

# Protein Functional Annotation

Michelle Giglio

1

## Protein Functional Annotation

- Experimental knowledge of function
  - Literature curation
  - Perform experiment
  - Not possible for all proteins in most organisms (not even close in most)
- Sequence similarity
  - Protein not DNA
  - Shared sequence can imply shared function
    - but there are examples of single aa change leading to change in function
    - All sequence-based annotations are putative until proven experimentally
- Two steps
  - Collect sequence-based information by performing searches
  - Evaluate the search results
    - Automatically or manually

2

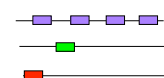
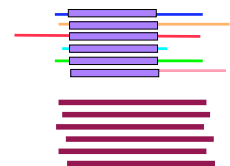
## Basic Set of Protein Annotations

- **protein name**
  - descriptive common name for the protein
    - e.g. “ribokinase”
- **gene symbol**
  - mnemonic abbreviation for the gene
  - e.g. “recA”
- **EC number**
  - only applicable to enzymes
    - e.g. 1.4.3.2
- **Role**
  - what the protein is doing in the cell and why
  - e.g. “amino acid biosynthesis”
- **Supporting Evidence**
  - accession numbers of BER and HMM matches
  - TmHMM, SignalP, LipoP
  - whatever information you used to make the annotation
- **Unique Identifier**
  - e.g. locus ids

3

## Collecting Sequence Similarity Evidence

- pairwise alignments
  - two protein’s amino acid sequences aligned next to each other so that the maximum number of amino acids match
- multiple alignments
  - 3 or more amino acid sequences aligned to each other so that the maximum number of amino acids match in each column
  - more meaningful than pairwise alignments since it is much less likely that several proteins will share sequence similarity due to chance alone, than that 2 will share sequence similarity due to chance alone. Therefore, such shared similarity is more likely to be indicative of shared function.
- protein families
  - clusters of proteins that all share sequence similarity and presumably similar function
  - may be modeled by various statistical techniques
- motifs
  - short regions of amino acid sequence shared by many proteins
    - transmembrane regions
    - active sites
    - signal peptides



4

# Protein Alignment Tools

- Local pairwise alignment tools do not worry about matching proteins over their entire lengths, they look for any regions of similarity within the proteins that score well.
  - BLAST
    - fast
    - comes in many varieties (see NCBI site)
  - Smith-Waterman
    - finds best out of all possible local alignments
    - slow but sensitive
- Global pairwise alignment tools take two sequences and attempt to find an alignment of the two over their full lengths.
  - Needleman-Wunsch
    - finds best out of all possible alignments
- Multiple alignments tools try to align 3 or more proteins so that the maximal number of amino acids from each protein are matched in the alignment - this may or may not include the full length of some or all of the proteins
  - clustalW
  - muscle

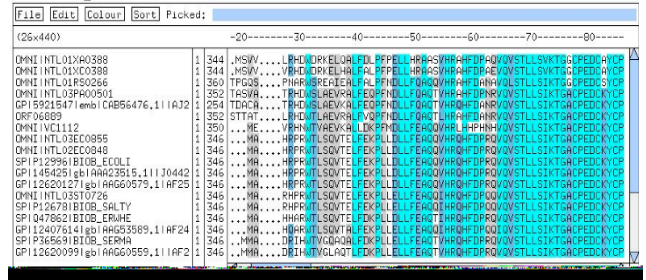
# Sample Alignments

## Pairwise



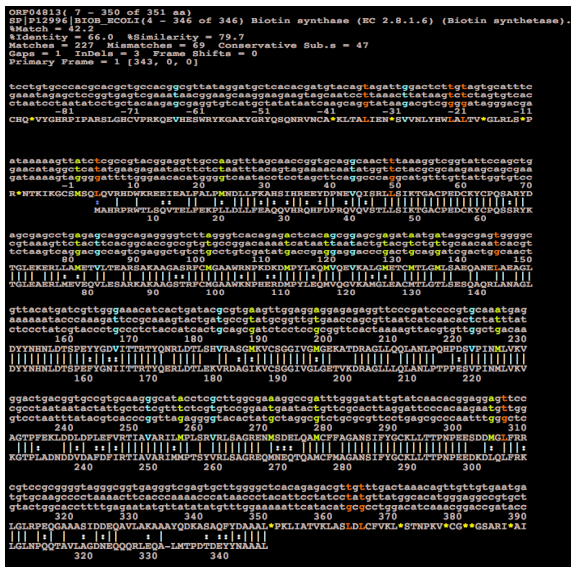
- two rows of amino acids compared to each other, the top row is the search protein and the bottom row is the match protein, numbers indicate amino acid position in the sequence
- solid lines between amino acids indicate **identity** (the same amino acid in the corresponding position in the alignments)
- dashed lines (colons) between amino acids indicate **similarity** (amino acids with similar chemical structure in corresponding positions in the alignments, these may be able to perform the same role in the protein).

