

Genome annotation of the soil bacterium *Pedobacter heparinus*

Background

Many bacterial genomes have now been completely sequenced. To date, this amounts to over 2,000 bacterial species! Hundreds more are still being sequenced. This has opened the door to an in depth understanding of many bacterial species which play critical roles in certain ecological niches as well as those which are pathogenic to humans and animals. Recently, in a new initiative started at the Joint Genome Institute (JGI), the genomes of relatively uncharacterized bacteria are being sequenced. The goal is to generate a Genome Encyclopedia of Archea and Bacteria (GEBA). Often these bacteria are founding members of a new genus or phylum and are potentially important in our understanding of the diversity of bacteria which inhabit unusual environments. Many of the bacteria in the GEBA project also contain enzymes which might be useful in the development of biofuels.

Having a complete genome sequence for an uncharacterized bacterium is very helpful! It allows scientists to identify the whole complement of organismal genes which in turn provides crucial insights into the physiology of that organism. For example, it would allow one to investigate which metabolic pathways are used to provide energy for the organism. In other words, did the bacterial genome possess genes which code for all the enzymes in glycolysis, Krebs cycle and electron transport? Did the bacterial genome contain genes which would code for proteins necessary for manufacture of a means of locomotion, e.g., a flagellum? In one particular spectacular example, a recent study isolated DNA from water at the bottom of a deep hot spring. When the DNA was sequenced and the genes predicted, a chromosome for a single species of novel bacterium was predicted. From this information, the scientists were able to infer a lot about this new species and its lifestyle without ever having seen it!

Annotation

To enable such genome wide surveys of genes in a bacterial genome, the genome must first be annotated. Annotation is the process by which a raw DNA sequence is converted into a chromosomal map of accurately predicted gene segments. There are various levels of annotation. We will focus on the first step which is to *predict* genes using bioinformatic programs. For example, computer algorithms predict “open reading frames” (**ORF's**) and uses this information to make a gene “call” or “prediction”. An ORF begins with a canonical start codon ATG in a particular reading frame. The ORF continues until a stop codon is encountered in the same reading frame. The DNA sequence of these ORFs can then be compared to a large database of previously sequenced bacterial genes to decide if the predicted ORF is in fact bona fide.

There are now a number of automated *pipeline* gene annotation computer programs which can literally predict all the genes in a genomic DNA sequence in one step. The DNA sequence is uploaded into the program (fed into a pipeline) which then outputs a map of the genome with thousands of predicted genes arranged in order on the chromosome. This allows for very efficient throughput, but can also have an error rate of about 30%. Common errors include mis-identification of the start codon resulting in a gene which is shorter or longer than the actual gene and errors in the identification of the gene product. Hence, in genomes for which genes are predicted by a pipeline program every gene is still a hypothesis!

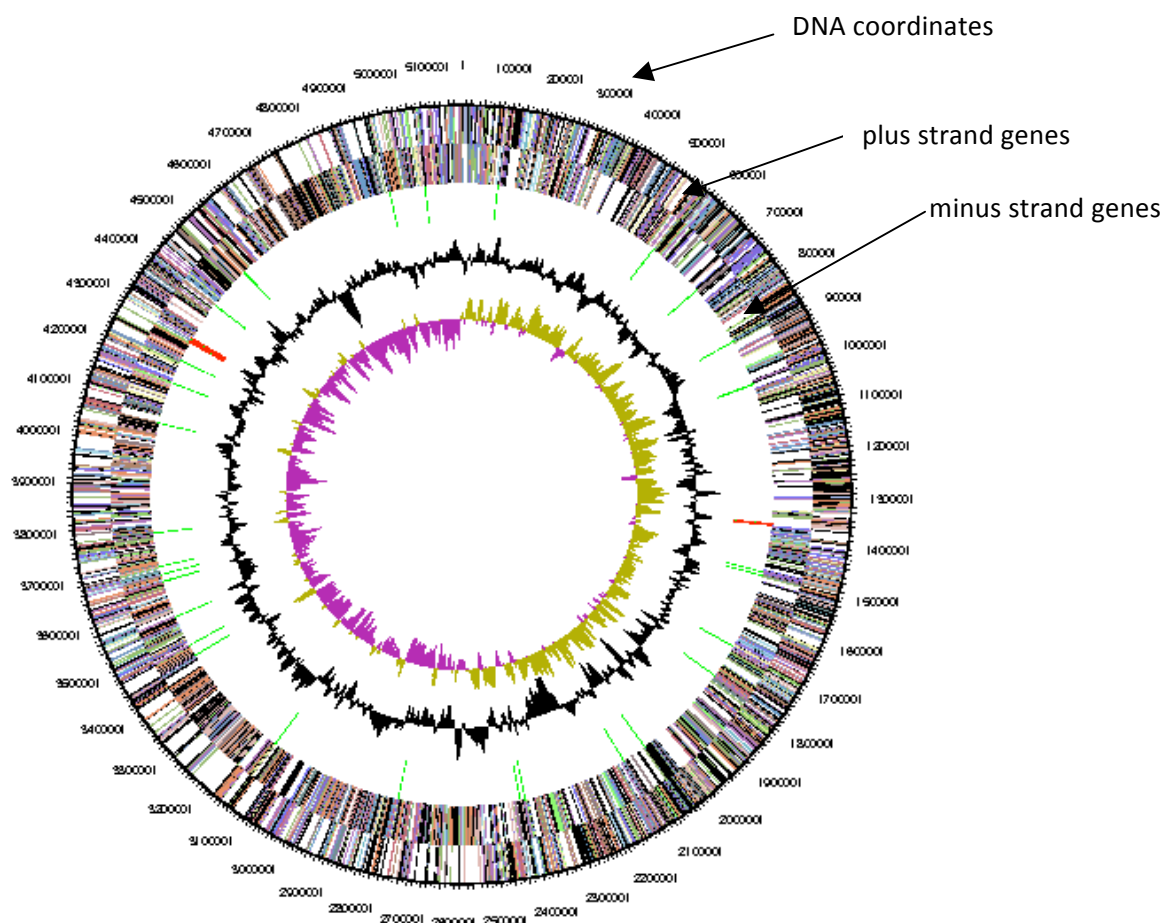
Our Annotation Project

Geneva College Biology and Chemistry Departments were selected by the JGI to help annotate the recently sequenced genome of the soil bacterium *Pedobacter heparinus* (*P. heparinus*). Its genome has been subjected to pipeline annotation. This is a GEBA bacterium and is a founding member of the genus *pedobacter*. It was isolated from dry soil and was studied primarily to learn more about some unique genes which code heparinase which degrade the polysaccharide heparin. Interestingly, *P. heparinus* has been found in close proximity to the roots of several plant species including rice and certain dune grasses and may have a symbiotic or pathogenic relationship with these plants. Dr. David Essig has organized a *P. heparinus* genome project to annotate all ~4,500 genes. A unique aspect of this genome project is that the annotation will be done by Geneva College undergraduate students in Biology and Chemistry.

Your task is to come behind these hypothetical predictions and to employ additional prediction programs and your biological expertise to verify or nullify the pipeline gene annotation hypothesis for genes in the novel bacterium *P. heparinus*. This allows the genome to be annotated to a higher degree of reliability for future research of the physiology of the organism.

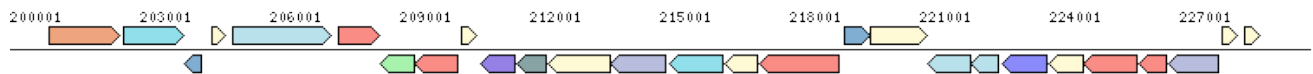
A whole genome map for *Pedobacter heparinus*

predicted by a "pipeline" annotation program.



The chromosome is circular and consists of ~4,300 predicted genes. The outer ring with numbers is the DNA coordinates (ranges from 1– 5,000,000) in base pairs. The next layer in represents genes on the plus strand of DNA. Each colored arrow shaped line represents a gene. The length corresponds to the number of base pairs in the gene. Absences of gene arrow corresponds to intergenic DNA. The next layer in represents the minus strand and its complement of genes. Notice that when there is an empty white space (no genes) on one strand, there are genes on the opposite strand. Most of the DNA is occupied by genes either on one or the other strand. There is relatively little intergenic DNA.

Below is a magnified section of the chromosome showing a subset of genes on both strands. Each gene is an arrowhead which points in the direction of transcription.



