Prokaryotic Annotation Overview

Michelle Gwinn
August, 2005

Annotation

- **dictionary definition of “to annotate”:**
  - “to make or furnish critical or explanatory notes or comment”
- **some of what this includes for genomics**
  - gene product names
  - functional characteristics of gene products
  - physical characteristics of gene/protein/genome
  - overall metabolic profile of the organism
- **elements of the annotation process**
  - gene finding
  - homology searches
  - functional assignment
  - ORF management
  - data availability
- **manual vs. automatic**
  - computers do a fair job at preliminary annotation
  - high quality annotation requires manual review

Finding Real Genes

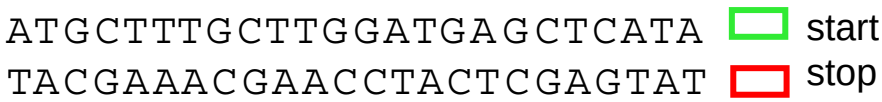
ORFs vs. Genes

- **ORF = open reading frame**
  - absence of translation “stop” codons (TAA, TAG, TGA)
    - an ORF goes from “stop” to “stop”
  - ORFs are found easily by one of many ORF finding tools
  - ORFs can easily occur by chance and since “stop” codons are AT rich:
    - GC rich DNA has, on average, more, longer ORFs
    - AT rich DNA has, on average, fewer, shorter ORFs

- **Gene**
  - requires translation “start” codon
    - bacterial starts = ATG, GTG, TTG
    - genes go from “start” to “stop”
  - has biological significance
    - catalytic or structural RNAs
    - protein coding regions

- **Telling the difference between random ORFs and genes is the goal in the gene finding process.**

A DNA sequence has 6 possible translation frames



Frame +1 codons = ATG CTT TGC TTG GAT GAG CTC ATA

M L C L D E L I

Frame +2 codons = TGC TTT GCT TGG ATG AGC TCA

M S S

Frame +3 codons = GCT TTG CTT GGA TGA GCT CAT

L L G *

Frame -1 codons = TAT GAG CTC ATC CAA GCA AAG CAT

Frame -2 codons = ATG AGC TCA TCC AAG CAA AGC

M S S S K Q S

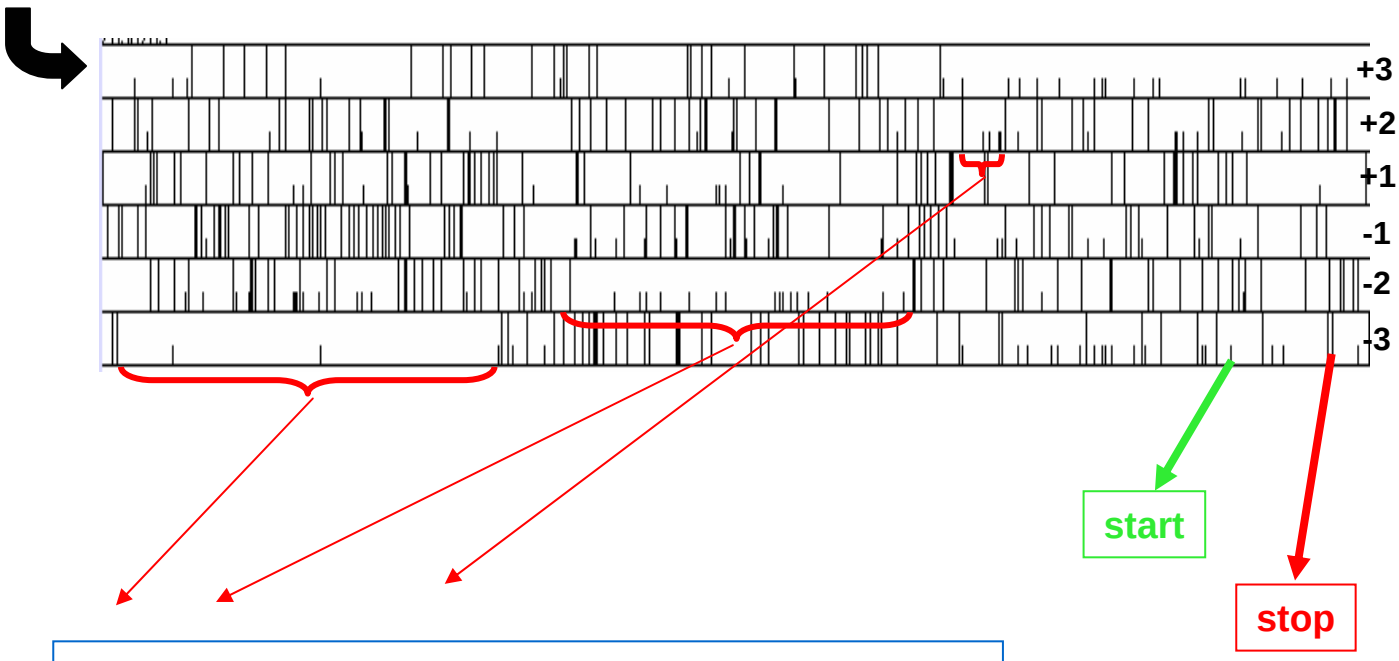
Frame -3 codons = TGA GCT CAT CCA AGC AAA GCA

*

6-Frame translations

```
>Shewanella oneidensis MR-1 1-2000
ACTAATAGATCTTAAGATCTTTATATAGATCTTCTTATTATTTTTACTATTAGGATCGCA
CTTTGCTGTGGGTAAGCCTATTTTCCCTTTTAGATCATTAGGTAACGTGTAATCAATCCT
GTGATCTACGCGAGATCTATTACCATAAAAGGTGGGGATAGATCCGCCTTTTATCCACAG
GGTGGATCTTTGAGCGGATCATGATGTGAATAGCACAGAGGTTGATCAGATCTAAATAAT
AGTTTATCCACAAAAATATACATAAAATTAGTTTTTTGTGGATAATTTGGATCTAAACTG
TGGGTTGTTTAAGGTATTTTCCAAAAGAAGATCTTAAATTTGATCTTGCCATTGTGCCAA
CCACGCTAAAGCAGGGTCTTCTGGCACAGGATCTTGTTGTACATCGATCTGGATCTTATC
CACAACGGCTTTTGCACCGCTATATTCAAGTGCTTCTATTAGCTTCTCTGGGCCTTGGCA
AAAGGTGTCATAACTTGAATCACCAATGGCACATAATGCGAATTTTACTTGGGTAAGATC
TGGCGTGTTGAGTAGAAGTTCTTTACAAAATGGCTGTAAATTATCGGGAAGATCACCTGC
GCCGTGGGTGGATGAAACAAGGATCCAGAGGGGATAGGGTTTTAGCTCATCTAGCGTTGG
ATGCAGGAAAGTATGTACCTCATGCCCTAACGGGGTTAGCTGCGCTTGCATTTCATCTGC
TATATATTCGCTGCTGCCGAGGGTGGTACCGACTAATATCGCAATCTTGGTCATTGTTAC
TGCTCTCTGTGGTCGTTAATTTTCTATCATTTTTGGGGCTTACATCTATTTACCGTGTTT
TGTTGTATAAATTTAGCCAATATTGATTCGCAGTCTGTTAGTATCTTTGGCTGCAATCTT
GGTTTTATACCTATAACCGTAAGTTTTTTATAACAATTATAATAGGATAACAGATGACTA
CACCTGTTGATGCGCCAAAATGGCCACGTCAAATCCCCTATATTATTGCCAGTGAAGCTT
GTGAGCGTTTCAGCTTTTACGGCATGCGCAATATTCTAACACCTTTCTTAATGACGGCAC
TTTTGCTGTCTATCCCTGAGGAGCTCCGCGGAGCAGTGGCAAAGGATGTGTTTCACTCCT
TTGTCATAGGTGTGTACTTTTTCCCATTACTGGGCGGTTGGATAGCTGATCGTTTTTTTG
GTAAATACAATACTATCTTATGGTTAAGTTTGATTTATTGTGTTGGCCATGCTTTCTTGG
CTATTTTTGAGCACAGTGTGCAAGGTTTTTATACTGGGTTATTTTTAATTGCTTTGGGTT
CTGGTGGGATAAAACCTTTAGTCTCATCCTTTATGGGCGATCAATTTGATCAGAGCAATA
AGTCCTTGGCTCAAAAGGCCTTTGATATGTTCTATTTTACGATCAACTTTGGTTCCTTCT
TCGCATCATTATCTATGCCACTGTTACTTAAGAACTTTGGTGCTGCCGTAGCTTTTGGGA
TCCCCGGTGTGCTGATGTTTGTCGCGACGGTATTTTTCTGGTTAGGTCGTAAGCGTTACA
TACATATGCCGCCAGAACCAAAAGATCCCCATGGTTTTTTACCGGTGATCCGCAGTGCAT
TGCTTACTAAGGTGGAAGGTAAGGGAAATATCGGTTTAGTGCTTGCACTCATCGGTGGCG
TTTCTGCGGCCTATGCGCTGGTTAATATTCCAACCCTTGGCATTGTCGCAGGCCTTTGTT
GTGCCATGGTGCTAGTGATGGGGTTTGTTGGAGCGGGCGCCTCATTACAACTTGAACGCG
CTAGAAAAAGTCATCCAGATGCTGCGGTAGATGGTGTACGTTCAGTTCTTCGAATTTTGG
TTTTATTTGCTTTGGTGACCCCATTTTGGTCTCTATTCGATCAAAAGGCTTCGACTTGGA
TTTTGCAGGCCAATGATATGGTTAAACCCCAGTGGTTTGAACCAGCGATGATGCAGGCAT
TAAACCCGCTGTTAGTAATG
```

In order to visualize the genes within the context of their neighbors along a DNA sequence, the sequence is often represented as a 6-frame translation. There are 6 possible frames for translation in every sequence of DNA, 3 in the forward (+) direction on the DNA sequence, and 3 in the reverse (-) direction on the DNA sequence. These are represented as horizontal bars with vertical marks for stops (long) and starts (short) in the order shown below.



These are some of the many ORFs in this graphic.

Gene Finding with Glimmer

- Glimmer is a tool which uses Interpolated Markov Models (IMMs) to predict which ORFs in a genome are real genes.

- Glimmer does this by comparing the nucleotide patterns of “known” real genes to the nucleotide patterns of the ORFs in the whole genome. ORFs with patterns similar to the real genes are considered real themselves.

- Using Glimmer is a two-part process:
  - Train Glimmer for the organism that was sequenced.
  - Run the trained Glimmer against the genome sequence.

Gene finding with Glimmer:
Gathering the Training Set

Gather published sequences from the organism sequenced

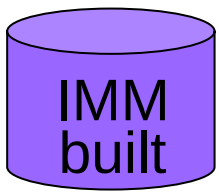
If you need more,
Find all ORFs

Two options to get additional training genes

long ORFs (500-1000 nucleotides depending on GC content of genome) that do not overlap each other

ORFs with a significant BLAST match to a protein from another organism (what we do at TIGR)

Training



computer algorithm in Glimmer

Gene Finding with Glimmer

What happens during training.

