

INTERNSHIP REPORT

*Roseobacter denitrificans*  
**genome annotation using  
Manatee**

An internship report presented in  
partial fulfillment of the requirement  
of the Professional Science Master's  
In Computational Biosciences

**Chaitanya R. Acharya**  
*Computational Biosciences Program  
Arizona State University*

**Dr. Rosemary Renaut, Ph.D.**  
**Dr. Robert E. Blankenship, Ph.D.**  
**Dr. Martin Wojciechowski, Ph.D.**  
*Internship Advisors  
Arizona State University*

NOT CONFIDENTIAL

Technical Report Number: 06- 02  
DATE: 04- 24- 06



## Abstract

Manatee is a web-based gene evaluation and genome annotation tool that can view, modify, and store annotation for prokaryotic and eukaryotic genomes. It is a stand-alone web application with a robustly designed relational annotation database that helps identify genes, thus making high quality functional assignments using a multitude of genome analyses tools. These tools consist of, but are not limited to Gene Ontology (GO) classifications, BLAST search data, paralogous families, metabolic pathways and annotation suggestions generated by automated analysis. *Roseobacter denitrificans* is a purple aerobic phototrophic bacterium (prokaryote), and it is a bacterium that performs photosynthesis in the presence of oxygen. The *R. denitrificans* genome sequence and annotation using the Manatee will greatly facilitate studies of the evolution photosynthesis, experiments on carbon dioxide fixation and production, and experiments to elucidate why the expression of photosynthetic genes in the aerobic phototrophic bacteria is independent of oxygen yet extremely sensitive to illumination. The genome of *R. denitrificans* is sequenced at the Translational Genomics Research Institute (TGen), and sent to the Institute for Genomic Research (TIGR), where the genome was automatically annotated and the resulting information will be returned to the annotation engine user. The information includes coordinates of open reading frames (ORFs) and RNAs; common name, gene symbol, EC numbers, TIGR roles, and GO terms for proteins; underlying search results including Blast-Extend-Repaze (BER), HMM, signalP, TMHMM, COGs, and paralogous families. Not all the information provided by the automated annotation is accurate; ORF start site, gene symbols, names and other attributes were manually checked for accuracy using the comparative tools developed by TIGR as well as other widely available bioinformatics tools. Entire *Roseobacter denitrificans* genome is now completely annotated and all the sequences were made available on the project web site (<http://genomes.tgen.org/>). *R. denitrificans* lacks key calvin cycle enzymes

(ribulose biphosphate, phosphoribulokinase) suggesting for a scattered loss of ancestral carbon- fixation in the  $\alpha$ -proteobacterial tree.

## **To My Beloved Parents**

*For your Eternal Love, Encouragement and Best Wishes.*

## ACKNOWLEDGEMENTS

I am extremely grateful to my internship advisor, Dr. Robert E. Blankenship for his support and valuable guidance throughout the project. I would also like to thank my mentor, Dr. Rosemary Renaut for her valuable guidance. I would also like to thank Dr. Martin Wojciechowski for kindly consenting to be on my committee and for his invaluable feedback. I can never forget the contributions made by my colleagues, Michael Lince, Wesley Swingley, Sumedha Gholba, Hector Ramos and Heather Matthies. My sincere thanks to Dr. Touchman for helping us out in sequencing the genome and answering few questions on annotation, and finally thanks to Jicheng Hao, who is a member of technical support at the Translational Genomics Institute (TGen).

This work was supported by a grant from the National Science Foundation (NSF), awarded to Dr. Blankenship, Dr. Jeff Touchman, Dr. Thomas Beatty, Dr. Carl Baver, and Dr. Michael Madigan. (NSF grant # MCB 0412824).

## TABLE OF CONTENTS

|   |           |
|---|-----------|
| <b>CHAPTER 1: Introduction</b>  | <b>6</b>  |
| 1. <i>Roseobacter denitrificans</i><br>genome – a brief overview                            | <b>6</b>  |
| 2. Significance of <i>Roseobacter</i><br><i>denitrificans</i> Genome Sequence               | <b>8</b>  |
| <b>CHAPTER 2: Glimmer System</b>  | <b>10</b> |
| <b>CHAPTER 3: Genome Annotation</b>   | <b>18</b> |
| 1. Basic Rules  | <b>18</b> |
| 2. Finding Real Genes –<br>Manatee’s perspective  | <b>18</b> |
| 3. Determining functions of<br>newly discovered proteins                                    | <b>20</b> |
| 4. ORF Descriptors  | <b>27</b> |
| <b>CHAPTER 4: Annotation Process<br/>Using Manatee</b>                                      | <b>29</b> |
| 1. Annotation Engine (AE)<br>project  | <b>29</b> |
| 2. Accessing Manatee  | <b>30</b> |
| 3. Role Category Breakdown  | <b>34</b> |
| <b>CHAPTER 5: Conclusions and<br/>Future Directions</b>                                     | <b>35</b> |
| 1. Biological Importance of<br>annotating <i>Roseobacter</i><br><i>denitrificans</i> genome | <b>42</b> |
| 2. Future directions  | <b>43</b> |
| <b>BIBLIOGRAPHY</b>   | <b>44</b> |
| <b>GLOSSARY</b>   | <b>46</b> |
| <b>APPENDIX: TIGR role categories</b>   | <b>51</b> |

## CHAPTER 1

### INTRODUCTION

In today's world, it is important to not only sequence any genome but also attempt to understand the resulting sequenced genome. Identifying genes (both protein coding and non-coding) and other regulatory regions is important in order to understand the metabolic profile of that organism. This process of unraveling the genome by identifying various elements and their roles is called as annotation. In this project, the genome of *Roseobacter denitrificans* has been sequenced and annotated using a tool called Manatee. This project is accomplished in collaboration with the Translational Genomics Institute (TGen) and The Institute for Genomic Research (TIGR). TGen helped in sequencing the entire genome, while TIGR helped in providing resources and infrastructure to annotate the raw sequence data. TIGR played a significant role in providing the annotation tool, Manatee to annotate the entire genome (See Fig 1).

#### ***1. Roseobacter denitrificans* genome – a brief overview**

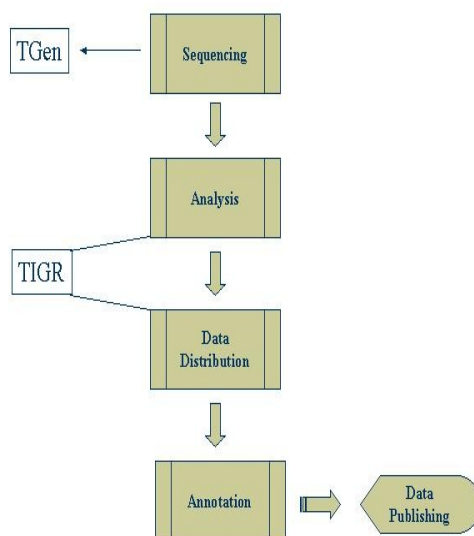
Annotation of *Roseobacter denitrificans* is a large and challenging task that cannot be accomplished without collaboration. TIGR offers a service to all prokaryotic sequencing centers, which is called the Annotation Engine (AE). Annotation is done at no cost with the help of AE. More on AE is discussed in the later sections of the paper. Using AE is very advantageous to scientists working on one or two genome projects. AE first executes the annotation in a format that promotes consistency of data types across all genomes. It also allows straightforward reincorporation of annotation data back into the Comprehensive Microbial Resource (CMR) data management system for display at the CMR at TIGR. This improves dissemination of the genome sequence data. Preliminary annotation is

allowed to be displayed locally on our web site. AE also aids in quality control of the sequence. The final annotation data set is provided to the annotators along with a tool (Manatee, here) for viewing and analyses. All of the proteins are later searched against hidden Markov Models (HMMs) using the ‘hmmpfam’ program. Two sets of HMMs used in the search are Pfam HMMs (Bateman, et al., 2000) and TIGRFAMs (Haft, et al., 2001). HMMs are useful for annotation since they are more sensitive and accurate than any pair wise alignment. HMM searches result in a score measuring the probability that the query protein belongs to the group of proteins used to build the model. Each HMM has an associated cut off score above which hits are known as significant.

Naming of the predicted proteins is mostly based on the HMM scores and BER matches. Proteins with a pair wise match to a hypothetical protein from another species, but no HMM hit, are named ‘conserved hypothetical proteins’ while proteins with no HMM or BER matches are named ‘hypothetical proteins’. In order to accurately annotate the genome, a protocol has been created and followed. This protocol will be discussed later in the forthcoming sections.

The Glimmer system, discussed in detail later, is used by the Annotation Engine to find genes in bacterial genomes. However, a summary on what happens when a raw genomic sequence is sent to AE, geared with Glimmer system is presented. Glimmer relies solely on the raw DNA sequence itself since it can be trained only on a raw DNA sequence

and consistently finds about 99% of all genes present in a genomic data, which are presented in a fully automated form. Once the open reading frames (ORFs) that are candidate genes have been established by Glimmer, several types of searches are performed



on the set of hypothetical proteins coded by these ORFs. BLAST-Extend-Repeat (BER) is the algorithm that is employed in these exhaustive searches. The results of this search are viewed both as pair wise and multiple alignments of the top scoring matches.

**Fig 1:** This is an overview of the project undertaken. First, *Roseobacter denitrificans* genome is sequenced at the Translational Genomics Institute (TGen) followed by sequence analysis and data distribution performed in collaboration with The Institute of Genomic Research (TIGR). The data is represented as open reading frames (ORFs) and predicted protein sequences, which are annotated by Manatee. The annotated sequences are made available for public view in the later stages.

## 2. Significance of *Roseobacter denitrificans* Genome Sequence

*Roseobacter denitrificans* is a representative aerobic phototrophic bacterium, and its genome sequence is important for three compelling reasons:

1. ***To understand the evolutionary genesis of photosynthetic genes:***  
A Genome sequence-based approach may provide insights by expanding recent bioinformatics approaches to include the aerobic phototrophic bacterium (Raymond et al. 2002, 2003). Whole genome comparisons could also help clarify the true evolutionary position of the aerobic phototrophic bacteria.
2. ***To understand the pathways of carbon dioxide fixation and production:*** The genome sequence will reveal the genetic potential for autotrophic fixation of CO<sub>2</sub> via the Calvin- Benson- Bassham cycle (Tabita 1995) or, conceivably, the reverse citric acid cycle (Ormerod 2003). The genomic sequence would also help construct metabolic pathways *in silico* (Larimer et al., 2004), which could be tested in biochemical and molecular biology experiments. The question of autotrophy is central to our understanding of the significance of the vast numbers of aerobic phototrophic bacteria as either net consumers or producers of CO<sub>2</sub> in the global carbon cycle.

3. *To understand the light and oxygen signal transduction in gene expression:* Although high light intensity represses photosynthesis in a variety of organisms, the extreme repression found in the aerobic phototrophic bacteria offers an opportunity to use this exaggerated response to facilitate an understanding of pathways and mechanisms that are subtle, and hence difficult to study effects in other species. The genome sequence could be used to identify potential new light sensors and signal transducers. Also, because the aerobic phototrophic bacteria produce photosynthetic apparatus under aerobic conditions, their study would contribute to a general understanding of oxygen regulation. The genome sequence would also help understand both the light- and oxygen-responsive pathways by cloning and over-expressing genes for biochemical and biophysical analysis of purified proteins, as well as by gene disruption.

*R. denitrificans* is easy to cultivate in the laboratory, and it is a model phototrophic organism. More work can be done on the respiratory and photosynthetic electron transfer pathways in this species. Also, it is the only aerobic phototrophic bacterium that is capable of anaerobic growth, by the use of nitrate as terminal electron acceptor (Yurkov and Beatty 1998), which will facilitate subsequent studies on the effects of oxygen on photosynthetic and other metabolic processes. The GC content of this organism is estimated at 59% and the genome size is predicted to be approximately 4 Mb (Mega base pairs). Thus, there should not be any significant technical obstacles to obtaining and annotating its sequence.

The genome sequencing and the processing of *R. denitrificans* will facilitate studies of the evolution of photosynthesis, experiments on carbon dioxide fixation and production, and experiments to elucidate why the expression of photosynthesis genes in the aerobic phototrophic bacteria is independent of oxygen yet extremely sensitive to illumination.

Understanding fundamental aspects of the origin and dispersion of photosynthetic genes, the global carbon cycle, and general biological mechanisms for sensing and transducing the environmental signals of oxygen and light is imminent in the near future.

