
Roseobacter denitrificans genome
annotation using Manatee

Chaitanya R. Acharya
Computational Biosciences
Arizona State University
May 3, 2006

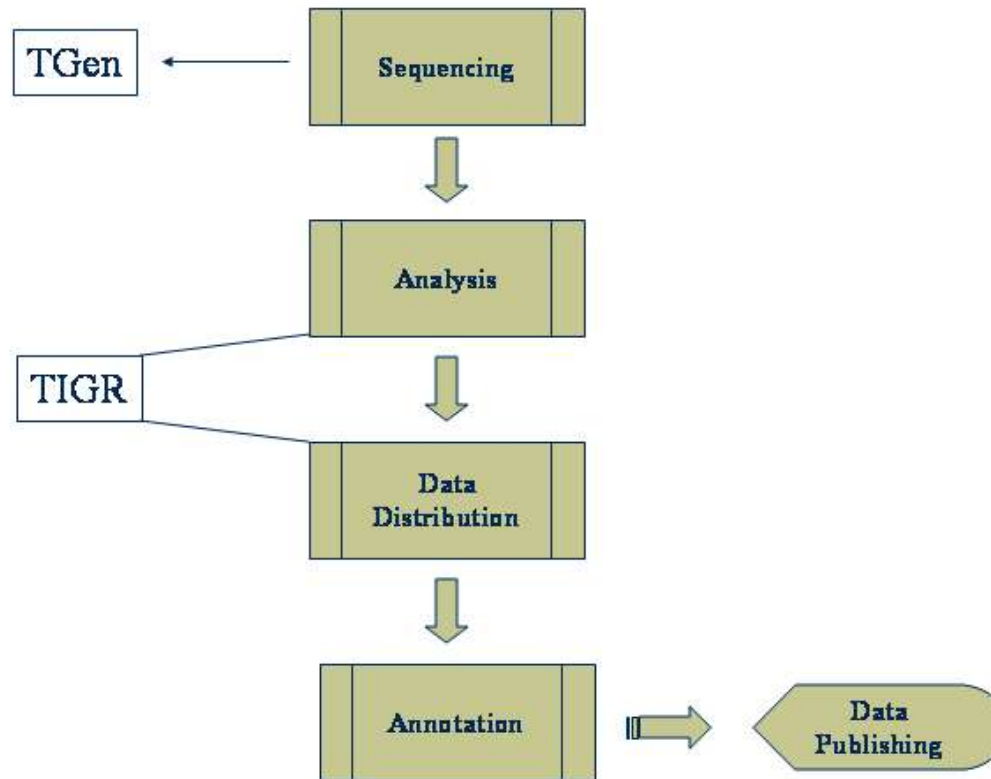
Acknowledgements

- Thanks to my project advisor Dr. Robert Blankenship, who proposed this internship project and provided guidance throughout.
 - Thanks to my team: Sumedha Gholba, Hector Ramos, Wesley Swingley, Heather Matthies and Michael Lince, for their help and encouragement throughout.
 - Thanks to Dr. Jeffrey Touchman and his crew (at TGen) for their help in understanding the sequence 'deformities'.
 - Thanks to my committee: Dr. Rosemary Renaut, Dr. Blankenship and Dr. Martin Wojciechowski, for the knowledge they have shared throughout my program and for review of this applied project.
 - Thanks to Loretta Goldberg, for her help in putting together this presentation.
-

This internship project is presented in partial fulfillment of the requirements of the Professional Science Master's in Computational Biosciences,
Arizona State University

Research conducted from
June 2005 – August 2005

Project Overview



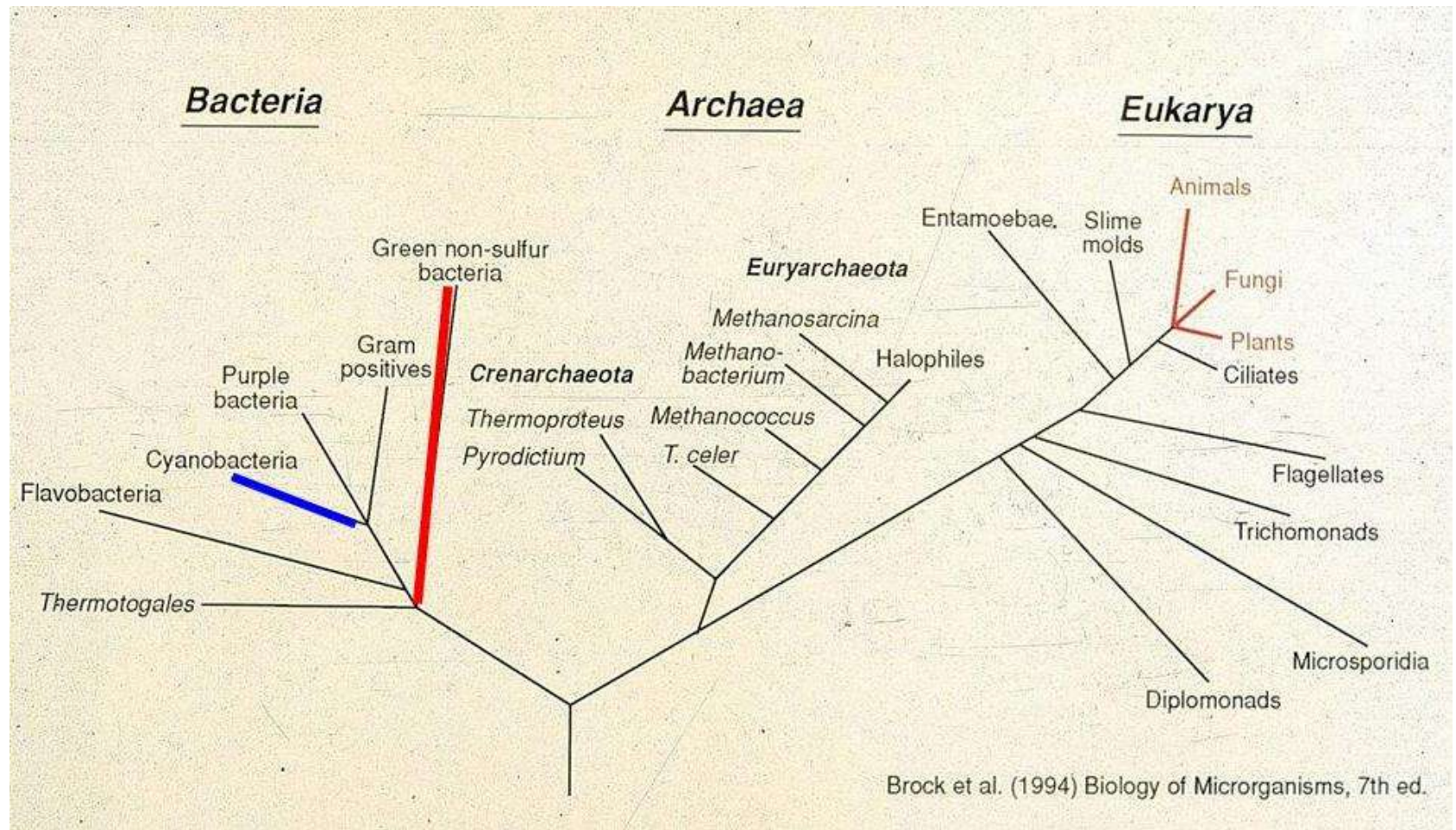
Annotation

- Putative genes are located and identified on a newly sequenced genome
 - Homology search is also used to understand the physical and functional characteristics of the gene products
 - Identify the role played by that specific gene product, if it is a protein, in a metabolic pathway (Metabolic profile)
 - The elements of the annotation process include gene finding, homology searches, functional assignment, ORF management and finally making this data available to the public.
-