Glimmer moves sequentially through each sequence in the training set, recording the nucleotide that occurs after each possible oligomer up to oligomers of length eight

Example for a 5-mer:

ATGCGTAAGGCTTTCACAGTATGCGAGTAAGCTGCGTCGTAA
GG

ATGCGTAAGGCTTTCACAGTATGCGAGTAAGCTGCGTCGTAA
GG

ATGCGTAAGGCTTTCACAGTATGCGAGTAAGCTGCGTCGTAA
GG

ATGCGTAAGGCTTTCACAGTATGCGAGTAAGCTGCGTCGTAA
GG

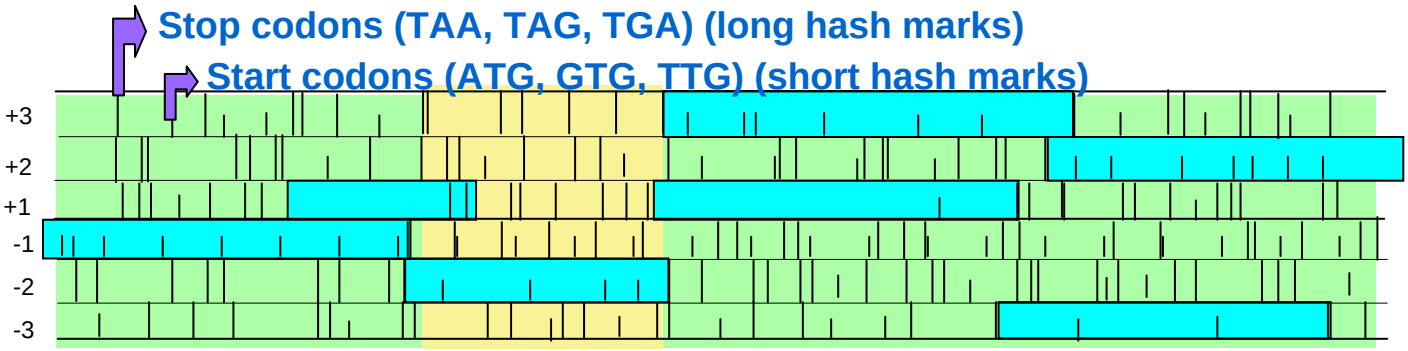
ATGCGTAAGGCTTTCACAGTATGCGAGTAAGCTGCGTCGTAA
GG

Glimmer then calculates the statistical probability of each pattern appearing in a real gene. These probabilities form the statistical model of what a real gene looks like in the given organism.

This model is then run against the complete genome sequence. All ORFs above the chosen minimum length (99 bp at TIGR) are scored according to how closely they match the model of a real gene.

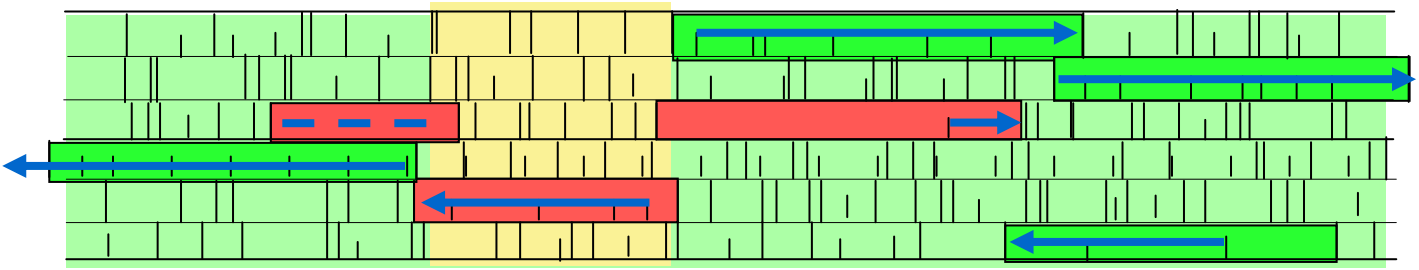
Candidate ORFs

6-frame translation map for a region of DNA

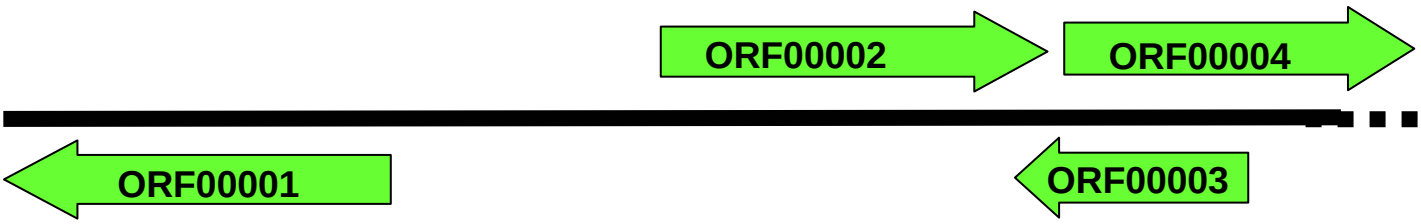


= ORFs meeting minimum length = laterally transferred DNA

The coding sequencse resulting from the candidate ORFs are represented by arrows, going from start to stop, the dotted line represents an ORF with no start site. which therefore can not be a gene. A long ORF does not necessarily result in a long putative gene.



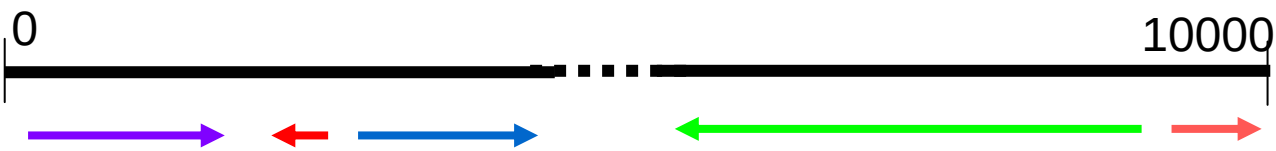
Green ORFs scored well to the model, red ORFs scored less well. The green ORFs are chosen by Glimmer as the set of likely genes and numbered sequentially from the beginning of the DNA molecule on which they reside. ORFs in the area of lateral transfer, although real genes, often will not be chosen since they don't match the model built from the patterns of the genome as a whole. Often when viewing a 6-frame translation, the genes are represented as arrows drawn above (or, as in this slide, below) the 6-frame translation.



Coordinates

Genes are mapped to the underlying genome sequence via coordinates. Each gene is defined by two coordinates: end5 (the 5 prime end of the gene) and end3 (the 3 prime end of the gene).

Nucleotide #1 for each molecule in the genome is the beginning of each final assembled molecule. Some genomes have just one DNA molecule, some have several (multiple chromosomes or plasmids).



gene	end5	end3
purple	12	527
red	802	675
blue	927	1543
green	9425	7894
pink	9575	9945

Note that for forward genes have end5<end3, while reverse genes have end5>end3.

Determining How The New Proteins Function

Finding the function of a new protein

- Experimental characterization
  - mutant phenotype
  - enzyme assay
  - difficult on a whole-genome scale
    - microarrays
      - expression patterns
    - large-scale mutant generation
      - done in yeast
- Homology searching
  - comparing sequences of unknown function to those of known function

Homology searching

- shared sequence implies shared function
  - binding sites
  - catalytic sites
  - full length match with significant identity between amino acids (>35% minimum)
- but beware
  - there are occurrences of proteins where one amino acid substitution changes the function of an enzyme
  - all functional assignments made by sequence similarity should be considered putative until experimental characterization confirms them
- identity vs. similarity
  - identity means amino acids match exactly
  - similarity means the amino acids share similar structure and thus could carry out the same or similar roles in the protein

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

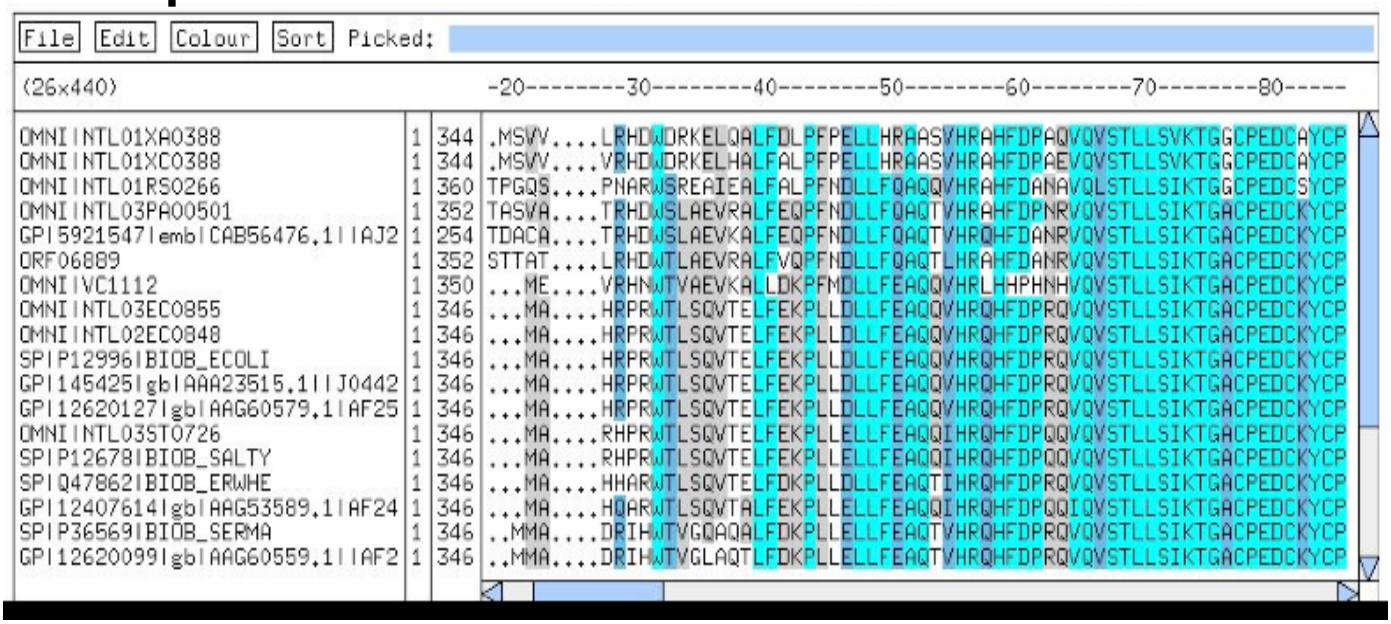
- **Multiple alignments** are more meaningful than pairwise alignments since it is much less likely that several proteins will share sequence similarity due to chance alone, than that 2 will share sequence similarity due to chance alone. Therefore, such shared similarity is more likely to be indicative of shared function.
  - HMMs
  - motifs

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
-dashed lines (colons) between amino acids indicate similarity

Multiple



Different shadings indicate amount of matching

Useful Databases Slide 1

- NCBI
  - National Center for Biotechnology Information
  - protein and DNA sequences
  - taxonomy resource
  - many other resources
- Omnium
  - database that underlies TIGR's CMR
  - contains data from all completed sequenced bacterial genomes
  - data is downloaded from the sequencing center
- Enzyme Commission
  - not sequence based
  - catagorized collection of enzymatic reactions
  - reactions have accession numbers indicating the type of reaction
    - Ex. 1.2.1.5
- KEGG, Metacyc, etc.