## **CHAPTER 2**

### **GLIMMER SYSTEM**

The first major step after the completion of genome sequencing is to identify probable open reading frames or genes. 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. Glimmer is trained on a raw DNA sequence since it relies on nothing but DNA sequence itself. Glimmer was used on many microbial genomes, and without any surprise, consistently finds over 99% of the genes in a fully automated fashion. Glimmer is available in many versions; the latest being Glimmer 3.01 Beta Version is the latest one. Not all features are yet fully implemented and tested yet. The algorithm at the core of the Glimmer is

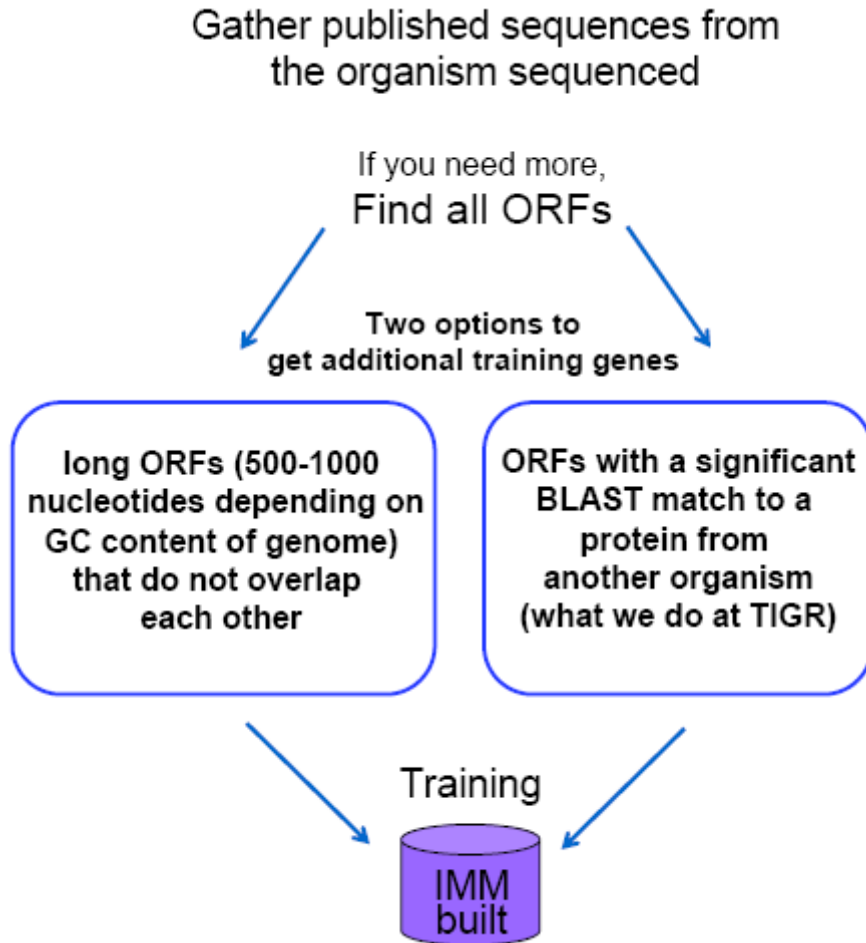
an Interpolated Markov Model (IMM), which is a special kind of Markov chain. A Markov chain calculated the statistical information about any sequence by computing the conditional probability  $P(x|S)$  that a nucleotide 'x' appears after a sequence 'S' (Eddy S 1996). The most effective Markov chains are those that look at triplet codons or multiples of triplets because DNA encodes proteins with triplets. So, second- or fifth- or eighth- order Markov chains work well because they are computing statistics based on codons, dicodons and tricodons respectively. There are some limitations to computing statistics using Glimmer software. As the order of the model increases so does the amount of data to compute. This takes up a lot of virtual memory and time since huge data sizes contain exponentially more probabilities that need to be estimated. A typical bacterial genome contains enough data sufficient to train a fifth- order model, but no larger. IMMs get around this limitation by considering much longer sequences, but only use statistics where sufficient data are available. All probabilities associated with a fixed length sequence will not be estimated. Even in the presence of insufficient data to train an eighth- order model, there may be some short eight base pairs that occur sufficiently frequently that some computational system could be used to estimate the probability of the next base. The IMM implemented by Glimmer calculates all the probabilities for all Markov chains starting from the zeroth- order through eighth- order, but only those for which sufficient data is available.

The gene- density in bacterial genomes is very high with about 90% coding regions on an average. Any trivial gene finder can quickly find at least half the genes in no time. The genomic sequence on which Glimmer is trained will find an open reading frame that has at least a pre- selected length. All those ORFs that are over- lapping will be eliminated. The remaining long ORFs are likely to be genes, since they are too long to occur by chance and since they do not overlap any other ORF. These remaining long ORFs are then subjected to other statistical analyses including finding

GC-content (high GC-content is a flag for presence of an ORF), and homology searches using BLAST against a protein database. All the homologous sequences can then be used to again train Glimmer. Even though this is a very effective system, it takes up a lot of CPU memory (Arthur et al 1999).

Using Glimmer is a two-part process; Glimmer has to be trained first on an organism's sequence that has been already annotated followed by running the trained Glimmer on the sequence to be annotated. Choosing a training set is the most important step before they are 'Glimmered'. After obtaining the genomic sequence, performing prior homology searches would give us more training sets. However, there should be a threshold level after which ORFs should not be selected into the training set. Versions 2 and 3 of Glimmer (called Glimmer2 and Glimmer3 respectively, for simplicity) scored ORFs based on the homology, and for those scoring above the threshold value for the candidate training set. Glimmer2 examines pair wise overlaps between candidates, and based on some set rules ORFs are either eliminated or their start sites are adjusted. This process iterates continuously until no further changes occur. The rules set by Glimmer2 cannot finally resolve an overlap between two ORFs, and the final prediction list contains these with comment tags concerning ORFs which are not fully resolved. This increases the possibility of having more false-positives in the final list. However, Glimmer3 applies an HMM-like algorithm that selects high scoring ORFs and their start sites. The final list of predictions does not contain any overlaps, and thus is devoid of any comment tags. In general, the number of predictions is fewer when compared to those of Glimmer2 thus reducing the number of false-positives. Glimmer3 scores ORFs in reverse direction, i.e. from 3' to 5' direction, which improves the accuracy of the scores near the start codon of genes because the trailing context of the interpolated context model (ICM) is in the coding region of the gene on which it has been trained (See Fig 2) (Arthur et al 1999).

Once Glimmer identifies the candidate genes, several types of searches are performed on the set of hypothetically encoded proteins. Each hypothetical protein is compared to known proteins sequences (homology search) in an internal non-redundant amino acid database that is made of all proteins available at GenBank (<http://www.ncbi.nlm.nih.gov>), SWISS-PROT (<http://www.expasy.ch/sprot>) and TIGR's internal protein database, EGAD (<http://www.tigr.org>). BLAST-Extend- Repraze is the search algorithm employed. A BLAST search (Altschul, et al., 1990) (<http://blast.wustl.edu>) is done first on each protein against a non-redundant amino acid database and the results are stored in a mini-database. A modified Smith-Waterman alignment (Smith, 1981) is then performed on the protein against the mini-database of BLAST hits. The gene is extended 300 base pairs upstream and downstream of the predicted coding regions in order to identify potential frame shifts or point mutations. The program will continue alignment if there exists a significant homology to a match protein, and extends into a different frame from that predicted, or extends through a stop codon. The results can be viewed as pair-wise and as multiple alignments of the top scoring matches.



**Fig 2:** Comprehensive computer algorithm in the Glimmer system

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 (See Fig 3). There are six possible frames for translation in every sequence of DNA, three in the forward direction and three in the reverse direction on the DNA sequence.

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

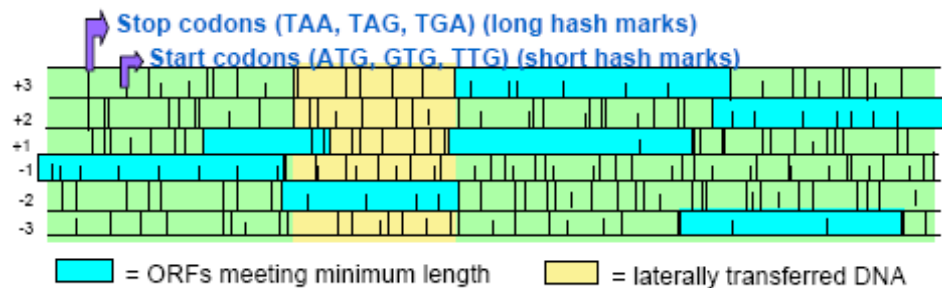
---

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

---

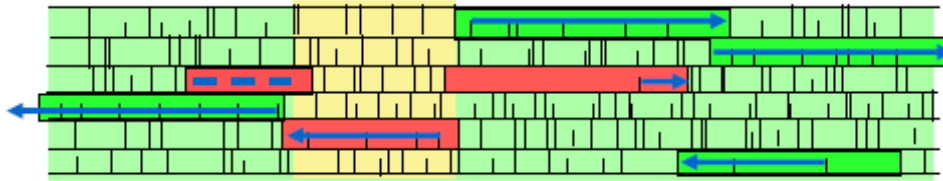
**Fig 3:** An example of six reading frames in a DNA sequence with the DNA sequence given at the top followed by the six frames.

All of the proteins from the genome sequences are also searched against hidden Markov models (HMM) using program ‘hmmpfam’. So, a two-fold homology search is performed to identify the probable ORFs. The Glimmer program used by TIGR (Glimmer 3.01 Beta Version) comes up with a six-frame translation map for a region of DNA. This translation map represents the start and stop codon, effectively giving the length of the ORF (See Fig 4).



**Fig 4:** Six-frame translation map for a region of DNA

The coding sequences resulting from the candidate ORFs are represented by the arrows, going from start to stop, the dotted line representing an ORF with no start site. An ORF with no possible start site is not recognized as a gene. A long ORF doesn't necessarily result in a long putative gene (See Fig 5). The probabilities scored by Glimmer form the statistical model of what a real gene looks like in the given organism. The scoring of the ORFs plays an important role in forming this statistical model.



**Fig 5:** Scoring of different ORFs in the genomic sequence; green regions represent high scoring ORFs, red regions correspond to the ORFs that are not scored well, and yellow regions represent laterally transferred DNA.

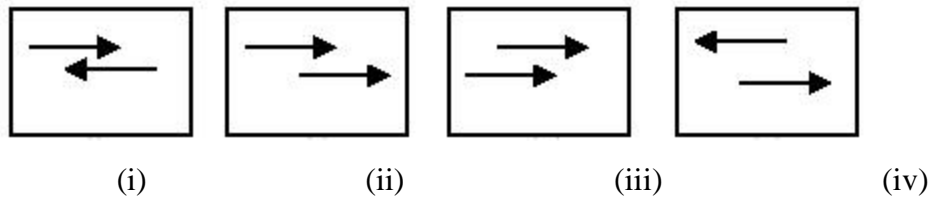
Based on the scoring of the ORFs, they are color coded for easy identification. Green ORFs are scored well to the model, while red ORFs scored less well. The Green ORFs are chosen by Glimmer as the set of likely genes and numbered sequentially from the beginning of the DNA molecule on which they reside. ORFs in the region of lateral transfer, often real genes, often will not be chosen since they do not match the model built from the patterns of the genome as a whole. While viewing a six-frame translation, the genes are represented as arrows drawn either above or below the six-frame translation (See Fig 6).



**Fig 6:** Arrows representing the orientation of the ORFs after the six-frame translation. In the figure above, ORF00004 overlaps ORF00003.

Finally, genes are mapped to the underlying genome sequence via coordinates. Each gene is identified by two coordinates: end5 (the 5 prime end of the gene) and end3 (the 3 prime end of the gene). Forward genes have  $\text{end5} < \text{end3}$  while reverse genes have  $\text{end5} > \text{end3}$ .

From the figure above, it is evident that the scoring ORFs/genes sometimes overlap. When two potential genes A and B overlap, the overlap region is scored. If A is longer than B, and if A scores higher on the overlapping region, and if moving B's start site will not resolve the overlap, then B is rejected (See Fig 7). The system attempts to move the locations of the start codons much more aggressively in the following way. Four different cases of gene-orientations are considered if gene A scores higher (when compared to gene B in this example).



**Fig 7:** Different cases of gene overlaps tested by the Glimmer system. The top and the bottom arrows represent gene A and gene B. The head of the arrow is the 3' end while the tail corresponds to the 5' end. For example, in (i) gene A is oriented from 5'-3' direction and gene B is anti-parallel to gene A.

In the first case, it is assumed that postponing the start site does not remove any overlap. If A is significantly longer than B (as determined by a program parameter) then B is rejected. Otherwise, both A and B are called genes, with an annotation that there was a doubtful overlap. In the second case it is assumed that only moving the start site of B can resolve the overlap, if it can be moved. If not, and if B is significantly shorter than A, then B is rejected. Otherwise, both A and B will be listed as genes with a note indicating the overlap. When a start site is moved, the system shortens the predicted gene by shifting the start location to the next available start codon. This is performed in continuous iterations until the

overlap is resolved, and the resulting gene is longer than the minimum gene length (a parameter during Glimmer analysis). In the third case it is assumed that only moving the start site of A can resolve the overlap. Since A is scored higher than B, the start site is only moved when the overlap is a relatively small fraction of A's length. If adjusting A is not successful, B is rejected. In the final case, it is assumed that both start sites can be adjusted. The start site of B is first moved until the overlapping region is scored higher for B. The start site of A is then moved until it scores higher. This process is continued until either overlap is eliminated or no further start site adjustments can be made (Arthur et al 1999).

An additional step is taken by the Glimmer to help find genes that were missed due to high scores resulting from the independent probability model. The independent probability model is used by the Glimmer system to compete against the IMMs used to score all the six reading frames. Its main purpose is to serve as a model of non-coding DNA. In order to be called a gene, an ORF must score higher than the independent model as well as the other five reading frames. The process of evaluating the overlaps is performed in an iterative fashion in order to avoid rejecting genes unnecessarily. For example, in cases where ORF B is rejected due to ORF A, and B in turn causes C to be rejected, only B is rejected and not C and A. Thus, the rejection phase is performed in multiple stages, checking for overlaps over and over again. Even after the Glimmer analysis some overlaps remain. They have to be manually checked by the 'annotators'. All the information that is obtained from the Glimmer is now stored in a database that can be accessed by Manatee (Arthur et al 1999).

## **CHAPTER 3**

### **GENOME ANNOTATION \***

#### **1. Basic rules**

Genome Annotation is the process by which putative genes are located and identified on a newly sequenced genome. Annotation is a multi-step process; the first step is to find all the open reading frames (ORF), which are RNA- and protein-coding regions. Identifying the gene products by a robust homology search follows this. The homology search is also used to understand the physical and functional characteristics of the gene products. After identifying the above-mentioned features, the role played by that specific gene product, if it is a protein, would be evident. Thus, understanding and identifying the overall metabolic profile of the organism would complete the overall annotation process. The elements of the annotation process include gene finding, homology searches, functional assignment, ORF management and finally making this data available to the public.

