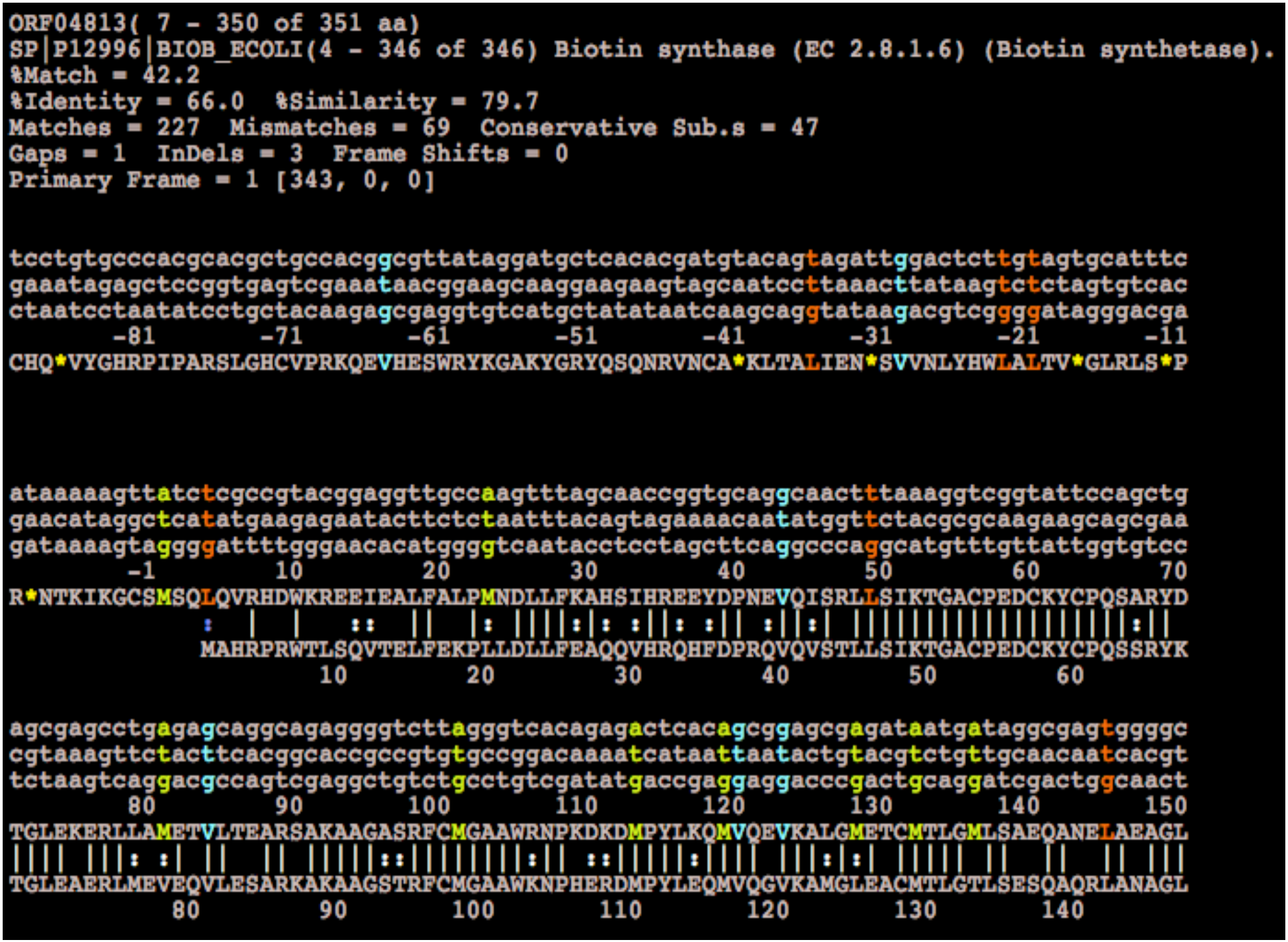
Gather Evidence

Pairwise Alignments – IGS uses BER (blast_extend_repraze)