Phototrophic Genome Project

- This internship project is a part of phototrophic genome project
 - Obtaining genomic sequences of four representative bacterial species: Heliobacteria (*Heliobacterium modesticaldum*), Aerobic Phototrophic bacterium (*Roseobacter denitrificans*), Alpha Proteobacteria (*Rhodocista centenaria*) and Cyanobacteria (*Acaryachloris marina*). Please visit <http://genomes.tgen.org>
-

16S rRNA diversity analysis



Roseobacter denitrificans – a model aerobic phototrophic bacteria

- Depend of the respiration of organic compounds for growth
 - Cessation of photosynthetic pigment synthesis upon illumination
 - Requires a respiratory terminal electron acceptor (invariably oxygen)
 - Questions are raised on the evolution and genetic regulation of photosynthesis
-

Significance *Roseobacter denitrificans* genome

- The evolutionary genesis of photosynthetic genes
 - True evolutionary positions of aerobic phototrophic bacteria would be clarified by whole genome comparisons
- Pathways of carbon dioxide fixation and production
 - Constructing the metabolic profile is very important which could be tested in biochemical and molecular biology experiments
- Light and oxygen signal transduction in gene expression
 - Study both oxygen- and light-responsive pathways by cloning, gene-disruption methods and over-expressing genes for biochemical and biophysical analysis of purified proteins

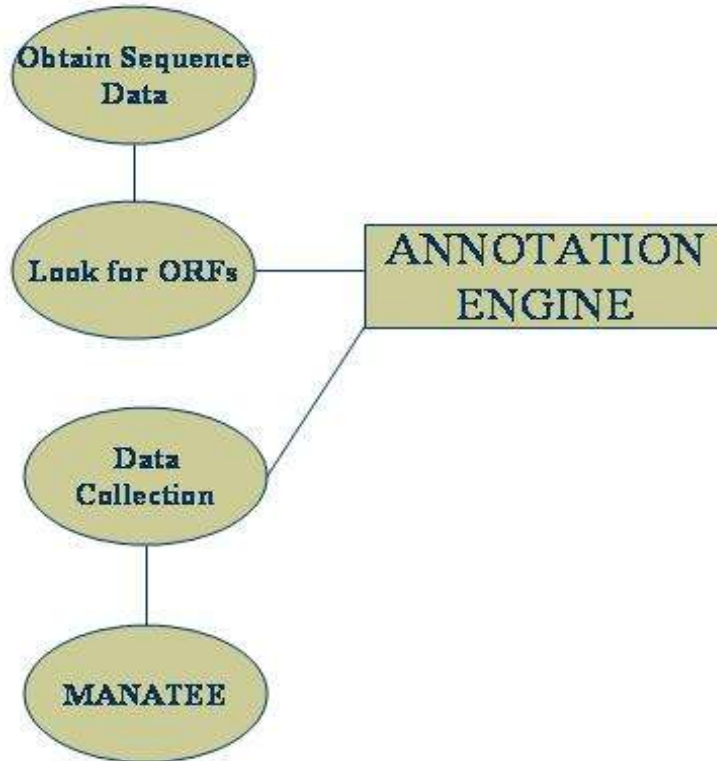
Why *Roseobacter denitrificans*?

- Readily cultivated in the laboratory
 - Electron transfer pathways establish that this is the model aerobic phototrophic bacterium
 - Only bacterium that is capable of anaerobic growth (nitrate as terminal electron acceptor)
 - Marine bacterium
 - High GC content (~59%) and relatively small genome size (~4 Mb)
-

Sequencing *R. denitrificans*: TGen's role

- Sequencing *R. denitrificans* was accomplished at Translatory Genomics Institute (TGen)
 - *R. denitrificans* contains a primary circular chromosome of 4,133,097 base pairs and four plasmid containing a total of 4,403 predicted coding sequences
-

TIGR's role



‘Annotation Engine (AE)’ at TIGR

- First executes the annotation in a format that promotes consistency of data types across all genomes
 - Allows straightforward reincorporation of annotation data back into the Comprehensive Microbial Resource (CMR) data management system
 - Aids in quality control of the sequence
 - Preliminary annotation is displayed on the web
-

Components of AE

- Production of output from TIGR's automated annotation pipeline - includes search results and automatically generated annotation in a MySQL database and associated files
 - The manual annotation tool 'Manatee' - an open source web based interface for interacting with and editing annotation data
-