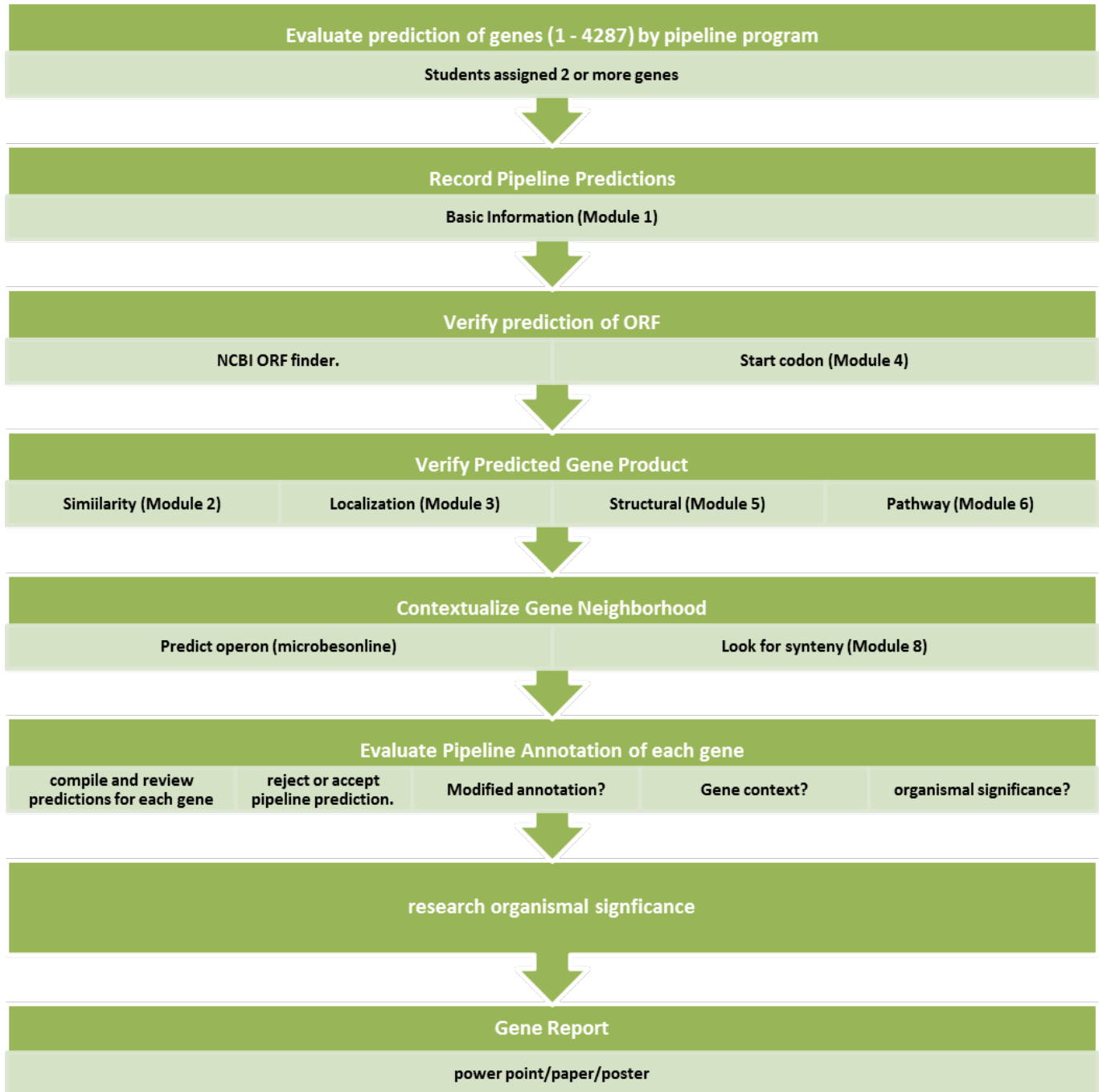
In the succeeding pages, instructions are provided for you to complete the annotation of one or more genes in the *P. heparinus* genome.

**Gene Annotation Manual
for
IMG-ACT**

**Dr. David A. Essig
Geneva College
2014**

With generous assistance by Hannah Dell (class of 2013)

Overall Process of Annotation



Gene Assignment and Lab Notebook Management

Enrollment in the course

1. You will receive an email from your instructor with a course token.
2. Go to the geni-act webpage <http://geni-act.org/>
3. Create an account (Sign Up button)
4. Once enrolled, your instructor will create a gene assignment for you. The assignment will include a link to a notebook wikipage for each gene.

Browser and Tab Management

1. The IMG ACT notebook application will **not** run in the Explorer browser. Make sure you are running either Firefox or Chrome.
2. To access your gene notebook, log in to your account on the main geni-act webpage.
3. Click on the hyperlink corresponding to your course.
4. Click on the hyperlink with your gene assignment.
5. Click on the hyperlink under the lab notebook for your gene.
6. Choose the gene you wish to annotate and click on link to access the lab notebook.
7. Within the notebook are a variety of hyperlinks which work best if you right click to open in a new tab. This will allow you to keep your notebook open and copy and paste from the analysis website.

The Notebook


1. The lab notebook is a wikipage (or like a Google Docs page) which means you and your instructor both have access to the page.
2. The lab notebook is a toolbox (10 modules) of state of the art online bioinformatic prediction tools which can be used to address a variety of questions regarding genome annotation.
3. The notebook is also a place to record the data that you collect from the various bioinformatic investigations.
4. Within each module (eg. **Sequence-based Similarity Data**, see below) are
 - a. hyperlinked Module Instructions
 - b. One or more headings which indicate a particular type of analysis (eg., **BLAST**)
 - c. Below each heading is a hyperlink which allows you to upload data to a website, have it analyzed and then the data sent back to you.
 - d. Beneath each hyperlink is one or more interactive fields which can be enacted by clicking on the notebook icon. This allows for recording data by copy and paste, making notes or for uploading of saved images.


[-] Sequence-based Similarity Data

[Module Instructions](#)

BLAST

go to BLAST at <http://www.ncbi.nlm.nih.gov/blast>

Gene product name (top hit) 

Organism 

Recording the Pipeline Predictions

Basic Information Module

Purpose

The purpose of this phase is to collect data from the Gene Detail Page for your gene. This page is found in the Integrated Microbial Genomes (IMG) database for Education (IMG/EDU) from the Joint Genome Institute. The Gene Detail page contains the predicted DNA and protein sequences as called by the Pipeline Prediction Program in a file format called FASTA. This FASTA format will be recognized by the various bioinformatic prediction programs. There is also access to one or more bioinformatic programs on this page as well and so you will want to keep this page open as a tab in your browser.

Instructions for Accessing Gene Detail Page for your Gene

1. Open the IMG/EDU website <http://img.jgi.doe.gov/cgi-bin/edu/main.cgi>
2. Click on the Find Genes tab and select Gene Search. Enter your gene locus tag (Phep_XXXX) in the keyword and select locus tag as the filter.

Gene Search

Find genes in selected genomes by keyword. It's required to add selections into "Selected Genomes" unless blocked.

Keyword:	Phep_0555
Filters:	Locus Tag (list)

3. Hit Go and your Gene Detail page will load in a new tab. Keep this tab open during your annotation.

Instructions for Entering Pipeline Predictions

1. Open the Basic Information module in the notebook in one tab. (Note: We will not be using the hyperlink for Gene Page associated with geni-act for this particular module.)
2. From the IMG Gene Detail page in another tab, copy and paste the left most DNA coordinate and paste into the lab notebook where indicated. Repeat for the right most DNA coordinate. This is the boundaries of your gene in the genome. Hit Save!
3. Go back to the IMG Gene Detail page. Next to the DNA coordinates on the Gene Detail page is the number of base pairs (sequence length). Right click on the hyperlink a new window will appear with the FASTA DNA sequence. Copy and paste this entire sequence (including the blue text on the first line) into the lab notebook text box. Type in the total length of the gene in the sequence length box. Hit Save! (**Tip:** FASTA format includes the >genenumber name of gene [organism]).
4. Go back to the IMG Gene Detail page and right click on the amino acid sequence length hyperlink. Copy and paste the amino acid sequence (including FASTA formatting again). Also, type in the sequence length of the protein.
5. Make sure you have saved your work!
6. Leave the Basic Information Module open since you will need to copy the protein sequence to paste/upload into various bioinformatic programs.

Confirm ORF Prediction for Assigned Gene

Alternative Open Reading Frame Module

During this phase of analysis you will decide if the DNA coordinates (beginning and ending base pair) for your gene are correct. In other words, can we trust that a gene's start and stop codon were predicted correctly by the pipeline annotation. To establish this we use a program which predicts the start codon based on its proximity to the consensus ribosome binding site in bacteria DNA (Shine Dalgarno sequence).

Background

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

The diagram illustrates the processes of transcription and translation. The top part shows a DNA double helix with a pink top strand (non-template) and a blue bottom strand (template). The DNA is divided into regions: Promoter (containing RNA polymerase recognition sites at -35 and -10, and the RNA polymerase binding site/Pribnow box at +1), Antileader, Coding region, Antitrailer, and Terminator. The bottom part shows an mRNA strand (yellow) with a 5' end and a 3' end. The mRNA has a Leader region (containing a Shine-Dalgarno sequence and a 30S ribosome binding site) and a Trailer region (containing a UAA stop codon). The AUG start codon is marked as the translation start (initiation codon). An arrow indicates the direction of transcription from left to right.

RNA polymerase recognition site

RNA polymerase binding site (Pribnow box)

Nontemplate strand

Template strand

DNA

5'

3'

Promoter

Antileader

Coding region

Antitrailer

Terminator

Shine-Dalgarno sequence

G or A

30S

AUG

mRNA

5'

3'

Leader

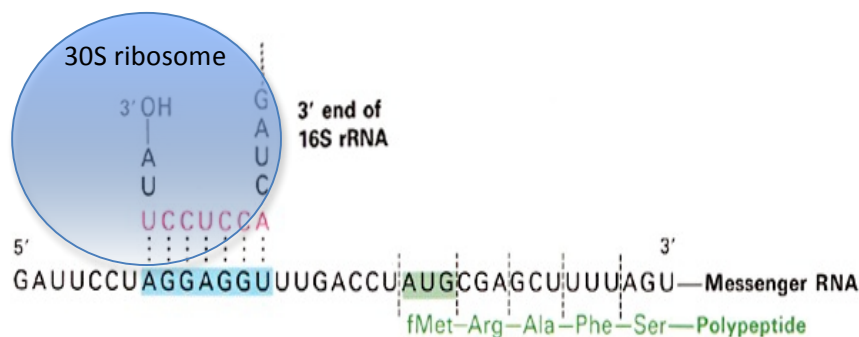
UAA

Trailer

Direction of transcription

Transcription start

Translation start (initiation codon)



Instructions

1. Go to the bottom of the Gene Detail page under the heading Evidence for Function Prediction.
2. Click on the hyperlink Sequence Viewer for Alternate ORF Search.
3. In the Select Gene Neighborhood, type in 100 in the bp upstream and in the bp downstream.
4. Enter 100 aa as minimum ORF
5. Select graphic view.
6. Click Submit
7. The output shows the predicted gene (start and stop codons in red) along with 100 bp of upstream and downstream DNA. The flanking DNA sequence is in green with the DNA sequence of the ORF in black.
8. Also displayed are the plus and minus strand each with the possible 3 reading frames. Make sure you look at the correct strand either plus or minus (refer to your DNA coordinates from module 1).
9. Besides the predicted start codon in red, other possible start codons are highlighted in yellow. Remember that these start codons are in all possible reading frames on the plus and minus strand.
10. Shaded in blue is a predicted Shine Dalgarno (SD) sequence which encodes the binding site for the small ribosomal subunit.