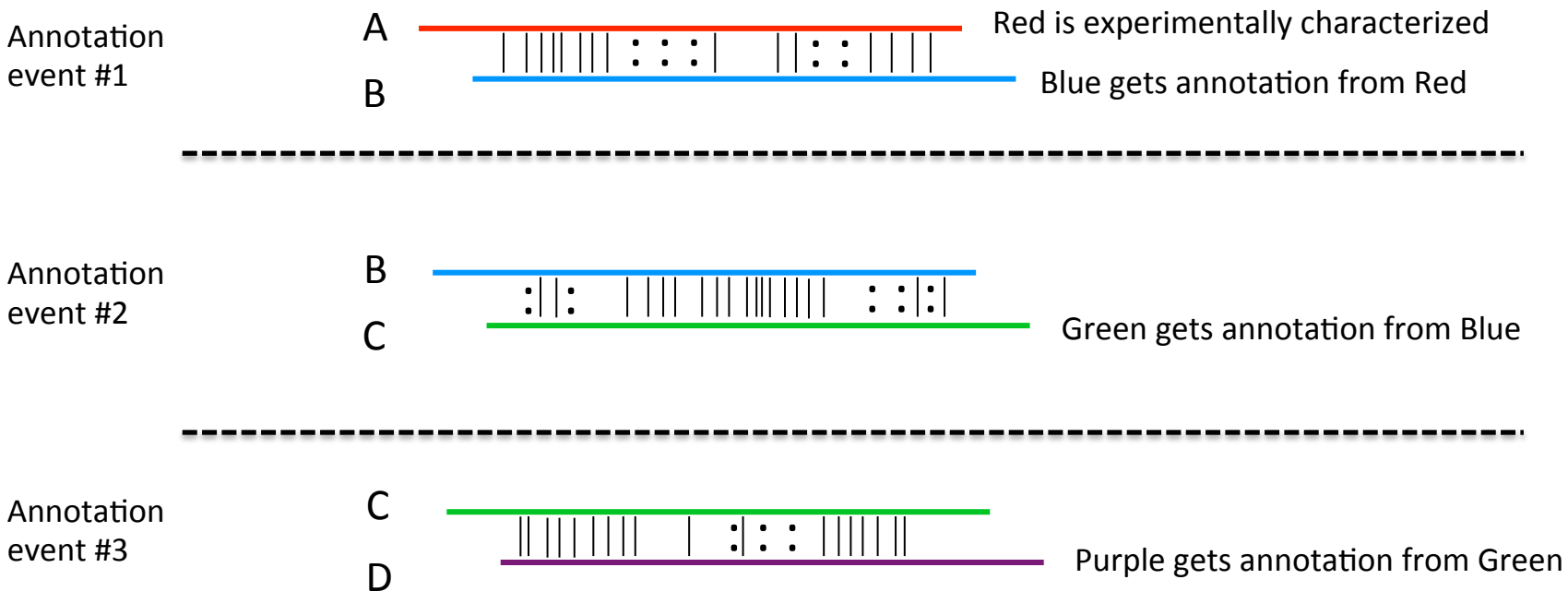
BER Alignment




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]
tcctgtgcccacgcacgctgccacggcgttataggatgctcacacgatgtacagtagattggactcttgtagtgcatttc
gaaatagagctccggtgagtcgaaataacggaagcaaggaagaagtagcaatccttaaacttataagtctctagtgtcac
ctaatcctaatatcctgctacaagagcgaggtgtcatgctatataatcaagcaggtataagacgtcgggatagggacga
-81 -71 -61 -51 -41 -31 -21 -11
CHQ*VYGHRPIPARSLGHCVPRKQEVHESWRYKGAKYGRYQSQNRVNCA*KLTALIEN*SVVNLYHWLALTV*GLRLS*P
ataaaaagttatctcgccgtacggaggttgccaagtttagcaaccggtgcaggcaactttaaaggtcggtattccagctg
gaacataggctcatatgaagagaatacttctctaatttacagtagaaaacaatatggttctacgcgcaagaagcagcgaa
gataaaagtagggggattttgggaacacatgggtcaatacctcctagcttcaggcccaggcatgtttgttattggtgtcc
-1 10 20 30 40 50 60 70
R*NTKIKGCSMSQLQVRHDWKREEIEALFALPMNDLLFKAHSIHREEYDPNEVQISRLLSIKTGACPEDCKYCPQSARYD
MAHRPRWTLSQVTELFEKPLLDLLFEAQQVHRQHFDPRQVQVSTLLSIKTGACPEDCKYCPQSSRYK
10 20 30 40 50 60
agcgagcctgagagcaggcagaggggtcttagggtcacagagactcacagcggagcgagataatgataggcgagtggggc
cgtaaagttctacttcacggcaccgccgtgtgccggacaaaatcataattaatactgtacgtctgttgcaacaatcacgt
tctaagtcaggacgccagtcgaggctgtctgcctgtcgatatgaccgaggaggacccgactgcaggatcgactggcaact
80 90 100 110 120 130 140 150
TGLEKERLLAMETVLTEARSAKAAGASRFCMGAAWRNPKDKDMPYLKQMVQEVKALGMETCMTLGMLSAEQANELAEAGL
TGLEAERLMEVEQVLESARKAKAAGSTRFCMGAAWKNPHERDMPYLEQMVQGVKAMGLEACMTLGTLSESQAQRLANAGL
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: Swiss-Prot, which is manually annotated and reviewed. TrEMBL, which is automatically annotated and is **not** reviewed. Includes complete and reference proteome sets.
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords, subcellular locations, cross-referenced databases and more.