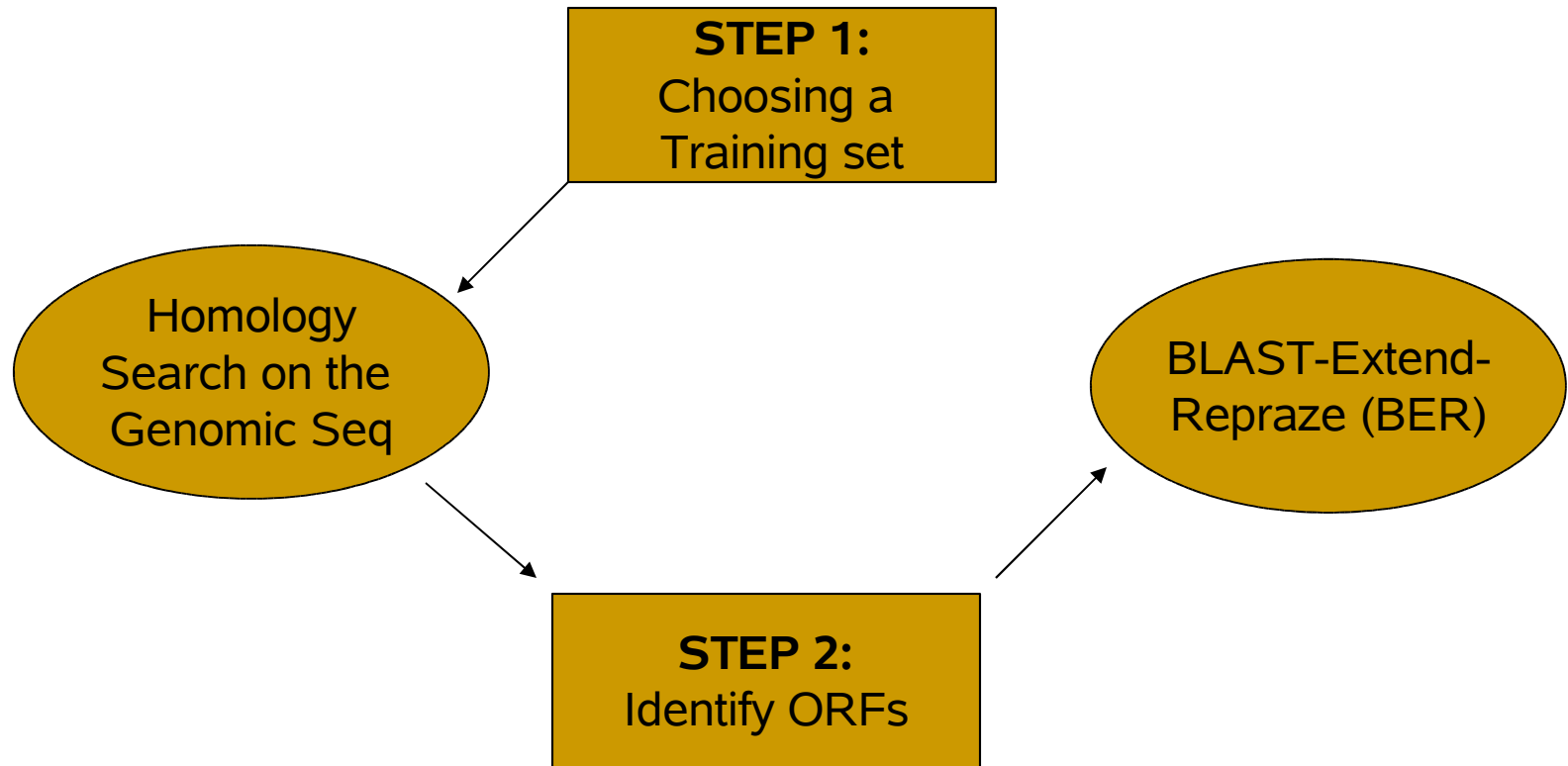
Elements of AE

- Gene Finding using the ‘Glimmer system’
 - Predicting candidate genes and the proteins they code for
 - Functional assignment of all predicted proteins
 - Search against an internal non-identical amino acid (NIAA) database
 - Data supplied to the AE user or annotator
 - Data is available as a MySQL database
 - Providing annotation tool, Manatee.
 - TIGR's manual annotation tool
-

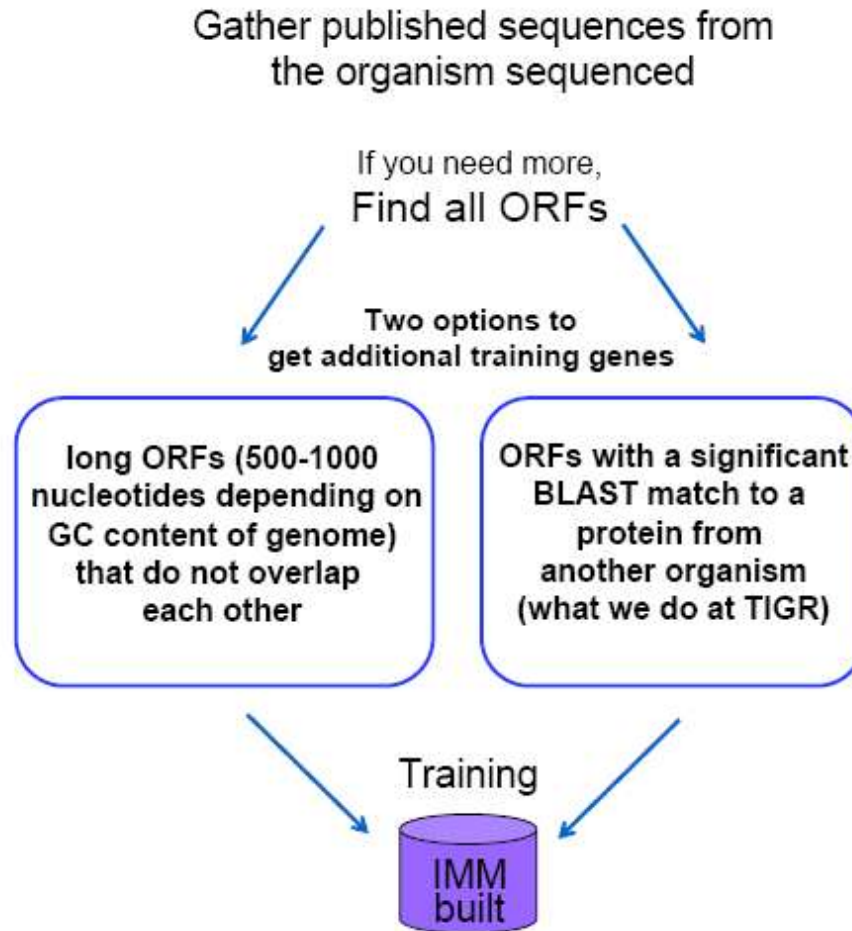
Glimmer System

- Identifies probable open reading frames (ORFs)
 - The Glimmer software system developed by Salzberg and Delcher *et al.* is used to find genes in many prokaryotic genomes such as bacterial, viral and archaeal genomes.
 - The algorithm at the core of the Glimmer is an Interpolated Markov Model (IMM), which is a special kind of Markov chain.
 - A Markov chain calculated the statistical information about any sequence by computing the conditional probability $P(x|S)$ that a nucleotide 'x' appears after a sequence 'S'.
 - Second- or fifth- or eighth-order Markov chains work well because they are computing statistics based on codons, dicodons and tricodons respectively.
-

'Glimmering'