## Multiple



Different shadings indicate amount of matching

# Sample full-length protein alignment



# Some terms to understand

- homologs**
  - two sequences have evolved from the same common ancestor
  - they may or may not share the same function
  - two proteins are either homologs of each other or they are not. A protein can not be more, or less, homologous to one protein than to another.
- orthologs**
  - a type of homolog where the two sequences are in different species that arose from a common ancestor. The fact of the speciation event has created the two copies of the sequence.
  - orthologs often, but not always, share the same function
- paralogs**
  - a type of homolog where the two sequences have arisen due to a gene duplication within one species
  - paralogs will initially have the same function (just after the duplication) but as time goes by, one copy will be free to evolve new functions, as the other copy will maintain the original function. This process is called "neofunctionalization".
- xenologs**
  - a type of ortholog where the two sequences have arisen due to lateral (or horizontal) transfer
  - lateral transfer is when genes move from one species to another by a means other than reproduction



# UniProt continued

References Hide | Top

[Hide 'large scale' references](#)

- "Nucleotide sequence of the *lig* gene and primary structure of DNA ligase of *Escherichia coli*."**  
Ishino Y, Shiragami H, Makino K, Tsunashima S, Sakiyama T, Nakata A.  
*Mol. Gen. Genet.* 204:1-7(1988) [PubMed: 3018436] [Abstract]  
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA], PROTEIN SEQUENCE OF 1-13 AND 666-671.  
Strain: K12 / CG501 / ATCC 25725 / DSM 3025 / LMG 3041 / NCIB 10222.
- "Cloning of the *lig* gene and primary structure of DNA ligase of *Escherichia coli* K-12."**  
O'Connor M.J., Alty A., Alty D., Zhang X., Robinson M., Backman K.  
Submitted (APR-1989) to the EMBL/GenBank/DBJ databases  
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA].
- "Construction of a contiguous 874-kb sequence of the *Escherichia coli*-K12 genome corresponding to 50.0-68.8 min on the linkage map and analysis of its sequence features."**  
Yamamoto Y, Aiba H, Baba T, Hayashi K, Inada T, Isono K, Itoh T, Kimura S, Kitagawa M, Makino K, Miki T, Mitsuhashi N, Mizobuchi K, Mori H, Nakade S, Nakamura Y, Nishimoto H, Oshima T, Horuchi T.  
*DNA Res.* 4:511-519(1997) [PubMed: 9228020] [Abstract]  
Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].  
Strain: K12 / W3110 / ATCC 27325 / DSM 5911.
- "The complete genome sequence of *Escherichia coli* K-12."**  
Bathner F.B., Parkhill G.R., Rhee C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F., Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J., Mau B., Shaq Y.  
*Science* 277:1453-1474(1997) [PubMed: 9273303] [Abstract]  
Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].  
Strain: K12 / MG1655 / ATCC 47076.
- "Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110."**  
Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, Ohtsubo E, Baba T, Wanner B.L., Mori H, Horuchi T.  
*Genome Res.* 15:1053-1061(2005) [PubMed: 15832268] [Abstract]

Sequence databases Family and domain databases

EMBL GenBank DDBJ

PIR

RefSeq

3D structure databases

Protein-protein interaction databases

Proteomic databases

Genome annotation databases

Organism-specific databases

Entry information

Entry name: DNLJ\_ECOLI

Accession: Primary (stable) accession number: P19042  
Secondary accession number(s): P78197, P78198

Entry history: Integrated into UniProtKB/Swiss-Prot: April 1, 1990  
Last sequence update: November 1, 1997  
Last modified: March 23, 2010  
This is version 110 of the entry and version 2 of the sequence. [Complete history]

Entry status: Reviewed (UniProtKB/Swiss-Prot)

Annotation project: HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes)

# Other Useful Databases

## • other NCBI resources

- National Center for Biotechnology Information
- DNA sequences
- taxonomy resource
- many other resources/tools
- <http://www.ncbi.nlm.nih.gov/>

## • Enzyme Commission

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

## • metabolic pathways

- KEGG
  - <http://www.genome.jp/kegg/>
- MetaCyc/BioCyc
  - <http://metacyc.org/>
  - <http://www.biocyc.org/>
- BRENDA
  - <http://www.brenda-enzymes.info/>

14

## What is the Enzyme Commission and an EC number?

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 Catalyse

<http://www.chem.qmul.ac.uk/iubmb/enzyme/>

### 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>

not the complete list....