---

All the images in this section are obtained from Manatee's web site at TIGR  
<http://manatee.sourceforge.net/pdf/overview.pdf>

Annotation can be done both manually and by using high performance computers that implement sophisticated algorithms. Manual annotation would take up a lot of manpower and time. Computers do a fair job at preliminary annotation. However, high quality annotation requires manual review of the preliminary annotation.

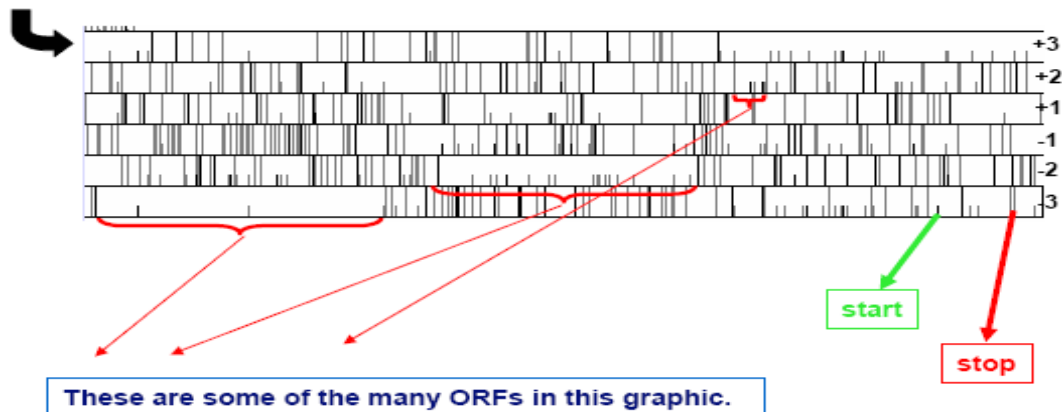
## **2. Finding Real Genes – Manatee’s perspective**

In order to understand the concept of genes, one should understand an open reading frame or ORF. During the annotation process, the computer algorithm is designed to find a candidate region that extends from one stop codon to another. This ‘inter- stop- codon’ sequence or the candidate region is called an open reading frame (ORF). Many ORF- finding tools can identify ORFs as they can occur easily by chance. Stop codons are usually AT-rich regions. GC-rich DNA has on average more, longer ORFs while AT-rich DNA has on average fewer, shorter ORFs. In order to obtain genes from ORFs, start codons must be identified. In eukaryotes, the start codon is almost always ATG that codes for Methionine. However, prokaryotes could have any of the three codons, ATG, GTG (codes for Valine) and TTG (codes for Leucine). A gene is identified by a start and a stop codon; it has a biological significance, be it a protein or RNA. One should remember that a gene product is not always a protein. There are some non- coding RNAs that have different roles in the genome (TIGR; Michelle Gwinn).

Glimmer is the gene- finding tool that is most commonly used by the Institute for Genomic Research (TIGR). The Glimmer system for microbial gene identification finds ~97- 98% of all genes in a genome when compared with the published annotation. A significant proportion of the genes missed by the system appear to be hypothetical proteins whose existence is only supported by the predictions of other programs. Please refer to the section in this paper that describes the algorithm.

Distinguishing a gene from a random ORF is the goal of the annotation process (TIGR; Michelle Gwinn).

In order to visualize the genes within the context of their neighbors along a DNA sequence, the sequence is often represented as a six-frame translation with six possible frames in every sequence of DNA, three in the forward (+) direction and three in the reverse (-) direction. These six frames of translations are represented as horizontal bars with long vertical (stop codons) or short vertical (start codons) on them. An ORF is shown existing between two long vertical bars (See Fig 8). The Glimmer system finds possible gene candidates and their coordinates for further investigation.



**Fig 8:** Image showing six translation frames in six horizontal bars and long & short vertical bars representing stop and start codons respectively.

### 3. Determining functions of newly discovered proteins

As new ORFs are discovered by the Glimmer system, they are stored in the database along with the protein sequences. Determining the functions of these new proteins is a challenging task. There are two main ways to accomplish this task- 1. Experimental characterization of the proteins, and 2. Homology searching.

Experimentally, one can look at mutant phenotypes and perform enzyme assays to determine the function of a protein. On a large scale, expression

of the mRNAs could be checked using complimentary DNA microarrays or by checking for protein expression using protein arrays. However, these are complicated methods that demand more time and money. Homology searching is also done by comparing sequences of unknown function to the sequences of known function. High identities (at least >35%) prove that the sequences share their functions. However, there are occurrences of proteins where one amino acid substitution could lead to change in function. All the functional assignments made after sequence alignments should be considered 'putative' (suggested) until confirmed by experimental characterization. One should never be misled by 'identity' and 'similarity'. Identity means exact amino acid matches; similarity means the amino acids share similar structure and thus could carry out similar roles in the protein (TIGR; Michelle Gwinn).

It is also important to know what proteins in the Manatee's search database are characterized. All the characterized proteins are stored in a "Characterized table" within the Manatee database, and they are also designated a confidence status. Annotators see this information as a color-coded output, where every color has a distinct meaning.

A green color means full experimental characterization, red color means characterized by Swiss-Prot (<http://ca.expasy.org/sprot/>) by an automated process, a sky blue color means partial characterization, an olive color means trusted to be characterized (it is used when multiple extremely good lines of evidence exist for function but no experiment has been done), a blue-green color means only a fragment/domain has been characterized, a fuzzy gray color means void (used to indicate that something that was originally thought to be characterized is not really characterized), and a gray color means that the sequence exists in the 'OmniUM' (database that underlies TIGR's CMR) only, and therefore represents automated annotation. The characterized protein table in the Manatee database does not contain all characterized proteins. Blast Extend- Repraze (BER) searches are performed on all the candidate genes



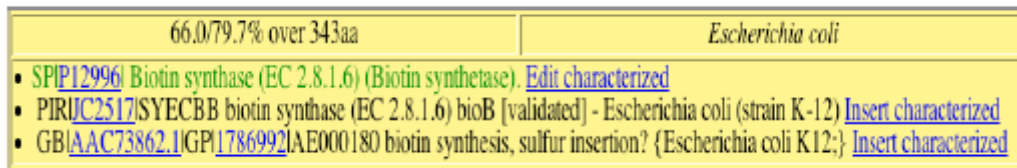
(or proteins) and the resulting alignments are stored in the Manatee database.

**Fig 9:** This is the display of a BER alignment\* web page of a candidate ORF.

The BER alignments are stored in a mini-database and the closest matches are displayed in a table. This table is present in the Gene Curation Page (GCP)

of an ORF. More information on GCP is in the ‘Annotation- process using Manatee’ section of this paper.

All the information regarding the alignment and the sequences aligned with (called query and subject sequences) are listed on this page. The alignment page shows statistics, start codons and similarities/identities with the aligned protein.



**Fig 10:** The background color of the top box is gold if the protein is characterized. Percent similarity and identity and the organism from which the protein comes (*Escherichia coli*) are also listed along with the sequence accession ids, sequence names and their Enzyme Commission (EC) numbers.

The background of the alignment page is black with the top box colored gold if the known protein is characterized (See Fig 9). 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

Only one specific BER alignment is shown. A host of them (based on the homology) would be present in a table on the ‘Gene Curation Page (GCP)’.

names for all the instances of the match protein from the source databases (See Fig 10). 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 (depending on the characterization status) if that is the accession number that is entered into the characterized table of the Manatee database. This tells the annotators which link they should follow to find experimental characterization information. If any one accession number for the match protein is in the characterized table, the header turns gold. There are links at the end of each line to enter the accession number into the characterized table or to edit an already existing entry in the characterized 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]
```

**Fig 11:** Alignment header – more header information here like gaps, insertions and deletions, frame- shifts and primary frame of translation.

It is most important to look at the range over which the alignment stretches and the percent identity before one makes a decision over a ‘good match’ (See Fig 11). The top line shows the amino acid coordinates over which the match extends for the newly discovered protein. The second line shows the amino acid coordinates over which the match extends for the match protein, and 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/ORF. The primary frame is the one starting with the nucleotide one of the gene. If there are no frame shifts in the protein, all of the matching amino acids should be in the first frame.

```

CRF04313( 3 - 260.3 of 260.3 aa)
CMNI|NPL03PA02011(2 - 265 of 267) probable transcriptional regulator (Pseudomonas aeruginosa)
%Match = 28.7
%Identity = 59.5 %Similarity = 75.2
Matches = 156 Mismatches = 51 Conservative Subs = 41
Gaps = 4 InDels = 15 Frame Shifts = 1
Primary Frame = 1 [162, 96, 0]

```

**Fig 12:** Any Frame shifts in the amino acid sequence are listed for further investigation.

Any presence of frame shifts would be highlighted as “Frame Shifts = #” to indicate how many frame shifts are present (See Fig 12).

Annotator should also pay attention to the alignment of the amino acid sequence because it is imperative to identify the correct start site for the newly discovered protein.