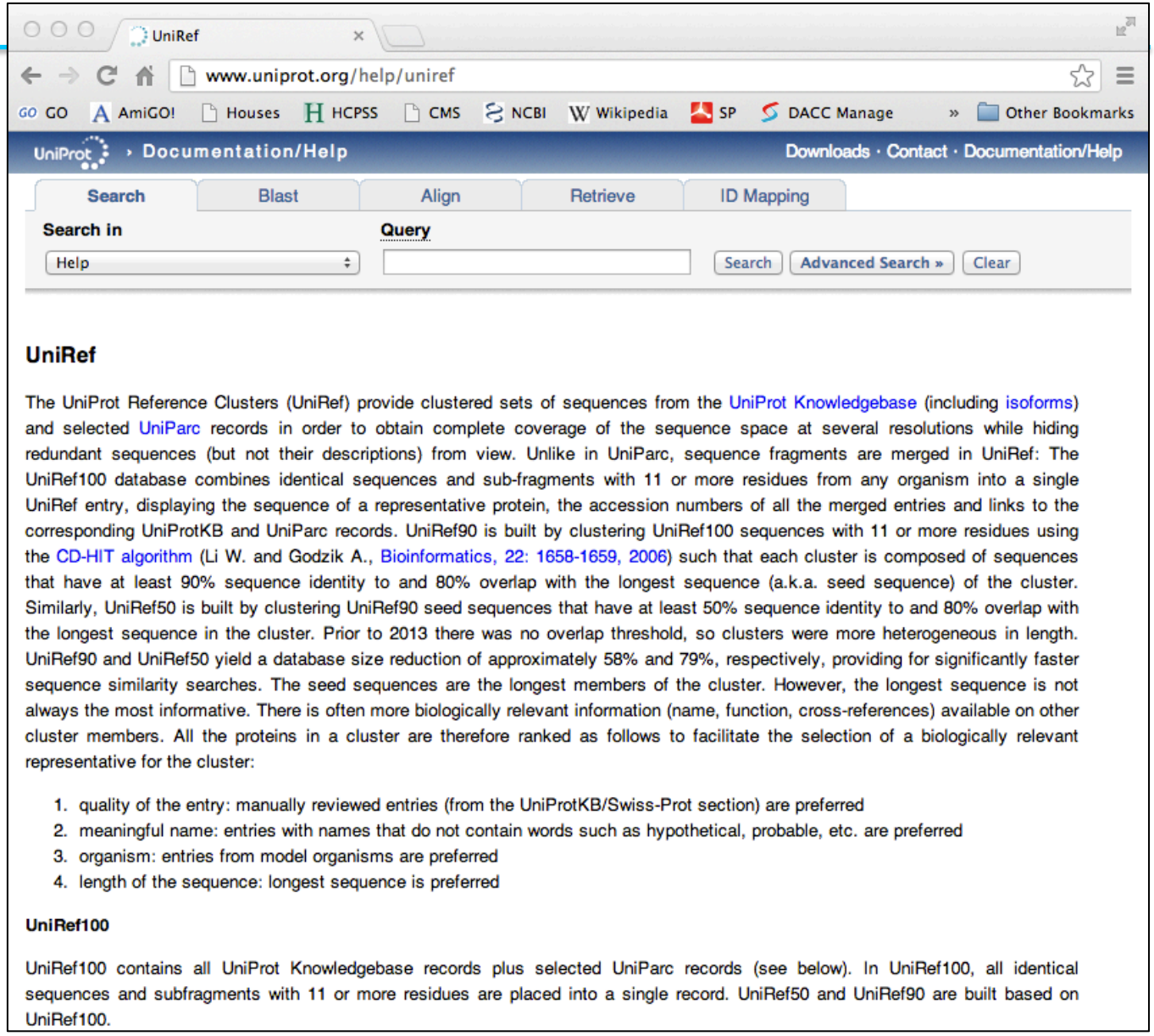
Getting started

- 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

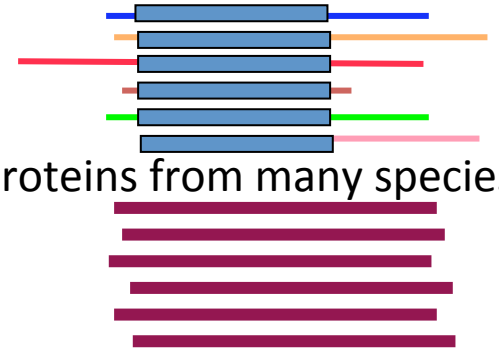
- UniRef 100
- UniRef 90
- UniRef 50



HMM databases

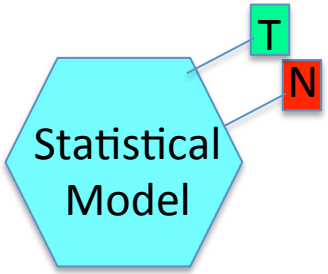
Statistical models of patterns of amino acids within an alignment of proteins – can be used to predict membership of a protein in a functional family. Some families are functionally specific, some are more general.

Pfam and TIGRFAM