Glimmer Algorithm



Visualizing genes

ATGCTTTGCTTGGATGAGCTCATA start
TACGAAACGAACCTACTCGAGTAT stop

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
*

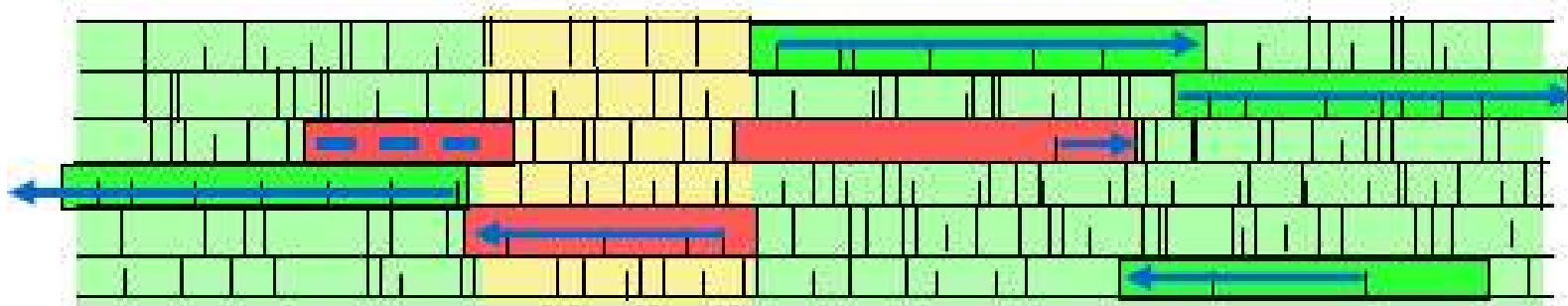
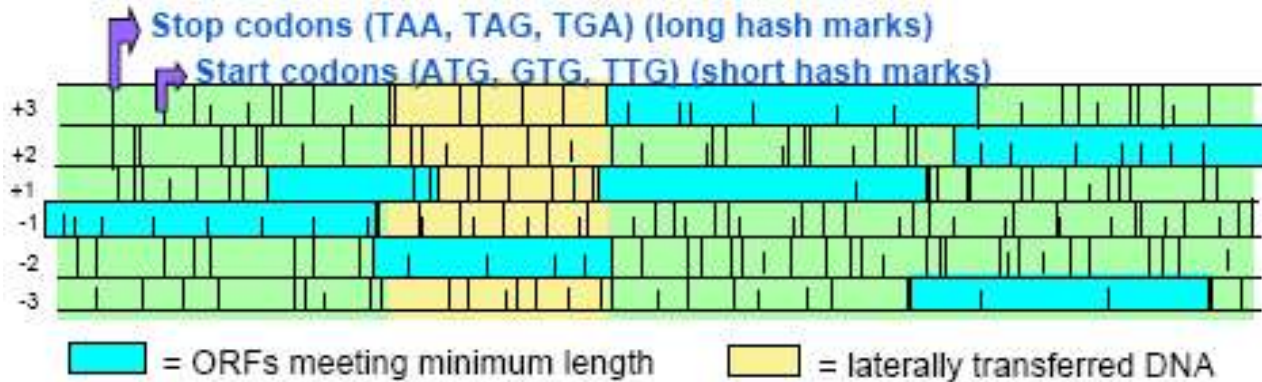
Proteins searched against HMMs using 'hmmpfam' program.

Two-fold homology search is performed to identify probable ORFs.

ORFs visualized in six-frame translations maps

HMMs used here are Pfam HMM and TIGRFAM

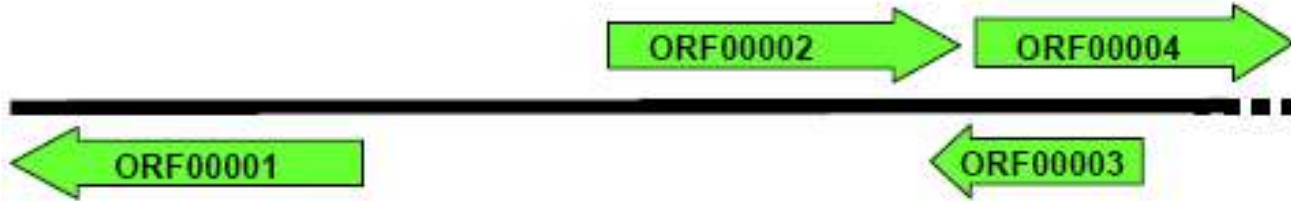
Translation Maps



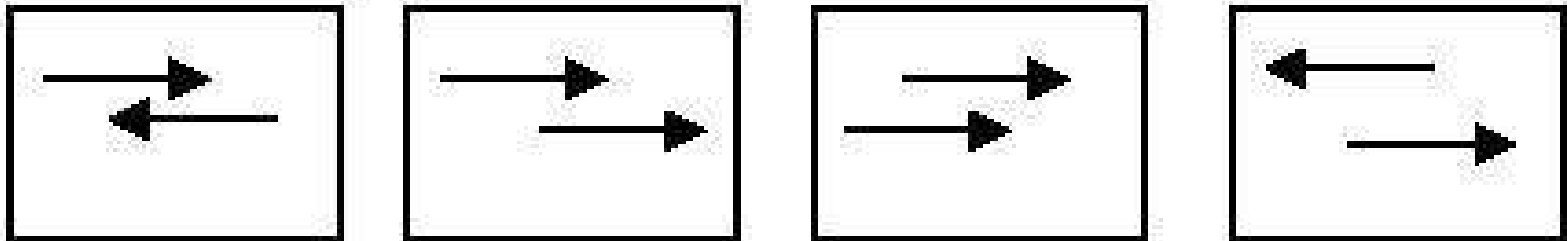
Green regions: High scoring ORFs
Red regions: Low scoring ORFs