```

ataaaaagttatctcgccgtacggaggttgccaggttagcaaccgggtgca
gaacataggctcatatgaagagaatacttctctaatttacagttagaaaaca
gataaaaagtaggggattttgggaacacatgggtcaatacctcctagcttc
-1      10      20      30      40
R*NTKIKGCSMSQLQVRHDWKREEIEALFALPMDLLFKAHSIHREEYDPN
| | | | | | | | | | | | | | | | | | | | | | | | | | | |
MAHRPRWTLSQVTELFKPLDLLFEAQQVHRQHFDPF
10      20      30

```

**Fig 13:** Alignment of amino acids is displayed here. It should be read ‘AGATAA...’ every three amino acids coding for an amino acid. This alignment

shows that the first codon is ATG (coding for Methionine) at position # 1. ‘TTG’ is also seen as a start codon since that codes for ‘Leucine’ (As this is a microbial genome). The vertical bars represent identical amino acids while dotted lines represent similar amino acids.

In these alignments, the codons of the DNA sequences are read down in the columns with the corresponding amino acid underneath. The numbers listed by the alignment correspond to the positions of the amino acid. Position 1 is the first amino acid of the protein. Negative amino acid numbers indicate positions upstream of the predicted start of the protein. Vertical lines between amino acids represent identical amino acids whereas dotted lines indicate similar amino acids (See Fig 13). The start sites are color coded- ATG is green, GTG is blue and 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. When the similarities extend outside the coordinates of the protein coding sequence, a frame shift or a point mutation is detected (TIGR; Michelle Gwinn).

**Fig 14:** This is a display of BER skim as seen in the GCP. The matches are all listed in this table with a certain p- value that determines the authenticity of the match.

| BER SKIM                         |       |                                   |  |  |
|----------------------------------|-------|-----------------------------------|--|--|
|                                  |       | <a href="#">View BER Searches</a> | search date: Wed Oct 23 12:59:20 2002                        | <a href="#">submit</a>   <a href="#">E</a> |
| <a href="#">Refresh Searches</a> |       |                                   |  |  |
| 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) (Emerica)   | 2.5e-119                                   |
| SP:PI2966                        | 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-Saka)/CGP1.3 | 5.1e-119                                   |
| OMNI:NTL01YPI094                 | 81.0  | 340                               | biotin synthase (Yersinia pestis CO92)/OMNI:NTL02YP2986 hiot | 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? (Escherichia coli O157:H | 2.2e-118                                   |
| SP:Q47862                        | 79.2  | 339                               | Biotin synthase (EC 2.8.1.6) (Biotin synthetase) (Emericia h | 3.6e-118                                   |
| SP:PI2678                        | 78.6  | 344                               | Biotin synthase (EC 2.8.1.6) (Biotin synthetase) (Salmonell  | 5.1e-119                                   |
| OMNI:VCI112                      | 81.8  | 348                               | biotin synthase (Vibrio cholerae El Tor N16961)/CGP9655583lg | 5.1e-119                                   |

‘Non- identical Amino Acid (NIAA)’ is TIGR’s protein file used for searching. 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. If the score is between the trusted cutoff and the noise cutoff of an HMM, it deserves closer examination before excluding the HMM from consideration; if the HMM is used as the basis of annotation, a score between trusted and noise provides sufficient evidence to append ‘putative’ to the name of the ORF. HMM evidence is observed on GCP. Ideally, a match should extend across the entire HMM. However, matches having a >20% length discrepancy compared to the HMM are flagged in red. The GCP displays hits to HMMs built by TIGR (HMM accession numbers begin with ‘TIGR’) and Pfam (HMM accession numbers start with ‘PF’). TIGR classifies TIGR and Pfam HMMs into fifteen isologies. Each isology defines a specific database match.

| ISOLOGY   | DEFINITION   |
|-----------|--|
| Equivalog | a collection of proteins that share function back through their last common ancestor |

|                         |   |
|-------------------------|---|
| Equivalog_Domain        | a region with an assignable conserved function that with some regularity shows up in different protein architectures. It can be the sole functional domain in a protein or it can be one of several functional domains in a longer, multifunctional protein |
| Hypoth_Equivalog        | a family of uncharacterized proteins hypothesized to be equivalogs  |
| Hypoth_Equivalog_Domain | a region with a hypothesized conserved function that, with some regularity, shows up in different uncharacterized proteins architectures. It can be the sole domain in one of these proteins or it can be one of several domains in longer proteins         |
| Paralog                 | a family whose members are all drawn from the same (or very closely related) genome   |
| Paralog_Domain          | a region shared by members of a family that are all drawn from the same (or very closely related) genome  |
| Subfamily               | formally, a branch of a superfamily. Subfamilies often include fairly   |

|             |  |
|-------------|--|
|             | <p>closely related proteins with functional heterogeneity. In practical terms, this iso_type is warning against trying to interpret the set of proteins as equivalogs</p>  |
| Superfamily | <p>a collection of proteins with the same domain structure, encompassing all homologs, usually including proteins with at least two different functions</p>  |
| Domain      | <p>This is the broadest isology; it connotes a region of similarity shared by proteins homologous over portions of their length, encompassing all homologs, usually including proteins with at least two different functions. The domain itself is not presumed to have the same function in all instances; the HMM describes only a sequence similarity. Contrast to 'equivalog_domain' HMMs, where conserved domain function is presumed</p> |
| Repeat      | <p>a region that is found in multiple copies in members of the HMM</p>   |
| Pfam        | <p>a Pfam model not yet preliminarily classified by isology type at TIGR.</p>  |

|                       |  |
|-----------------------|--|
|                       | These should be approached with skepticism   |
| Pfam_Equivalog        | a Pfam model that appears to find only equivalogs, but cutoffs are probably too lenient for automated annotation. Beware of 'false positive' hits              |
| Pfam_Equivalog_Domain | a Pfam model that appears to find only equivalog_domains, but again cutoffs are probably too lenient for automated annotation. Beware of 'false positive' hits |

**Table 1:** Table showing different isologies with their definitions. Source: Genome Annotation and Naming Guidelines ([http://manatee.sourceforge.net/pdf/Gene\\_Annotation\\_Naming\\_Guidelines.pdf](http://manatee.sourceforge.net/pdf/Gene_Annotation_Naming_Guidelines.pdf))

The Pfam HMM ‘gathering’ score is analogous to a TIGRFAM ‘trusted’ score, but is less rigorously assigned. An ORF that scores above gathering but below trusted to a Pfam HMM should be treated with caution. Annotators should always be on the lookout for new characterized matches by checking the literature on candidates that are in the BER Skim (whether flagged by a DB\_PARSE, flagged as ‘experimental=1’ or ‘experimental=- 1’, or chosen by the annotator). SwissProt accessions in the BER alignment link to a SwissProt page of relevant literature for that protein sequence. Annotators can also scan PubMed (<http://www.pubmed.com>) for literature related to particular gene name using appropriate query in NCBI’s Entrez PubMed search field (TIGR; Michelle Gwinn).

#### 4. ORF Descriptors

Finally, each gene is annotated by assigning as many descriptors as are relevant to each gene. In a practical sense, annotation amounts to populating database tables with descriptions of the gene. 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.

The most specific name of a predicted protein sequence is justified by the evidence. For example, enzymes are assigned the standard enzyme commission (IUBMB) name. Gene symbol would be a three or four letter word (for example, `ftsZ`). For duplicate genes, hyphenated numbers are not used, since this form is classically reserved for alleles. Instead, the specific number is added to the `gene_sym`. However, this is not done by the annotator. The Enzyme Commission (EC) number is a four part numbering scheme for representing specific enzyme activity. This system is curated by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) in consultation with the IUPAC/IUBMB Joint Commission on Biochemical Nomenclature (JCBN). Multiple EC numbers are separated by spaces.

The Gene Curation Page (GCP) provides space for comments about any protein. Caution has to be exercised when writing these comments as anything written in these boxes would be read by the public. TIGR roles describe the biological role of the protein. Assignment of a protein to a TIGR role is achieved by assigning the id number of the TIGR role. The gene ontology (GO) classification system classifies a protein sequence based on their molecular functions, the biological processes in which they are involved, and the cellular components in which they live and act. These three aspects of a protein are captured in the three controlled ontologies of the GO system.

## **CHAPTER 4**

### **ANNOTATION PROCESS USING MANATEE**

“Manatee is a web-based gene evaluation and genome annotation tool that can view, modify, and store annotation for prokaryotic and eukaryotic genomes” (<http://manatee.sourceforge.net/>). All the information (or data) that is obtained after running Glimmer system and various other programs is stored in a database by TIGR. 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. The Manatee project, which was created by the bioinformatics department at TIGR in Rockville, MD., 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.

#### **1. Annotation Engine (AE) project**

The process of annotation begins with the genome being sequenced and provided by TIGR’s ‘Annotation Engine (AE)’ project, funded by the Department of Energy (DOE). The AE project was made available by TIGR in order to bring various genomic tools of genomic science to the modern researchers in need to annotating prokaryotic sequences. Any prokaryotic genome can be submitted free-of-cost to the AE. The AE project has two components (TIGR; <http://www.tigr.org>):

1. Production of the search results and automatically generation of annotation in a MySQL database and associated files from TIGR’s automated pipeline

2. An open web-based interface for viewing and editing the annotation data.

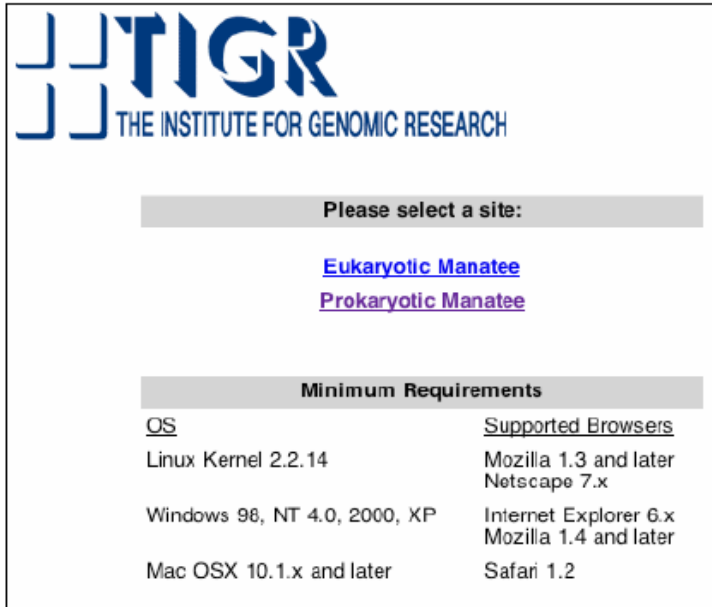
It has become easy to sequence an organism's genome because the cost of sequencing has dropped and many companies are now offering on-demand sequencing of genomes. Scientists who have modest annotation goals do not wish to go through the manual annotation process, which is cumbersome and requires significant infrastructure and tools. So, TIGR, with the help of a three-year grant from the DOE, assists by providing access to its infrastructure which saves a lot of time and the expense of reproducing similar systems at their sites.

Once the automated annotation process is complete, a MySQL database and related search files will be generated, packed and placed on a private FTP site from where Manatee can be downloaded and installed to manipulate the data.

## **2. Accessing Manatee**

Manatee can be downloaded and installed from its official web site <http://manatee.sourceforge.net/downloads.shtml> . Manatee draws information from its underlying database for its displays, and it sends information entered by the annotators to the same underlying database for storage. Multiple users (or annotators) can access the same database from different computers. In this paper, users and annotators will be used interchangeably. In order to log into the Manatee within TIGR one must have a Sybase account and password. This particular genome annotation project is being done through Translatory Genomics Institute (TGen), and TGen issues the user an account and a password to work with Manatee. In this case, Sybase account and password is not needed, and Manatee is being accessed through TGen. Just by typing <http://manatee.TGen.org> the first web-page can be reached. The users are directed to choose a site

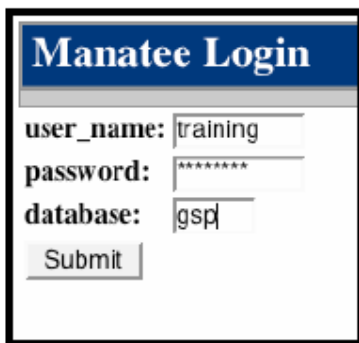
based on the type of the genomes. *Roseobacter denitrificans* is a prokaryote, and so 'Prokaryotic Manatee' is selected.



**Fig 15:** The first web page of the Manatee web site is displayed here (<http://manatee.TGen.org>). The annotators can choose a site based on the type of the genome- eukaryotic or prokaryotic. *Roseobacter denitrificans* is a prokaryote, hence 'Prokaryotic Manatee' is chosen.

This takes the user to a login page. When logging into Manatee, one must enter a

username, password and the appropriate database name. TIGR database names tend to be three to five letter codes.



**Fig 16:** The Manatee login page is displayed here. The database accessed in this case is 'gsp', which stands for genome *Shewanella putrefaciens*\*. This image is obtained from the demo version of Manatee that is available online at TIGR's Manatee web site, <http://manatee.sourceforge.net/demos.shtml>

The databases assigned during this annotation project were 'ntrd01' and 'ntrd02'. Successful login will lead to 'Welcome to the Manatee' web page. Here several menu options and some search options are available for probing into the genome. This Welcome page is a portal from where a user can go to many nooks and corners of the Manatee web site. 'Annotation Tools' web page allows a user to access the ORFs on the entire chromosome or a specific plasmid. In this project, there were four

---

\*The species name '*putrefaciens*' was changed to '*oneidensis*' by the research community during annotation of this genome.

plasmids, including pTB1, pTB2, pTB3 and pTB4. This is an easy way to look at an ORF, if the user knows the ORF identifier within the database. Otherwise, there are other ways to get to the ORF. The ‘Genome Summary’ web page is the most essential page for the project. The next two web pages, ‘Genome Ontology’ and ‘Genome Properties’ are not used in the project to keep the annotation process simple. ‘Genome Viewer’ web page gives a snapshot of all the ORFs, their orientations on the Chromosome, their locations and any overlaps. Genome viewer is often visited in order to identify insignificant ORFs.

**Fig 17:** “Welcome to Manatee” web page provides many options to the users to choose from.

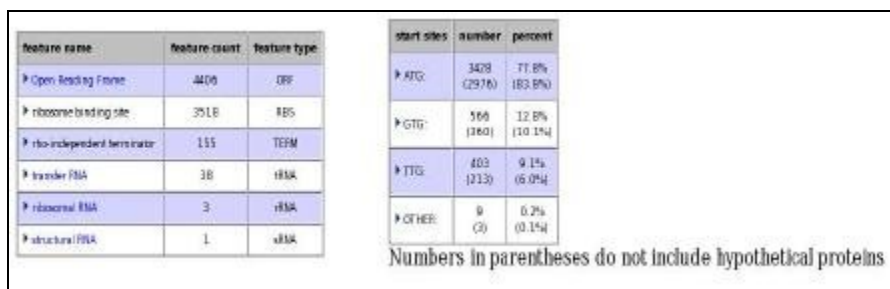


The upper right hand corner of the every Manatee page has a web frame that has a button (hyperlink) “Home” that takes the user back to the Welcome page. Beside the home button, there is an important piece of

information that tells the user the database being accessed and the username (login name). The username is an active link that leads to the login page. There are other search options like, ‘Access Gene Curation Page’ that allows a user to access the Gene Curation Page (GCP) of a specific ORF (if its identifier is known prior search). The GCP displays almost everything known about a protein. This GCP can be reached either by entering a feat\_name or locus id into the search box and then clicking