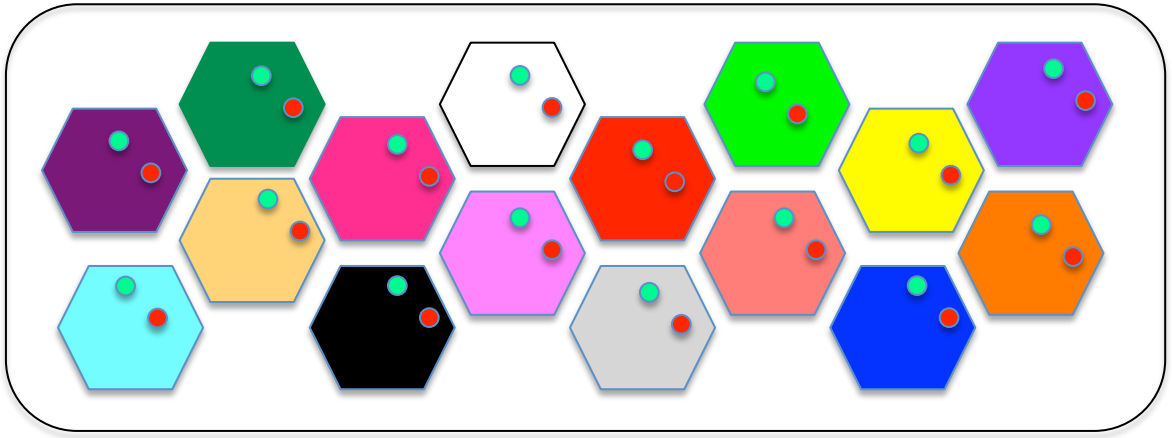
Proteins from many species

Alignments of functionally related proteins act as training sets for HMM building



Model specific to a family of proteins, generally found across many species

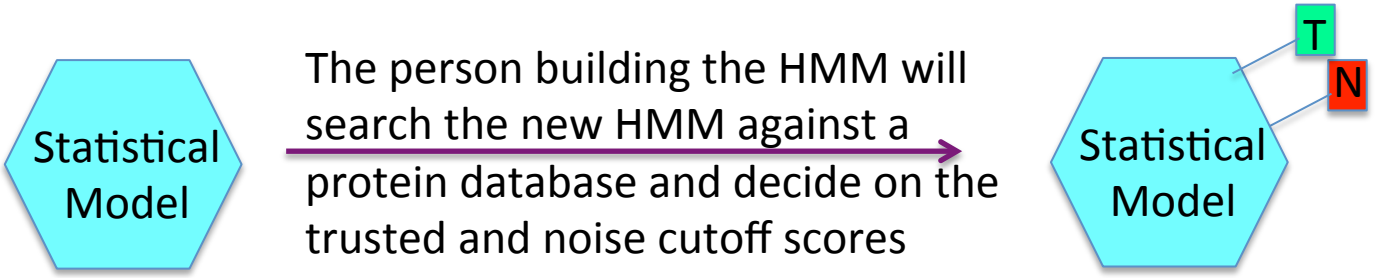
Add this model to the database



Database of HMM models, each specific to one protein family and/or functional level

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 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)