Useful Databases Slide 2

- Swiss-Prot
  - European Bioinformatics Institute (EBI) and Swiss Institute of Bioinformatics (SIB)
  - all entries manually curated
  - annotation includes
    - links to references
    - coordinates of protein features
    - links to cross-referenced databases
      - HMMs
      - Enzyme Commission
- TrEMBL
  - EBI and SIB
  - entries have not been manually curated
  - once they are accessions remain the same but move into Swiss-Prot
- PIR (at Georgetown University)
- UniProt
  - Swiss-Prot + TrEMBL + PIR

NIAA

- Non-Identical Amino Acid
- TIGR's protein file used for searching
- File composed of protein sequences from several source databases
  - Swiss-Prot
  - Omnium
  - NCBI
  - PIR
- The file is made non-redundant
  - identical protein sequences from the same gene in the same organism that came into the file from different source databases are collapsed into one entry
  - all of the protein's accession numbers from the various source databases where it is found are stored linked to the protein
  - users can always view the protein at the source database

NIAA entry

>biotin synthase, Escherichia coli
MAHRPRWTLSQVTELFEKPLLDLLFEAQQVHRQHFDPRQVQVSTLLSIKTGACPED
PQSSRYKTGLEAERLMEVEQVLESARKAKAAGSTRFCMGAAWKNPHERDMPYLE
KAMGLEACMTLGTLSESQAQRLANAGLDYYNHNLDTSPEFYGNIITTRTYQERLDT
RDAGIKVCSGGIVGLGETVKDRAGLLLQLANLPTPPESVPINMLVKVKGTPLADNDD
FDFIRTIAVARIMMPTSYVRLSAGREQMNEQTQAMCFMAGANSIFYGCKLLTTPNPE
DLQLFRKLGLNPQQTAVLAGDNEQQQRLEQALMTPDTDEYYNAAAL

Source databases where this protein is found:

-Swiss-Prot, accession # SP:P12996
-Protein Information Resource, accession # PIR:JC2517
-NCBI's GenBank, accession # GB:AAC73862.1

All of these are collapsed into one entry in NIAA that is linked to all three accessions.

Experimentally Characterized Proteins

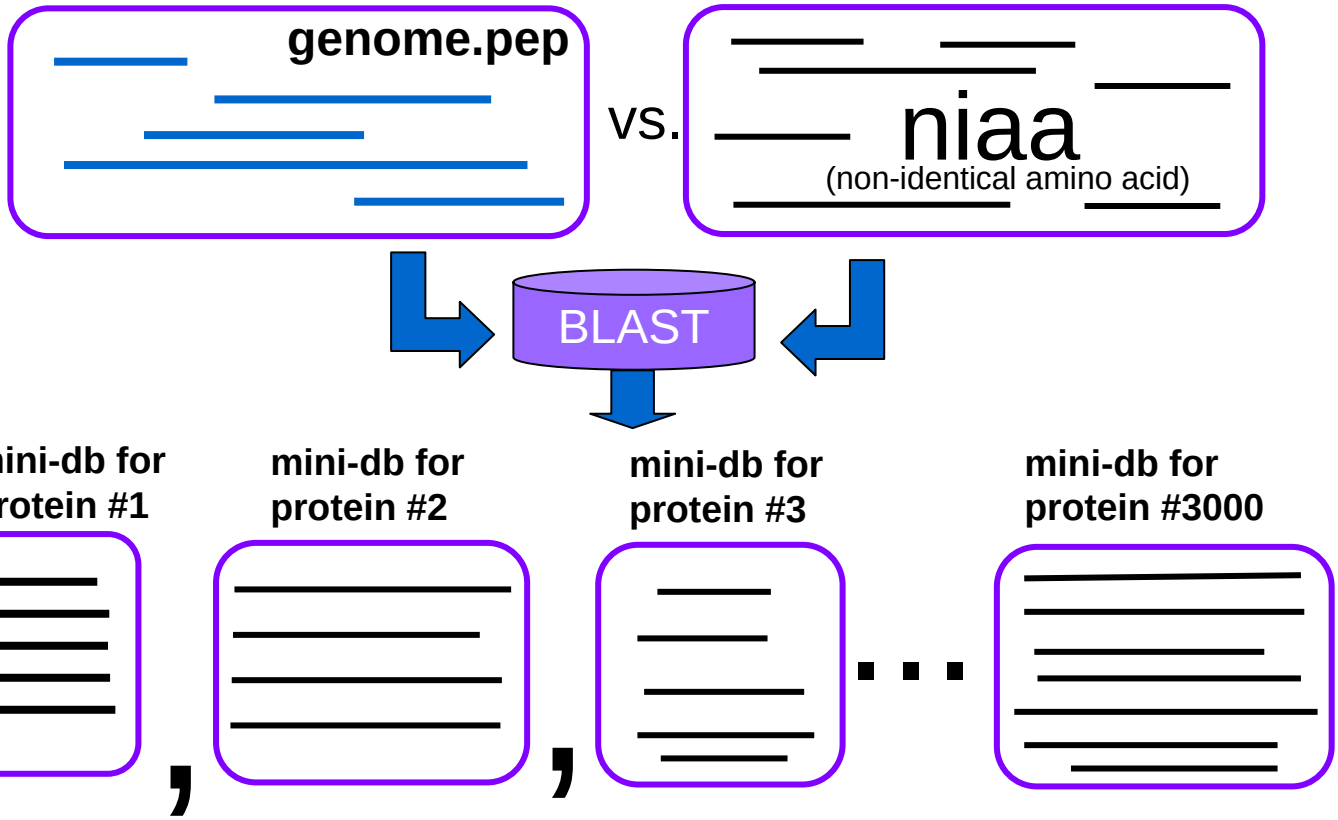
- It is important to know what proteins in our search database are characterized.
  - We store the accessions of proteins known or suspected to be characterized in the "characterized table" in our database
  - A confidence status is assigned to each entry in the characterized table.
- Annotators see this information in the search results as color coded output:
  - **green** = full experimental characterization
  - **red** = automated process (Swiss-Prot parse)
  - **sky blue** = partial characterization
  - **olive** = trusted, used when multiple extremely good lines of evidence exist for function but no experiment has been done (rarely used)
  - **blue-green** = fragment/domain has been characterized
  - **fuzzy gray** = void, used to indicate that something that was originally thought to be characterized really is not (rarely used)
  - **gray** = accession exists in the omnium only - therefore represents automated annotation
- Our table does not contain all characterized proteins, not even close.
  - Any time additional characterized proteins are found it is important that they be entered into the table

BLAST-extend-repraze (BER)

TIGR's pairwise protein search tool

- Initial BLAST search
  - against NIAA
  - finds local alignments
  - stores good hits in mini-database for each protein
- Protein sequence is extended by 300 nucleotides on each end and translated (see later slide)
- A modified Smith-Waterman alignment is generated with each sequence in the mini-database
  - extends local alignments as far as homology continues over lengths of extended proteins
  - produces a file of alignments between the query protein and the match protein for each match protein in the mini-database
  - as the alignment generating algorithm builds the alignment, if the level of similarity falls below the necessary threshold, the program looks in different frames and through stop codons for homology to continue - this similarity can continue into the upstream and/or downstream extensions

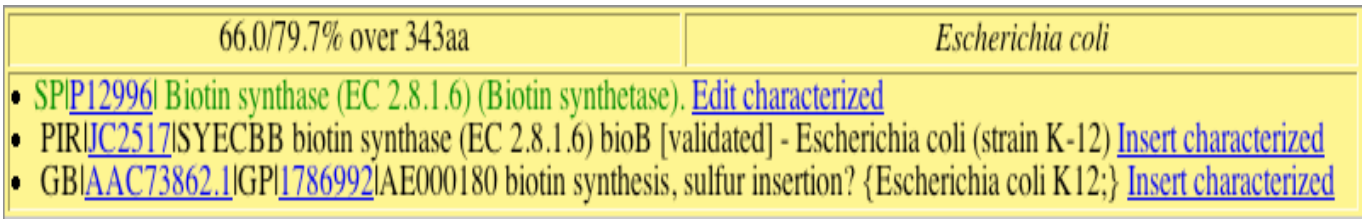
BLAST-extend-repraze (BER)



BER Alignment

An alignment like this will be generated for every match protein in the mini-database. In the next slides we will look closely at the types of information displayed here.

BER Alignment detail: Boxed Header



-The background color of this box will be gold if the protein is in the characterized table and grey if it is not.

-The top bar lists the percent identity/similarity and the organism from which the protein comes (if available).

-The bottom section lists all of the accession numbers and names for all the instances of the match protein from the source databases (used in building NIAA for the searches.)

-The accession numbers are links to pages for the match protein in the source databases.

-A particular entry in the list will have colored text (the color corresponding to its characterized status) if that is the accession that is entered into the characterized table - this tells the annotators which link they should follow to find experimental characterization information. Only one accession for the match protein need be in the characterized table for the header to turn gold.

-There are links at the end of each line to enter the accession into the characterized table or to edit an already existing entry in the characterized table.

BER Alignment detail: alignment header

```
ORF04813( 7 - 350 of 350 aa)
SP|P12996|BIOB_ECOLI(4 - 346 of 346) Biotin synthase (EC 2.8.1.6)
%Match = 42.3
%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]
```

-It is most important to look at the range over which the alignment stretches and the percent identity

-The top line show the amino acid coordinates over which the match extends for our protein

-The second line shows the amino acid coordinates over which the match extends for the match protein, along with the name and accession of the match protein

-The 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. The primary frame is the one starting with nucleotide one of the gene. If all is well with the protein, all of the matching amino acids should be in frame 1.

-If there is a frameshift in the alignment (see later slide) the phrase “Frame Shifts = #” will flash and indicate how many frameshifts there are.

BER Alignment detail: alignment of amino acids

-In these alignments the codons of the DNA sequence read down in columns with the corresponding amino acid underneath.

-The numbers refer to amino acid position. Position 1 is the first amino acid of the protein. The first nucleotide of the codon coding for amino acid 1 is nucleotide 1 of the coding sequence. Negative amino acid numbers indicate positions upstream of the predicted start of the protein.

-Vertical lines between amino acids of our protein and the match protein (bottom line) indicate exact matches, dotted lines (colons) indicate similar amino acids.

-Start sites are color coded: ATG is green, GTG is blue, TTG is red/orange

-Stop codons are represented as asterisks in the amino acid sequence. An open reading frame goes from an upstream stop codon to the stop at the end of the protein, while the gene starts at the chosen start codon.

BER Skim

A list of best matches from niaa to the search protein with statistics on length of match and BLAST p-value. Colored backgrounds indicate presence in characterized table and corresponding status.