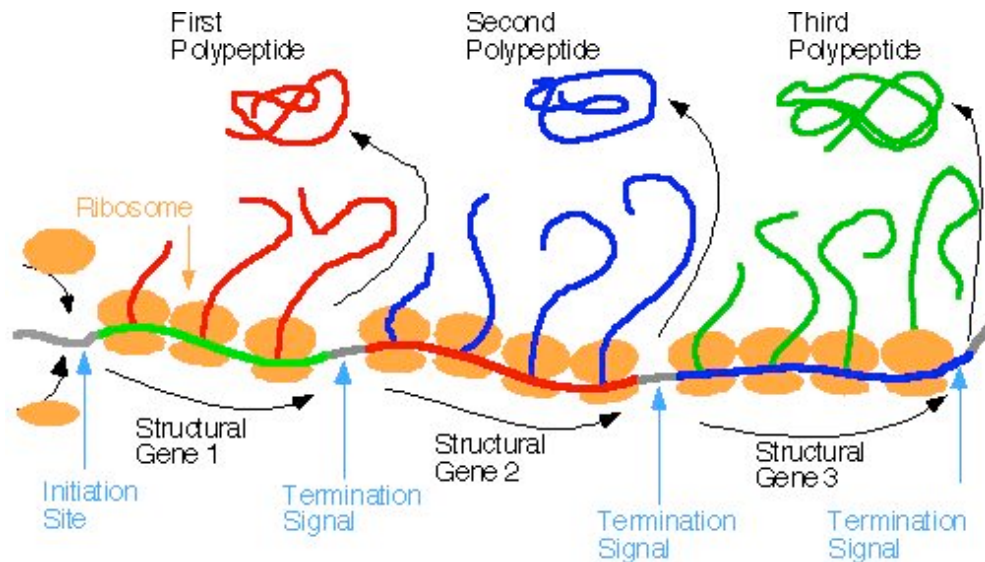
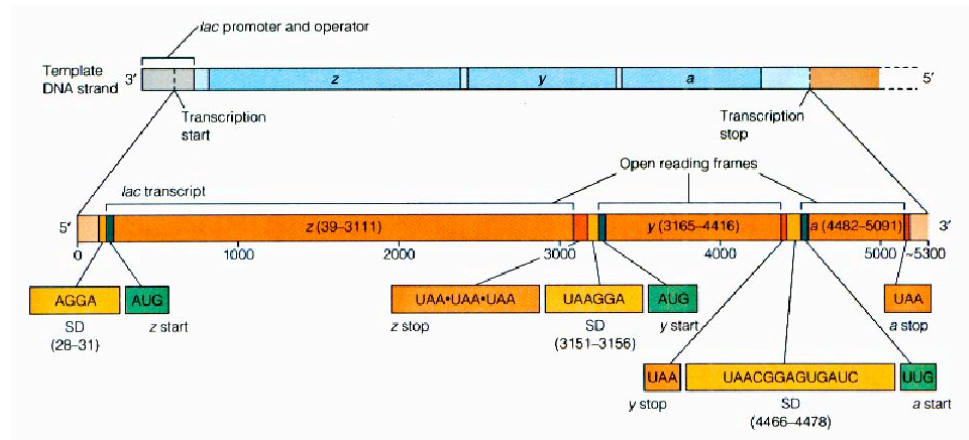
[For review see figure below: Part A. This sequence is the coding strand. Shaded in blue is the sequence **AGGAGGT** just upstream of the ATG which encodes the SD. Part B. The anticoding strand of the gene. Part C. When the mRNA is transcribed from the gene, the SD sequence is included and is part of the leader portion of the mRNA upstream from the AUG. The SD sequence is typically located 5 – 13 bp upstream of the bona fide start codon.]

A.	DNA	5'-GATTCCT AGGAGGT TTGACTTAACCGCACCT ATG CGA
B.	DNA	3'-CTAAGGTTCTCCAAACTGAATTGGCGTGGATACGCT
C.	mRNA	5'-GAUUCCA AGGAGGU UUCACUUAACCGCACCU AUG CGA
D.	Protein	met – arg –

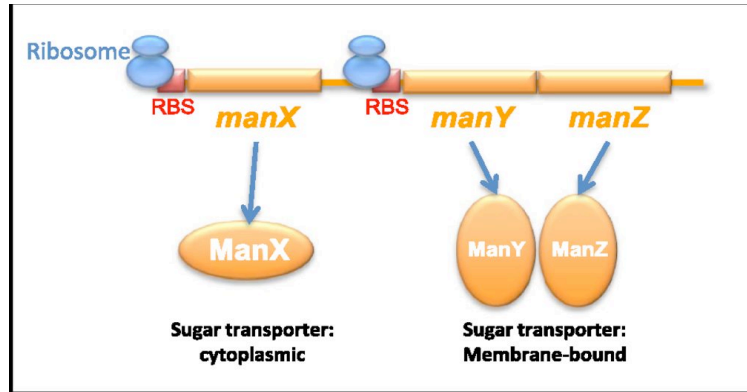
11. Your job is to assess whether there is a Shine Dalgarno sequence and if it is located 5 – 15 bp (or close to that range) upstream of the start codon. If the Shine Dalgarno is 50 bp upstream, then perhaps you should look for an alternative start codon. **You should assume that the pipeline start codon prediction is in the correct reading frame.** Hence, if there is a possible *alternative* start codon (i.e., the correct start codon) it will only be found in the same reading frame. In other words, it will either be a short distance upstream (perhaps closer to the Shine Dalgarno) or possibly downstream of the “called” start codon, but in the same reading frame. Alternative start codons in a different reading frame that look to be in the correct location relative to the Shine Dalgarno can be disregarded.
12. Sometimes you may not find a predicted Shine Dalgarno sequence.
 - a. One reason for this is that this sequence was based on work done in *E. coli* and may not apply to other microbial species.
 - b. A second reason is that your gene may be in an operon. An operon is composed of several genes in close proximity to each other (often only a few bp apart). This set of

genes is co-transcribed into one mRNA (polycistronic mRNA). This polycistronic mRNA is also translated with each gene's mRNA coding sequence producing its individual protein (see figures below). In most cases, each gene's mRNA has a SD sequence in its leader so the translational initiation occurs 5' of each mRNA coding region.

Multiple Shine-Dalgarno sequences in the polycistronic *lac* mRNA



However, some genes are so close together that no SD is present and the translational initiation in the preceding gene allows the ribosome to simply pass to the neighboring gene (as in gene *manY* to *manZ* below)



- Record your findings in the text boxes in Module 4 even if it simply confirms the pipeline prediction. Give a brief summary of how you verified your prediction even if it was the same as the pipeline.

Verification of Predicted Gene Product

Sequence Similarity Suite of Programs

Cellular Localization Program

Structural Similarity Suite of Programs

Pathway Prediction

General Introduction to Sequence Similarity Prediction Programs (Module 2)

During this phase of annotation we will utilize several programs which help identify the protein product of your gene. This is crucial since often the gene identity is based on the product identity or function. All the programs you will use have their basis in the comparison the amino acid sequence of your gene (the query gene) to a database of subject or reference protein amino acid sequences in bacteria and other organisms. The amino acid sequence of a particular class or type of protein is often highly similar between species. Statistical analysis can help us to decide if the similarity is real or due to random chance. Defining the proteins which are homologous to your gene's predicted amino acid sequence helps to confirm the name of the gene and insights into the function of the protein product.

The sequence similarity programs are in Module 2 and include

1. BLASTp
2. Conserved Domains Database (CDD)
3. T-COFFEE
4. WebLogo

BLASTp Alignment Prediction Program

Introduction

The first program we will utilize is NCBI BLASTp (Basic Local Alignment Search Tool protein). This program seeks to create an alignment between your protein (query) and a subject protein. Only amino acids which are identical are aligned. This creates a maximum score. However, amino acids which have the similar side chain chemistry are scored positively though less than identical amino acids. Insertions and deletions are made to maximize the alignment but reduce the overall score.

Instructions

1. Copy the FASTA protein sequence from the Basic Information Module.
2. Go to the Sequence Similarity Module and right click on the link to BLAST program.
3. Paste in the FASTA sequence into the box in the BLAST page. This is referred to as the **query** sequence. Note that there is a drop-down box with several database options. We will be using the Non-redundant protein sequences (nr), but there are other options that you can use.
4. Click the big blue BLAST button at the bottom of the screen. It will take a few seconds to load completely. Just be patient.
5. Scroll down to the "sequences producing significant alignments" section. Note that the very top one (1st hit) is a perfect match; do not use this one.
6. Click the hyperlink description of the subject gene in the 2nd hit and this will take you to the actual alignment of the query and subject proteins.
7. Look at the statistics and names at the top of the alignment.
 - a. The *score* is a similarity score. The higher the score the greater the percentage of identical amino acid sequences that can be aligned. The score depends upon the length of alignment and whether the program had to introduce insertions or deletions to obtain alignment with the subject protein.
 - b. Examine the E-value (expect value). Values less than 10^{-3} indicate that the alignment of your query sequence with a subject protein is statistically significant (not due to random

chance). The score is sensitive to the number of subject proteins available in the database.

- c. Determine the alignment.
 - i. Calculate the numeric alignment of the query with the subject protein. This is simply a measurement of the total length of alignment. The length can be calculated by subtracting the position number of the first amino acid (N terminus) aligned minus the last amino acid aligned (C terminus) in the query protein plus 1.
 - ii. Also, record the alignment as a percentage of the total number of amino acids in the protein or *query coverage*.. This number can be found in the table of alignments in the 3rd column.
- d. Record in the lab notebook each of the following: Gene product name, organism, alignment length and query coverage, Score, and E-value. If you are working with a hypothetical gene, did any of the homologous genes give an actual product name? If so, this is worth noting in the concluding comments of the sequence similarity module.
- e. Copy and paste the alignment into the lab notebook (starting with "query 1"). Repeat steps 2e - 2k for the next hit .