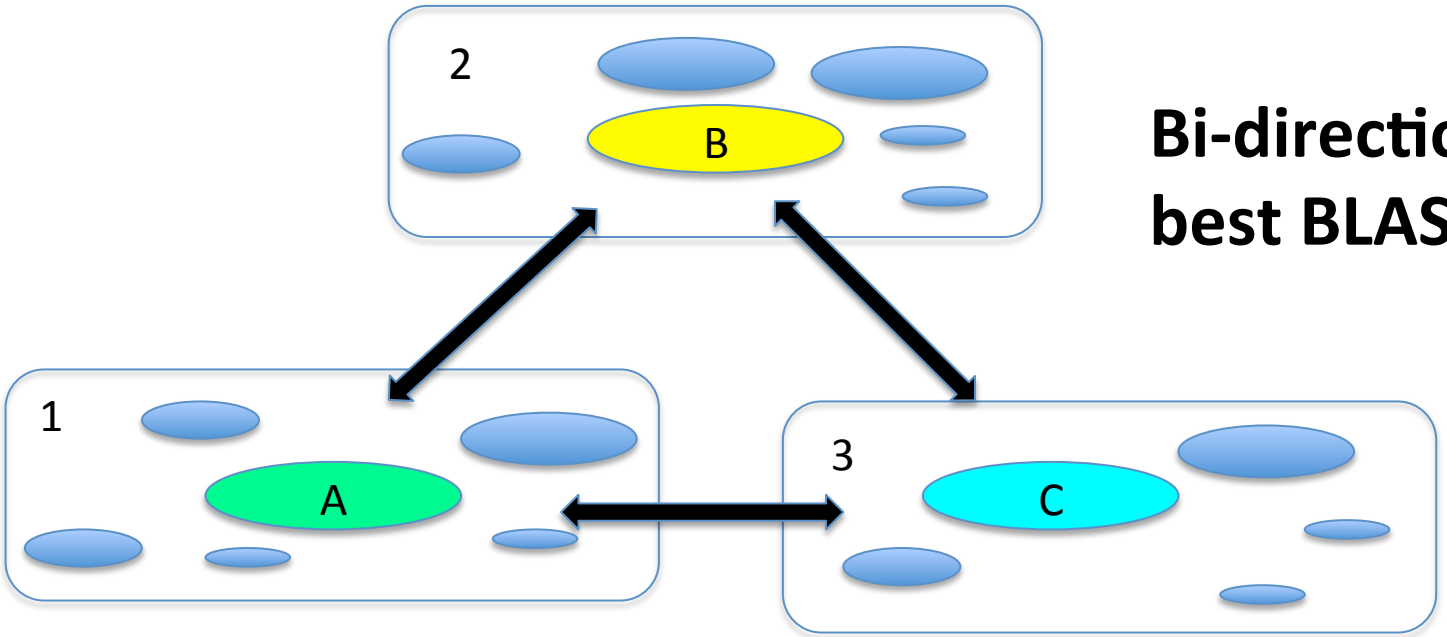
- Functionally general HMMs have general annotations
  PF04055
  - name: radical SAM domain protein
  - category: domain
  - EC: not applicable
  - gene symbol: not applicable
  - Roles:
    - enzymes of unknown specificity (TIGR role 703)
    - catalytic activity (GO:0003824)
    - metabolism (GO:0008152)

tigr fams
tigr protein families
J. Craig Venter INSTITUTE

Pfam
EMBL-EBI

Orthologous Groups

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



Bi-directional best BLAST

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)

The CBS tools cannot be included in CloVR

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!**

Alex756, https://commons.wikimedia.org/wiki/File:Blacks-Law-Dictionary.jpg

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[5]

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

http://en.wikipedia.org/wiki/Enzyme_Commission_number

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	
EC 1.1	Acting on the CH-OH group of donors	separate	up to 50
EC 1.1.1	With NAD^+ or $NADP^+$ as acceptor	separate	up to 50
EC 1.1.2	With a cytochrome as acceptor	separate	up to 50
EC 1.1.3	With oxygen as acceptor	separate	up to 50
EC 1.1.4	With a disulfide as acceptor	separate	up to 50
EC 1.1.5	With a quinone or similar compound as acceptor	separate	up to 50
EC 1.1.99	With other acceptors	separate	up to 50
EC 1.2	Acting on the aldehyde or oxo group of donors	separate	up to 50
EC 1.2.1	With NAD^+ or $NADP^+$ as acceptor	separate	up to 50
EC 1.2.2	With a cytochrome as acceptor	separate	up to 50
EC 1.2.3	With oxygen as acceptor	separate	up to 50
EC 1.2.4	With a disulfide as acceptor	separate	up to 50
EC 1.2.7	With an iron-sulfur protein acceptor	separate	up to 50
EC 1.2.99	With other acceptors	separate	up to 50
EC 1.3	Acting on the CH-CH group of donors	separate	up to 50
EC 1.3.1	With NAD^+ or $NADP^+$ as acceptor	separate	up to 50
EC 1.3.2	With a cytochrome as acceptor	separate	up to 50
EC 1.3.3	With oxygen as acceptor	separate	up to 50
EC 1.3.5	With a quinone or related compound as acceptor	separate	up to 50
EC 1.3.7	With an iron-sulfur protein as acceptor	separate	up to 50
EC 1.3.99	With other acceptors	separate	up to 50
EC 1.4	Acting on the CH-NH_2 group of donors	separate	up to 50
EC 1.4.1	With NAD^+ or $NADP^+$ as acceptor	separate	up to 50
EC 1.4.2	With a cytochrome as acceptor	separate	up to 50
EC 1.4.3	With oxygen as acceptor	separate	up to 50
EC 1.4.4	With a disulfide as acceptor	separate	up to 50
EC 1.4.7	With an iron-sulfur protein as acceptor	separate	up to 50
EC 1.4.99	With other acceptors	separate	up to 50

