Roseobacter denitrificans genome annotation using Manatee
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

- TGen
  - Sequencing
  - Analysis
  - Data Distribution
  - Annotation
  - Data Publishing

- TIGR
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.
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

Bacteria
- Green non-sulfur bacteria
- Gram positives
- Purple bacteria
- Cyanobacteria
- Thermotogales

Archaea
- Crenarchaeota
- Thermoproteus
- Pyrodictium
- Methanococcus
- Methanobacterium

Eukarya
- Euryarchaeota
- Entamoeba
- Slime molds
- Methanosarcina
- Animals
- Fungi
- Plants
- Ciliates
- Flagellates
- Trichomonads
- Diplomonads
- Microsporidia

Courtesy http://landresources.montana.edu/dward/pages/diversity_ecology_evolution/molecular_analysis.htm
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

1. Obtain Sequence Data
2. Look for ORFs
3. Data Collection

MANATEE

ANNOTATION ENGINE
‘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
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.
STEP 1: Choosing a Training set

Homology Search on the Genomic Seq

STEP 2: Identify ORFs

BLAST-Extend-Repraze (BER)
Glimmer Algorithm

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

IMM built
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

These are some of the many ORFs in this graphic.
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.
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.
<table>
<thead>
<tr>
<th>accession</th>
<th>% sim</th>
<th>length</th>
<th>description</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMN1SO2740</td>
<td>100.0</td>
<td>349</td>
<td>biotin synthase ( \text{Shewanella oneidensis MR-1} )</td>
<td>1.5e-176</td>
</tr>
<tr>
<td>SP:PI36569</td>
<td>80.7</td>
<td>340</td>
<td>biotin synthase ( \text{Biotin synthase (EC 2.8.1.6) (Biotin synthetase) (Serratia} )</td>
<td>2.5e-119</td>
</tr>
<tr>
<td>SP:PI12996</td>
<td>79.7</td>
<td>342</td>
<td>biotin synthase ( \text{Biotin synthase (EC 2.8.1.6) (Biotin synthetase) (Escherichia} )</td>
<td>7.2e-120</td>
</tr>
<tr>
<td>GP:145425</td>
<td>79.7</td>
<td>342</td>
<td>biotin synthetase ( \text{Escherichia coli} )</td>
<td>1.5e-119</td>
</tr>
<tr>
<td>GP:12620127</td>
<td>79.4</td>
<td>342</td>
<td>biotin synthase BioB ( \text{(uncultured bacterium pCosHE2)} )</td>
<td>1.5e-119</td>
</tr>
<tr>
<td>OMN1NT03EC0855</td>
<td>79.4</td>
<td>342</td>
<td>biotin synthetase ( \text{(Escherichia coli O157:H7 VT2-Sakai)CGP13} )</td>
<td>5.1e-119</td>
</tr>
<tr>
<td>OMN1NT01YP1094</td>
<td>81.0</td>
<td>340</td>
<td>biotin synthese ( \text{(Yersinia pestis CO92)OMN1NT02YP2986 biot} )</td>
<td>8.3e-119</td>
</tr>
<tr>
<td>GP:12620099</td>
<td>79.5</td>
<td>340</td>
<td>BioB-\text{like protein (uncultured bacterium pCosFS1)}</td>
<td>9.5e-118</td>
</tr>
<tr>
<td>OMN1NT02EC0848</td>
<td>79.1</td>
<td>342</td>
<td>biotin synthesis,\text{ sulfur insertion? (Escherichia coli O157:H7}</td>
<td>2.2e-118</td>
</tr>
<tr>
<td>SP:Q47862</td>
<td>79.2</td>
<td>339</td>
<td>Biotin synthase ( \text{EC 2.8.1.6) (Biotin synthetase) (Ero.} )</td>
<td>3.6e-118</td>
</tr>
<tr>
<td>SP:PI12678</td>
<td>78.6</td>
<td>344</td>
<td>Biotin synthase ( \text{EC 2.8.1.6) (Biotin synthetase) (Salmonell} )</td>
<td>5.1e-119</td>
</tr>
<tr>
<td>OMN1VC1112</td>
<td>81.8</td>
<td>348</td>
<td>biotin synthase ( \text{(Vibrio cholerae El Tor N16961)GPI9655583} )</td>
<td>5.1e-119</td>
</tr>
</tbody>
</table>
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...

There are two parts:
1. Top gold box
2. Bottom alignment
One look at a match...

<table>
<thead>
<tr>
<th>66.0/79.7% over 343aa</th>
<th>Escherichia coli</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SPI12996</strong> Biotin synthase (EC 2.8.1.6) (Biotin synthetase), <em>Edit characterized</em></td>
<td></td>
</tr>
<tr>
<td><strong>PIRJC2517</strong>/<strong>SYECBB</strong> biotin synthase (EC 2.8.1.6) bioB [validated] - Escherichia coli (strain K-12) <em>Insert characterized</em></td>
<td></td>
</tr>
<tr>
<td>**GB</td>
<td>AAC73862.1</td>
</tr>
</tbody>
</table>

**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)

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>LacI</td>
<td>~~~C A A U C A G G G U G G U G A A U G U G A A A C C A G U A ~~~</td>
</tr>
<tr>
<td>Ribosomal L10</td>
<td>~~~C A U C A A G G A G C A A A G G C U A U G G C U U U A A A U ~~~</td>
</tr>
<tr>
<td>Ribosomal L7/L12</td>
<td>~~~U A U U C A G G A A C A A U U U A A A U U G U C U A U C A C U ~~~</td>
</tr>
</tbody>
</table>

(b)

- 3' end of 16S rRNA
- 5' end of mRNA
- Shine-Dalgarno sequence
- Anticodon of fMet-tRNA$_{f}^{Met}$
- fMet, Thr, Met, Ile

Genome View
Start site edited based on the Genome View

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.
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

Please go to:
http://manatee.tgen.org
Observations

- *R. denitrificans* lacks key Calvin cycle enzymes ribulose bisphosphate 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


[Arth 05] Arthur L. Delcher; GLIMMER Release Notes, Version 3.01 (Beta) 10 October, 2005


References


[Ed 99] Eddy S; Profile hidden Markov models; Bioinformatics; 14(9):755-763


[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 July 13, 2005


[1] Phototrophic Genome Project web site; http://genomes.tgen.org