Conserved Domain Database (CDD)

Introduction

The next program will look for predicted "Conserved Domains" in the protein product. A hit means this domain (defined as a segment or module of the amino acid sequence found in a specific region of the protein) has been aligned in a particular database . These domains are indicative of function for a protein. For more background on protein domains refer to the introduction for Pfam below.

Instructions

1. Scroll to the top of the BLAST results page. If there is a graphic that says "putative conserved domains have been detected," Click the graphic. (If there is no graphic, record in the lab notebook that "no putative conserved domains found").
2. Under "list of domain hits" search for domains which have the prefix COG or CDD followed by a number. Other domain databases are also listed in particular Pfam or Tigrfam but are not to be recorded here. We will search these databases in a separate module.
3. Record the COG or CDD number in the lab notebook.
4. Record the name of the domain (listed under description)
5. Record the E-value.
6. Repeat steps 3b – 3e for the second hit.
7. **Leave your BLAST results page open for your next analysis!**

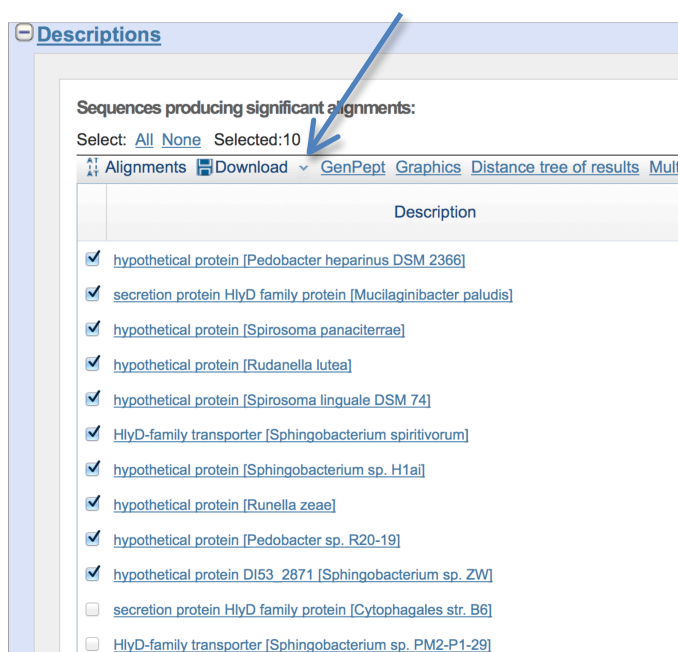
T-COFFEE Multi-alignment Prediction Program

Introduction

Our next analysis will be to see if the predicted protein can be aligned with their closest homologous proteins. Multiple alignments such as T-COFFEE attempt to align the entire protein amino acid sequences in a graphical manner. Successful alignments gives powerful evidence as to the identity and function of the predicted protein. Sometimes these are helpful to begin to identify domains within the protein.

Instructions

1. Return to your BLAST output and scroll down the page to the Descriptions section as shown below. Select the top 10 hits.



2. Click on carrot next to Download button to download all 10 protein sequences. A dialog box will open with a list of options. Make sure FASTA sequence is selected and click Continue.
3. A dialog box will open and ask if you want to open or save file. Click open with text editor.
4. You will now have all the FASTA sequences for each of the 10 proteins which you wish to align. This is known as a sequence dump.
5. Create a space between each sequence by clicking in front of each >gi and hitting return. This will format the file.
6. Copy the entire file and paste it into the field entitled "Sequences used for alignment" in your notebook.
7. Right click on the T-coffee link in the notebook
8. Paste the amino acid sequences you just copied into the T-coffee search box and click submit.
9. Copy and paste the resulting multiple sequence alignment into the lab notebook. Save the notebook page.
10. Note the symbols at the bottom of row of the alignment. The * indicates when all aligned proteins possess the identical amino acid at a particular amino acid position. The : indicates

high frequency of identical amino acids and conserved (similar side chain chemistry) at that position and the . indicates some identical amino acids and somewhat less frequent conservative substitutions at that position. What can you learn about your protein from this information??

WebLogo Multi-alignment Display Program

Introduction

On the notebook page click the link for WebLogo or copy and paste the URL given. WebLogo creates a “consensus” histogram of the 10 protein multiple alignment created in T-COFFEE. This is essentially the same sort of graphical analysis as T-COFFEE but somewhat easier to see patterns of sequence conservation between the aligned proteins due to the use of color symbols.

Instructions for WebLogo

1. Paste the sequence alignment from the T-coffee search (do not include the header that starts with “clustal”)
2. Click “create” at the top of the page. Wait for settings page to open.
3. Settings:
 - a. logo size per line = 18 X 10 cm
 - b. multiline logo (symbols per line) = 32 (check box)
4. At bottom right, click Create Logo.
5. The image created will appear de-magnified. Click to magnifier icon to enlarge.
6. Right click to copy image and paste in the Sequence Logo field in your notebook.
7. The output shows the consensus amino acid sequence based on the multiple sequence alignment. A single tall amino acid symbol letter indicates that the same amino acid was present in all 10 homologs. Some positions have more than one amino acid and the frequency is inferred from the relative height of each amino acid symbol letter stacked on each other.
8. Like with the T-COFFEE, you can easily identify places where there are clusters of sequence conservation which might indicate regions of possible functional significance. Make notes concerning the relative positions (coordinates) where there are a series of tall letters. Limit your investigation to segments where you find a minimum of 15 – 20 amino acids where the majority are tall letters.

Instructions for Recording Comments/Observations for Sequence Similarity

1. Summarize what you found for BLASTp, Conserved Domain, T-COFFEE/BioLogo.
2. Be aware of the species compared. Note if the top hits for the query protein is in the same genus i.e., pedobacter or in the same phylum which bacteroidetes. You can Google any species to find its taxonomy or go to this link http://img.jgi.doe.gov/cgi-bin/w/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=644736398
3. Pay particular attention if neither top hit is in the same genus or phylum. Make note what species the two hits were and look up its phylum.
4. BLASTp - significant alignments mean evolutionary relationship and more importantly shared structure and function → identify protein

5. T-COFFEE and WebLogo – you can now begin to see how much the newly identified protein resembles its closest relatives. In particular where is most of the amino acid sequence shared? How might this help understand function?
6. SAVE MODULE

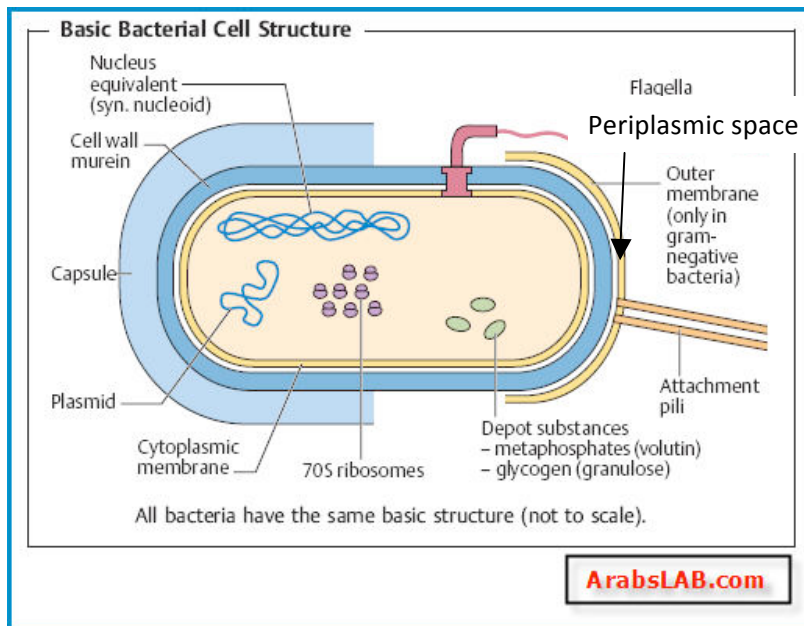
General Introduction to Cellular Localization Prediction Programs (Module 3)

This is series of programs which will allow you to predict the Cellular localization of your predicted protein. Knowing where a protein might be located provides additional insight into its function. For example, knowing that protein is found in the cytoplasmic membrane helps to narrow down its possible function to something like a receptor or ion channel. In contrast, a protein which is predicted to be found in the periplasm is secreted and may be an enzyme which degrades material in the extracellular environment of the bacterium

We utilize the following programs found in **Module 3** to predict cellular localization of the gene product.

1. TMHMM
2. Signal P
3. PSORT
4. Phobius

Before you start, review the basic anatomy of a bacterial cell. Pay particular attention to the cell wall/outer membrane (since we are dealing with a gram – bacteria), periplasmic space, cytoplasmic membrane, and cytoplasm.