## An example EC entry

IUBMB Enzyme Nomenclature

### EC 1.2.3.4

**Accepted name:** oxalate oxidase

**Reaction:** oxalate + O<sub>2</sub> + 2 H<sup>+</sup> = 2 CO<sub>2</sub> + H<sub>2</sub>O<sub>2</sub>

**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:

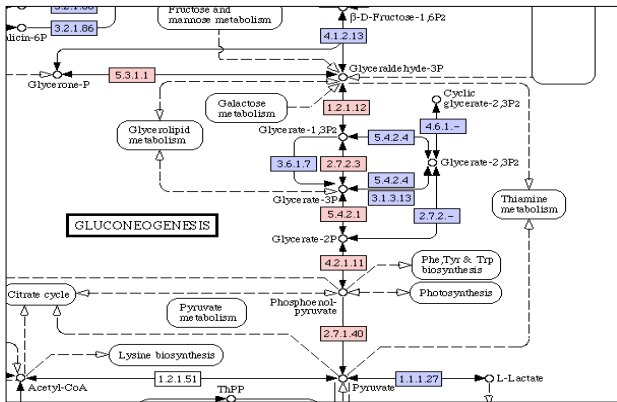
- Datta, P.K., Meeuse, B.J.D., Engstrom-Heg, V. and Hilal, S.H. Moss oxalic acid oxidase - a flavoprotein. *Biochim. Biophys. Acta* 17 (1955) 602-603. [PMID: 13250021]
- Kotsira, V.P. and Clonis, Y.D. Oxalate oxidase from barley roots: purification to homogeneity and study of some molecular, catalytic, and binding properties. *Arch. Biochem. Biophys.* 340 (1997) 239-249. [PMID: 9143327]
- 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]

# KEGG

MODULE: M10002 Help

<b>Entry</b>	M10002	Pathway	Module
<b>Name</b>	Glycolysis, core module		
<b>Definition</b>	K01803 K00134 K00927 K01834 K01689 K00873		
<b>Pathway</b>	ko00010 Glycolysis / Gluconeogenesis		
<b>Orthology</b>	K01803 triosephosphate isomerase [EC:5.3.1.1] [RN:R01015] K00134 glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [RN:R01061 R01063] K00927 phosphoglycerate Kinase [EC:2.7.2.3] [RN:R01512] K01834 phosphoglycerate mutase [EC:5.4.2.1] [RN:R01518] K01689 enolase [EC:4.2.1.11] [RN:R00658] K00873 pyruvate Kinase [EC:2.7.1.40] [RN:R02320]		
<b>Reaction</b>	R01015 C00111 → C00118 R01061,R01063 C00118 → C00236 R01512 C00236 → C00197 R01518 C00197 → C00631 R00658 C00631 → C00074 R02320 C00074 → C00022		
<b>Compound</b>	C00111 Glycerone phosphate C00118 (2R)-2-Hydroxy-3-(phosphonoxy)-propanal C00236 3-Phospho-D-glyceroyl phosphate C00197 3-Phospho-D-glycerate C00631 2-Phospho-D-glycerate C00074 Phosphoenolpyruvate C00022 Pyruvate		
<b>LinkDB</b>	All DBs		



# BioCyc



## E. coli K-12 Pathways Class: Alanine Biosynthesis

Login (Optional): [Why Login?](#) [Create New Account](#) [Help](#)

### Summary:

This class contains pathways of the biosynthesis of L-alanine de novo (from intermediates of central metabolism) and from other amino acids or their biosynthetic intermediates; L-alanine is a constituent of protein and peptidoglycan. It is a source of D-alanine, which is also a constituent of peptidoglycan.