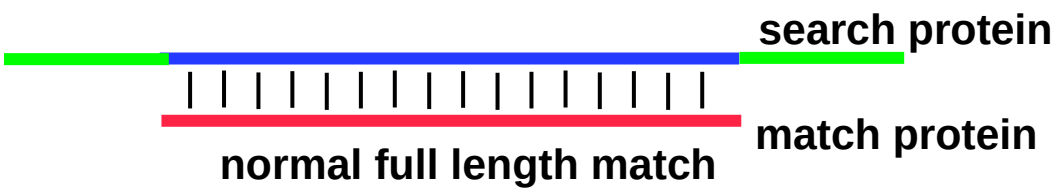
BER SKIM submit

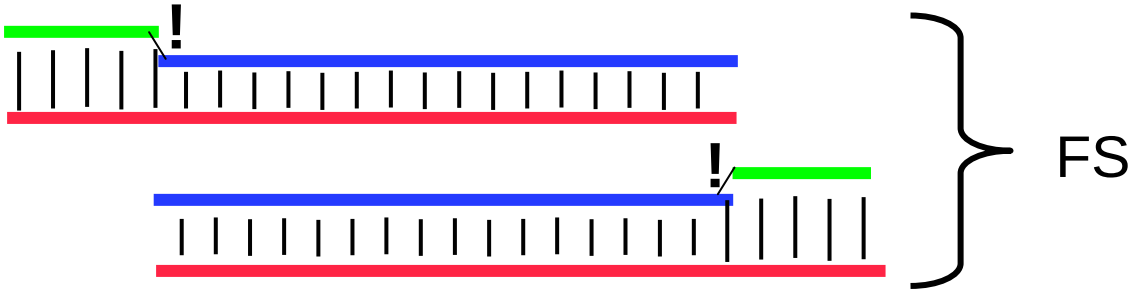
Belvu | View BER Searches | search date: Wed Oct 23 12:59:20 2002 | Refresh Searches

accession	%sim	length	description	p-value
OMNI:SO2740	100.0	349	biotin synthase {Shewanella oneidensis MR-1}	1.5e-176
SP:P36569	80.7	340	Biotin synthase (EC 2.8.1.6) (Biotin synthetase). {Serratia	2.5e-119
SP:P12996	79.7	342	Biotin synthase (EC 2.8.1.6) (Biotin synthetase). {Escherich	7.2e-120
GP:145425	79.7	342	biotin synthetase {Escherichia coli}	1.5e-119
GP:12620127	79.4	342	biotin synthase BioB {uncultured bacterium pCosHE2}	1.5e-119
OMNI:NTL03EC0855	79.4	342	biotin synthetase {Escherichia coli O157:H7 VT2-Sakai}□GPI13	5.1e-119
OMNI:NTL01YP1094	81.0	340	biotin synthase {Yersinia pestis CO92}□OMNI|NTL02YP2986 biot	8.3e-119
GP:12620099	79.5	340	BioB-like protein {uncultured bacterium pCosFS1}	9.5e-118
OMNI:NTL02EC0848	79.1	342	biotin synthesis, sulfur insertion? {Escherichia coli O157:H	2.2e-118
SP:Q47862	79.2	339	Biotin synthase (EC 2.8.1.6) (Biotin synthetase). {Erwinia h	3.6e-118
SP:P12678	78.6	344	Biotin synthase (EC 2.8.1.6) (Biotin synthetase). {Salmonell	5.1e-119
OMNI:VC1112	81.8	348	biotin synthase {Vibrio cholerae El Tor N16961}□GPI9655583|g	5.1e-119
OMNI:NTL03ST0726	78.6	344	biotin synthetase {Salmonella enterica serovar Typhi CT18}□G	1.1e-118
OMNI:NTL03PA00501	78.9	348	biotin synthase {Pseudomonas aeruginosa PAO1}□GPI9946364|gbl	7.7e-116
GP:12407614	76.8	339	biotin synthase BioB {uncultured bacterium pCosAS1}	9.1e-113
OMNI:NTL01XC0388	79.2	311	biotin synthase {Xanthomonas campestris pv. campestris ATCC3	2.8e-111
OMNI:NTL01XA0388	78.5	311	biotin synthase {Xanthomonas axonopodis pv. citri 306}□GPI21	6.6e-110
OMNI:NTL02BA0265	77.0	340	biotin synthase {Buchnera aphidicola Sg}□GPI21623185|gb|AAM6	1.4e-109
OMNI:NTL01XF00065	79.4	309	biotin synthase {Xylella fastidiosa 9a5c}□GPI9104834|gb|AAF8	8.4e-110
OMNI:NTL01RS0266	79.5	306	PROBABLE BIOTIN SYNTHASE PROTEIN {Ralstonia solanacearum GMI	4.7e-109
SP:P57378	77.3	342	Biotin synthase (EC 2.8.1.6) (Biotin synthetase). [Buchnera	1.1e-107
GP:15419053	79.1	328	biotin synthase {Acinetobacter calcoaceticus}	1.6e-106
OMNI:CC3521	76.2	339	biotin synthase {Caulobacter crescentus CB15}□GPI13425251|gb	3.0e-105
OMNI:NTL01BMA0776	79.8	311	BIOTIN SYNTHASE {Brucella melitensis 16M}□GPI17984969|gb|AAL	6.3e-105

Extensions in BER

The extensions help in the detection of frameshifts (FS) and point mutations resulting in in-frame stop codons (PM). This is indicated when similarity extends outside the coordinates of the protein coding sequence



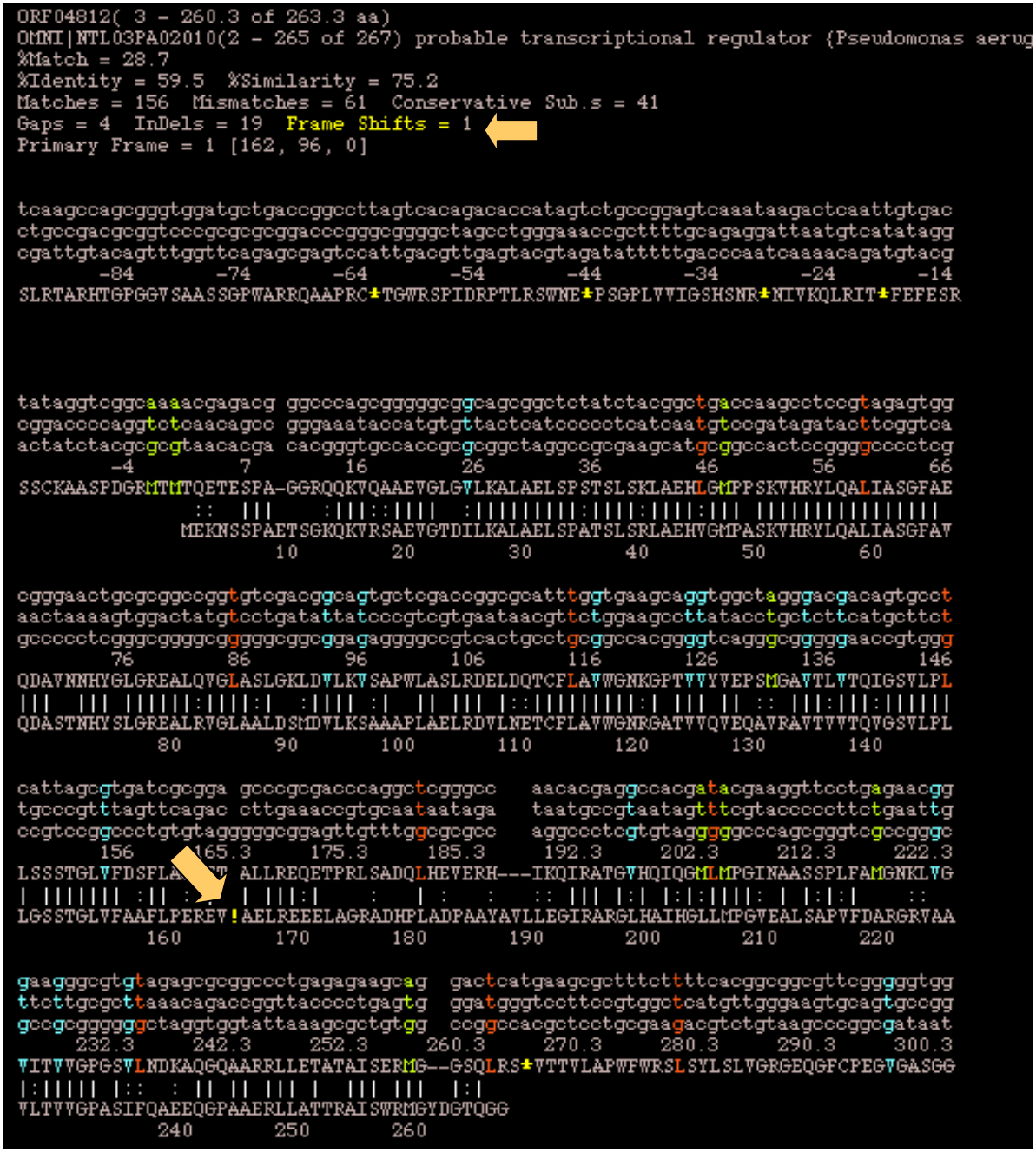


similarity extending through and frameshift upstream or downstream into extensions

similarity extending in the same frame through a stop codon

two functionally unrelated genes from other species matching one of our proteins could indicate incorrectly fused ORFs

Frameshifted alignment



Hidden Markov Models - HMMs

- HMMs are statistical models of the patterns of amino acids in a group of functionally related proteins found across species.
  - this group is called the “seed”
  - HMMs are built from multiple alignments of the seed members.
- Proteins searched against an HMM receive a score indicating how well they match the model.
  - Proteins scoring well to the model can be expected to share the function that the HMM represents.
- HMMs can be built at varying levels of functional relationship.
  - The most powerful level of relationship is one representing the exact same function.
  - It is important to know the kind of relationship an HMM models to be able to draw the correct conclusions from it
- Annotation can be attached to HMMs
  - protein name
  - gene symbol
  - EC number
  - role information

TIGR's HMM Isology Types

Equivalog: The supreme HMM, designed so that all members and all proteins scoring above trusted share the same function.

Superfamily: This type of HMM describes a group of proteins which have full length protein sequence similarity and have the same domain architecture, but which do not necessarily have the same function.

Subfamily: This type of HMM describes a group of proteins which also have full length homology, which represent more specific sub-groupings with a superfamily.

Domain: These HMMs describe a region of homology that is not required to be the full length of a protein. The function of the region may or may not be known.

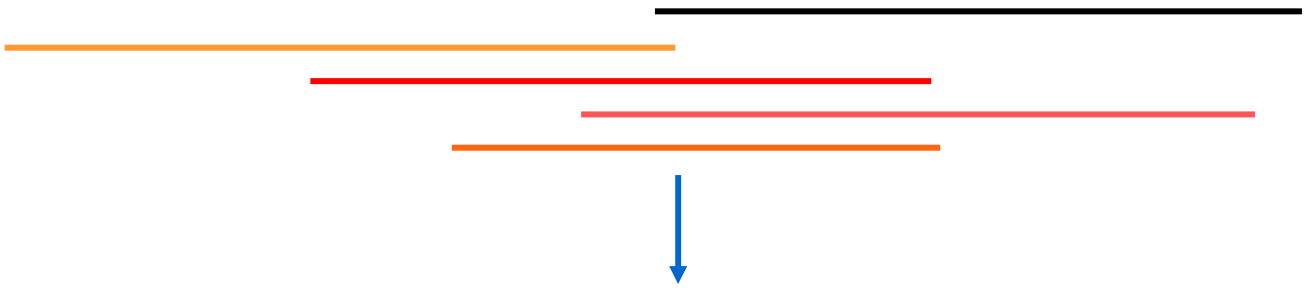
Equivalog_domain: Describes a protein region with a conserved function. It can be found as a single function protein or part of a multifunctional protein.