Specific Example

IUBMB Enzyme Nomenclature

EC 1.2.3.4

Accepted name: oxalate oxidase

Reaction: oxalate + O_2 + **2** H^+ = **2** CO_2 + H_2O_2

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

Systematic name: oxalate:oxygen oxidoreductase

Comments: Contains Mn^{2+} 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
- 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

Gene Ontology

Gene Ontology Consortium

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[a]

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 [b]	53 042 843
Manually annotated (experimental) gene products	311 335
Manually annotated (phylogenetic) gene products	79 839
Total annotations	4 185 487

[a]As of September 2014.

[b]Includes isoforms.

Nucleic Acids Res. 2015 Jan;43(Database issue):D1049-56.

Effective capture of information

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

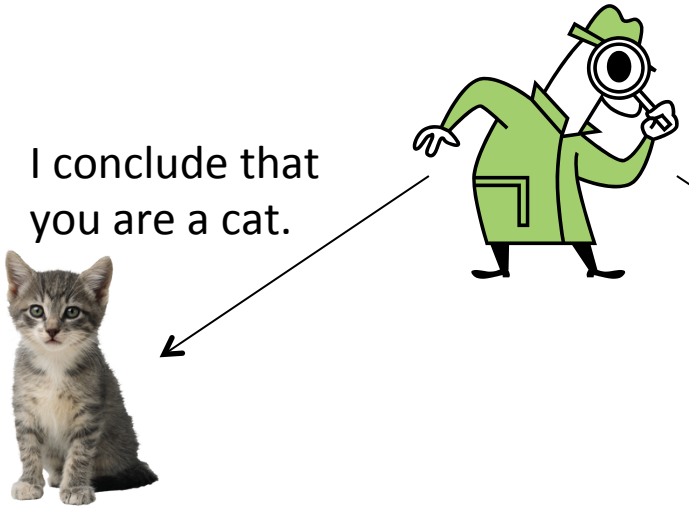
A protein is integral to the plasma membrane and is part of an ATP-binding cassette (ABC) transporter complex. It functions as part of a transporter to accomplish the transport of sulfate across the plasma membrane using ATP hydrolysis as an energy source.

=

- 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

- Classes
  - assertion method
    - automatic assertion
    - manual assertion
  - evidence
    - author statement
      - **author statement used in manual assertion**
      - non-traceable author statement
      - traceable author statement
    - combinatorial analysis
      - **combinatorial analysis used in automatic assertion**
      - **combinatorial analysis used in manual assertion**
      - computational combinatorial analysis
        - **computational combinatorial analysis used in aut**
        - **computational combinatorial analysis used in mar**
    - curator inference
      - **curator inference used in manual assertion**
      - inference from background scientific knowledge
      - no biological data found
    - experimental evidence
      - animal model system study data
      - clinical study data
      - direct assay result
      - **experimental evidence used in manual assertion**
      - experimental genomic evidence
      - experimental phenotypic evidence
      - expression pattern evidence
      - physical interaction
      - sequencing assay result
    - genomic context
    - similarity
      - genetic similarity
      - phenotypic similarity
      - phylogenetic evidence
      - sequence similarity

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.