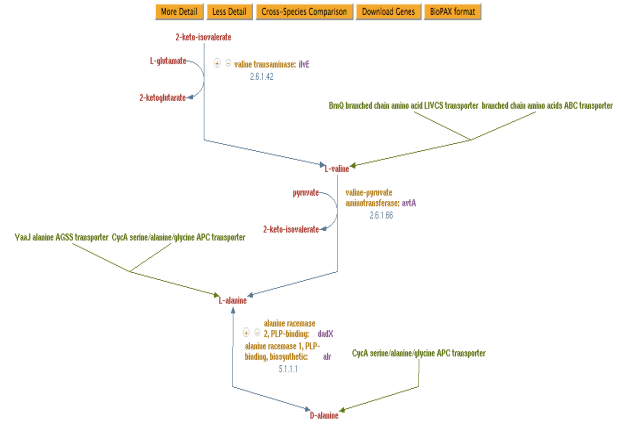
### Parent Classes:

[Individual Amino Acids Biosynthesis](#)

Note: This class is a variant class, i.e. its purpose is to group together a set of variant pathways. Variant pathways are those that accomplish roughly the same biological function, such as degradation of a given starting material, or biosynthesis of an end product. The variant pathways may or may not share any common reactions.

### Instances:

[alanine biosynthesis I](#),  
[alanine biosynthesis II](#),  
[alanine biosynthesis III](#),  
[superpathway of alanine biosynthesis](#)



# Searches

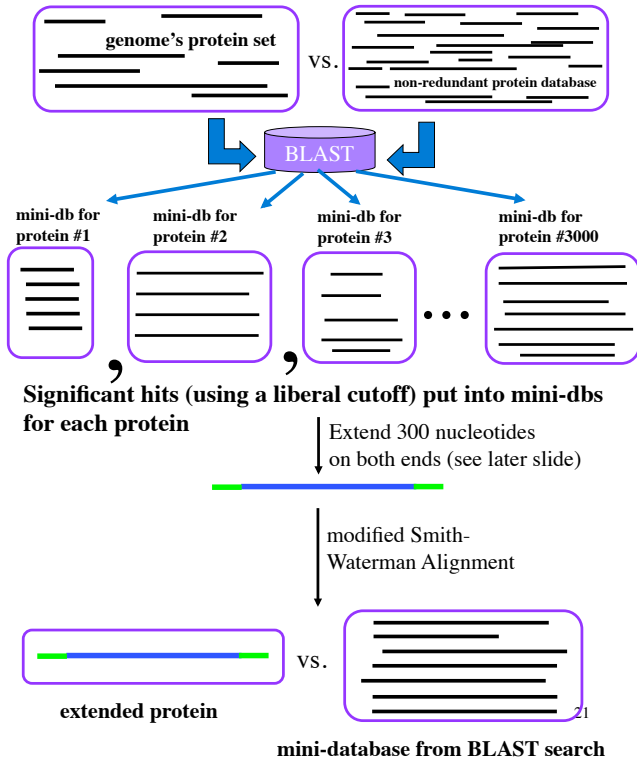
## BLAST-extend-repraze (BER)

### A pairwise protein search tool

- Initial BLAST search of new query proteins from the genome sequence
  - search against non-redundant protein database
  - stores hits in mini-database for each query protein
- Query protein sequence is extended by 300 nucleotides on each end (and translated)
  - this helps in finding frameshifts and in-frame stops and will be described in more detail later
- A modified Smith-Waterman alignment is generated between the query and each match sequence in the mini-database
  - produces a file of pairwise alignments between the query protein and the match protein - one alignment for each match protein in the query's mini-database
- BER tool has the ability to look through in-frame stop codons and across frameshifts to determine if similarity continues
  - This is why we provide the extensions
  - Very useful for finding sequencing errors or potential mutations



# BLAST-extend-repraze (BER)



# BER Alignment

An alignment like the one shown here will be generated for every match protein in the mini-database.

# BER Alignment detail

- The boxed header gives a brief summary of alignment statistics, the species that match comes from and a list of synonymous accessions for the match protein. Background is gold for experimentally characterized proteins.

- 1st/2nd line indicate range of match over query/match
- third line shows percent identity
- last line indicates the number of amino acids in the alignment found in each forward frame for the sequence as defined by the coordinates of the gene.

- BER alignments include nucleotide sequence which reads down in columns of 3 which correspond to codons
- The amino acid for each codon is placed under that codon
- Starts are color-coded: ATG=green, GTG=blue, TTG=red
- Numbers indicate position in amino acids, where position 1 is the start of the protein sequence
- asterisks indicate stop codons

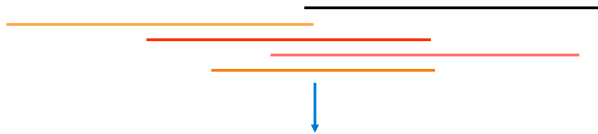
# BER Skim

A list of matches to the query protein with statistics on length of match and BLAST p-value. Colored backgrounds indicate match protein is likely to be experimentally characterized.