Hypothetical_equivalog: These are built in the same way as equivalogs except they are made from only conserved hypothetical proteins. Therefore, although the function is not known, it is believed that all proteins that score well to the HMM share the same function.

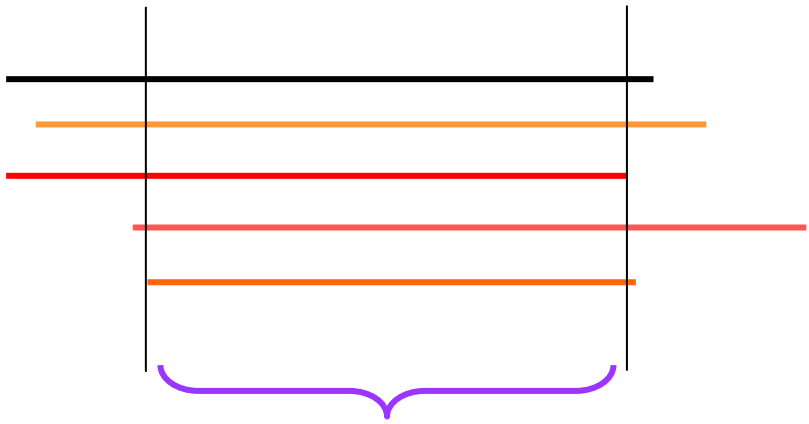
Pfam: Indicates that no TIGR isology type has yet been assigned to the Pfam HMM.

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. HMM is ready to go!

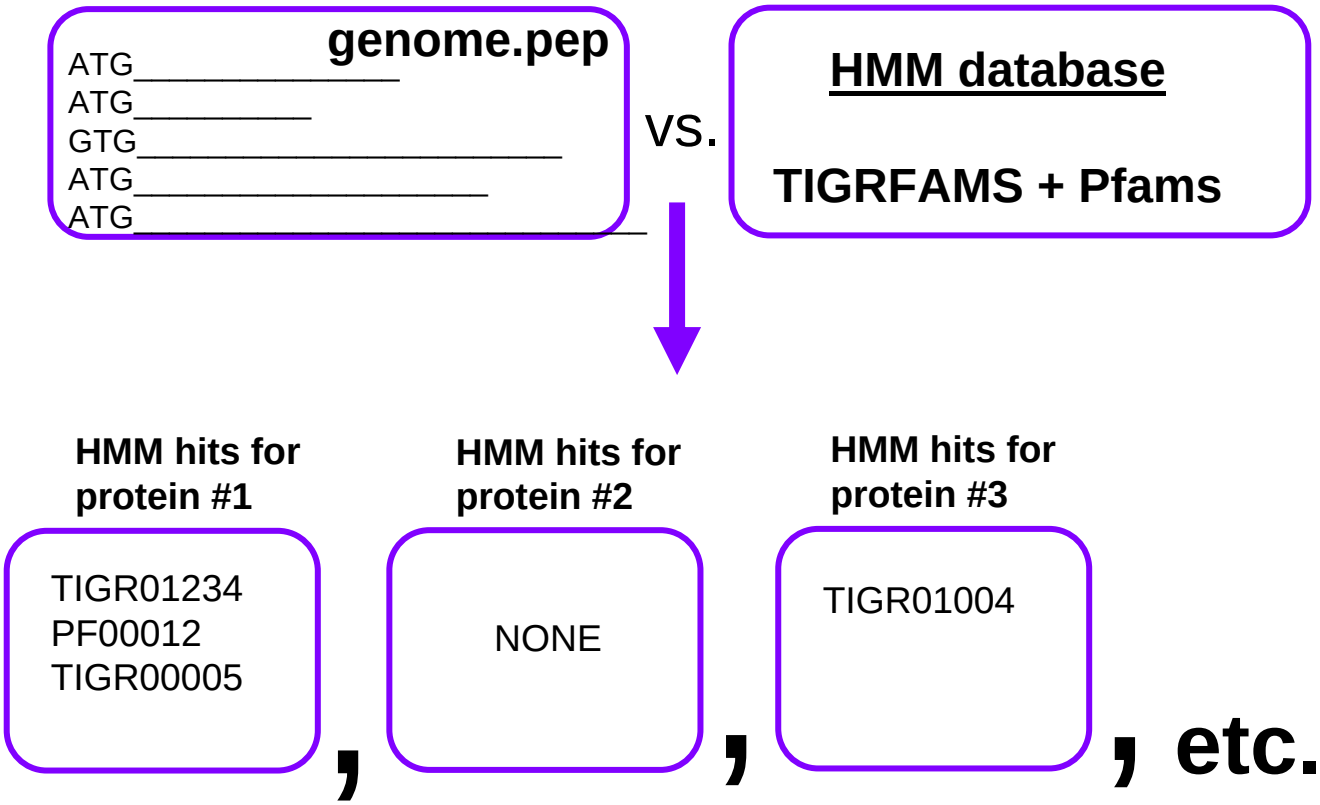
Choosing cutoff scores

=search the new HMM against NIAA
=see the range of scores the match proteins receive
=do analysis to determine where known members score
=do analysis to determine where known non-members score
=set the cut-offs accordingly

matches (seed members bold)	score	
protein "definitely"	**547**	
protein "absolutely"	**501**	
protein "sure thing"	487	
protein "confident"	**398**	
protein "safe bet"	376	
protein "very confident"	**365**	
protein "has to be one"	**355**	250
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

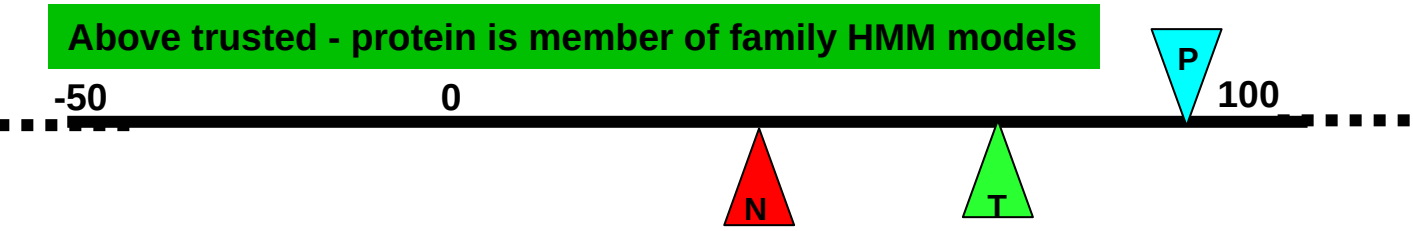
HMM Searches



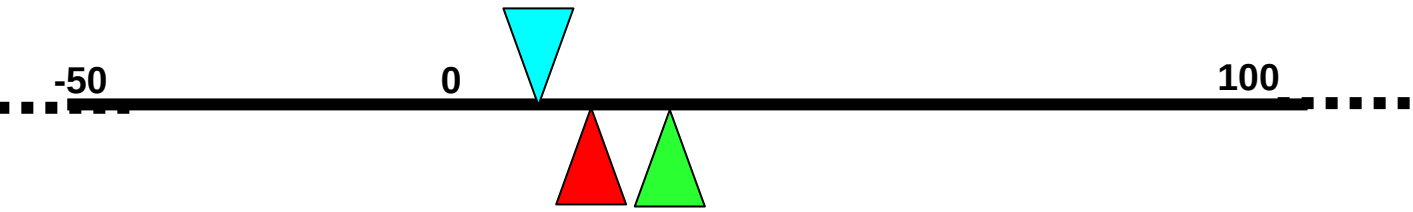
Each protein in the genome is searched against all HMMs in our db. Some will not have significant hits to any HMM, some will have significant hits to several HMMs. Multiple HMM hits can arise in many ways, for example: the same protein could hit an equivalog model, a superfamily model to which the equivalog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

Evaluating HMM scores

Above trusted - protein is member of family HMM models



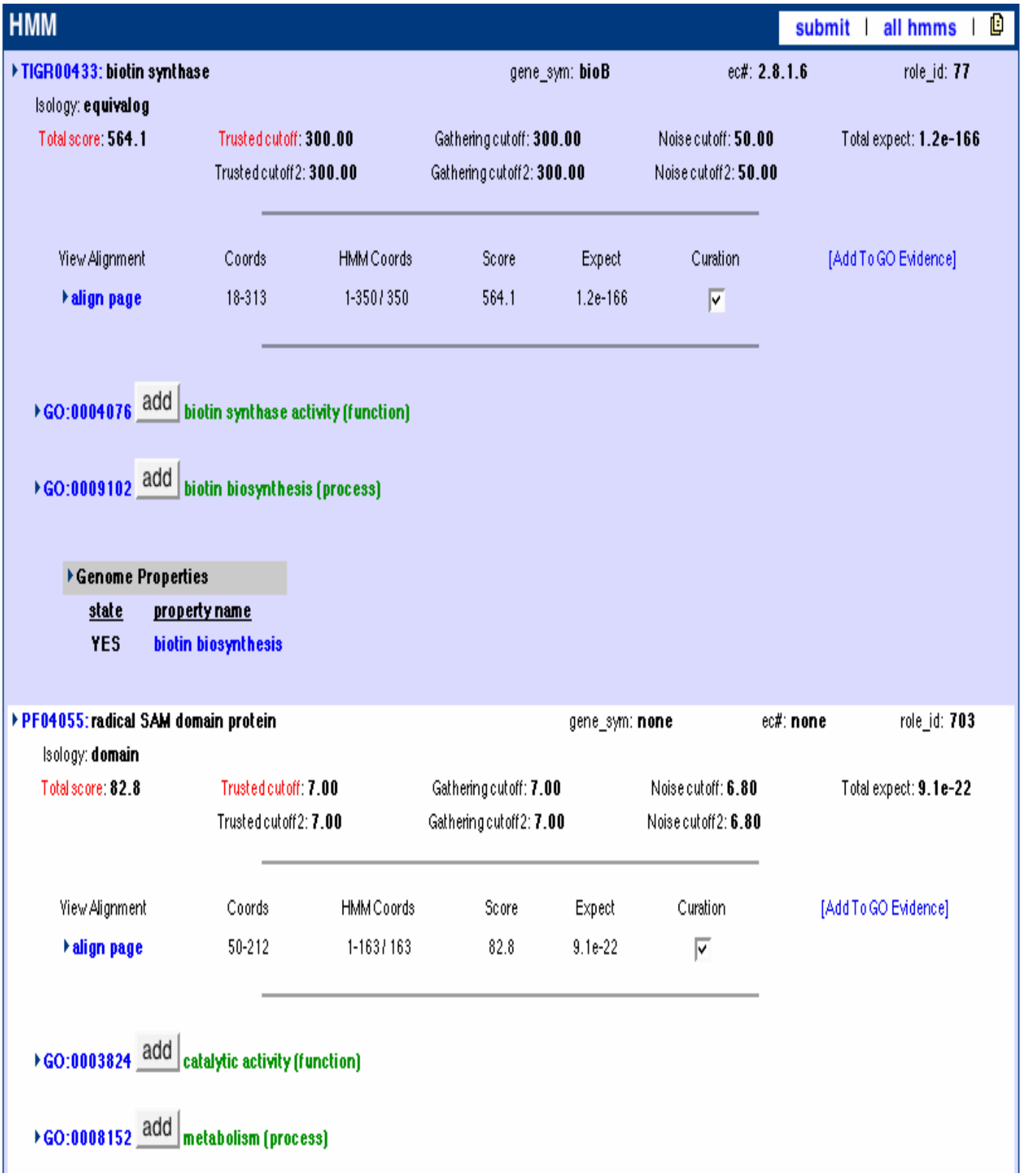
Below noise - protein is not member of family HMM models