Gram stain

Introduction

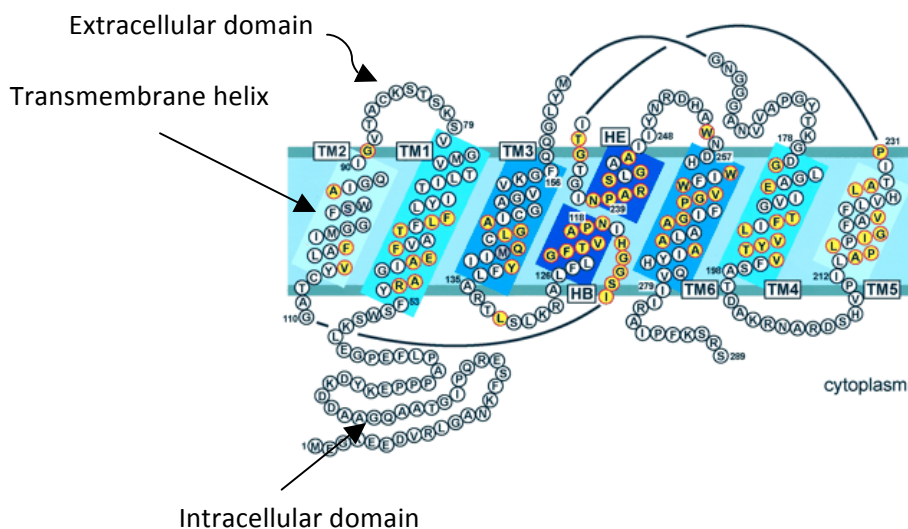
Bacteria are classified based on the properties of their outer membrane. A gram positive bacteria stains positively with a gram stain and has a thicker capsular layer while bacteria which stain negative have a thinner fibrous outer membrane like a cell wall in plant cells. The type of outer membrane (gram – or + bacteria) must be taken into account by the localization prediction programs.

Instructions

P. heparinus is a gram negative bacteria and therefore enter “negative”

Background Biology on Cellular Localization Process

This program predicts the presence of transmembrane helices of a membrane protein. These helices are secondary structures which thread in and out of the membrane as shown below. Numbers of helices vary from 1 to 7 and sometimes more.



Proteins which localize in the membrane are directed to the membrane by a signal peptide. The signal peptide is the first ~30 amino acids at the N terminus. The signal peptide contains + charged amino acids and hydrophobic amino acids which allow it to insert in the lipid bilayer. The signal peptide is cleaved off with a peptidase enzyme once the protein inserts in the membrane.

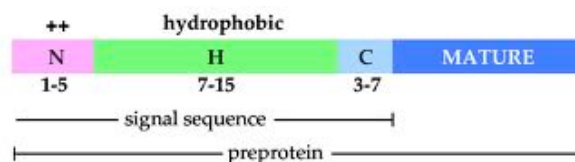
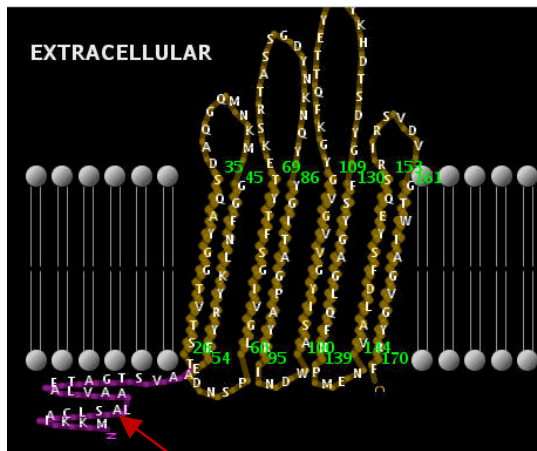
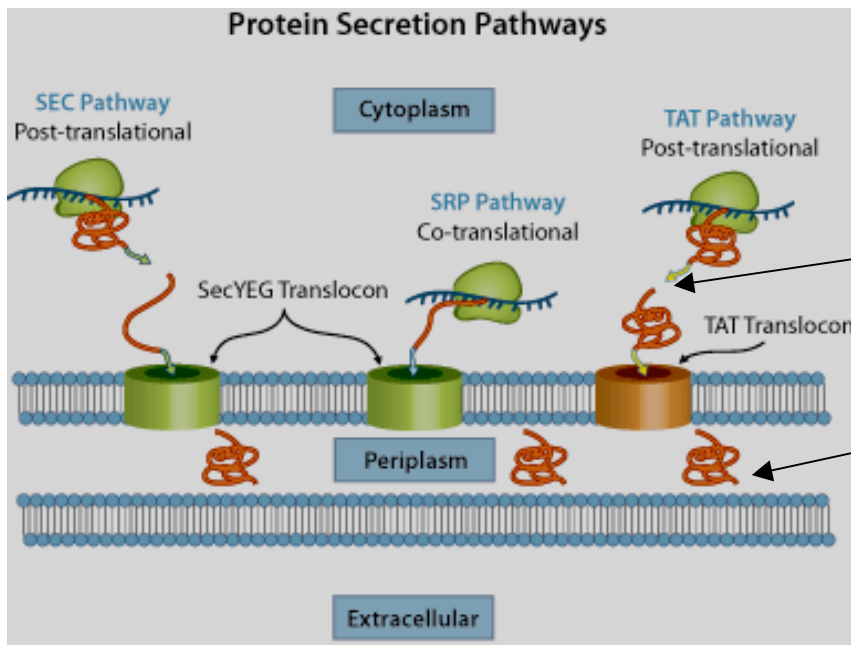


Figure 1. 3: The tripartite pattern of signal sequence



Membrane protein with 7 TM helices

- a. Proteins which are secreted out of cells are also directed to the membrane by a signal peptide. The signal peptide is cleaved once the protein traverses the membrane.



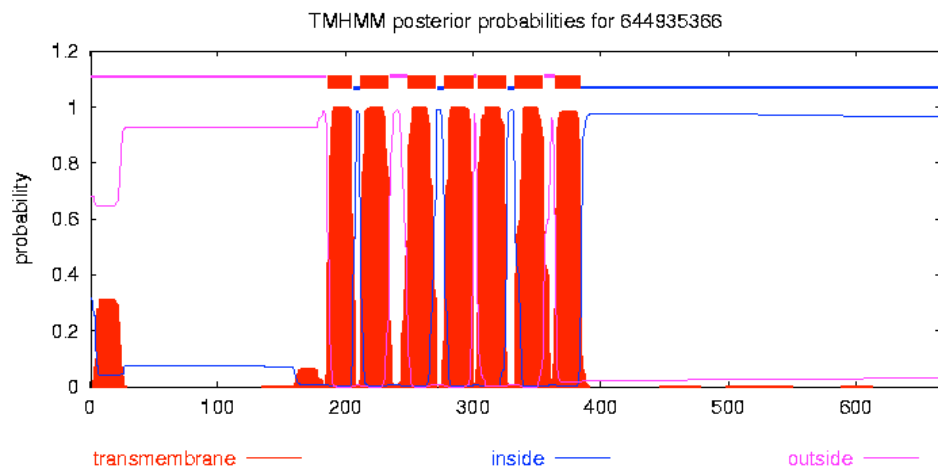
TMHMM

Introduction

The TMHMM program predicts the membrane spanning domains based on their high level of hydrophobicity. It also tends to predict signal peptides which have similar levels of hydrophobicity and thus signal peptides may be mistaken for a transmembrane helices.

Instructions

1. Copy and paste the FASTA amino acid sequence found on the gene detail page or from your locus gene page. Click submit.
2. The output will consist of a list of data and a graph.
3. Record the number of transmembrane helices in the lab notebook.
4. Look at the graph:
 - a. On the Y axis is the probability of TM helices (0 to 1). On the x axis is the amino acid positions along the length of your protein (N terminus to C terminus).
 - b. Should there be predicted TM helices, one or more red hills will appear in the graph. These are rise and decline for the probability of TM helices in this particular segment. Between TM helices the protein will either be inside the cell (blue line) or outside the cell (purple). See diagram below:



1. If there is no TM helices predicted (no red hills), **stop!** No information about cellular location can be inferred from the graph.
 2. If there is a TM helix, near the beginning of the protein, take note. This may be a signal peptide!
- c. Right click on the graph and save as a .PNG file in your image folder. Upload it to the notebook using the image icon as with the BioLogo image.

Signal P Signal Peptide Prediction

Introduction

This program predicts the presence of a signal peptide using two criterion. The first is the whether the N terminal region where signal peptides are found contains the expected positively charged and/or hydrophobic side chains characteristic of the first 30 amino acids. The second criterion is whether there is a consensus signal peptide cleavage site in the vicinity towards the end of the signal peptide at amino acid position 30, for example.

Instructions

1. Open a new link under SignalP
2. Paste the FASTA format amino acid sequence, select gram negative bacteria, and click submit
3. The output will consist of a list of data and a graph.
 - a. On the graph as with THHMM the Y axis is probability and the X axis is amino acid sequence.
 - b. The red vertical lines represent the C score or probability of a signal peptide cleavage site. Most are at 0.1. If there is a signal peptide, the height of the line or lines will rise to and perhaps exceed 0.5.
 - c. The green line is the S score and is the probability for the presence of the signal peptide itself (hydrophobic amino acid side chains).
 - d. The blue line (Y score) combines the S and C scores to predict a signal peptide. If the blue line is above 0.5 then there is good probability for a signal peptide.
 - e. There is also another score not on the graph but listed below the graph. This is the D score. It takes into account the mean S score and the Y score. It is the most reliable predictor of the signal peptide.
4. Record in your notebook
 - a. The most likely cleavage site (position for max C)
 - b. The D value. Note if it is greater than the cutoff, record the probability in the lab notebook.
 - c. Save the graph and upload it to the lab notebook as for THHMM.

PSORT-B Cell Localization Prediction

Introduction

This program predicts the likelihood that your query protein amino acid sequence will align to the amino acid sequence patterns characteristic of proteins found in various locations in bacterial cells.

Instructions

1. Open the link under PSORT-B
2. Paste the FASTA format amino acid sequence **from the IMG Gene Detail Page (for some reason the amino acid sequence from the Geni Act Gene Page will not work)**, indicate gram negative, and select "via web" under the Show results option. Click submit.
3. Look at the localization scores. Record these in the lab notebook, as well as the "final prediction".

Phobius TM and Signal Peptide Prediction

Introduction

Like TMHMM, Phobius predicts membrane spanning domains but also predicts and distinguishes between a signal peptide and transmembrane helices.