## Building HMMs

**Collect proteins to be in the “seed”**  
(same function/similar domain/ family membership)



**Generate and Curate Multiple Alignment of Seed proteins**



Region of good alignment and closest similarity

**Run HMM algorithm**  
Computes statistical probabilities for amino acid patterns in the seed

this step may need a few iterations

**Search new model against all proteins**

**Choose “noise” and “trusted” cutoff scores based on what scores the “known” vs. “unknown” proteins receive.** <sup>29</sup>  
HMM is ready to go!

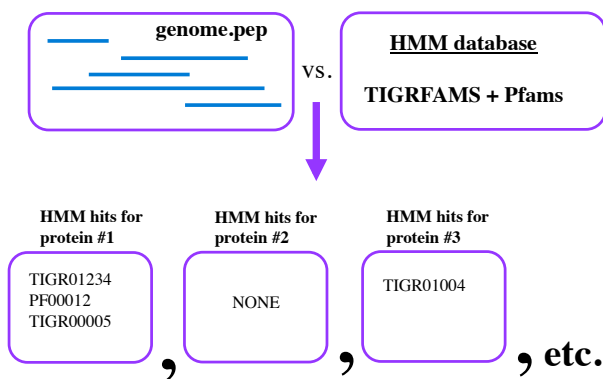
## Choosing cutoff scores

- search the new HMM against a protein database
- see the range of scores the match proteins receive
- do analysis to determine where confirmed members score
- do analysis to determine where confirmed non-members score
- set the cut-offs accordingly

matches (seed members bold)	score	
<b>protein “definitely”</b>	<b>547</b>	
<b>protein “absolutely”</b>	<b>501</b>	
protein “sure thing”	487	
<b>protein “confident”</b>	<b>398</b>	
protein “safe bet”	376	
<b>protein “very confident”</b>	<b>365</b>	
<b>protein “has to be one”</b>	<b>355</b>	300
protein “could be”	210	
protein “maybe”	198	
protein “not sure”	150	100
protein “no way”	74	
protein “can’t be”	54	
protein “not a chance”	47	

- proteins that score above trusted can be considered part of the protein family modeled by the HMM
- proteins that score below noise should not be considered part of the protein family modeled by the HMM
- usefulness of an HMM is directly related to the care taken by the person building the HMM since some steps are subjective

## HMM Searches



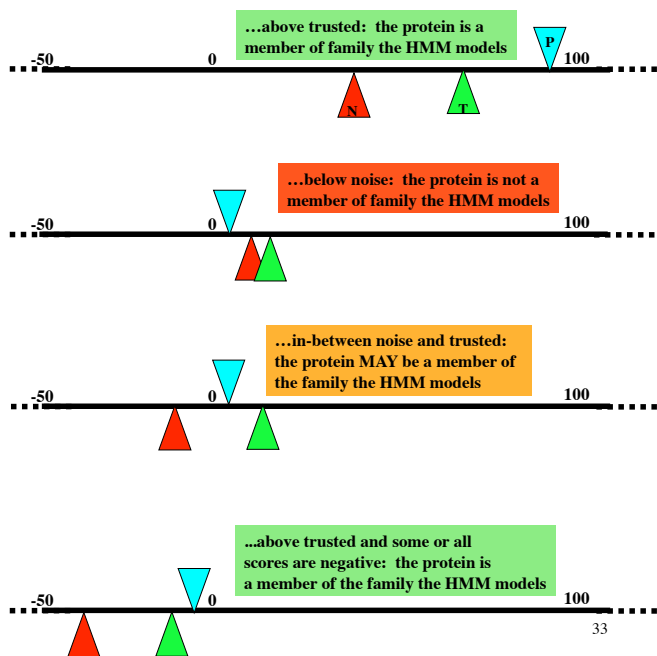
Each protein in the genome is searched against all HMMs in our db. Some will not have significant hits to any HMM, some will have significant hits to several HMMs. Multiple HMM hits can arise in many ways, for example: the same protein could hit an equivalog model, a superfamily model to which the equivalog function belongs, and a domain model representing the catalytic domain for the particular equivalog function. There is also overlap between TIGR and Pfam HMMs.

## Searching with HMMs

- query proteins searched against HMM database with HMMER software package
  - [www.psc.edu/general/software/packages/hmmer/manual/main.html](http://www.psc.edu/general/software/packages/hmmer/manual/main.html)
- query proteins receive a score that measures how closely the patterns of amino acids in the protein match those captured with the model
- the 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.