Testing Overlaps

Mapping the ORFs



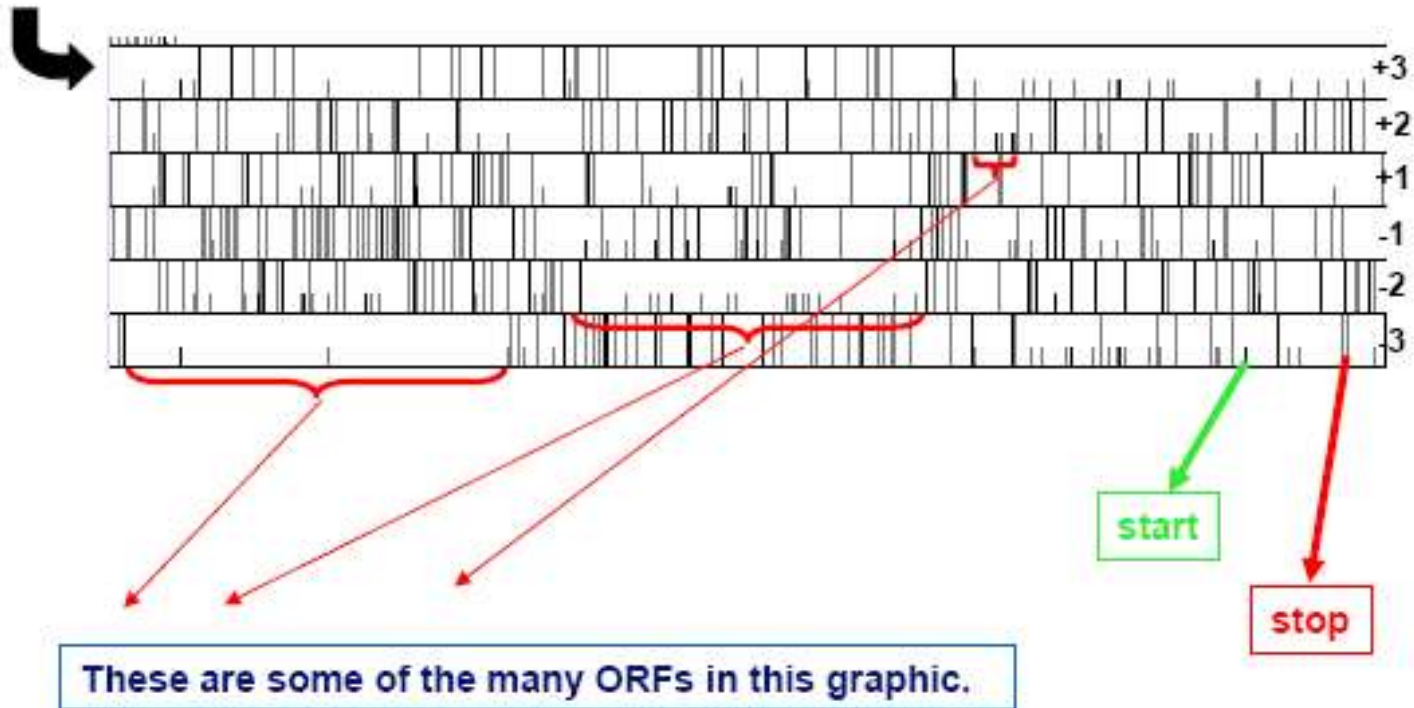
Scoring of the overlap region



Manatee

- Manatee is a web-based gene evaluation and genome annotation tool that can view, modify, and store annotation for prokaryotic and eukaryotic genomes
 - Manatee consists of a ‘suite of programs’, including Gene Ontology (GO) classifications, Blast-Extend Repraze (BER), Blast search data, paralogous families, and annotation suggestions generated from the automated analysis.
 - It is an open-source initiative that was developed for two main reasons: 1) to help biologists annotate their genomes using a powerful, stand-alone web application with a robustly designed relational annotation database, and 2) to invite developers from all over the world to enhance Manatee’s ability to completely accomplish biological goals.
-

ORF Management



Determining Functional Roles of Predicted Proteins

- Determined by homology searching and/or experimental characterization of proteins
 - High identities (at least >35%) prove that the sequences share their functions.
 - All the functional assignments made after sequence alignments should be considered 'putative' (suggested) until confirmed by experimental characterization.
-

Determining Functional Roles of Predicted Proteins

- Characterized proteins are stored in “Characterized Table” within the database; they are designated a confidence status
 - They are all color coded:
 - Green color = full experimental characterization
 - Red color = characterized by Swiss-Prot (by an automated process)
 - Sky blue color = partial characterization
 - Olive color = trusted to be characterized
 - Blue-green color = only a fragment/domain has been characterized
 - Fuzzy gray color = void
 - Gray color = sequence exists in the ‘Omnium’ (database that underlies TIGR’s CMR)
-


Non-identical amino acid (NIAA) database

- This file is composed of protein sequences from many protein translations of all ORFs searched against hidden Markov models (HMM) built from the multiple amino acid sequence alignments.
 - Each HMM is associated with a 'noise' cutoff score and a 'trusted' cutoff score. ORFs are considered to be members of the HMM model if they score higher than the trusted cutoff.
 - TIGR classifies TIGR and Pfam HMMs into fifteen isologies. Each isology defines a specific database match
-

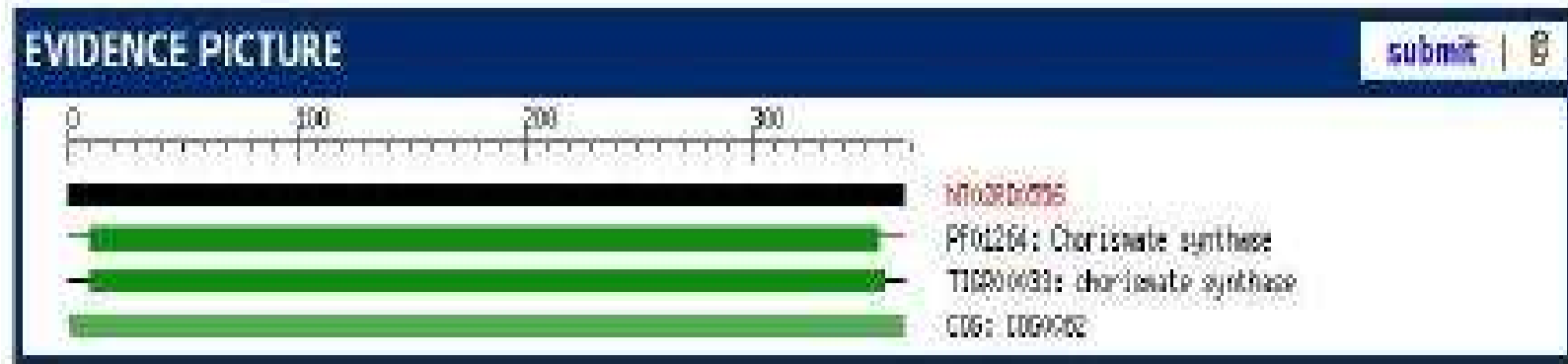
BLAST-Extend Repraze (BER)

- The BER alignments are stored in a mini-database and the closest matches are displayed in a table. They are assigned scores.
 - Table displayed in 'Gene Curation Page (GCP)' of an ORF as 'BER Skim'.
 - The matches are color coded (as mentioned earlier) depending on their characterization
-

BER Skim

BER SKIM				
		View BER Searches	search date: Wed Oct 23 11:59:20 2002	<input type="button" value="Refresh Searches"/>
accession	% sim	length	description	p-value
OMNI:SO2740	100.0	349	biotin synthase { <i>Stewanella oneidensis</i> MR-1}	1.5e-176
SP:P36569	80.7	340	Biotin synthase (EC 2.8.1.5) (Biotin synthetase). (<i>Serratia</i>	2.5e-119
SP:PL2906	79.7	342	Biotin synthase (EC 2.8.1.5) (Biotin synthetase). (<i>Escherich</i>	7.2e-120
GP:145425	79.7	342	biotin synthetase { <i>Escherichia coli</i> }	1.5e-119
GP:12620127	79.4	342	biotin synthase BioB {uncultured bacterium pCosHE2}	1.5e-119
OMNI:NTL03EC0855	79.4	342	biotin synthetase { <i>Escherichia coli</i> O157:H7 VT2-Sakai} [EGP13	5.1e-119
OMNI:NTL01YP1094	81.0	340	biotin synthase { <i>Yersinia pestis</i> CO92} [OMNI:NTL02YP2986 biot	8.3e-119
GP:12620099	79.5	340	BioB-like protein {uncultured bacterium pCosPS1}	9.5e-118
OMNI:NTL02EC0848	79.1	342	biotin synthesis, sulfur insertion? { <i>Escherichia coli</i> O157:H	2.2e-118
SP:Q47862	79.2	339	Biotin synthase (EC 2.8.1.5) (Biotin synthetase). (<i>Erwinia ch</i>	3.6e-118
SP:PL2678	78.6	344	Biotin synthase (EC 2.8.1.5) (Biotin synthetase). (<i>Salmonell</i>	5.1e-119
OMNI:VC1112	81.8	348	biotin synthase { <i>Vibrio cholerae</i> El Tor N16961} [EGP9655583g	5.1e-119