OBO Foundry ontologies

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

OBO Foundry candidate ontologies and other ontologies of interest

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

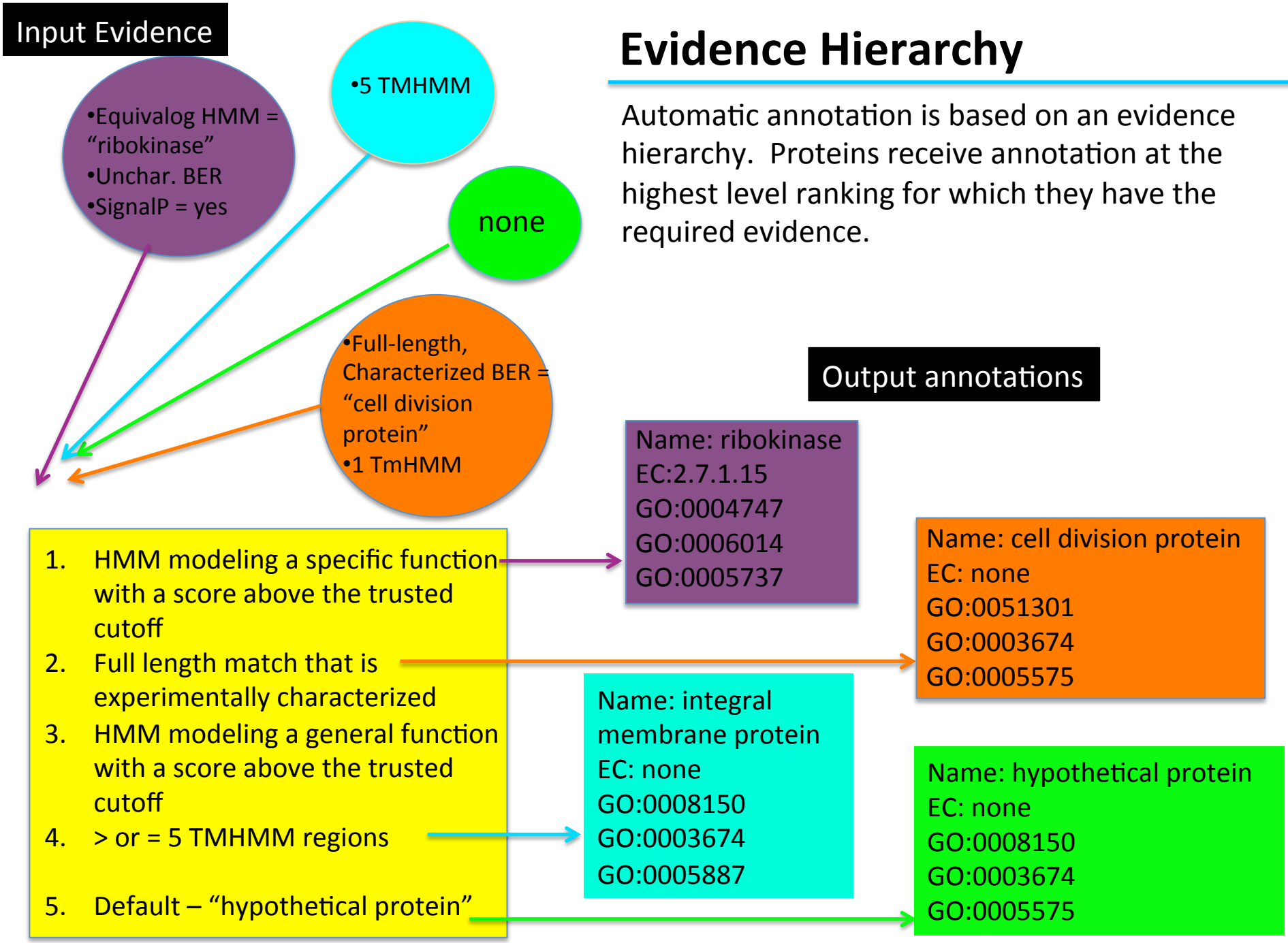
Making the Assignments

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.

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.