‘submit’. A feat\_name is an internal identifier given to each gene (or ORF) in a genome. They are initially assigned by the Glimmer and generally are numbered sequentially from the beginning of the genomic sequence given to the Glimmer. They have the format ‘ORF#####’ where ##### is a five-digit zero padded number. Locus ids are assigned to the proteins at the end of the annotation process. These ids are numbered sequentially from the origin of replication of the genome. Loci are unique accession identifiers that are used during public release and display of the proteins (TIGR; Michelle Gwinn). ‘Search Genes by Gene Name’ is self-explanatory. This is a keyword-based search for the common names that are given to the ORFs or the protein after the annotation process (both automatic and manual). A gene name can be searched against a database. This keyword (gene or protein name here) would be treated as though it has wildcards flanking it. The results would include the names containing the keyword as an individual word or names that contain words that contain the keyword. A keyword search for ‘Kinase’ would yield many protein kinases in the results including ‘protein kinase’ and ‘glutamate 5-kinase’. There is an option for switching the organism database, if the user knows the database name. In this project, two databases- ntrd01 and ntrd03 - were frequently accessed. In this entire paper, the terms gene, ORF and proteins are used interchangeably even though protein translations of the predicted genes are actually annotated.

Upon clicking the ‘Genome Summary’ button, a user will be introduced to a host of hyperlinks that form the heart of the annotation process. Genome calculations lead to a web page that gives statistical information on the genomic sequence.



**Fig 18:** ‘Genome Calculations’ page gives the statistical information on the entire genomic

sequence. The tables displayed here talk about the different aspects such as 'feature count' and 'feature type'. There is another table that shows statistical information on the availability of the three start codons, ATG, GTG and TTG (in a prokaryote there are more than one start codons).

| Chromosome Information Table              |  |
|---|--|
| Sequence id:                              | 1  |
| Type:                                     | chromosome                                 |
| Prokaryotic length:                       | 4134008 bp                                 |
| GC content:                               | 58%  |
| Base frequencies:                         | (A) (C) (G) (T)<br>20.5% 29.3% 29.3% 21.5% |
| Flanking characters:                      |  |
| ORF count:                                | 4713                                       |
| Average gene length:                      | 888 nt                                     |
| Percent coding:                           | 90.5%                                      |
| Percent coding ORF, RNA, rRNA, or repeat: | 74.21%                                     |

**Fig 19:** This table found on the same page talks about different aspects of a Chromosomal sequence such as GC content, number of ORFs, average ORF length, percent coding region, percent non-coding region and other plasmid

information.

In this page, the information on all the four plasmids can be obtained. The plasmid length, the number of observed ORFs on that plasmid, the GC content, average ORF length, percent coding region and non-coding region and base frequencies. For example, the lengths of plasmids generated after sequencing the genome are 106469 bp (pTB1), 69269 bp (pTB2), 16575 bp (pTB3) and 5824 bp (pTB4).

### 3. Role Category Breakdown

“Role Category Breakdown” web page shows a summary of ORFs found in various broad categories based on TIGR roles and then a breakdown TIGR sub category. TIGR has classified the roles of genes in to many categories (See Appendix for more information of TIGR role categories). Every role category has a specific role id. After ‘Glimmering’ the genomic sequence, some of the rejected ORFs will be stored in the ‘Glimmer Rejects’ role category. The best part of Glimmer system is identifying the frame-shift mutations. Any minute frame-shifts would be

stored, flagged reported in ‘Disrupted reading frame’ category. Other ORFs would resemble many other proteins existing in public databases like NCBI and PROSITE but their functions would be unknown. Those ORFs would be stored under ‘Unknown functions’ role category. Some of the ORFs stored would be reported as ‘Hypothetical’, since there are no matches to any know genes. These hypothetical genes or ORFs have all the qualities of a gene and are sometimes considered to be unique in that organism.

| Role Breakdown |  |        |          |   |
|----------------|--|--------|----------|---|
| role id        | name   | number | complete | % |
| main           | Unclassified                                     | 0      | 0        | 0 |
| 185            | Role category not yet assigned                   | 0      | 0        | 0 |
| main           | Amino acid biosynthesis                          | 251    | 149      | 0 |
| 70             | Aliphatic amino acid family                      | 39     | 39       | 0 |
| 71             | Aspartic family                                  | 33     | 33       | 0 |
| 73             | Glutamate family                                 | 32     | 32       | 0 |
| 74             | Pyruvate family                                  | 21     | 22       | 0 |
| 75             | Serine family                                    | 33     | 33       | 0 |
| 161            | Histidine family                                 | 11     | 22       | 0 |
| 68             | Other  | 2      | 2        | 0 |
| main           | RNAse, pyrimidines, nucleosides, and nucleotides | 76     | 76       | 0 |
| 123            | 2'-Deoxyribose nucleotide metabolism             | 5      | 5        | 0 |
| 124            | Nucleosides and nucleotide interconversions      | 20     | 20       | 0 |
| 125            | Purine ribonucleotide biosynthesis               | 16     | 16       | 0 |
| 126            | Pyrimidine ribonucleotide biosynthesis           | 20     | 20       | 0 |
| 127            | Salvage of nucleosides and nucleotides           | 20     | 20       | 0 |
| 128            | Salvage nucleotide biosynthesis and conversions  | 4      | 4        | 0 |
| 122            | Other  | 8      | 8        | 0 |

**Fig 20:** Role Category Breakdown of all the ORFs is displayed here. The blue box has the name of the TIGR Role and the ‘Main’ link takes to that specific role category web page with a list of all the ORFs.

‘Annotation Notebook’ web page includes a set of text fields

associated with each TIGR role category. Annotators or users store some information regarding a specific ORF which they feel the investigators of the project should know for purposes of publishing their manuscripts, or performing experiments to characterize that protein. The notebook often consists of items of particular biological interest involving the presence or absence of certain genes, pathways, gene order etc. In this project, the annotation notebook has little role to play. Other links on the Genome Summary page, including ‘Project Administration’, ‘Frame shift Status’, ‘Annotation progress report’, ‘Interpro Domain’, and ‘Genome Properties’ were not accessible to those working outside of TIGR. The user can also choose to view the genes based on one of several ‘attributes’ they might have. On the Genome Summary page, there is a search option for looking up ORFs based on attributes such as molecular weight, GC content etc. All

these attributes are listed on the Genome Calculations page. Genes can also be selected based on one of several types of clustering evidence found for them as a result of extensive search performed by the Hidden Markov Model (HMM) associated with the Glimmer system. ‘Paralogous Families’ is also one of the options on the Genome Summary page, where the gene will be looked up based on the membership in paralogous families, ordering done either by number of family members or by family name. Paralogous families are built by first searching all of the proteins within a genome against themselves and against a HMM database. If there is a positive match with an HMM, the family would be named based on the matched HMM. Further searches are performed based on the regions that did not match any HMM. Those regions would be given a numerical value but no description. Proteins can be viewed based on a predicted location in a membrane (looking for SignalP here). “Membrane Proteins” link would allow an user to choose a particular SignalP cutoff values, number of predicted transmembrane regions, proteins that have OMP signal, or lipid attachment site.

Annotation process begins as soon as the user clicks on the role ids. Clicking on the role ids takes a user to a web page that contains a list of all the ORFs, which are currently assigned to that specific TIGR microbial role category. Annotators can now choose to work on all the ORFs present in that particular role category.

| Amino acid biosynthesis    |   |               |            |         |         |   |                          |          |             |
|----------------------------|---|---------------|------------|---------|---------|---|--------------------------|----------|-------------|
| Aromatic amino acid family |   |               |            |         |         | Role id: 70                                   | Edit Annotation Notebook |          |             |
| A                          | C | gene id       | locus      | end1    | end3    | gene name                                     | gene symbol              | ec       | other roles |
|                            |   | ORF02623 (GV) | NT02RD2631 | 2330716 | 2331139 | 3-dehydroquinate dehydratase, type II         | aroD                     | 4.2.3.10 |             |
|                            |   | ORF03052 (GV) | NT02RD3087 | 2962263 | 2962332 | 3-dehydroquinate synthase                     | aroB                     | 4.2.3.4  |             |
|                            |   | ORF03073 (GV) | NT02RD3034 | 3176656 | 3176804 | 3-phosphohydroxymethyl-3-carboxyglutamate     | aroA                     | 2.5.3.19 |             |
|                            |   | ORF03173 (GV) | NT02RD3210 | 3076784 | 3077793 | anthranilate phosphoribosyltransferase        | trpD                     | 2.4.2.18 |             |
|                            |   | ORF03168 (GV) | NT02RD3205 | 3069243 | 3070776 | anthranilate synthase component I             | trpE                     | 4.1.3.27 |             |
|                            |   | ORF03172 (GV) | NT02RD3209 | 3076334 | 3076373 | anthranilate synthase component III           | trpD                     | 4.1.3.27 |             |
|                            |   | ORF00926 (GV) | NT02RD0233 | 902707  | 904892  | aromatic amino acid aminotransferase          | trpB                     | 2.6.1.27 |             |
|                            |   | ORF01296 (GV) | NT02RD2212 | 3267039 | 3266763 | chorismate mutase, putative                   |                          | 5.4.99.5 |             |
|                            |   | ORF00531 (GV) | NT02RD0296 | 330307  | 339207  | chorismate synthase                           | aroC                     | 4.2.3.3  |             |
|                            |   | ORF01266 (GV) | NT02RD2323 | 3188633 | 3187939 | glutamine amidotransferase, putative          |                          | 4.1.3.27 |             |
|                            |   | ORF03176 (GV) | NT02RD3212 | 3078857 | 3079250 | indole-3-glycerol phosphate synthase          | trpC                     | 4.1.3.48 |             |
|                            |   | ORF03866 (GV) | NT02RD3907 | 3773662 | 3773334 | N-(5'-phosphoribonyl)anthranilate isomerase   | trpF                     | 5.3.1.24 |             |
|                            |   | ORF02607 (GV) | NT02RD0628 | 2515210 | 2516660 | phospho-2-dehydro-3-deoxyphosphonate aldolase |                          | 2.5.1.58 |             |
|                            |   | ORF01247 (GV) | NT02RD1261 | 1222772 | 1221933 | prephenate dehydratase, putative              | preA                     | 4.2.3.51 |             |
|                            |   | ORF03340 (GV) | NT02RD3887 | 3255234 | 3254934 | prephenate dehydrogenase                      | trpC                     | 1.3.1.32 |             |
|                            |   | ORF00436 (GV) | NT02RD0438 | 428041  | 428874  | shikimate 5-dehydrogenase, putative           | aroE                     | 1.3.1.25 |             |
|                            |   | ORF03053 (GV) | NT02RD3088 | 2962890 | 2962257 | shikimate kinase, putative                    | aroK                     | 2.7.1.71 |             |
|                            |   | ORF03654 (GV) | NT02RD3895 | 3763301 | 3764092 | tryptophan synthase, alpha subunit            | trpA                     | 4.2.3.20 |             |
|                            |   | ORF03645 (GV) | NT02RD3896 | 3772820 | 3773597 | tryptophan synthase, beta subunit             | trpB                     | 4.2.3.20 |             |

**Fig 21:** This is the image of a list of ORFs present under a specific role category. In this case, the role category chosen was “Amino acid biosynthesis”, whose role id is listed as 70. In the left hand corner of the table, there are two columns: ‘A’ means automatically annotated

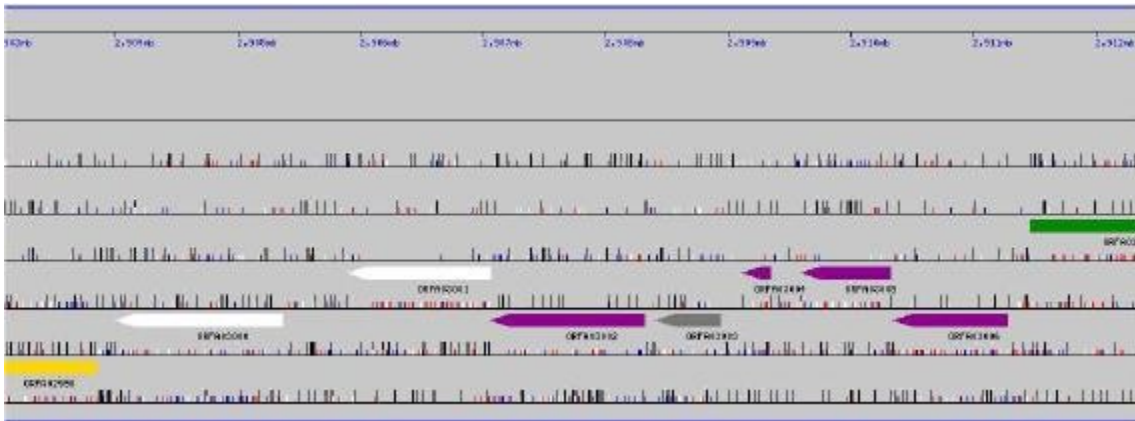
ORFs\* (appearance of a green ball suggests that it had been completely annotated by Manatee) and 'C' means confirmation of the annotation process (appearance of a pink ball suggests that the ORF is completely annotated). This is the most important step during the completion of the annotation process.

If 'Main' is clicked, all the ORFs present in that role category are listed. There are however, many sub role categories that can be individually accessed by clicking on their role ids. Annotating based on role category is very useful and easy because the ORFs are already grouped based on the homology with other known proteins. The list of all the ORFs in that particular role category also included the position of that ORF on the plasmid, starting and ending points, possible gene names and symbols, Enzyme Commission (EC) numbers and other significant roles. There are two columns (to the left of the ORF) that denote whether the ORF is automatically annotated by Manatee (denoted by A) or the annotation of the ORF is approved by the user or annotator. Even if the Manatee has automatically annotated an ORF, the user still needs to check for correct annotation of the ORF. Appearance of green and pink balls in those columns suggests that the ORFs are annotated automatically and have been checked for correctness by an annotator.

ORF id (in the list) is followed by a link 'GV', which means 'Genome View'. Genome view of any ORF shows the orientation and position of that ORF on the genomic sequence. This is very helpful because any overlaps would be checked for based on this genome view. Genome viewer is a tool, which allows one to view the genes in context with their neighboring genes in the genome. It displays a graphic showing the six-frame translation of a region of DNA sequence, where each horizontal bar is a different frame. Colored arrows represent the genes according to the TIGR main-role assignment. There are many viewing and editing options available from this web page.