In-between noise and trusted - protein may be in family HMM models

Above trusted with negative scores - protein is in family HMM models

HMM Output in Manatee



HMM

submit | all hmms |

TIGR00433: biotin synthase gene_sym: bioB ec#: 2.8.1.6 role_id: 77

Isology: equivalog

Total score: 564.1 Trusted cutoff: 300.00 Gathering cutoff: 300.00 Noise cutoff: 50.00 Total expect: 1.2e-166

Trusted cutoff2: 300.00 Gathering cutoff2: 300.00 Noise cutoff2: 50.00

View Alignment	Coords	HMM Coords	Score	Expect	Curation	[Add To GO Evidence]
align page	18-313	1-350 / 350	564.1	1.2e-166	☑	

GO:0004076 add biotin synthase activity (function)

GO:0009102 add biotin biosynthesis (process)

Genome Properties

state	property name
YES	biotin biosynthesis

PF04055: radical SAM domain protein gene_sym: none ec#: none role_id: 703

Isology: domain

Total score: 82.8 Trusted cutoff: 7.00 Gathering cutoff: 7.00 Noise cutoff: 6.80 Total expect: 9.1e-22

Trusted cutoff2: 7.00 Gathering cutoff2: 7.00 Noise cutoff2: 6.80

View Alignment	Coords	HMM Coords	Score	Expect	Curation	[Add To GO Evidence]
align page	50-212	1-163 / 163	82.8	9.1e-22	☑	

GO:0003824 add catalytic activity (function)

GO:0008152 add metabolism (process)

Genome Properties

- Used to get “the big picture” of an organism. Specifically to record and/or predict the presence/absence of:
  - metabolic pathways
    - biotin biosynthesis
  - cellular structures
    - outer membrane
  - traits
    - anaerobic vs. aerobic
    - optimal growth temperature
- Particular property has a given “state” in each organism, for example:
  - YES - the property is definitely present
  - NO - the property is definitely not present
  - Some evidence - the property may be present and more investigation is required to make a determination
- The state of some properties can be determined computationally
  - metabolic pathway
    - the property is defined be several reaction steps which are modeled by HMMs
    - HMM matches to steps in pathway indicate that the organism has the property
- Other property’s states must be entered manually (growth temp, anaerobic/aerobic, etc.)
- data for a particular genome viewable in Manatee
  - links from HMM section
  - links from gene list for role category
  - entire list of properties and states can be viewed
- Searchable across genomes on the CMR site
  - covered in the CMR segment of the course

“Biotin Biosynthesis” Genome Property

biotin biosynthesis (GenProp0036, PATHWAY)

Description

Biotin is an essential cofactor for many carboxylation (addition of C02) reactions. This property reflects biosynthesis from pimeloyl-CoA. The source of pimeloyl-CoA may vary. BioF (EC 2.3.1.47, 8-amino-7-oxononanoate synthase, also called 7-keto-8-aminopelargonic acid synthetase) converts pimeloyl-CoA to 8-amino-7-oxononanoate. BioA (EC 2.6.1.62, adenosylmethionine-8-amino-7-oxononanoate aminotransferase) converts the product to 7,8-diaminononanoate, from which BioD (EC 6.3.3.3, dethiobiotin synthase) makes dethiobiotin. BioB (EC 2.8.1.6, biotin synthase) then makes biotin itself. Enzymes such as BioH involved in pimeloyl-CoA biosynthesis typically receive biotin-related annotations but may also appear in genomes in which biotin is not synthesized and pimeloyl-CoA is used for something else.

Literature References

No References Found

Web References

KEGG: Biotin Metabolism

Associated Gene Ontology (GO) terms

regulation of transcription, DNA-dependent	process	GO:0006355
biotin biosynthesis	process	GO:0009102

Components and evidence

8-amino-7-oxo-nonanoate synthase(2)

Required	Branch	Evidence	
YES	1	HMM: TIGR00858	8-amino-7-oxononanoate synthase

adenosyl methionine 8-amino-7-oxononanoate transaminase(3)

Required	Branch	Evidence	
YES	1	HMM: TIGR00508	adenosylmethionine-8-amino-7-oxononanoate aminotransferase

dethiobiotin synthase(4)

Required	Branch	Evidence	
YES	1	HMM: TIGR00347	dethiobiotin synthase

biotin synthase(5)

Required	Branch	Evidence	
YES	1	HMM: TIGR00433	biotin synthase

BioC(bioC)

Required	Branch	Evidence	
NO	1	HMM: TIGR02072	biotin biosynthesis protein BioC

bioH protein(bioH)

Required	Branch	Evidence	
NO	1	HMM: TIGR01738	bioH protein

biotin repressor(represso)

Required	Branch	Evidence	
NO	1	HMM: TIGR00122	biotin operon repressor

“Cell Shape” Genome Property

cell shape
(GenProp0173, PHENOTYPIC)

Description

This property holds descriptions of the cellular shape of unicellular organisms, typical values are ROD-SHAPED, COCCI and SPIRAL.

Literature References

[1] David R. Boone, Richard W. Castenholz, editors *Bergeys manual of systematic bacteriology* New York : Springer, 2001 PMID:

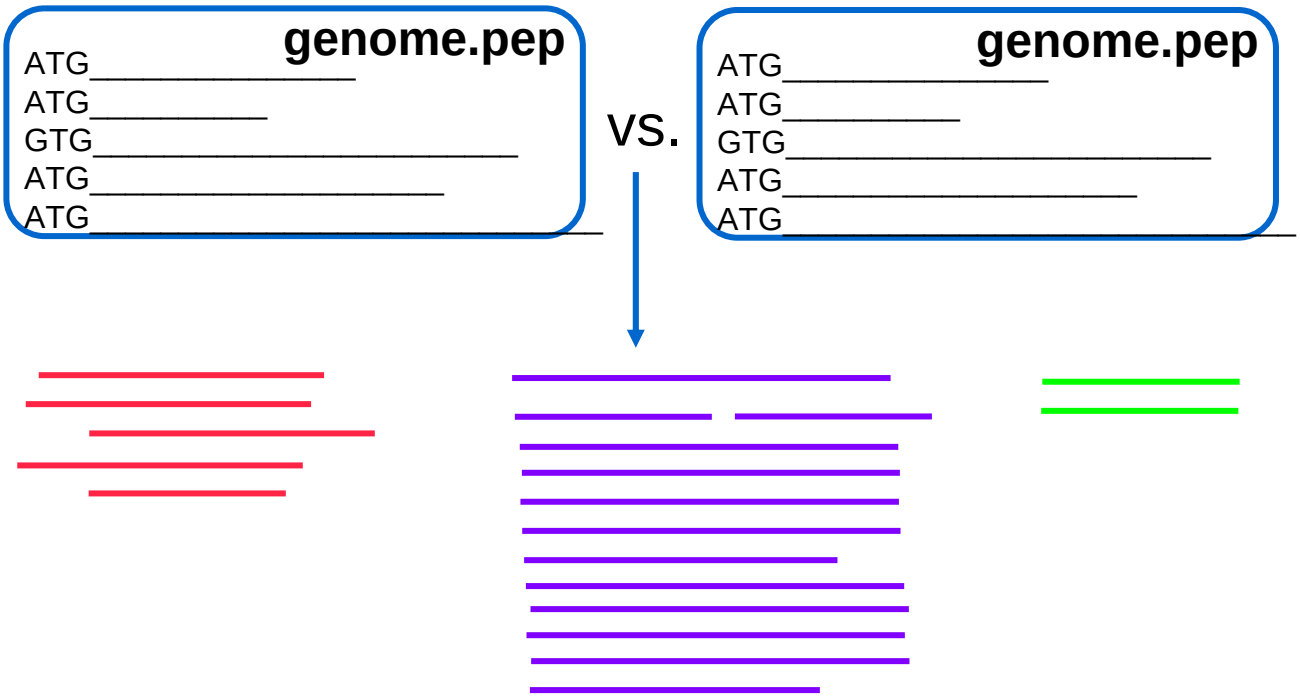
Web References

No References found

Associated Gene Ontology (GO) terms

No GO terms found

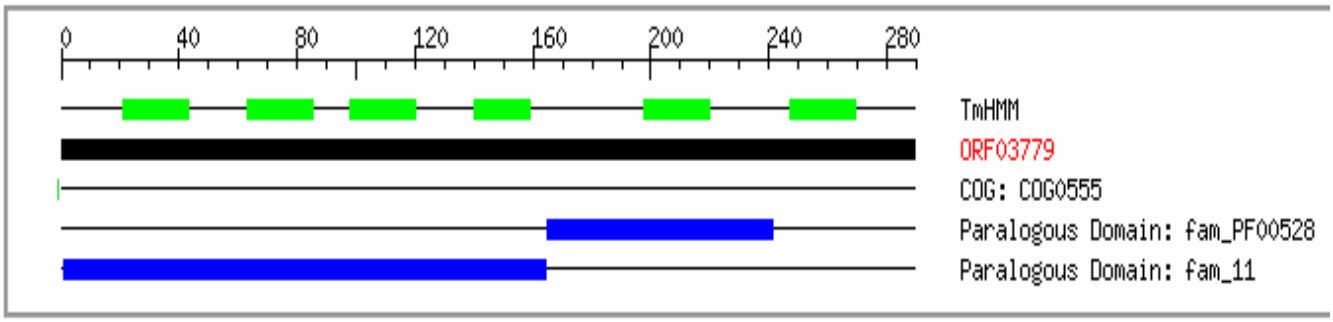
Paralogous Families



Groups proteins from within the same genome into families (minimum two members) according to sequence similarity. First, proteins are clustered according to HMM hits, second, other regions of the proteins, not found in HMM hits, are searched and clustered.

- Reveals expansion/contraction of various families of proteins in one genome verses another.
- Helps in annotation consistency, frameshift detection, and start site editing.