Evidence Picture



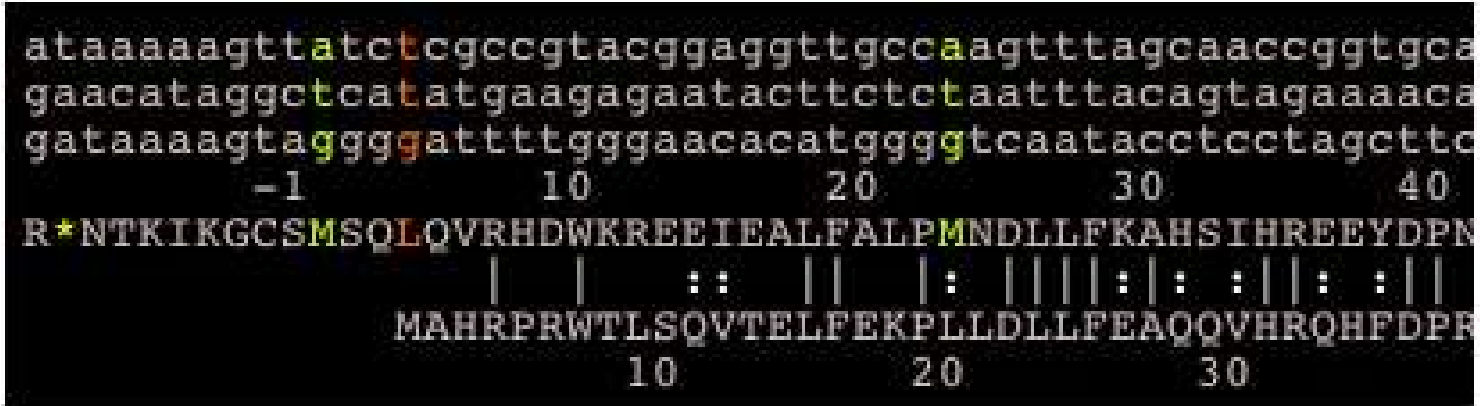
- Green colored matches are all characterized...
- Clusters of Orthologous groups (COGs) also play an important role in predicted protein identification.

One look at a match...

66.0/79.7% over 343aa	<i>Escherichia coli</i>
<ul style="list-style-type: none">• SPIP12996 Biotin synthase (EC 2.8.1.6) (Biotin synthetase). Edit characterized• PIRJC2517/SYECBB biotin synthase (EC 2.8.1.6) bioB [validated] - Escherichia coli (strain K-12) Insert characterized• GBAAC73862.1/IGP1786992/AE000180 biotin synthesis, sulfur insertion? {Escherichia coli K12;} Insert characterized	

```
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]
```

Alignments



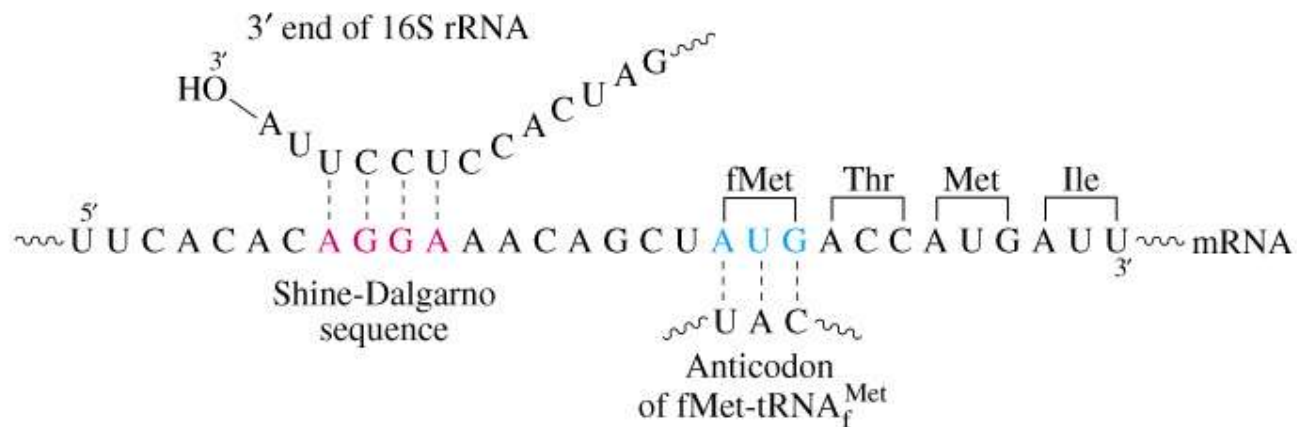
- Codons read from top to bottom
- Asterisk indicates a stop codon
- Negative numbers indicate downstream sequence from the putative start site
- Start sites are all color coded
- There are three start sites- ATG (Methionine), TTG (Leucine) and GTG (Valine)
- Solid line indicates identical amino acid, broken line indicates similar amino acid