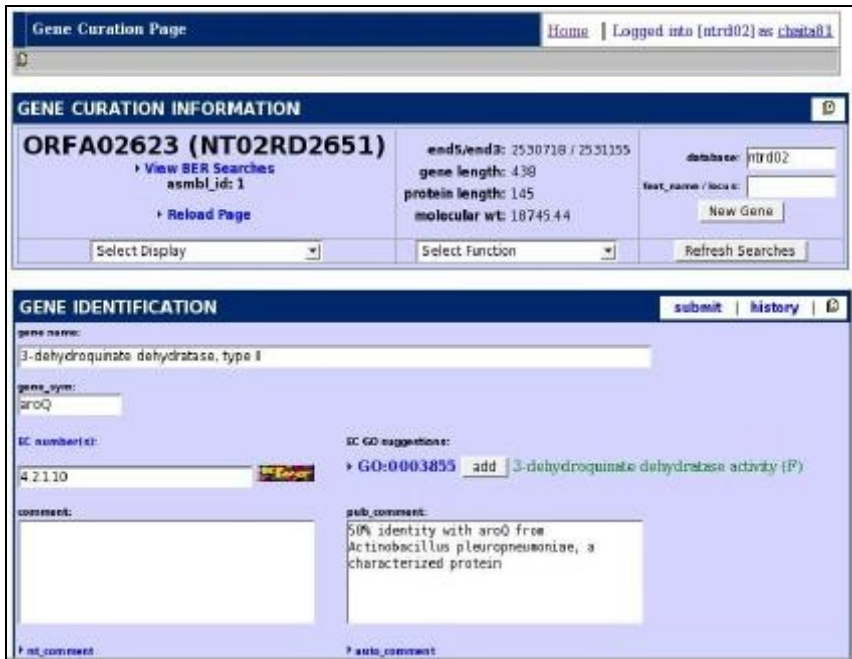
---

Even if the ORFs are automatically annotated by Manatee, annotators should confirm it by manually checking the annotation.



**Fig 22:** Genome view of an ORF is being shown here. ORFs are shown as arrows in different colors on a scale that can be seen at the top of the viewer. Each horizontal bar is a frame of translation. Six bars are shown here that represent six frames of translation. On the horizontal bar long lines represent stop codons and short lines represent the start codons. An ORF is located between any two stop codons. If the user is looking for the pink colored ORF (in the middle of the viewer), it is clear from the viewer that there are no overlaps, and it is the longest ORF in the given window view.

The presence of any overlaps has to be recognized and taken care of. The genome viewer shows any overlaps. Just by clicking on the ‘Genome View (GV)’ link of the ORF takes a user to genome view that will show the overlaps. In case of overlaps, one must look at the sequence start site and predict the wrong alignment on that sequence. Clicking on the ORF takes a user to the ‘Gene Curation Page (GCP)’. The GCP web page has all the necessary information on that specific ORF (See Fig 24). It begins with ORF id, ORF sequence statistics (like length of the predicted amino acid sequence), ORF length, molecular weight and the database in which that ORF is observed. The ORF id is followed by a link ‘View BER Searches’ that leads to BER Skim web page. This information is followed by the ‘Gene identification’, where a possible name, symbol and EC number are available.



**Fig 23:** Gene Curation Page (GCP) is displayed here. It contains all the necessary information on the gene including the gene identification, GO classification, evidence picture, and BER skim. In this case the ORF id is ORFA02623.

There is a link by the EC number box that connects to

Kyoto Encyclopedia of Genes and Genomes (KEGG) web site where many interesting things like ‘biochemical pathways’ in which the ORF is involved would be displayed. This information is followed by the Gene Ontology (GO) evidence. In this project, GO evidence was skipped due to the complications it presents. More importance is given to the ‘Evidence Picture’ (See Fig 25).



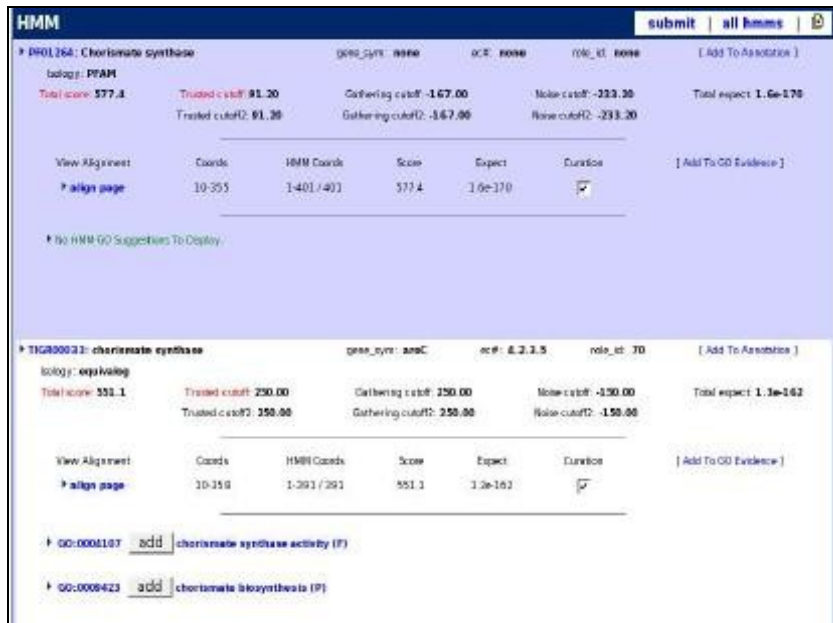
**Fig 24:** Evidence picture of an ORF shows various conserved regions or domains (found in known

proteins). Clusters of Orthologous (COG) were also looked at. It is based on these domain matches, identification and naming of the ORF is done.

Evidence Picture of any ORF shows the presence of any transmembrane helices based on ‘TmHMM’ scores. TmHMM is a HMM specific for transmembrane regions, built by the Center for Biological Sequence Analysis, Denmark. It is in the evidence picture that any paralogous family relationship is displayed. If a protein is a member of a paralogous family, it will be represented with a blue bar; clicking on the bar takes the annotator

to a web page listing all the family members. Presence of characterized matches is shown by green colored bars. Additional evidence types that are shown include signal P, lipoprotein predictions, and PROSITE hits. Signal P and PROSITE information are both displayed in the evidence picture and in the sections of their own on the GCP.

This evidence picture is followed by a listing of the best HMMs (See Fig 26). This is a list of the best matches with their cut off and noise values, with the best result published first.

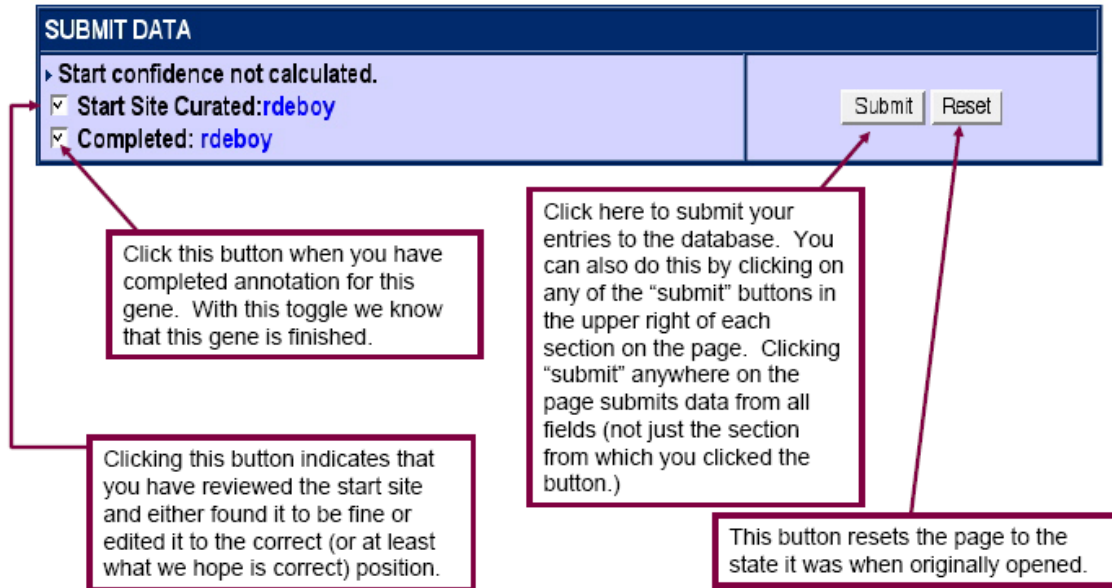


**Fig 25:** This is a display of HMM matches of a predicted protein sequence which assist with determining the name of the predicted protein.

Along with the HMM, there are buttons that allow us to add the HMMs to the GO evidence. If the

score of the HMM is above the cut off value, it is added to the GO evidence. This HMM list is followed by the most important thing, ‘Submit Data’ box (See Fig 27). There are two check boxes, ‘Start site curates’ and ‘Completed’. The annotation is not complete until the start site is curated. Start sites can be looked at by clicking on the best matches in the BER skim. An alignment of the predicted protein and the best match is shown in BER alignment page (click on the match). This alignment also shows the start sites of the predicted protein sequence. As earlier mentioned, there could be three different start sites in prokaryotes. However, in prokaryotes, any start site is preceded by a AT-rich region called the ‘Shine-Dalgarno’ sequence. As soon as the start site is curated, the name, symbol, and EC number (if possible) is assigned. Following this, clicking

on the 'Completed' button on the GCP completes the annotation of that ORF.



83

**Fig 26:** This is the SUBMIT DATA section of the GCP. The name 'rdeboy' represents the annotator's username. Manatee would allow completion of annotation only after start site is curated.

Start sites, if incorrectly annotated by Manatee, should be changed manually. This page to change the start site can be reached from "Gene Options" menu on the Genome Viewer web page. This page can also be reached from the GCP. In the GCP web page click on "Select Function" pull-down menu and select "Edit Start Site" (See Fig 28).

| Start Edits |  |            |
|-------------|--|------------|
| 2836723     | TTCCCGCTTCCAATCATGACGAACCTGCAACTGCGACATTGAACACCCCTTTTATTTTGT | Nucleotide |
|             | E P L P I M T N L O L R H * T E F Y E C                      | Frame1     |
|             | E R E Q R * E T S N S Q I E H E K I E Y                      | Frame2     |
|             | P A S N H D B L A T A T L N T L L E L Y                      | Frame3     |
|             | E G S G I M V E K C S R C Q V S E * K Q                      | Frame4     |
|             | B G A R L * S S S A V A V H F V R K N K                      | Frame5     |
|             | K R K M D H R V O L O S M S C G K I K T                      | Frame6     |
| 2836785     | ATTTTACCTGGCTAGGATAACCTCAGCCCTTAAACTGTCACGCCAACCCAGTGATACAG  | Nucleotide |
|             | I L P W L G * P Q P L N C O R O P V I O                      | Frame1     |
|             | E X L G * R N L S P * T V N A N Q * Y R                      | Frame2     |
|             | F T L A B I T S A L K L S T F T S D T G                      | Frame3     |
|             | I K G Q S P Y G * G K P Q * R W G T I C                      | Frame4     |
|             | V K V K A L I V E A R L S D V G V L S V                      | Frame5     |
|             | N * E P * S L B L G * Y T L A L M H Y L                      | Frame6     |
| 2836843     | GTTTACCAGTGATTAATTTCAATCAACGCTGTGAGCTTTTATGCGCAATTTACTCGATT  | Nucleotide |
|             | V Y M * L I E N O B C K L V S A I Y S I                      | Frame1     |
|             | F T T D * F S I N A V E 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 D F                      | Frame3     |
|             | T * W Q N I K L * R O S S K H A I * E I                      | Frame4     |
|             | E K G S I L K * D V S H A K I H L K S S                      | Frame5     |
|             | N V V S * N E I L A T L K * A C N V R N                      | Frame6     |
| 2836905     | TTGACTTTGATAGCGCCATATTTGGCACCCCTTATACTCCATGACTCGTGCACTTCCTG  | Nucleotide |
|             | L T L I A E I E G T L I E P * L V H E L                      | Frame1     |
|             | * L * E P Y L A F L L H P S C T E C                          | Frame2     |
|             | D F D S A H I W H F Y T S M T R A L F V                      | Frame3     |
|             | E V K I A G M N P V R I G G H S T C K R                      | Frame4     |
|             | K S K S L A W I Q C G * V E M V R A S G                      | Frame5     |
|             | Q S Q Y R G Y K A G K Y R W S E H V E Q                      | Frame6     |

**Fig 27:** This is the web page for changing the start site of an ORF. Green nucleotides represent ribosome binding sites, purple nucleotides represent amino acids of the query gene and blue nucleotides represent amino acids of one of genes in the

region other than the query gene.

Purple text represents the gene of interest. Blue text represents other genes in the region. It is important to know if any changes in the start site leads to an overlap. Often editing a start site will remove or introduce overlap between genes. Occasionally an annotator may want to extend a start into an upstream gene and will find that the upstream gene in question is a small hypothetical gene with no homology to anything. In such cases the annotator should consider deleting the short hypothetical gene, since it is likely that it is not a real gene. To edit a start site, click on the appropriate start site in the appropriate frame of translation. The new coordinate for the selected start site will appear in the “New End5” box. To save the changes to the database, “Submit” button has to be clicked.

Based on the observations, a protocol on annotating an ORF has been generated and strictly followed. Here are the steps involved in annotating an ORF:

- 1) Genome View – check for overlap
- 2) HMM – evidence picture
- 3) BER SKIM
- 4) Edit start site
- 5) Check links

- 6) Naming
  - a) Characterized match (p-value <  $10^{50}$  or 35% identity)
  - b) PFAM
- 7) Gene Symbol
- 8) EC number
- 9) Comment
- 10) TIGR roles
- 11) Gene Ontology terms (*Can be skipped*)
- 12) Submit Data
- 13) Report Frame shifts, start errors & overlaps

In this way, almost all of the ORFs have been annotated in 'ntrd01' and 'ntrd02' databases.

## CHAPTER 5

### Conclusions and Future Directions

#### 1. Biological Importance of annotating *Roseobacter denitrificans* genome

*Roseobacter denitrificans* belongs to the purple aerobic phototrophic bacteria (APB) family. They are the only bacteria that perform photosynthesis in the presence of oxygen, but do not produce oxygen gas as a result. Typically, APBs grow photoheterotrophically by the respiration of the organic substrates resulting in the release of CO<sub>2</sub>, counter to the traditional "CO<sub>2</sub>-sink" of the upper ocean. The marine ABP species *Roseobacter denitrificans* grows photoheterotrophically in the presence of

oxygen and light; it also grows anaerobically in the dark using nitrate or trimethylamine *N*-oxide as electron acceptors, making *R. denitrificans* an interesting subject of research. Now, with the unraveling of the genomic sequence of *R. denitrificans*, the study of a variety of unique features can be pursued.

*R. denitrificans* contains a primary circular chromosome of 4,133,097 base pairs and four plasmids containing a total of 4,403 predicted coding sequences. The presence of many DNA integration and transfer proteins, plasmid replication proteins and inverted repeats imply that plasmid pTB1 might promote lateral gene transfer.

*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. This suggests evidence for a scattered loss of ancestral carbon-fixation in the  $\alpha$ -proteobacterial tree. With the advent of an aerobic lifestyle in these proteobacteria, the competition of O<sub>2</sub> with CO<sub>2</sub> for RubisCO catalysis was likely overwhelming. Thus, many  $\alpha$ -proteobacteria might have lost the Calvin cycle genes. Even though these important enzymes are absent in *R. denitrificans*, some putative genes related to the C<sub>4</sub> sequestration, Crassulacean acid metabolism (CAM), and anaplerotic carbon-fixation enzymes such as pyruvate-orthophosphate dikinase and phosphoenolpyruvate (PEP) carboxylase, are present.

Based on the genome sequence of *R. denitrificans*, it can be hypothesized that APBs fix CO<sub>2</sub> heterotrophically using a C<sub>4</sub> sequestration pathway supplemented by additional CO<sub>2</sub> provided by CO oxidation and heterotrophic respiration. Future biochemical studies are absolutely required to determine whether *R. denitrificans* and other bacteria lacking Calvin enzymes assimilate CO<sub>2</sub> through the hypothesized pathway.

## 2. Future directions

The genome analysis of *R. denitrificans* (the first APB sequenced) has opened the door to understanding their physiological capacity to fill a broad niche in the ocean ecosystem. The numerous mechanisms that members of *Roseobacter* genus have evolved for chemotrophic and phototrophic energy metabolism may explain their wide dispersal and high abundance in the biosphere. However, it is still unclear how *R. denitrificans* adapted to the changing atmospheric conditions, and how RubisCO evolved in the APBs such as *R. denitrificans*. It is really important to also understand the contributions of the *Roseobacter sp.* to the carbon cycle. Genome sequencing and annotation are only the first steps towards understanding a primitive, yet sophisticated organism. Some of the genes found in the organism that are unique to the species should be biochemically characterized. 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 APBs should be sequenced and annotated. Comparative sequence (or genome) analysis of the sequenced APBs should show the evolutionary relationships of certain genes or photosynthetic gene clusters. More sophisticated sequence analysis programs (like MEGA and ClustalX) should be employed to understand the phylogeny of photosynthetic genes.

The role of transcriptional regulator PpaA is not defined. It has been proved that over expression of PpaA in *R. sphaeroides* activates transcription of photosynthetic proteins under aerobic conditions. Another transcription factor PpsR that activates photosynthetic gene transcription under aerobiosis is observed in *R. denitrificans*. However, its role is

unclear. Further studies on these transcriptional regulators should throw some light on the functions.

Finally, genome sequencing is only a way to decipher the genetic code and unravel the mystery surrounding the metabolic profile of any organism.

## BIOBLOGRAPHY

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.

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.

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

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

Bateman A., et al.; The Pfam protein families database; *Nucleic Acid Res.* 1990; 28(1):263- 266

Bauer CE and Bird TH; Regulatory circuits controlling photosynthesis gene expression; *Cell* 1996; 85: 5- 8

Beatty JT; On the natural selection and evolution of the aerobic phototrophic bacteria; *Photosynth Res* 2002; 73: 109- 114.

Blankenship RE (2002) Molecular Mechanisms of Photosynthesis, *Blackwell Science*, Oxford, UK.

Blankenship RE and Hartman H (1998); The origin and evolution of oxygenic photosynthesis. *Trends Biochem Sci* 23: 94- 97.

Eddy S; Profile hidden Markov models; *Bioinformatics*; 14(9):755- 763

Gish W, et al.; Identification of protein coding regions by database similarity search; *Nat. Gent* 1993; 3(3):266- 272

Haft D, et al.; TIGRFAMs: A protein family resource for the functional identification of proteins; *Nucleic Acids Res* 2001; 29(1):41- 3

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

Larimer FW, et al.; The genome sequence of the metabolically versatile photosynthetic bacterium *Rhodospseudomonas palustris*. *Nature Biotechnol* 2004; in press

Oh J-I and Kaplan S; Generalized approach to the regulation and integration of gene expression; *Molec Microbiol* 2001; 39: 1116- 1123  
Ormerod J; Every dogma has its day: a personal look at carbon metabolism in photosynthetic bacteria; *Photosyn Res* 2003; 76: 135- 143