**HMM scores:** The cutoff scores attached to HMMs, are sometimes high and sometimes low and sometimes even negative. There is no inherent meaning in how high or low a cutoff score is, the only thing to worry about is whether or not your protein scores above trusted or below noise.



## HMM Output in Manatee

**HMM** submit | all hms |

**TIGR00433: biotin synthase** gene\_sym: bioB ecf: 2.0.1.6 role\_id: 77

Isology: **equivalog**

Total score: <b>584.1</b>	Trusted cutoff: <b>300.00</b>	Gathering cutoff: <b>300.00</b>	Noise cutoff: <b>50.00</b>	Total expect: <b>1.2e-166</b>
	Trusted cutoff2: <b>300.00</b>	Gathering cutoff2: <b>300.00</b>	Noise cutoff2: <b>50.00</b>	

View Alignment	Coords	HMM Coords	Score	Expect	Curation	[Add To GO Evidence]
<a href="#">align page</a>	18-313	1-350/350	584.1	1.2e-166	<input checked="" type="checkbox"/>	

[GO:0004076](#) [add](#) [biotin synthase activity \(function\)](#)

[GO:0009102](#) [add](#) [biotin biosynthesis \(process\)](#)

**PF04055: radical SAM domain protein** gene\_sym: none ecf: none role\_id: 703

Isology: **domain**

Total score: <b>82.8</b>	Trusted cutoff: <b>7.00</b>	Gathering cutoff: <b>7.00</b>	Noise cutoff: <b>6.00</b>	Total expect: <b>9.1e-22</b>
	Trusted cutoff2: <b>7.00</b>	Gathering cutoff2: <b>7.00</b>	Noise cutoff2: <b>6.00</b>	

View Alignment	Coords	HMM Coords	Score	Expect	Curation	[Add To GO Evidence]
<a href="#">align page</a>	50-212	1-163/163	82.8	9.1e-22	<input checked="" type="checkbox"/>	

[GO:0003824](#) [add](#) [catalytic activity \(function\)](#)

[GO:0008152](#) [add](#) [metabolism \(process\)](#)

## Things to ask yourself when annotating with HMMs

- Does my protein score above the trusted cutoff?
- What category is the HMM in?
  - domain, superfamily, equivalog, etc.
- What annotation on the HMM can I use for my protein?

## Metabolic Reconstruction

- find pathways, protein complexes, biological systems, etc. that are present in your genome
- many tools which store/predict pathways
  - **Pathway Tools (SRI)**
    - [bioinformatics.ai.sri.com](http://bioinformatics.ai.sri.com)
  - SEED
    - [www.theseed.org](http://www.theseed.org)
  - KEGG
    - <http://www.genome.ad.jp/kegg/pathway.html>
  - HAMAP
    - [www.expasy.org/sprot/hamap](http://www.expasy.org/sprot/hamap)
  - Genome Properties
    - [cmr.jvci.org/cgi-bin/CMR/shared/GenomePropertiesHomePage.cgi](http://cmr.jvci.org/cgi-bin/CMR/shared/GenomePropertiesHomePage.cgi)
  - all look for the “steps” or “components” of the various systems or structures
- **PRIAM**
  - **Enzyme-specific profiles for metabolic pathway prediction**
  - **Assigns EC numbers to proteins**

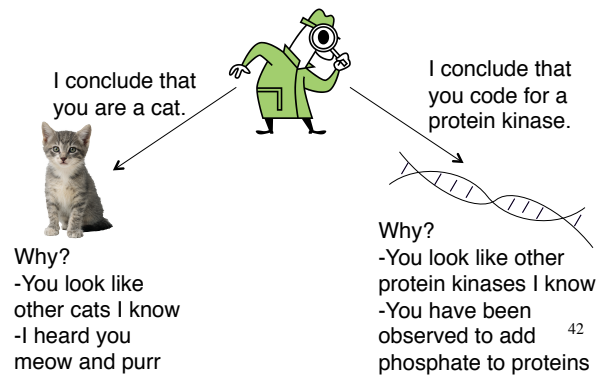


## Applying Search Information to Make Functional Annotations

41

### Annotation must be grounded with supporting evidence

- 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.
- Functional annotations should only be as specific as the supporting evidence allows
- All evidence that led to the annotation conclusions that were made must be stored.
- In addition, detailed documentation of methodologies and general rules or guidelines used in any annotation process should be provided.