Identifying a start site

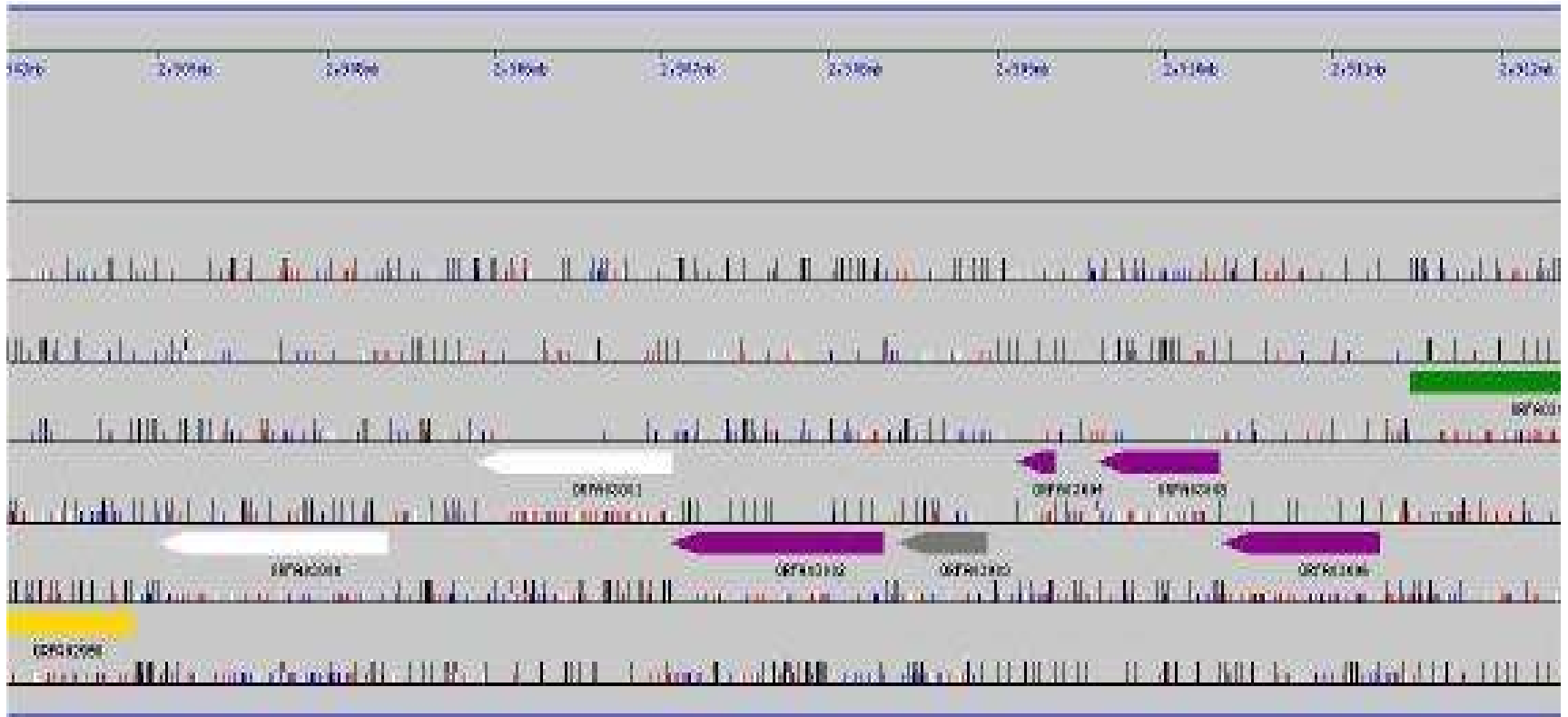
(a)

Lipoprotein	~AUCUA GAGG UAUUAUA AUG AAAGCUACU~
RecA	~GGCAUGAC AGG AGUAAAA AUG GCUAUCG~
GalE	~AGCCUAAU GAGG CGAAUU AUG AGAGUUCUG~
GalT	~CCCGAU UAGGA ACGACCA AUG ACGCAAUUU~
LacI	~CAAUUCAG GGUG GUGAAU UGUG AAACCAGUA~
LacZ	~UUCACAC AGGA AACAGCU AUG ACCAUGAUU~
Ribosomal L10	~CAUCA AGGAG CAAAGCUA AUG GCUUUAAAU~
Ribosomal L7/L12	~UAUUC AGGA ACAAUUUAA AUG UCUAUCACU~

(b)



Genome View



Editing Start site

Start Edits

Accession	Nucleotide	Frame
2816723	TTCCCGGTTTCCAATCATGACGAACTTGCAACTGGGACATTGAACACCCTTTTATTTTGT	Nucleotide
	E P L P I N T N L Q L R H * T P F Y E C	Frame1
	E E E Q S * E I C N E P I E H E F I E Y	Frame2
	P A S N H D E L A T A T L N T L L E L Y	Frame3
	E G S G I M V E E K C S R Q Q V G E * E Q	Frame4
	R G A E L * S S S A V A V N F V R X N K	Frame5
2816785	ATTTTACCTTGGCTAGGATAACCTCAGCCCTTAAACTGTCAACGCCAACCAAGTACATACAG	Nucleotide
	I L P W L G * P O P L N C O R O P V I O	Frame1
	E Y L G * R M L E P * T Y N A N Q * Y E	Frame2
	E T L A B I T E A L E K L E T F T S D T G	Frame3
	I K G Q S P Y G * C E F Q * H W G T I C	Frame4
	Y E V E A L I V E A B L S D V Q V L S V	Frame5
2816843	GTTTACCCTGATTAATTTTCAATCAACGCTGTGAGCTTTTATGGCGCAATTTACTCGATT	Nucleotide
	V Y H * L I E H Q R E E L L L C A I Y E I	Frame1
	F T T D * F S I N A V S F Y A Q F T R P	Frame2
	L P L I N E Q S T L * A E M R N L L L D F	Frame3
	T * K Q N I K L * R Q S E K H A I * E I	Frame4
	E E G S I L E * D Y S H A K I H L K S S	Frame5
2816903	TTGACTTTGATACCGCCCATATTTGGCACCCCTTATACCTCCATGACTCGTGCACTTCTCTG	Nucleotide
	L T L I A E I E G T L I E E * L V H E L	Frame1
	* L * * E E Y L A P L Y L H R E C T E C	Frame2
	D P S A H I H E F Y T S H T R A L F V	Frame3
	H V K I A S H N H V R I Q S H S T C K R	Frame4
	K E K E L A S I Q C G * V K M V E A S G	Frame5
O S Q Y R G Y K A G K Y R W S E H V E Q	Frame6	

Green nucleotides represent ribosome binding sites, purple nucleotides represent amino acids of the query gene and blue nucleotides represent amino acids of one of the genes in the region other than the query gene.