Paralogous family output in Manatee



A	C	gene id	gene length	gene name	role id	11	PF00528	other fams
		ORF00271	343 aa	peptide ABC transporter, permease protein	142		+	
		ORF00272	296 aa	peptide ABC transporter, permease protein	142		+	
		ORF00367	738 aa	phosphate ABC transporter, permease protein, putative	143		+	
		ORF00902	271 aa	polyamine ABC transporter, permease protein	142		+	
		ORF00903	301 aa	polyamine ABC transporter, permease protein	142		+	
		ORF01167	226 aa	amino acid ABC transporter, permease protein	142		+	
		ORF01529	544 aa	iron(III) ABC transporter, permease protein	145		+	
		ORF02439	235 aa	ABC transporter, permease protein	141		+	
		ORF00364	552 aa	phosphate ABC transporter, permease protein, putative	143		+	38
		ORF02958	290 aa	phosphate ABC transporter, permease protein	143		+	38
		ORF02959	278 aa	phosphate ABC transporter, permease protein	143		+	38
		ORF02518	283 aa	sulfate ABC transporter, permease protein	143	+	+	
		ORF02519	293 aa	sulfate ABC transporter, permease protein	143	+	+	
		ORF02772	226 aa	molybdenum ABC transporter, permease protein	143	+	+	
		ORF03459	245 aa	molybdenum ABC transporter, permease protein	143	+	+	
		ORF03779	289 aa	sulfate ABC transporter, permease protein	143	+	+	
		ORF03783	281 aa	sulfate ABC transporter, permease protein	143	+	+	

Other searches

- PROSITE Motifs
  - collection of protein motifs associated with active sites, binding sites, etc.
  - help in classifying genes into functional families when HMMs for that family have not been built
- InterPro
  - Brings together HMMs (both TIGR and Pfam) Prosite motifs and other forms of motif/domain clustering (Prints, Smart)
  - Useful annotation information
  - GO terms have been assigned to many of these
- TmHMM
  - an HMM that recognizes membrane spans
  - a product of the Center for Biological Sequence Analysis (CBS), Denmark
- Signal P
  - potential secreted proteins
  - another CBS product
- Lipoprotein
  - potential lipoproteins
  - this is actually a specific Prosite motif

Other Searches/Information

- Molecular Weight/pI
- DNA repeats
- RNAs
  - tRNAs are found using tRNAscan (Sean Eddy)
  - structural RNAs are found using BLAST searches
  - We are starting to implement Rfam, a set of HMMs modeling non-coding RNAs (Sanger, WashU)
- GC content
  - for the whole genome and individual genes
- terminators
- operons

Making the annotations:
Assigning names and roles to the proteins

Functional Assignments:

What we want to accomplish.

Name and associated info

Descriptive common name for the protein, with as much specificity as the evidence supports; gene symbol. EC number if protein is an enzyme

Role

Both TIGR and Gene Ontology, to describe what the protein is doing in the cell and why.

Supporting evidence:

HMMs, Prosite, InterPro
Characterized match from BER search
Paralogous family membership.

Functional Assignments:
What we want to avoid!

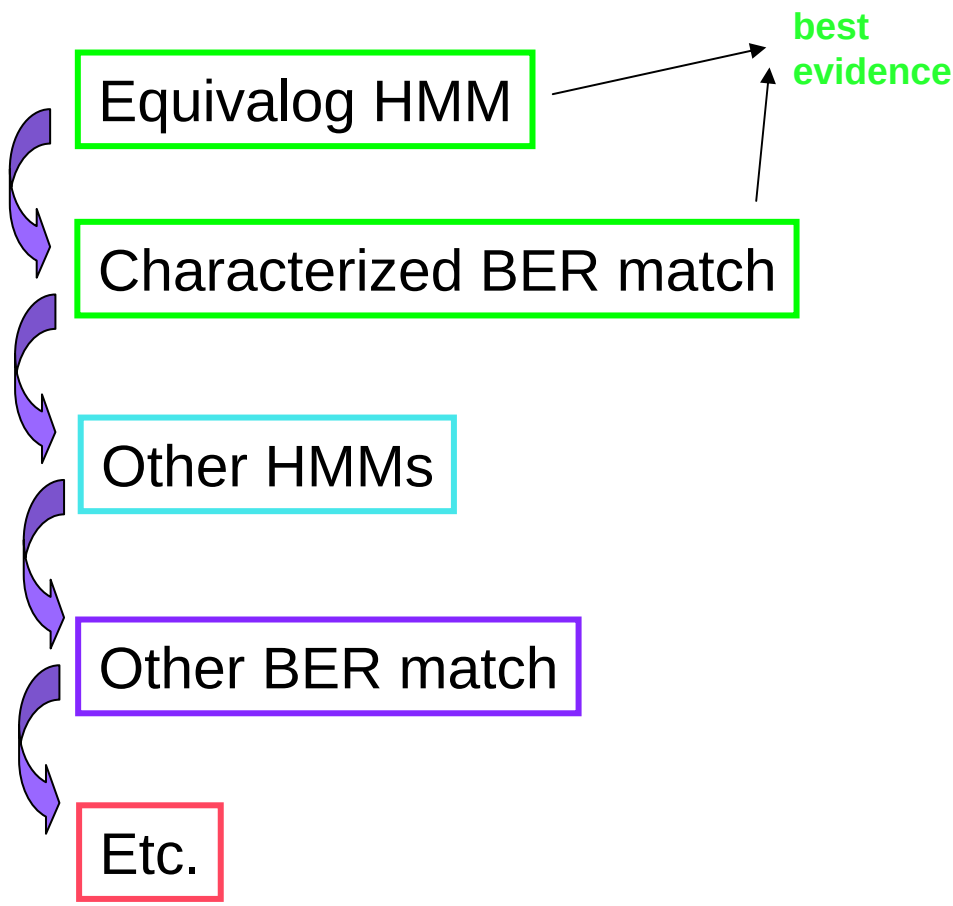
Genome Rot!

- Transitive Annotation: A is like B, B is like C, C is like D, but A is not like D
- We take a very conservative approach and err on the side of missing homology rather than stretching weak data.
- Increasingly, the BER search results are filled with sequences from genome projects, the names of those proteins can not be considered reliable.

AutoAnnotate

Software tool which gives a preliminary name and role assignment to all the proteins in the genome.

Makes decisions based on ranked evidence types



Manual Annotation:
Assigning Names to Proteins

Functional Assignments:
High Confidence in Precise Function

Criteria:

-At least one good alignment (minimum 35% identity, over the full lengths of both proteins) to a protein from another organism that has been experimentally characterized, preferably multiple such alignments.
-Above trusted cutoff hits to any HMMs for this gene.
-Conservation of active sites, binding sites, appropriate number of membrane spans, etc.

Give the protein a specific name and accompanying gene symbol, this is the only confidence level where we assign gene symbols. We default to E. coli gene symbols when possible, for Gram positive genes we use B. subtilis gene symbols.
Example:

name: “adenylosuccinate lyase”
gene symbol: purB
EC number: 4.3.2.2

Functional Assignments: High Confidence in Function, Unsure of Specificity

A good example of this is seen with transporters, what you'll see:

-Multiple hits to a specific type of transporter

-Hits to appropriate HMMs

-The substrate identified for the proteins your protein matches may not all be the same, but may fall into a group, for example they are all sugars.

The name for a specific substrate:

"ribose ABC transporter, permease protein"

The name for specific function but a more general substrate specificity:

"sugar ABC transporter, permease protein"

Sometimes it will not be possible to identify particular substrate group, in that case:

"ABC transporter, permease protein"

Another example of known function but not exact substrate:

"carbohydrate kinase, FGGY family"

Functional Assignments:
Function Unclear

The “family” designation:
-No matches to specific characterized protein
-score above trusted cutoff to an HMM which defines a family, but not a specific function.

“CbbY family protein”

The “homolog” designation:
-if match to a characterized protein is not good enough to say for sure that the two proteins share function (in general, less than 35% id)
-HMM match might be below trusted and above noise
-some active sites missing
OR
-good match to a function not expected in the organism (like a photosynthesis gene in a non-photosynthetic bug)

“galactokinase homolog”

The “putative” designation is used when data is very close to being enough for actual functional assignment:
-has been largely replace by “homolog” and “family”

“putative galactokinase”

Functional Assignments: Hypotheticals

If a protein has no matches to any protein from another species, HMM, Prosite, or InterPro it is called:

“hypothetical protein”

If a “hypothetical protein” from one species matches a “hypothetical protein” from another, they both now become:

“conserved hypothetical protein”

Functional Assignments:
Frameshifts and Point Mutations

Possible sequence errors detected in the BER alignments are sent back to the lab for checking.

Sometimes an error in the sequence is found and corrected.

In others the sequence is shown to be correct and the protein is annotated to reflect the presence of a disruption in the open reading frame:

“great protein, authentic frameshift”

“fun protein, authentic point mutation”

Functional Assignments:
Other ORF disruptions

-many FS/PM,
"degenerate"

-"truncation"

-"deletion"

-"insertion"
(~20-50aa)

-interruption
"interruption-N"
"interruption-C"

-"fusion"

-"fragment"

These are given descriptive terms in the common name and all are put into a "Disrupted reading frame" role category to make them easy to find.

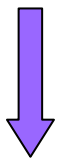
Manual Annotation:
Assigning Roles to Proteins

TIGR Roles

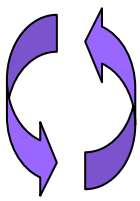
TIGR bacterial roles were first adapted from Monica Riley's roles for E. coli - both systems have since undergone much change.

Unclassified (not a real role)
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

AutoAnnotate makes a first pass at assigning role, based on roles associated with HMMs or with match proteins.



Human annotator checks and adjusts as necessary.



Role Notes:
Notes written by annotators expert in each role category to aid other annotators in knowing what belongs in that category and the TIGR naming conventions for it.

Gene Ontology (GO) Consortium

- **began as collaboration between the databases for mouse, fly, and yeast, but has grown considerably and TIGR is now a member**
- **three controlled vocabularies for the description of:**
  - molecular function (what a gene does)
  - biological process (why a gene functions)
  - cellular component (where a gene acts/lives)
- **can reflect annotation with assignment of GO terms**
  - GO terms exist at many levels of specifiicy (granularity)
  - assign GO terms with specificity appropriate for what is known about the function of the protein in question
  - provides mechanism for storing evidence for GO terms and thus annotation
- **can be easily searched by a computer and allows searches/comparisons across species, kingdoms**
- **if everyone uses the same system, it allows greater exchange of data**

GO term composition

- **GO terms have 3 required parts**
  - ID number – unique stable ids
  - Name
  - Definition – the part of the term that the id number actually refers to, if the name is changed but the definition remains the same – the id stays the same, but if the definition changes – a new GO term must be made
- **Other info connected to GO term**
  - comment
    - gives additional information for proper annotation, tells users why terms were obsoleted
  - cross reference
    - ex. EC numbers
  - synonyms
    - alternate enzyme names
    - abbreviations (TCA)

Example GO term

- ID number: GO:0004076
- Name: biotin synthase activity
- Definition: Catalysis of the reaction\: dethiobiotin + sulfur = biotin.
- comment: none
- cross reference: EC:2.8.1.6
- synonyms: none
- parent term: sulfurtransferase activity (GO:0016783)
- relationship to parent: “is a”

GO term relationships

- **Each GO term has a relationship to at least one other term**
  - process/function/component are roots
  - terms always have at least one parent
  - a term may have children and/or siblings, siblings share the same parent
  - as one moves down the tree from parents to children, the functions, processes, and structures become more specific (or granular) in nature
- **GO is a DAG (directed acyclic graph)**
  - a term can have many parents (as opposed to a hierarchical structure)
- **relationship types**
  - is a (most terms)
    - ribokinase “is a” kinase
  - part of (generally found mostly in component)
    - periplasm is “part of” a cell
  - (regulates - arriving soon)