42

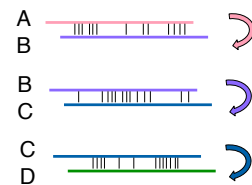
## Evaluate the Evidence

- **Visually inspect alignments**
  - they should be full length
    - or partial matches to identified domains
  - at least 40% identity
  - look for conservation of protein features or catalytic residues
  - avoid the pitfalls of transitive annotation (see next slide)
- **Check HMM scores**
  - need to be above trusted to be considered part of the family modeled by the HMM
  - how specific is the HMM?
  - what annotation on the HMM is appropriate for the query?
- **Look at any available metabolic analysis**
  - pathways, complexes?
- **Check for operon structure or other information from neighboring genes.**
  - presence of a gene in an operon can supplement weak similarity evidence
- **Are there transmembrane regions?**
- **Is there a signal peptide?**
- **Are there any motifs that might give a clue to function?**

43

## The Pitfalls of Transitive Annotation

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



- A's name has now passed to D from A through several intermediates.
  - This is fine if A is also similar to D
  - This is NOT fine if A is NOT similar to D, which can happen easily and happens often
- Current public datasets full of such errors
- To avoid transitive annotation errors we require that in a pairwise match, the match **must** be experimentally characterized
- Be conservative
  - err on the side of not making an annotation, when possibly you should, rather than making an annotation when probably you shouldn't.

44

## Experimentally Characterized Proteins

- It is important to know what proteins in our search database are characterized.
  - proteins marked as characterized from public databases
    - Gene Ontology repository (more on this later)
    - GenBank (only recently began)
  - UniProt proteins at “protein existence level 1”
- Annotators see this information in the search results in Manatee as color coded output:
  - **green** = full experimental characterization
  - **sky blue** = partial characterization
  - **red** = automated process (Swiss-Prot parse)
- This set does not contain all characterized proteins, not even close.

45

## Knowledge about function reflected in specificity of protein names

- **specific function**
  - Required evidence:
    - At least one good alignment (minimum 40% identity, over the full lengths of both proteins) to a protein from another organism that has been experimentally characterized, preferably multiple such alignments.
    - Hits to appropriate HMMs above the trusted cutoff.
    - Conservation of active sites, binding sites, appropriate number of membrane spans, etc.
  - Example: “adenylosuccinate lyase”, purB, 4.3.2.2
- **varying knowledge about substrate specificity**
  - A good example: ABC transporters
    - ribose ABC transporter
    - sugar ABC transporter
    - ABC transporter
  - choosing the name at the appropriate level of specificity requires careful evaluation of the evidence looking for specific characterized matches and HMMs.
- **family designation** - no gene symbol, partial EC
  - “Cbby family protein”
  - “carbohydrate kinase, FGGY family”
- **hypotheticals**
  - “hypothetical protein”
  - “conserved hypothetical protein”

46

## Functional Assignments: Frameshifts and Point Mutations

When a frameshift or in-frame stop codon is seen within a region of alignment in our BER search results, two possible things may have occurred: a sequencing error, or a mutation in the gene.










In the past when most genomes were closed (that is every base is known at high confidence) the underlying sequence was checked. Most were true mutations.

Proteins are flagged as having problems:  
frameshift  
point mutation (for in-frame stop codons)

We do NOT use the term “pseudogene” as this implies that some experiment was done to confirm the lack of function of the gene. We have done no such experiments and indeed it may be that the protein may be functional in some truncated form, or that some read-through mechanism allows expression. Also, it may be that mutations we see only occurred in the cells grown for the DNA sample used for the sequencing project.

47

## ORF disruptions

frameshift	
point mutation	
degenerate	
truncation	
deletion	
insertion	
interruption	
fusion	
fragment	

These would be classed as “pseudogenes” if lack-of-function is experimentally verified, otherwise we just tag these with labels

48

## Basic Set of Protein Annotations

- **protein name**
  - descriptive common name for the protein
    - e.g. “ribokinase”
- **gene symbol**
  - mnemonic abbreviation for the gene
  - e.g. “recA”
- **EC number**
  - only applicable to enzymes
    - e.g. 1.4.3.2
- **Role**
  - what the protein is doing in the cell and why
  - e.g. “amino acid biosynthesis”
- **Supporting Evidence**
  - accession numbers of BER and HMM matches
  - TmHMM, SignalP, LipoP
  - whatever information you used to make the annotation
- **Unique Identifier**
  - e.g. locus ids