Start site edited based on the Genome View

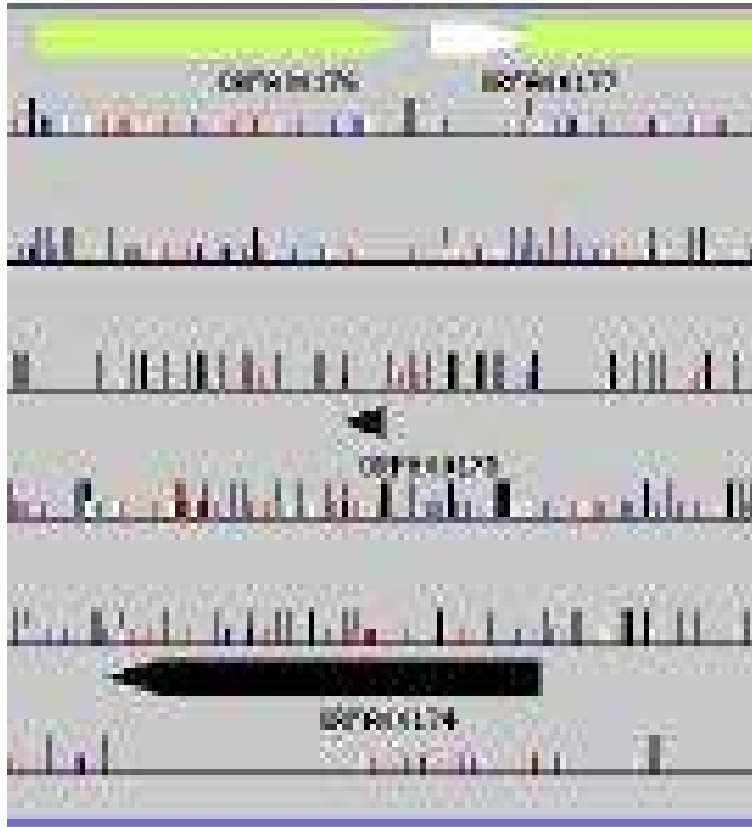
Gene Curation Page

- Has ORF descriptors
 - The annotator has the option of populating six fields: Common name (com_name), Gene Symbol (gen_sym), Enzyme Commission number (ec_num), comments, TIGR roles (role_id) and GO terms
 - HMMs are all shown along with the BER Skim
 - Annotation begins here...
-

Annotation Protocol

- Genome View – check for overlap
 - HMM – evidence picture
 - BER SKIM
 - Edit start site
 - Check links
 - Naming
 - Characterized match (p-value < 1050 or 35% identity)
 - PFAM
 - Gene Symbol
 - EC number
 - Comment
 - TIGR roles
 - Gene Ontology terms (*Can be skipped*)
 - Submit Data
 - Report Frame shifts, start errors & overlaps
-

Overlaps in Genome View



Overlaps have to be checked for before being discarded

Look at the sequence and check for any frame shift mutations

Demo



A screenshot of a web login form titled "Manatee Login". The form has a blue header with the title in white. Below the header, there are three input fields: "user_name:" with the value "training", "password:" with a masked value of "*****", and "database:" with the value "gsp". A "Submit" button is located below the input fields.

Manatee Login	
user_name:	<input type="text" value="training"/>
password:	<input type="password" value="*****"/>
database:	<input type="text" value="gsp"/>
<input type="button" value="Submit"/>	

Please go to:
<http://manatee.tgen.org>

Observations

- *R. denitrificans* lacks key Calvin cycle enzymes ribulose biphosphate carboxylase (RubisCO), phosphoribulokinase (PRK), and other proteins typically coded by the Calvin cycle operons in closely related anaerobic purple bacteria.
- Suggests evidence for a scattered loss of ancestral carbon-fixation in the α -proteobacterial tree.
- Some putative genes related to the C4 sequestration, Crassulacean acid metabolism (CAM), and anaplerotic carbon-fixation enzymes such as pyruvate-orthophosphate dikinase and phosphoenolpyruvate (PEP) carboxylase, are present.
- Hypothesized that APBs fix CO₂ heterotrophically using a C4 sequestration pathway supplemented by additional CO₂ provided by CO oxidation and heterotrophic respiration.

Future Directions

- Unclear how *R. denitrificans* adapted to the changing atmospheric conditions, and how RubisCO evolved in the APBs such as *R. denitrificans*.
 - Functional characterization of proteins in labs would allow scientists to further probe the organism and understand various other metabolic pathways associated with the overall metabolic profile of *R. denitrificans*.
 - More comparative sequence (or genome) analysis of the sequenced APBs
-

References

- [Alts 90] Altschul S F, Gish W, Miller W, Myers E W, Lipman D J; Basic Local Search Alignment Tool; *Journal of Molecular Biology* 1990; 215:403-410.
- [Alts 97] Altschul S F, Madden T L, Schaffer A A, Zhang J, Zhang Z, Miller W, Lipman D J; Gapped BLAST and PSI_BLAST: a new generation of protein database search programs; *Nucleic Acid Res* 1997; 25:3389-3402.
- [Arth 99] Arthur L. Delcher, Douglas Harmon, Simon Kasif, Owen White and Steven L. Salzberg; Improved microbial gene identification with GLIMMER; *Nucleic Acid Res* 1999; 27:4636-4641
- [Arth 05] Arthur L. Delcher; GLIMMER Release Notes, Version 3.01 (Beta) 10 October, 2005
- [Bate 90] Bateman A., et al.; The Pfam protein families database; *Nucleic Acid Res.* 1990; 28(1):263-266
- [Bea 02] Beatty JT; On the natural selection and evolution of the aerobic phototrophic bacteria; *Photosynth Res* 2002; 73: 109-114.
- [Bla 02] Blankenship RE (2002) Molecular Mechanisms of Photosynthesis, *Blackwell Science*, Oxford, UK.
-

References

- [Bla 98] Blankenship RE and Hartman H (1998); The origin and evolution of oxygenic photosynthesis. *Trends Biochem Sci* 23: 94-97.
- [Ed 99] Eddy S; Profile hidden Markov models; *Bioinformatics*; 14(9):755-763
- [Gis 93] Gish W, et al.; Identification of protein coding regions by database similarity search; *Nat. Genet* 1993; 3(3):266-272
- [Haf 01] Haft D, et al.; TIGRFAMs: A protein family resource for the functional identification of proteins; *Nucleic Acids Res* 2001; 29(1):41-3
- [Jean 98] Jeanmougin F, Thompson J D, Gouy M, Higgins D G, and Gibson, T J; Multiple Sequence Alignments with Clustal X; *Trends Biochem Sci* 1998; 23:403-5
- [Sal 98] Salzberg S., et al.; Microbial gene identification using Interpolated Markov Models; *Nucleic Acid Res* 1998; 26(2):544-548
- [Smi 81] Smith T F, et al.; Identification of common molecular subsequences; *J Mol Biol* 1981; 147(1):195-197
- [Tigr 04] TIGR; Michelle Gwinn; Prokaryotic Annotation Overview, October 2004
- [Tigr 04] TIGR; Michelle Gwinn; A guide to Manatee, October 2004
-

References

- [Tigr 05] TIGR; Michelle Gwinn, William Nelson, Robert Dodson, Steven Salzberg, Owen White; Small Genome Annotation and Data Management at TIGR
- [Tigr 05] TIGR; Domain Based Paralogous Protein Families; www.tigr.org; Annotation Workshop July13, 2005
- [TGen] Translatory Genomics Institute; <http://www.tgen.org>
- [Wes 06] Wesley D S, et al; A ubiquitous pathway marine phototroph with a novel carbon-fixation pathway; submitted to PNAS, 2006
- [Man 05] Manatee web site; <http://manatee.tgen.org> ; edit version active from June 2005 – August 2005;
- [1] Phototrophic Genome Project web site; <http://genomes.tgen.org>
-