Raymond J et al; Whole genome analysis of photosynthetic prokaryotes; *Science* 2002; 298: 1611- 1620

Salzberg S., et al.; Microbial gene identification using Interpolated Markov Models; *Nucleic Acid Res* 1998; 26(2):544- 548

Smith T F, et al.; Identification of common molecular subsequences; *J Mol Biol* 1981 ; 147(1):195- 197

Tabita FR; The biochemistry and metabolic regulation of carbon metabolism and CO<sub>2</sub> fixation in purple bacteria; *In* Blankenship RE, Madigan MT and Bauer CE (eds); *Anoxygenic Photosynthetic Bacteria* pp 885- 914; Kluwer Academic Publishers, Dordrecht.

TIGR; Michelle Gwinn; Prokaryotic Annotation Overview, October 2004

TIGR; Michelle Gwinn; A guide to Manatee, October 2004

TIGR; Michelle Gwinn, William Nelson, Robert Dodson, Steven Salzberg, Owen White; Small Genome Annotation and Data Management at TIGR

TIGR; Domain Based Paralogous Protein Families; [www.tigr.org](http://www.tigr.org); Annotation Workshop July13, 2005

TGen; Translational Genomics Institute; <http://www.tgen.org>

Wesley D S, et al; A ubiquitous pathway marine phototroph with a novel carbon- fixation pathway; submitted to PNAS, 2006

Manatee web site; <http://manatee.tgen.org> ; edit version active from June 2005 – August 2005;

Phototrophic Genome Project web site; <http://genomes.tgen.org>

## GLOSSARY \*

**Annotation Notebook** – a link from the gene list page that is associated with all TIGR role ids. The notebook is used by the annotator to note genes of interest for the Principle Investigator in charge of the project.

**assembler** – the program that “puts together” the raw sequence that is generated in the lab.

**assembly** – sequence that has been assembled together and loaded into the database.

**autoannotate** – the program that assigns annotation to newly defined open reading frames before they are manually curated.

**BER** – BLAST-extend- repace

**Characterized match** – a protein from the BER search results that is experimentally characterized and a good match to our query protein.

**Characterized table** – the table in the database that stores the accessions of the characterized proteins we have collected.

**CMR** – Comprehensive Microbial Resource.

**Codon** – the 3 base pair nucleotide sequence that codes for an amino acid.

**Codon usage** – The term describing the tendency of organisms to use certain codons and not others for a specific amino acid.

**COG** – clusters of orthologous groups.

**Common** – a table in the database that stores “common” information concerning genome projects, including GO ontologies, and Genome Property information

---

Obtained from ‘Bioinformatics Analyst Training’ at TIGR, created on 4/22/2005

**Contig** – the term given to a piece of assembled raw sequence.

**Coords** – the term describing the coordinate boundaries of an open reading frame.

**Database** - a systematically arranged collection of computer data, structured so that it can be automatically retrieved or manipulated.

**Database field** – an attribute of what the database table describes; a column of the table.

**Database table** – a unit of data storage in a database, it has a name and consists of one two or more defined fields which are populated with items of data, generally a particular table will store data relevant to a particular topic.

**Degenerate** – an open reading frame with multiple point mutations, frameshifts, deletions, etc.

**Domain** – the term describing homology to only part of a protein, and not the whole protein.

**EC** – Enzyme Commission

**End3** – a term used to describe the 3' end coordinate of an open reading frame.

**End5** – a term used to describe the 5' end coordinate of an open reading frame.

**Equivalog** – the term used to describe a hidden markov model that describes the exact function of a protein.

**Feat\_name** – the field in the database that holds the “feature name”.

**Feat\_type** – the field in the database that describes the type of “feature”.

**FS** – frameshift

**GC skew** – the tool that measures the 3rd base pair wobble for a particular reading frame.

**GCP** – Gene Curation Page

**GENBANK accession builder** – The tool used by an annotator to create files to be submitted to GenBank for a particular genome.

**Gene cluster** – A group of genes adjacent to one another, and often, but not always, oriented in the same direction, that have a similar function (e.g. transport) or are part of a biochemical pathway.

**Gene Symbol** – The abbreviation given to a particular gene. It is stored in the ident table.

**Genome Properties** - The Genome Properties system consists of a suite of "Properties" which are carefully defined attributes of prokaryotic organisms whose status can be described by numerical values or controlled vocabulary terms for individual completely sequenced genomes. Evaluation of these properties may take place via manual curation or by computer algorithms.

**Genome Viewer** – The graphical interface used by an annotator that allows the user to view the genomic organization in a selected genome.

**glimmer** – The program used at TIGR to find open reading frames in whole genomic sequence.

**GO** – gene ontology

**HMM** – hidden Markov model

**Homolog** - A gene similar in structure and evolutionary origin to a gene in another species.

**Hypotheticals with praze** – hypothetical proteins that have results in the BER table.

**ident** – a table in the database that stores “identity information” for an ORF.

**Intergenic region** – the region between two genes that does not have a coding ORF associated with it.

**InterPro** – a searchable database of protein families, domains, and functional sites that is maintained by the European Bioinformatics Institute.

**Isology** – sequence similarity of aligned nucleic acids or amino acids; the similarity may be due to homology or convergence.

**Library** – one of the stages of the genome sequencing process where a genomic DNA sample from the organism is fragmented and cloned into a plasmid or phage library for sequencing.

**Locus** – the final public identifier for a gene, has associated location and annotation

**Main Role** – the highest level of a TIGR role category, each TIGR main role has several sub roles

**MANATEE** – the web-based annotation tool used by all TIGR annotators.

**Multiple alignment** – an alignment of multiple DNA or protein sequences.

**niaa** – non-identical amino acid database; the internal TIGR protein database. It consists of all of the TIGR and public database protein information.

**nraa** – non-redundant amino acid database; this database was replaced by niaa.

**Omnium** – the name of the CMR database.

**ORF** – open reading frame

**ORF management** – the process that the annotators follow to resolve overlapping genes, find missing genes, and confirm start sites.

**orf\_attribute** – one of the tables in the microbial database that stores “attributes” associated with each ORF.

**ortholog** – genes in two different species that evolved from a common ancestor. Orthologs retain function over the course of evolution.

**Overlaps** – the term used to describe genes that “overlap” in a genome.

**Pairwise alignment** – the alignment of two DNA or amino acid sequences to each other.

**Pairwise match** – analogous to pairwise alignment, indicates that a protein matches another based on a pairwise alignment.

**Paralog** – proteins that are related by sequence similarity within a genome.

**Paralogous family** – a family of paralogs that is found in a genome. At TIGR we have a paralogous family program that generates these families.

**PFAM** – Protein families database of alignments and HMMs.

**PM** – point mutation

**PROSITE** - PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.

**RBS** – ribosome binding site

**Role id** – the identification number assigned to TIGR role categories.

**signalP** – the program used at TIGR that searches for signal peptides in a protein sequence. Each protein is scored, and based on these scores the annotator can predict whether the protein is secreted or not. This tool is built by CBS, [www.cbs.dtu.dk/index.shtml](http://www.cbs.dtu.dk/index.shtml).

**SQL** – structured query language; the programming language used to navigate and manipulate relational databases.

**start site** – the codon that corresponds to the beginning of the coding region of an open reading frame, the three start codons in bacteria are: ATG, GTG, and TTG

**stop site** – the codon that corresponds to the end of a gene or open reading frame, the three stop codons in bacteria are: TAA, TAG, TGA

**Subfamily** – the HMM isology that describes a type of TIGRFAM HMM. A subfamily HMM describes a specific family of proteins. This type of isology is more specific than the superfamily HMM isology.

**subrole** – the identification number that corresponds to roles that are subsets of a TIGR main role category.

**Superfamily** – a HMM isology that refers to a type of TIGRFAM HMM. A superfamily HMM describes a large family of related proteins. It is much more general than a subfamily isology HMM.

**Swiss-prot** – an external public database of DNA and proteins sequences that is maintained by the Swiss institute of Bioinformatics and the European institute of Bioinformatics.

**TIGRFAM** – The library of HMMs that are created and maintained here at TIGR.

**TMHMM** – the program used at TIGR to identify transmembrane spanning regions of a protein. This tool is built by CBS, [www.cbs.dtu.dk/index.shtml](http://www.cbs.dtu.dk/index.shtml).

**APPENDIX 1: TIGR Role Categories**



### ‘TIGR Role Categories’

Source: This table is obtained from TIGR Comprehensive Microbial Resource (CMR) web site, [http://210.218.199.237/tigr-scripts/CMR2/role\\_id\\_list.spl](http://210.218.199.237/tigr-scripts/CMR2/role_id_list.spl)

| MAINROLE   | SUBROLE                        | ROLE ID |
|--|--------------------------------|---------|
| Amino acid biosynthesis                                    | Aromatic amino acid family     | 70      |
| Amino acid biosynthesis                                    | Aspartate family               | 71      |
| Amino acid biosynthesis                                    | Glutamate family               | 73      |
| Amino acid biosynthesis                                    | Histidine family               | 161     |
| Amino acid biosynthesis                                    | Other                          | 69      |
| Amino acid biosynthesis                                    | Pyruvate family                | 74      |
| Amino acid biosynthesis                                    | Serine family                  | 75      |
| Biosynthesis of cofactors, prosthetic groups, and carriers | Biotin                         | 77      |
| Biosynthesis of cofactors, prosthetic groups, and carriers | Chlorophyll                    | 191     |
| Biosynthesis of cofactors, prosthetic groups, and carriers | Folic acid                     | 78      |
| Biosynthesis of cofactors, prosthetic groups, and carriers | Glutathione                    | 86      |
| Biosynthesis of cofactors, prosthetic groups, and carriers | Heme, porphyrin, and cobalamin | 79      |
| Biosynthesis of cofactors, prosthetic groups, and carriers | Lipoate                        | 80      |
| Biosynthesis of cofactors, prosthetic groups, and carriers | Menaquinone and ubiquinone     | 81      |
| Biosynthesis of cofactors, prosthetic groups, and carriers | Molybdopterin                  | 82      |
| Biosynthesis of cofactors,                                 | Other                          | 76      |

|  |   |     |
|--|---|-----|
| prosthetic groups, and carriers                            |   |     |
| Biosynthesis of cofactors, prosthetic groups, and carriers | Pantothenate and coenzyme A   | 83  |
| Biosynthesis of cofactors, prosthetic groups, and carriers | Pyridine nucleotides  | 163 |
| Biosynthesis of cofactors, prosthetic groups, and carriers | Pyridoxine  | 84  |
| Biosynthesis of cofactors, prosthetic groups, and carriers | Riboflavin, FMN, and FAD  | 85  |
| Biosynthesis of cofactors, prosthetic groups, and carriers | Thiamine  | 162 |
| Cell division  | DNA synthesis/replication   | 45  |
| Cell division  | NULL  | 8   |
| Cell division  | apoptosis   | 63  |
| Cell division  | cell cycle  | 36  |
| Cell division  | chromosome structure  | 27  |
| Cell envelope  | Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides | 90  |
| Cell envelope  | Biosynthesis of murein sacculus and peptidoglycan                               | 89  |
| Cell envelope  | Other   | 88  |
| Cell envelope  | Surface structures  | 91  |
| Cell signaling/cell communication                          | NULL  | 39  |
| Cell signaling/cell communication                          | cell adhesion   | 40  |
| Cell signaling/cell communication                          | channels/transport proteins   | 5   |
| Cell signaling/cell communication                          | effectors/modulators  | 2   |
| Cell signaling/cell  | hormone/growth factors  | 6   |

|   |   |     |
|---|---|-----|
| communication                                   |   |     |
| Cell signaling/cell communication               | intracellular transducers               | 7   |
| Cell signaling/cell communication               | metabolism                              | 33  |
| Cell signaling/cell communication               | protein modification                    | 1   |
| Cell signaling/cell communication               | receptors                               | 4   |
| Cell structure/motility                         | NULL                                    | 57  |
| Cell structure/motility                         | cytoskeletal                            | 28  |
| Cell structure/motility                         | extracellular matrix                    | 29  |
| Cell structure/motility                         | microtubule- associated proteins/motors | 62  |
| Cell/organism defense                           | NULL                                    | 58  |
| Cell/organism defense                           | homeostasis                             | 173 |
| Cell/organism defense                           | immunology                              | 34  |
| Cellular processes                              | Adaptations to atypical conditions      | 149 |
| Cellular processes                              | Cell adhesion                           | 701 |
| Cellular processes                              | Cell division                           | 93  |
| Cellular processes                              | Chemotaxis and motility                 | 706 |
| Cellular processes                              | Conjugation                             | 702 |
| Cellular processes                              | DNA transformation                      | 98  |
| Cellular processes                              | Detoxification                          | 96  |
| Cellular processes                              | Other                                   | 92  |
| Cellular processes                              | Pathogenesis                            | 187 |
| Cellular processes                              | Sporulation and germination             | 705 |
| Cellular processes                              | Toxin production and resistance         | 94  |
| Cellular structure, organization and biogenesis | Amyloplast                              | 220 |
| Cellular structure, organization and biogenesis | Cell wall                               | 204 |