Instructions

1. Open the link under Phobius.
2. Paste the FASTA format amino acid sequence and click submit.
3. Look at the graph.
 - a. This is very similar to TMHMM (probability vs. length of the protein) but with one exception. Phobius predicts both signal peptide and TM helices. The red line indicates probability of a signal peptide. It should be found only in the N terminal part of the protein. Why?
 - b. Count the number of TM helices predicted for this protein. How does the number compare to that predicted for TMHMM? Why might TMHMM have a greater number?
4. Save the graph and upload it to the lab notebook.
5. Using evidence from each of these searches, hypothesize where the protein is found in a cell, and some structural aspects of it.
6. SAVE MODULE BEFORE MOVING ON

General Introduction to Structural Similarity Programs (Module 5).

This is a series of programs which allow in depth analysis of your gene product in terms of its structural features. Structure can mean in this context either protein domains (analyzed by Pfam) or protein 3-D shape or tertiary structure (analyzed by PDB). ***We will not use TigrFam.***

Background on Protein Domains

Domains are modules or segments of the protein product. Proteins often have one or more domains which mediate the function or functions of that protein. For example, bacterial receptors belong to one family. Each member of the family will share a common domain and but also have one more different type of domains which lead to some specialization of function. For example, in the figure below, all members of the MCP receptor family have the MA domain in common but only a subset have a second LBR domain while even fewer have a third TM domain. Domains are kind of like Legos pieces. Which Legos pieces were used in the design of the protein determines the structure and function of that protein. Pfam will be used to compare the amino acid sequence of your query protein to a database of known domains. A significant score and E value for a domain or domains in your protein will indicate the level of confidence you can have that your protein has that particular domain(s).

Knowing what domains are present in your protein allows you to compare these Pfam predictions with the conserved domains you found with your multiple alignments (BioLogo). Are the domains predicted to be significant using Pfam the same ones found with the alignment in BioLogo?? Did BioLogo find additional conserved regions?

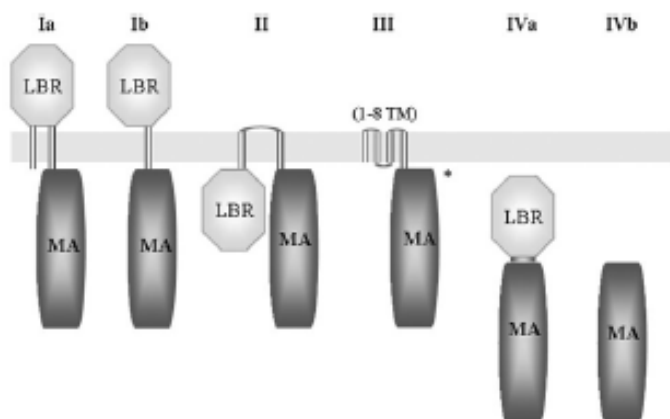


Fig. 4. Classification of MCPs into six different topologies. In total 3521 MCP sequences were analysed. Classification is based on domain annotation by SMART (Schultz *et al.*, 1998) and the prediction of TM regions by DAS (Cserzo *et al.*, 1997). The relative abundance of receptors with a given topology in bacteria and archaea is indicated. MA, methyl-accepting domains; LBR, ligand binding region. The asterisk "*" indicates that in a few cases only an LBR at the C-terminal extension of the TM regions is found.

Pfam Protein Domain Prediction

Introduction

This algorithm is based on high quality manually curated data base and alignment with an HMG logo. An HMG logo is a consensus sequence of a particular domain found in the data base. A consensus sequence is obtained by a BioLogos type analysis and represents the frequency of amino acid types at all the positions across the proteins sampled. The HMG logo represents each amino acid but weights the strength of consensus (shorter letters in Biologos) depending on how frequent an amino acid appears at a particular position. For example, in a 100 amino acid HMG, the amino acid tyrosine may always appear (100%) at position 10 in the HMG while at another position let's say 20, the amino acid phenylalanine occurs only in 60% of the proteins in the consensus database. So when Pfam compares your query protein amino acid sequence to a subject protein, it will calculate an alignment score and E value to allow you to decide if your query domain is homologous to the subject domain and to establish homology.

Instructions

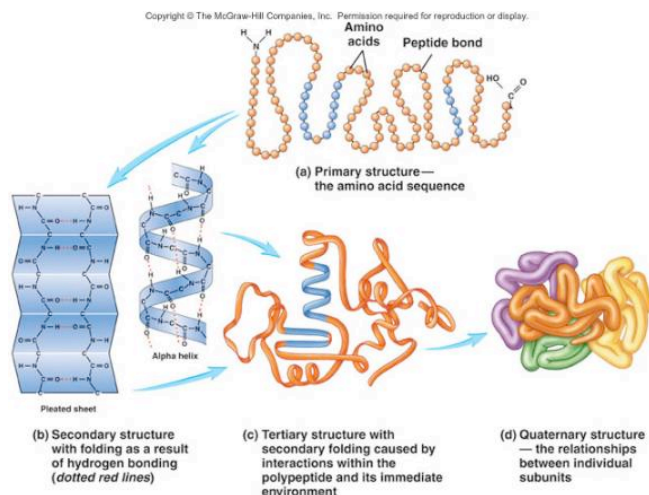
1. Within Module 5, click on link to Pfam. Make sure that it opens in a new tab.
2. Copy the FASTA amino acid sequence from the Gene Detail page into the Sequence Box and click submit.
3. The results include Pfam A and Pfam B searches either significant or insignificant. **Focus on the Pfam A significant matches which are located in a table.** The table reports the alignments of your query protein amino acid sequence with the domain HMM. From this table you can derive the requested to record in the notebook. [**Note: If you have more than 2 domains in your protein, create other sets of text boxes in the notebook to accommodate these extra domains. Every significant domain can be important!**]
4. To record the Pfam number click on the hyperlink listed under Family. This will take you to a page on which you can learn more about this particular domain. At the top left, you will see a number PFXXXXX). Copy and paste into the notebook.
5. Go back to the sequence search results tab and memorize the clan number. Type it in the notebook. The clan is basically a collection of related domain families. Kind of like Scottish clans!

6. Now go back to the sequence search Pfam matches and click on the clan number hyperlink. On the page that opens, copy the clan name and paste into the notebook.
7. Go back to the sequence search Pfam matches and copy and paste the Bit score and E values into the notebook text boxes.
8. Go back to the sequence search Pfam matches and click on the Show alignment button. Here is how to interpret the alignment data:
 - a. The alignment will be displayed in a manner somewhat reminiscent of a BLAST alignment. At the top of is the HMM with single letter abbreviations of for each amino acid. The ones in upper case signify amino acids which are identical across the homologs used to generate the HMM. The lower case letters indicate positions where the amino acid has been found to be of several types across the HMM.
 - b. The domain from your query protein is in the lowest row designed SEQ. All the letters are upper case and there may be insertions added (indicated by dashes) to improve alignment.
 - c. In the second row designated MATCH, is an indication of how well the query matches the HMM. The same letter indicates a perfect match whereas a "+" indicates a conserved amino acid substitution (amino acid substituted as same side chain chemistry).
 - d. The third row is PP which posterior probability. PP indicates is a measure of the confidence in the alignment. An "*" indicates a high probability. Numbers indicate lesser probability. If you mouse over the query sequence a color scale will appear which help you interpret the PP. Green shaded query sequence = strong PP. More green indicates a more valid alignment!
9. With the alignment graphic still displayed, take a screen shot and paste into the designated textbox in the note book. Click save.
10. Fill in the textbox asking for key functional residues with the amino acid letter(s) which appear as capital letters in both the HMM and query in the alignment. Indicate the relative position of the amino acid after the letter. For example, G17. Record if any of the key residues found in the HMM are also found in your query protein??
11. To display the HMM logo (just for reference) which is the top line of the alignment, you can click on the family name hyperlink. The screen that opens will provides a summary of that HMM. To the left, you will see a sidebar. On that list is a HMM log. Click on that hyperlink and you will find a biologo consensus sequence of the HMM. You can upload and insert this into the notebook where indicated.

PDB Tertiary Structure Prediction

Introduction

The amino acid sequence of a protein dictates the shape of the protein and its corresponding shape. The shape will then specify the function of your protein in its particular place inside or outside the bacterium. (For review, see the figure below). With Protein Data Base (PDB), your protein amino acid sequence will be compared to the amino acid sequence of a subject protein which has been studied in the lab to determine its 3-D shape. (This is often done by purifying the protein and then crystalizing it for X ray diffraction analysis.) If PDB can align your query gene with this subject gene, then the inference can be made that your query gene will assume a similar 3-D shape!



PDB is very similar to BLASTp. Query sequences are submitted to a data base to search for alignments with previously studied whole proteins (subjects) with a known amino acid sequence. Hence a query which aligns with the subject in PDB can now be predicted to assume a particular conserved 3-D shape found in the subject protein. As in BLASTp, there are bit scores and E values associated with each alignment.