Example Tree

Annotating with GO

- Decide what annotation the protein should have, find the corresponding terms
  - Your favorite tree viewing tool
    - Manatee GO viewer
    - AmiGO (on GO web page)
  - Mapping files
    - ec2go - a list of EC numbers and corresponding GO terms
    - Tigrfams2go - GO terms assigned to TIGR HMMs
  - Search against proteins already annotated to GO
    - GOst (at GO web page)
    - GO correlations
    - protein name search
    - TIGR GO Blast
- Try to get a term from every ontology at the level of specificity you are confidant of. Don't be afraid to use the "unknown" terms (there's one in each ontology).
- Assign as many terms as are appropriate to completely describe what is known about the protein (you can have multiple terms from each ontology)
- Send annotation to GO to be placed in the repository of annotated genes to be a resource to the community
  - currently 11 TIGR prokaryotic genomes at GO

Functional confidence captured with GO

Available evidence for
3 genes

#1
-HMM for “ribokinase’
-characterized match to ribokinase

#2
-HMM for “kinase”
-characterized matches to a “glucokinase’ AND a ‘fructokinase’

#3
-HMM for “kinase’

Function

- catalytic activity
  - kinase activity
    - carbohydrate kinase activity
      - ribokinase activity
      - glucokinase activity
      - fructokinase activity

Process

- metabolism
  - carbohydrate metabolism
    - monosaccharide metabolism
      - hexose metabolism
        - glucose metabolism
        - fructose metabolism
      - pentose metabolism
        - ribose metabolism

GO Evidence

- Just as we store evidence for our annotation, we must store evidence for GO term assignments:
  - Assign Evidence Code
    - Ev Codes tell users what kind of evidence was used
      - sequence similarity (99% of our work) - ISS
      - experimental characterization - IMP, IDA, etc.
    - IEA - code for electronic annotation - immediately allows users to tell manual curation from automatic
  - Assign “Reference”
    - PMID of paper describing characterization or method used for annotation
    - database reference (GO standards)
  - Assign “with” (when appropriate)
    - Used with ISS to store the accession of the thing the sequence similarity is **with**
  - Modifier column
    - “contributes to” - use this modifier when you assign a function term representing the function of a complex to proteins that are part of the complex but can not themselves carry out the function of the complex
- All accessions used as evidence must be represented according to GO’s format – “database:accession” (where “database” is the abbreviation defined at GO). Manatee knows these rules and automatically puts the accessions in the correct format.
  - Examples
    - TIGR_TIGRFAMS:TIGR01234
    - Swiss-Prot:P12345

GO Evidence codes

- IEA inferred from electronic annotation
- IC inferred by curator
- IDA inferred from direct assay - Enzyme assays
  - In vitro reconstitution (e.g. transcription)
  - Immunofluorescence (for cellular component)
  - Cell fractionation (for cellular component)
  - Physical interaction/binding
- IEP inferred from expression pattern
  - Transcript levels (e.g. Northerns, microarray data)
  - Protein levels (e.g. Western blots)
- IGI inferred from genetic interaction
  - "Traditional" genetic interactions such as suppressors, synthetic lethals, etc.
  - Functional complementation
  - Rescue experiments
  - Inference about one gene drawn from the phenotype of a mutation in a different gene.
- IMP inferred from mutant phenotype
  - Any gene mutation/knockout
  - Overexpression/ectopic expression of wild-type or mutant genes
  - Anti-sense experiments
  - RNAi experiments
  - Specific protein inhibitors
- IPI inferred from physical interaction
  - 2-hybrid interactions
  - Co-purification
  - Co-immunoprecipitation
  - Ion/protein binding experiments
- ISS inferred from sequence or structural similarity
  - Sequence similarity (homologue of/most closely related to)
  - Recognized domains
  - Structural similarity
  - Southern blotting
- NAS non-traceable author statement
- ND no biological data available
- TAS traceable author statement
- NR not recorded

http://www.geneontology.org/GO.evidence.html
http://www.geneontology.org/GO.annotation.html

Association files at GO

Current Annotations

What are IEA Codes? | View the Terms and Annotations

This table shows the number of gene products that have been annotated to the gene ontologies by each collaborating group. A gene product can have one or more molecular functions, be used in one or more biological processes and may be associated with one or more cellular components. Tab-delimited files of the associations between gene products and GO terms made by the member organizations are available from the FTP site or from the links in this table. The file format is described in the Annotation Guide. Any errors or omissions in annotations should be reported by writing to the GO mailing list: go@geneontology.org.

Notes:
1) The files are compressed using the UNIX gzip utility; use the "Download" link to download the compressed file to your disk.
2) Where available (e.g. for the Compugen and GO Annotations at EBI files), please also see the appropriate README file

	Biological Process		Molecular Function		Cellular Component		Total Gene Products Associated	Total References Included as Evidence	TAB Delimited File of Associations & Last Update
	All codes	non-IEA codes	All codes	non-IEA codes	All codes	non-IEA codes			
TIGR *Arabidopsis thaliana* README	9638	9638	24945	24945	5835	5835	25701	13653	Download Feb 10, 2004
TIGR *Bacillus anthracis* Ames	4414	4414	4416	4416	200	200	4417	6	Download Mar 12, 2004
TIGR *Coxiella burnetii* RSA 493	1359	1359	1349	1349	176	176	1365	4	Download Mar 12, 2004
TIGR Gene Index README	80031	0	100151	0	78400	0	126557	1	Download Apr 16, 2004
TIGR *Geobacter sulfurreducens* PCA	2800	2800	2800	2800	195	195	2800	5	Download Mar 12, 2004
TIGR *Listeria monocytogenes* 4b F2365	2680	2680	2680	2680	917	917	2681	4	Download Jun 29, 2004
TIGR *Pseudomonas syringae* DC3000	2941	2941	3101	3101	263	263	3137	4	Download Mar 12, 2004
TIGR *Shewanella oneidensis* MR-1	3696	3696	3696	3696	241	241	3696	5	Download Mar 12, 2004
TIGR *Trypanosoma brucei* chr 2 README	291	291	289	289	278	278	292	55	Download Apr 16, 2004
TIGR *Vibrio cholerae*	2923	2923	2728	2728	191	191	2924	9	Download Mar 12, 2004

GO is a work in progress

- **The GO actively requests user participation in ontology development**
  - new terms
  - changes to existing terms
  - SourceForge site
- **consortium meetings**
- **user meetings**
- **terms can become obsolete, but their ids are never used again and they remain in the ontolgies so people can track them**
- **as the ontologies change annotations must be changed too - in particular annotations to terms that have become obsolete (like “toxin activity”)**

ORF Management
and
Data Availability

ORF management: Start site edits

What to consider:

- Start site frequency: ATG >> GTG >> TTG
- Ribosome Binding Site (RBS): a string of AG rich sequence located 5-11 bp upstream of the start codon
- Similarity to match protein, both in BER and Paralogous Family - probably the most important factor.

(Remember to note, that the DNA sequence reads down in columns for each codon.)

-In the example below (showing just the beginning of one BER alignment), homology starts exactly at the first atg (the current chosen start, aa #1), there is a very favorable RBS beginning 9bp upstream of this atg (gagggaga). There is no reason to consider the ttg, and no justification for moving to the second atg (this would cut off some similarity and it does not have an RBS.)

3 possible start sites

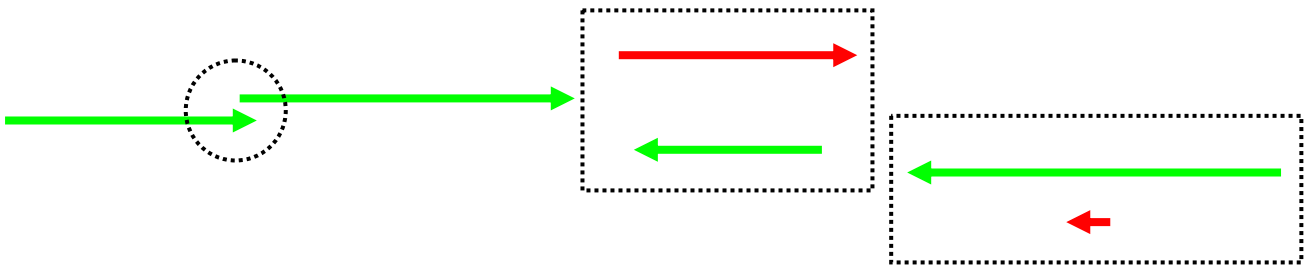
RBS upstream of chosen start

This ORF's upstream boundary

ORF management: Overlaps and Intergenics

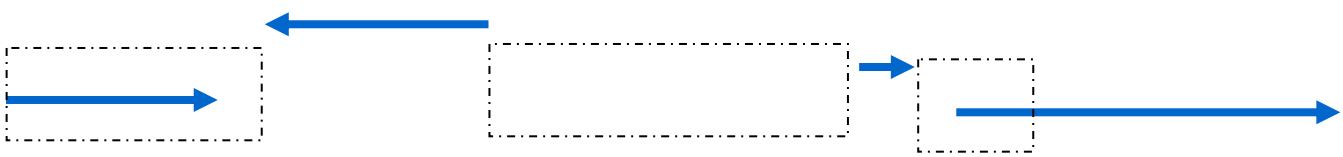
Overlap analysis

When two ORFs overlap (boxed areas), the one without similarity to anything (another protein, an HMM, etc.) is removed. If both don't match anything, other considerations such as presence in a putative operon and potential start codon quality are considered. This process has both automated (for the easy ones) and manual (for the hard ones) components. Small regions of overlap are allowed (circle).



InterEvidence regions

Areas of the genome with no genes and areas within genes without any kind of evidence (no match to another protein, HMM, etc., such regions may include an entire gene in case of "hypothetical proteins") are translated in all 6 frames and searched against niaa. Results are evaluated by the annotation team.



Data Availability

- **Publication**
  - TIGR staff/collaborators analysis of genome data
- **GenBank**
  - Sequence and annotation submitted to GenBank at the time of publication
  - Updates sent as needed
- **Comprehensive Microbial Resource (CMR)**
  - Data available for downloading
  - extensive analyses within and between genomes

Useful links

- CMR Home
  - http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl
- SIB web site (Swiss-Prot, Prosite, etc.)
  - http://www.expasy.org
- PIR
  - http://pir.georgetown.edu
- NCBI
  - http://www.ncbi.nlm.gov
- BLAST
  - http://blast.wustl.edu
- GO
  - http://www.geneontology.org
- TIGRFAM HMMs
  - http://www.tigr.org/tigr-scripts/CMR2/find_hmm.spl?db=CMR
  - OR
  - http://tigrblast.tigr.org/web-hmm/

Acknowledgements

leading the effort:
Owen White
Jeremy Peterson

Prokaryotic Annotation
Bill Nelson (Team leader)
Bob Dodson
Bob Deboy
Scott Durkin
Sean Daugherty
Ramana Madupu
Lauren Brinkac
Steven Sullivan
M.J. Rosovitz
Sagar Kothari
Susmita Shrivastava

CMR team:
Tanja Davidsen (Team leader)
Nikhat Zafar
Qi Yang

HMM team:
Dan Haft
Jeremy Selengut

Building the tools:
Todd Creasy (head Manatee developer)
Liwei Zhou
Sam Angiuoli (and his team)
Anup Mahurkar (and his team)

And the many other TIGR employees, present and past, who have contributed to the development of these tools and to the annotation protocols I have described.

Also thanks to the funding agencies that make our work possible including NIH, NSF, and DOE.