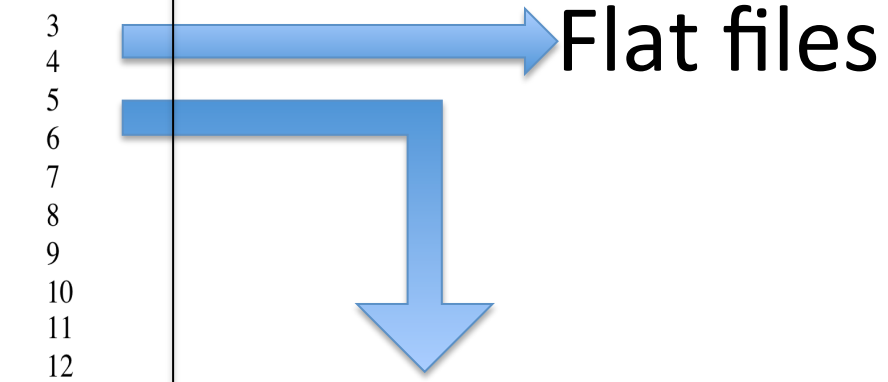


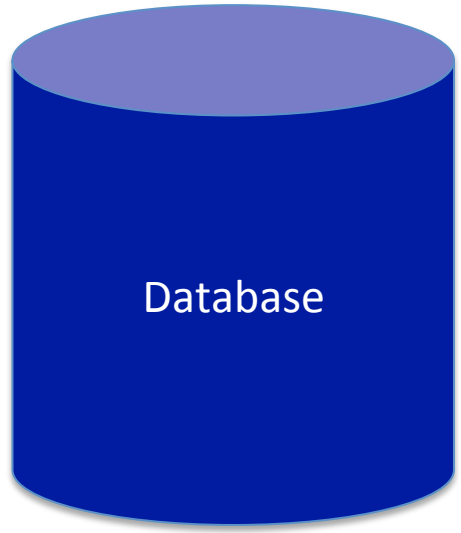
pFunc: the actual hierarchy in the pipeline

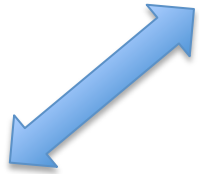
Evidence	Criteria	Query	Match	Rank
HMM	Equivalog	N/A	N/A	1
BER	Trusted	Full	Full	2
HMM	Equivalog Domain	Full	Full	3
BER	Trusted	Partial	Full	4
HMM	Subfamily	N/A	N/A	5
HMM	Superfamily	N/A	N/A	6
HMM	Subfamily Domain	N/A	N/A	7
HMM	Domain	Partial	Full	8
HMM	Pfam	Full	Full	9
BER	Trusted	Full	Partial	10
TMHMM	> 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

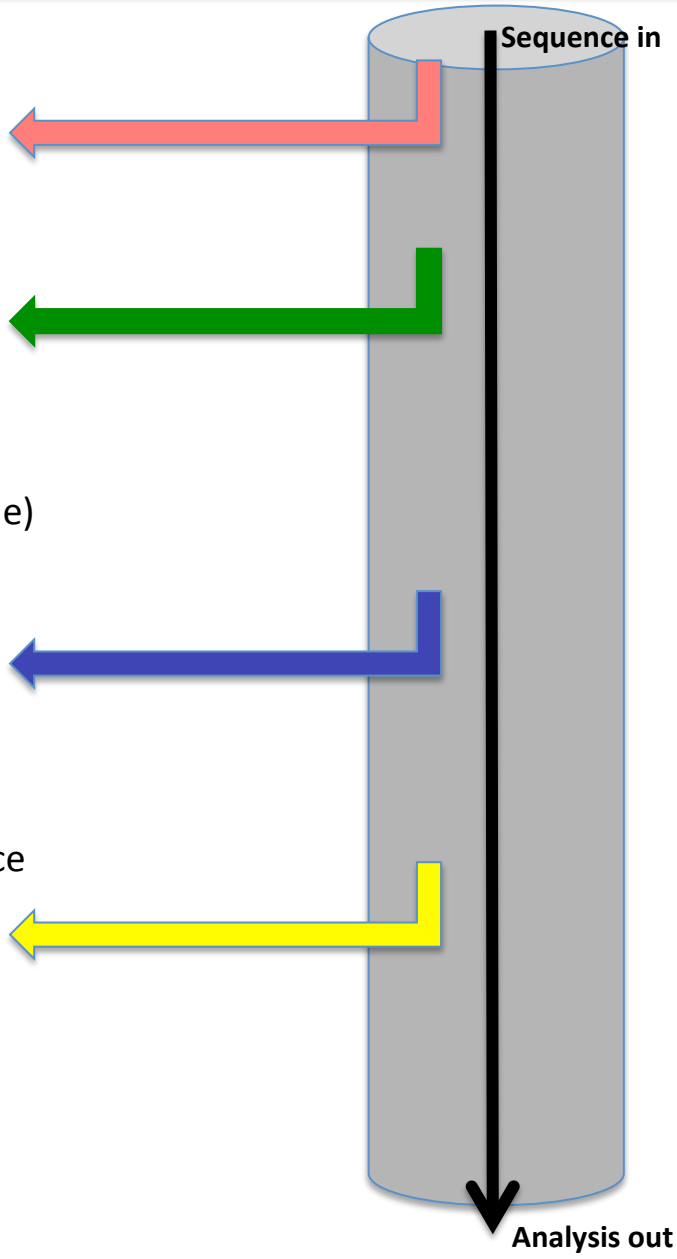
Visualization of the Data

- 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.