49

## TIGR Roles

**TIGR bacterial roles were first adapted from Monica Riley’s roles for E. coli - both systems have since developed independently. TIGR roles will soon be phased out of the IGS annotation system.**

### Main Categories:

Amino acid biosynthesis  
 Purines, pyrimidines, nucleosides, and nucleotides  
 Fatty acid and phospholipid metabolism  
 Biosynthesis of cofactors, prosthetic groups, and carriers  
 Central intermediary metabolism  
 Energy metabolism  
 Transport and binding proteins  
 DNA metabolism  
 Transcription  
 Protein synthesis  
 Protein Fate  
 Regulatory Functions  
 Signal Transduction  
 Cell envelope  
 Cellular processes  
 Other categories  
 Unknown  
 Hypothetical  
 Disrupted Reading Frame  
 Unclassified (not a real role)

### Role Notes:

Notes written by annotators expert in each role category to aid other annotators in knowing what belongs in that category and what the proper name format is.

50

## Publicly Available Annotation Tools

- **Manatee**
  - open source, Perl CGI web interface
  - first developed at The Institute for Genomic Research (TIGR), now under continuing development at IGS (here) and J. Craig Venter Institute (JCVI)
  - primarily designed to facilitate functional annotation
  - manatee.sourceforge.net
- **Artemis**
  - developed at The Sanger Institute
  - excellent for viewing features along a DNA sequence, fewer features designed to optimize functional annotation
  - www.sanger.ac.uk/Software/Artemis
- **Apollo**
  - developed jointly by the Berkeley Drosophila Genome Project and The Sanger Institute
  - the annotation tool of the Generic Model Organism Database (GMOD) project
  - www.apollo.berkeleybop.org/current/index.html

51

## Data availability and Analysis

- **Integrated Microbial Genomes (IMG)**
  - img.jgi.doe.gov
- **HMP DACC**
  - www.hmpdacc.org
- **Comprehensive Microbial Resource (CMR)**
  - cmr.jcvi.org
- **Bioinformatics Resource Centers**
  - PATRIC (bacteria)
    - http://patricbrc.vbi.vt.edu/portal/portal/patric/home
  - ViPR (viruses)
    - http://www.viprbrc.org/bre/home.do?decorator=vipr
  - VectorBase (invertebrate vectors of disease)
    - http://www.vectorbase.org/index.php
  - EuPathDB (eukaryotic pathogens)
    - http://eupathdb.org/eupathdb/
  - BRC Portal (links the 4 BRC’s together)
    - http://pathogenportal.net/
- **EcoliHub**
  - www.ecolihub.org

52



## Free Annotation with the IGS Annotation Engine

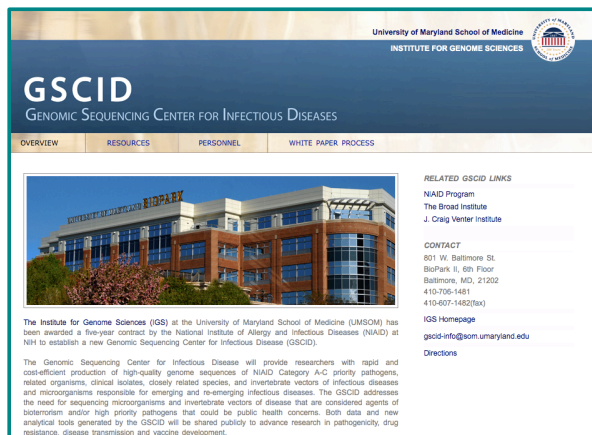
- a Free service for prokaryotic DNA sequence
- we run the sequence through our IGS prokaryotic pipeline
- We produce search data and automatically generated annotation
- Two choices for data access:
  - contained in MySQL database for download to user's local machine, users can also download Manatee
  - data continues to reside at IGS, user's have access via a password protected web interface, where users can access data and Manatee, no need to download or store anything
- We gratefully thank the



for funding this project.

53

## Free Sequencing with the IGS GSCID



- The IGS GSCID is one of three Genome Sequencing Centers funded by NIH to sequence organisms of relevance to Infectious Disease
  - IGS
  - Broad
  - JCVI
- Includes pathogens and disease vectors
- Anyone can submit a white paper for consideration
  - Info on the IGS GSCID can be found here: <http://gscid.igs.umaryland.edu/>
  - Info on the the application criteria and process can be found here: <http://www3.niaid.nih.gov/labsandresources/resources/gsc>

54