Instructions

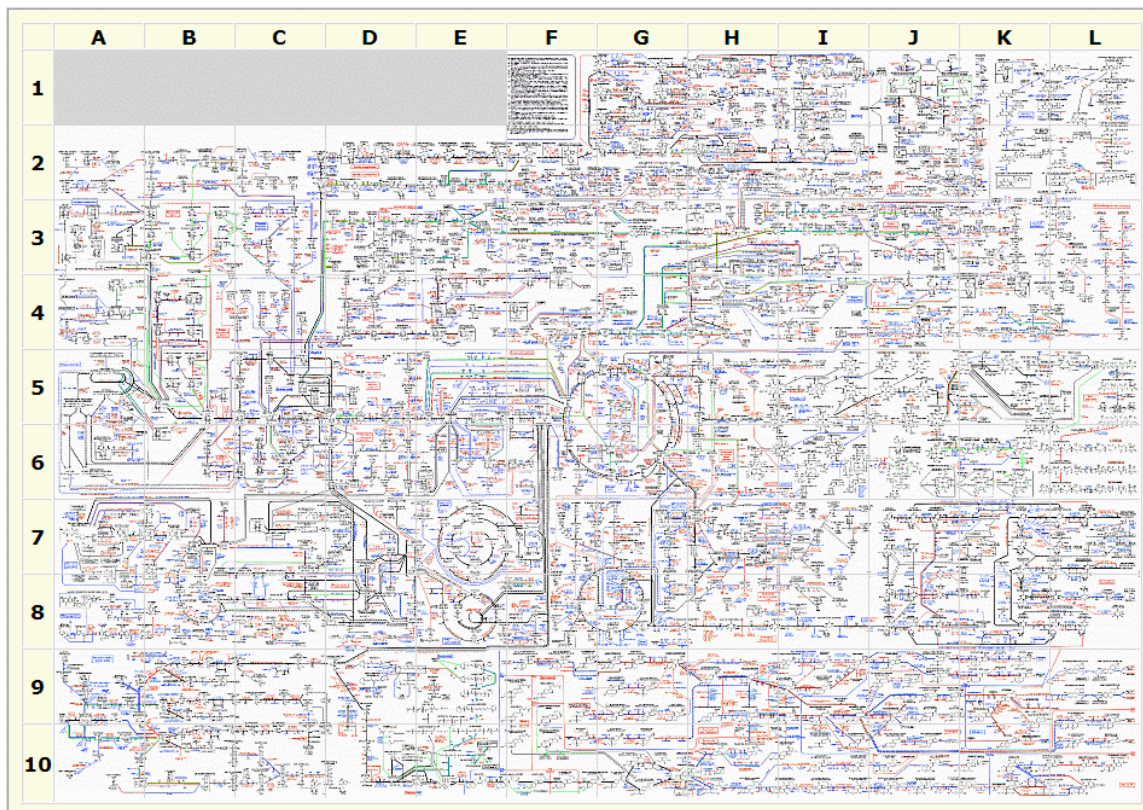
1. Copy the Fasta protein sequence and click on the link to the PDB website (in your notebook).
2. In the page that appears click on sequence located near the top center. Paste your sequence into the box and click search.
3. The results are alignments are similar to the Pfam output and are arranged in order of the lowest E score.
4. For the top hit, copy and paste the requested information (code is 4 digits upper left in large font) , name = title next to code, and E value in the specified text boxes.
5. Also copy and paste the alignment image in a manner similar to that used for the BLASTp and Pfam domain alignments.
6. Overall analysis: The strength of the alignments from Pfam and PDB will give you evidence as to the homology of your protein to well-studied proteins from other species. Long alignments and low E values will give you confidence in your ability to assess the validity of the pipeline annotation. Both these analyses give you insights into function of the *P. heparinus* gene products especially when you combine this. As a quick example, think about a protein which seem to have a domain near the N terminus which binds small MW ligands. The same protein is predicted to be a membrane protein and has a domain at its N terminus which is predicted to be in the periplasm. Starting to look like a receptor binding domain?? If it has such a domain, then maybe this protein is a receptor? This type of information will be used in your final annotation at the end of the notebook.

General Introduction to Pathway Prediction Programs (Module 6)

A pathway can be defined as “set of proteins with which your protein interacts either directly or indirectly”. There are literally hundreds of pathways in a cells (see figure below). Typically “pathway”

means an enzyme linked metabolic pathway where the product of one enzyme protein becomes the substrate for another enzyme protein. However, the term pathway can also describe the assembly of a multi-protein complex such as a ribosome or a motility apparatus (flagellum) or membrane protein (cytochrome oxidase or an ion channel). Pathway also means the interaction of transcription factor protein with DNA!

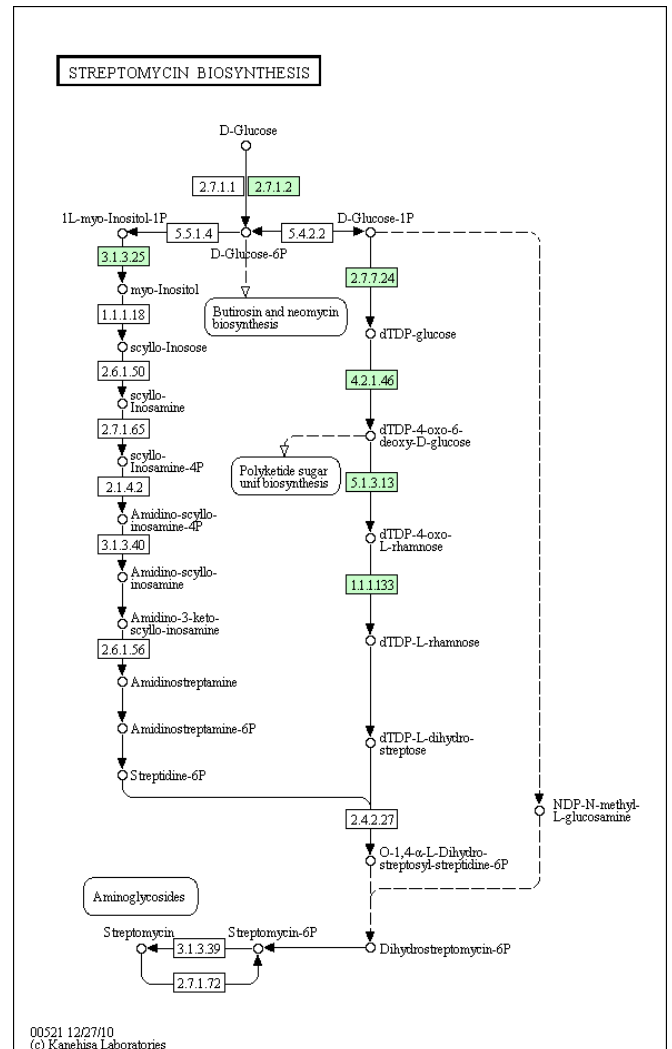
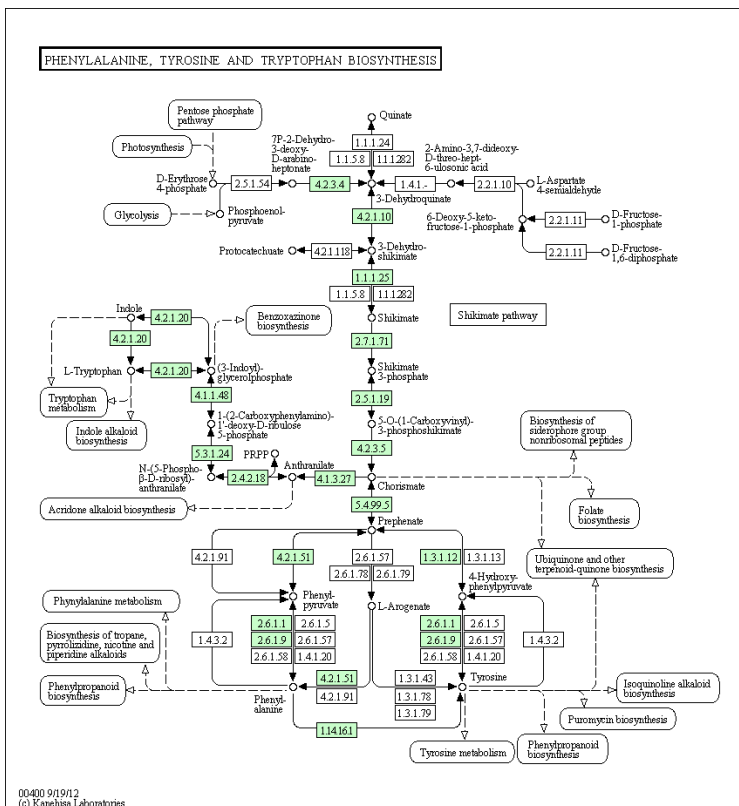
Note: In module 6, we will utilize only the KEGG prediction program.



KEGG Pathway Prediction

Introduction

The KEGG prediction program takes advantage of a huge database of pre-annotated pathways for bacterial cells using known experimentally verified pathways whose cognate proteins (and genes) are well established. KEGG can construct a *P. heparinus* specific pathways in which genes from the *P. heparinus* genome are screened via a BLASTp search and then placed into the proper positions within a particular pathway. In the examples given below, the presence of a green gene indicates that there is a homolog for that enzyme in the *P. heparinus* genome.



Instructions

1. To search KEGG, right click on link and open in new tab.
2. On the KEGG Pathway database page, enter "phe" in the prefix. Enter your gene locus tag in the form Phep_XXXX (see Gene Detail page for this) and click Go.
3. The next page will show the results of your search.
4. If the search returns a No Hits, then indicate this in the KEGG pathway ID textbox. This means your gene product has yet to associated with a pathway by the KEGG database. But see below for another method using Google.
5. If there is a pathway or pathways look over the thumbnail files and choose (if there are several) an appropriate pathway. In each diagram all the genes found in *P. heparinus* for this pathway will be shaded in green but your particular gene name will appear in red instead of black letters. If you have a choice of diagrams, choose the diagram which has only one pathway containing your gene product and a figure with several interconnecting pathways. Record the KEGG organism pathway number (pheXXXXX) in the notebook.

6. Upload the image file and insert into the notebook as with other images.
7. If you were not able to locate a pathway on KEGG, but you have a gene name for which you have gathered evidence, perform a Google search on the name. You should be able to locate a Wikipage or other source. From that source, try to infer the pathway.
8. If you can assign your gene (product) to a pathway, comment briefly on its role in the pathway. What does it do in that pathway? Also, what sort of insights might this provide about its possible significance to the organism.

Contextualization of Gene Neighborhood

Horizontal Gene Transfer

Operon Prediction

General Introduction to Horizontal Gene Transfer Prediction Programs (Module 8)

In this module there are several programs which allow us to look at the chromosomal organization of your gene. This is important for at least two reasons. First it can give insight as to a particularly unique mode by which bacteria can inherit genes from other bacterium. Through processes known as horizontal gene transfer, a recipient bacteria can take up DNA from a donor bacterium. The DNA is often a piece of DNA with several genes which is then incorporated into the recipient bacterial genome. The genes transferred sometimes code for useful proteins which may confer a selective advantage to that bacterium. Secondly, chromosomal organization reveals the presence of clustered genes whose transcription is controlled by a single promoter called operons. Hence your query gene may be part of a transcription unit and is coordinately expressed with partner genes. Knowing what those partner genes encode can often yield insight into the annotation of your gene.