| Cellular structure,                             | Centrosome                              | 208 |

|   |                                  |     |
|---|----------------------------------|-----|
| organization and biogenesis                     |                                  |     |
| Cellular structure, organization and biogenesis | Chloroplast                      | 218 |
| Cellular structure, organization and biogenesis | Chromosome structure             | 213 |
| Cellular structure, organization and biogenesis | Cytoplasm                        | 206 |
| Cellular structure, organization and biogenesis | Cytoskeleton                     | 207 |
| Cellular structure, organization and biogenesis | Endoplasmic reticulum            | 209 |
| Cellular structure, organization and biogenesis | Endosome                         | 216 |
| Cellular structure, organization and biogenesis | Etioplast                        | 219 |
| Cellular structure, organization and biogenesis | Extracellular matrix             | 268 |
| Cellular structure, organization and biogenesis | Extracellular/secreted proteins  | 222 |
| Cellular structure, organization and biogenesis | Golgi                            | 210 |
| Cellular structure, organization and biogenesis | Intracellular transport vesicles | 211 |
| Cellular structure, organization and biogenesis | Mitochondria                     | 214 |
| Cellular structure, organization and biogenesis | Nucleus                          | 212 |

|   |  |     |
|---|--|-----|
| Cellular structure, organization and biogenesis | Other                                      | 223 |
| Cellular structure, organization and biogenesis | Peroxisome                                 | 215 |
| Cellular structure, organization and biogenesis | Plasma membrane                            | 205 |
| Cellular structure, organization and biogenesis | Plasmodesmata                              | 221 |
| Cellular structure, organization and biogenesis | Vacuole and lysosome                       | 217 |
| Central intermediary metabolism                 | Amino sugars                               | 100 |
| Central intermediary metabolism                 | Nitrogen fixation                          | 179 |
| Central intermediary metabolism                 | Nitrogen metabolism                        | 160 |
| Central intermediary metabolism                 | One- carbon metabolism                     | 698 |
| Central intermediary metabolism                 | Other                                      | 102 |
| Central intermediary metabolism                 | Phosphorus compounds                       | 103 |
| Central intermediary metabolism                 | Polyamine biosynthesis                     | 104 |
| Central intermediary metabolism                 | Sulfur metabolism                          | 106 |
| DNA metabolism                                  | Chromosome- associated proteins            | 170 |
| DNA metabolism                                  | DNA replication, recombination, and repair | 132 |
| DNA metabolism                                  | Degradation of DNA                         | 131 |
| DNA metabolism                                  | Other                                      | 130 |
| DNA metabolism                                  | Restriction/modification                   | 183 |
| Disrupted reading frame                         | NULL                                       | 270 |
| Energy metabolism                               | ATP-proton motive force                    | 111 |

|  |   |     |
|--|---|-----|
|  | interconversion                                 |     |
| Energy metabolism                      | Aerobic   | 108 |
| Energy metabolism                      | Amino acids and amines                          | 109 |
| Energy metabolism                      | Anaerobic                                       | 110 |
| Energy metabolism                      | Biosynthesis and degradation of polysaccharides | 105 |
| Energy metabolism                      | Chemoautotrophy                                 | 180 |
| Energy metabolism                      | Electron transport                              | 112 |
| Energy metabolism                      | Entner- Doudoroff                               | 113 |
| Energy metabolism                      | Fermentation                                    | 697 |
| Energy metabolism                      | Glycolysis/gluconeogenesis                      | 116 |
| Energy metabolism                      | Methanogenesis                                  | 159 |
| Energy metabolism                      | Other   | 184 |
| Energy metabolism                      | Pentose phosphate pathway                       | 117 |
| Energy metabolism                      | Photosynthesis                                  | 164 |
| Energy metabolism                      | Pyruvate dehydrogenase                          | 118 |
| Energy metabolism                      | Sugars  | 119 |
| Energy metabolism                      | TCA cycle                                       | 120 |
| Environmental Response                 | Cold  | 250 |
| Environmental Response                 | Drought   | 252 |
| Environmental Response                 | Flooding  | 253 |
| Environmental Response                 | Heat  | 251 |
| Environmental Response                 | Heavy metal                                     | 259 |
| Environmental Response                 | Insects   | 255 |
| Environmental Response                 | Light   | 249 |
| Environmental Response                 | Osmotic   | 258 |
| Environmental Response                 | Other   | 260 |
| Environmental Response                 | UV  | 257 |
| Environmental Response                 | Wounding  | 256 |
| Environmental Response                 | Xenobiotics                                     | 254 |
| Fatty acid and phospholipid metabolism | Biosynthesis                                    | 176 |
| Fatty acid and phospholipid metabolism | Degradation                                     | 177 |

|  |                       |     |
|--|-----------------------|-----|
| Fatty acid and phospholipid metabolism | Other                 | 121 |
| Gene/protein expression                | NULL                  | 67  |
| Gene/protein expression                | RNA synthesis         | 17  |
| Gene/protein expression                | embryonic development | 172 |
| Gene/protein expression                | protein synthesis     | 174 |
| Growth and development                 | Abscission            | 233 |
| Growth and development                 | Embryogenesis         | 224 |
| Growth and development                 | Fertilization         | 237 |
| Growth and development                 | Floral development    | 229 |
| Growth and development                 | Flowering             | 228 |
| Growth and development                 | Fruit ripening        | 226 |
| Growth and development                 | Germination           | 225 |
| Growth and development                 | Gravitropism          | 235 |
| Growth and development                 | Juvenility            | 230 |
| Growth and development                 | Other                 | 239 |
| Growth and development                 | Photoperiodism        | 231 |
| Growth and development                 | Phototropism          | 234 |
| Growth and development                 | Phytohormones         | 238 |
| Growth and development                 | Senescence            | 227 |
| Growth and development                 | Thigmomorphogenesis   | 236 |
| Growth and development                 | Vernalization         | 232 |
| Hypothetical proteins                  | Conserved             | 156 |
| Hypothetical proteins                  | Domain                | 704 |
| Hypothetical proteins                  | No database match     | 190 |
| Metabolism                             | NULL                  | 37  |
| Metabolism                             | amino acid            | 19  |
| Metabolism                             | cofactor              | 41  |
| Metabolism                             | energy/TCA cycle      | 46  |
| Metabolism                             | lipid                 | 32  |
| Metabolism                             | nucleotide            | 23  |
| Metabolism                             | protein modification  | 25  |
| Metabolism                             | sugar/glycolysis      | 18  |
| Metabolism                             | transport             | 30  |
| No database match                      | NULL                  | 269 |

|  |  |     |
|--|--|-----|
| Other categories                                   | Plasmid functions                                    | 186 |
| Other categories                                   | Prophage functions                                   | 152 |
| Other categories                                   | Transposon functions                                 | 154 |
| Pathogen responses                                 | Bacteria   | 240 |
| Pathogen responses                                 | Cell death   | 247 |
| Pathogen responses                                 | Fungi  | 242 |
| Pathogen responses                                 | Hypersensitive response                              | 245 |
| Pathogen responses                                 | Nematodes  | 243 |
| Pathogen responses                                 | Other  | 248 |
| Pathogen responses                                 | Pathogenesis- related proteins                       | 244 |
| Pathogen responses                                 | Systemic acquired resistance                         | 246 |
| Pathogen responses                                 | Viruses  | 241 |
| Protein fate                                       | Degradation of proteins, peptides, and glycopeptides | 138 |
| Protein fate                                       | Other  | 189 |
| Protein fate                                       | Protein and peptide secretion and trafficking        | 97  |
| Protein fate                                       | Protein folding and stabilization                    | 95  |
| Protein fate                                       | Protein modification and repair                      | 140 |
| Protein synthesis                                  | Nucleoproteins                                       | 139 |
| Protein synthesis                                  | Other  | 136 |
| Protein synthesis                                  | Ribosomal proteins: synthesis and modification       | 158 |
| Protein synthesis                                  | Translation factors                                  | 169 |
| Protein synthesis                                  | tRNA aminoacylation                                  | 137 |
| Protein synthesis                                  | tRNA and rRNA base modification                      | 168 |
| Purines, pyrimidines, nucleosides, and nucleotides | 2'- Deoxyribonucleotide metabolism                   | 123 |
| Purines, pyrimidines, nucleosides, and nucleotides | Nucleotide and nucleoside interconversions           | 124 |

|  |  |     |
|--|--|-----|
| Purines, pyrimidines, nucleosides, and nucleotides | Other  | 122 |
| Purines, pyrimidines, nucleosides, and nucleotides | Purine ribonucleotide biosynthesis             | 125 |
| Purines, pyrimidines, nucleosides, and nucleotides | Pyrimidine ribonucleotide biosynthesis         | 126 |
| Purines, pyrimidines, nucleosides, and nucleotides | Salvage of nucleosides and nucleotides         | 127 |
| Purines, pyrimidines, nucleosides, and nucleotides | Sugar- nucleotide biosynthesis and conversions | 128 |
| Regulatory functions                               | DNA interactions                               | 261 |
| Regulatory functions                               | Other  | 129 |
| Regulatory functions                               | Protein interactions                           | 263 |
| Regulatory functions                               | RNA interactions                               | 262 |
| Regulatory functions                               | Small molecule interactions                    | 264 |
| Secondary metabolism                               | Alkaloids                                      | 194 |
| Secondary metabolism                               | Glucosinolates                                 | 197 |
| Secondary metabolism                               | Non- protein amino acids                       | 196 |
| Secondary metabolism                               | Other  | 198 |
| Secondary metabolism                               | Phenylpropanoids/phenolics                     | 192 |
| Secondary metabolism                               | Phytoalexins                                   | 195 |
| Secondary metabolism                               | Terpenoids                                     | 193 |
| Signal transduction                                | Channels                                       | 266 |
| Signal transduction                                | G-proteins                                     | 202 |
| Signal transduction                                | Hormones, cytokines                            | 265 |
| Signal transduction                                | Immunology                                     | 267 |
| Signal transduction                                | Kinases  | 200 |
| Signal transduction                                | Other  | 203 |
| Signal transduction                                | PTS  | 700 |
| Signal transduction                                | Phosphatases                                   | 201 |
| Signal transduction                                | Receptors                                      | 199 |

|                                |  |     |
|--------------------------------|--|-----|
| Signal transduction            | Two- component systems                     | 699 |
| Transcription                  | DNA- dependent RNA polymerase              | 135 |
| Transcription                  | Degradation of RNA                         | 134 |
| Transcription                  | Other                                      | 133 |
| Transcription                  | RNA processing                             | 166 |
| Transcription                  | Transcription factors                      | 165 |
| Transport and binding proteins | Amino acids, peptides and amines           | 142 |
| Transport and binding proteins | Anions                                     | 143 |
| Transport and binding proteins | Carbohydrates, organic alcohols, and acids | 144 |
| Transport and binding proteins | Cations                                    | 145 |
| Transport and binding proteins | Nucleosides, purines and pyrimidines       | 146 |
| Transport and binding proteins | Other                                      | 147 |
| Transport and binding proteins | Porins                                     | 182 |
| Transport and binding proteins | Unknown substrate                          | 141 |
| Unclassified                   | NULL                                       | 35  |
| Unclassified                   | Role category not yet assigned             | 185 |
| Unknown function               | Enzymes of unknown specificity             | 703 |
| Unknown function               | General                                    | 157 |
| Viral functions                | General                                    | 175 |