For **Module 8** we will utilize the Gene Context program (Gene Neighborhood) to infer horizontal gene transfer and will predict operons using an outside program from microbesonline.org. **We will not use the Phylogenetic Tree and Chromosomal GC Heat Map Programs.**

Gene Neighborhood Analysis

Introduction

The order and type of genes which form the immediate neighborhood of gene are often shared in related species. This similarity is called chromosomal *synteny*. This is in part due to the fact that related species of bacteria (those found in the *Pedobacter* genus, for example) often occupy the same niche and have similar functional roles in the ecosystem. (Some might also say that this similarity in gene neighborhoods is due to the evolution of the genome from a common ancestral species to match the same ecological niche). If on the other hand, your gene's neighborhood in *P. heparinus* is not at all similar to species in the *Pedobacter* genus, then this might indicate the inheritance of this segment of DNA from another bacteria species by a process of DNA transfer (lateral or horizontal transfer – see figure below).

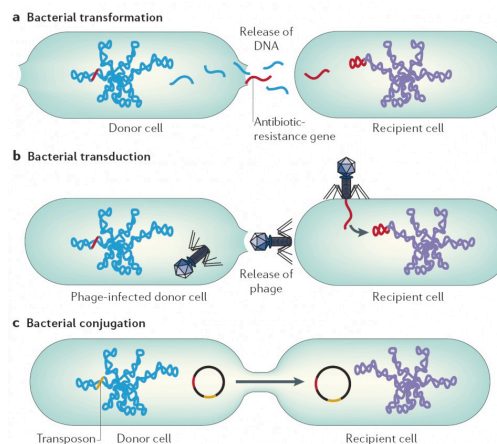


Figure 2 | **Horizontal gene transfer between bacteria.** a | Transformation occurs when naked DNA is released on lysis of an organism and is taken up by another organism. The antibiotic-resistance gene can be integrated into the chromosome or plasmid of the recipient cell. b | In transduction, antibiotic-resistance genes are transferred from one bacterium to another by means of bacteriophages and can be integrated into the chromosome of the recipient cell (lysogeny). c | Conjugation occurs by direct contact between two bacteria; plasmids form a mating bridge across the bacteria and DNA is exchanged, which can result in acquisition of antibiotic-resistance genes by the recipient cell. Transposons are sequences of DNA that carry their own recombination enzymes that allow for transposition from one location to another; transposons can also carry antibiotic-resistance genes.

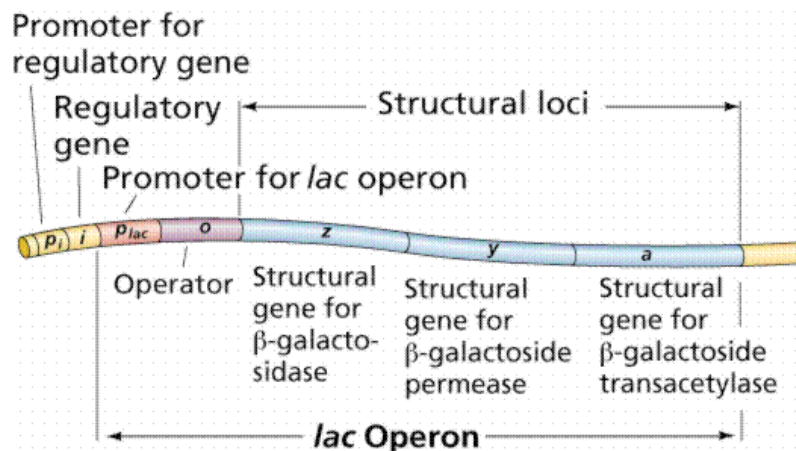
Instructions

1. Go to the bottom of IMG Gene Detail page and find heading Evidence for Function Prediction.
2. Just below the chromosomal map is a hyperlink "Show neighborhood regions with the same top COG hit". Click on the link and wait for the processing
3. Compare the chromosomal neighborhood (top row) of your gene (in red) in *P. heparinus* with the other species (rows below).
4. Analysis of gene neighborhoods
 - a. To what extent are the genes which surround your gene the same in other species? This can be assessed quickly by simply looking the gene sizes and colors to the left and right of your gene and seeing if the same pattern exists in other species. You can also just mouse over the gene and the gene detail page name of the gene will appear.
 - b. Try to assess how much synteny exists. Is it just with a small cluster (same operon?) or does the synteny go beyond the operon level?
 - c. Lastly, if the chromosomal neighborhood for your gene is shared with other species, are these species in the same genus or phylum? If *P. heparinus* does share similar operon structure to organisms not related to *P. heparinus*, this might indicate inheritance through horizontal gene transfer.

Operon Prediction

Introduction

Most bacterial genes are found in an operons with two or more genes separated whose reading frames are in close proximity (sometimes even overlapping) along the length on one strand of DNA. These genes are co-transcribed from a single promoter. This allows the cell to coordinate the synthesis of proteins which serve some common function. You may be familiar with the lac operon (shown below) which coordinates the synthesis of proteins which allow the transport of lactose into the cell and enzymes for its metabolism.



Instructions

1. Create two new textboxes in module 8. These will be used to record data obtain from a resource outside of IMG-ACT.
2. Connect to www.microbesonline.org in a new tab.
3. In the upper left corner, type in the first few letters of our organism "Pedo". Push return. Select "Pedobacter heparinus DSM 2366". Click Add →
4. The species name will now appear in the genomes selected window to the right.
5. Move to the right of the organism box and type the Phep_XXXX (gene locus tag – found on Gene Detail page) of your gene in the Search genes box. Click Find Genes.
6. On the results page, locate the hyperlinked acronym G O D H S T B. Click on the O = operon prediction.
7. A figure of the predicted operon is displayed. Your gene is in dark blue and is depicted as thick arrow with the arrow head pointing in the direction of transcription (5' - 3'). Other genes either upstream or downstream are in medium blue. If there is no operon your gene will appear as a single dark blue gene.



8. Information you can obtain from operon prediction:
 - a. The size of intergenic region between genes within the operon is given below the junction of the gene(s). These are typically less than 10 or less than zero (two genes overlap).
 - b. The intergenic region just upstream from the start of the operon (gene 1) is given as well. This region will contain the promoter where transcription starts and RNA polymerase binds as well as other binding sites for transcription factor proteins.
 - c. Also, record the names of the other genes in the operon with your gene. These can be seen by mouseing over each gene. Write these down. Often operons encode genes whose products might be found in the same pathway!
9. Upload and insert graphic into a textbox in your notebook.
10. Major questions to address in the second textbox.
 - a. Is your gene found in an operon?
 - b. If your gene is found in an operon, describe the operon in terms of other gene members, whether the genes might be in the same pathway, the size of the predicted promoter and if any genes contain overlapping reading frames (intergenic region is less than 0).

Annotation Reports
Proposed Identity of the Gene
Gene Report

Instructions for Proposed Annotation Module 10

Use this textbox to formally make a gene proposal hypothesis. Then proceed to present (summarize key findings) for just the following analyses:

1. DNA coordinates and start codon
2. Sequence similarity
3. Localization
4. Structural
5. Pathway

Accept or reject hypothesis of gene name (gene product name) from gene detail page. If you reject for example, a “hypothetical” hypothesis, then propose a new annotation.

If your data lead you to accept the hypothesis, then this needs to be explained as a confirmation of the pipeline.

Instructions for Annotation Gene Report

Audience: Other students in the class. Biology faculty members

Introduction

1. Gene Assigned
2. Hypothesis

Results (with figures and textual explanation of findings)

1. DNA coordinates and start codon
2. Sequence similarities
3. Localization
4. Structural similarities
5. Pathway
6. Gene Context
7. Organismal significance

Discussion

1. Gene Annotation – identity of gene product/pathway
2. Gene Expression – operon?
3. Gene Inheritance – gene neighborhood
4. Organismal significance – pathway or role of gene product

NCBI ORF Finder

Introduction

This program predicts all of the open reading frames (ORF) in a particular specified region of a chromosome. A ORF is a theoretical gene with an start and stop codon predicted for a one of six possible reading frames (3 from each strand of DNA). For an ORF to be considered as a gene it must be at least 100 amino acids long and can be in any reading frame. However, the longest ORF in a particular region of the DNA regardless of its reading frame always trumps the shorter ORFs. To further validate an ORF, it must also show significant alignment with other known proteins in the BLASTp database.

Instructions

1. Open new tab and connect to NCBI ORF finder program:
<http://www.ncbi.nlm.nih.gov/projects/gorf/>
2. Since this program is not found in your lab notebook, you will need to create a textbox to deposit the data. We will do this in Module 4 under the heading of DNA coordinates.
3. Scroll down to Module 4 and open.
4. Click on the space just in front of the text box "Explanation of Choice". Push return a few time to open up some space. Double click on a text box, right click, select copy and paste into the space created. Above the text box, type "DNA coordinates from the NCBI ORF reader.
5. Go back to NCBI ORF finder.
6. Use accession code CP001681. This will enable you to utilize the entire DNA sequence of the P. heparinus genome.
7. Enter a range of DNA coordinates (provided by your instructor) which spans the set of genes being annotated by the class.
8. For genetic codes, select 11 Bacterial Codes
9. Click ORF find.
10. To the left of the page, you can see all the ORFs predicted graphically.
11. Consult the list of predicted ORFs to the right which are arranged in descending order of length. The longer the ORF the more likely it is an actual gene.
12. Verify that your gene(s) has coordinates which match those in this ORF finder.
13. Using your newly created textbox, record the DNA coordinates in module 4. Be sure to note if the coordinates are different than the pipeline prediction. We will decide which coordinates are correct when we perform BLASTp analysis and T-COFFEE analysis in module 2.