SCIE2100 Practical 7, Week 10

Aim

In this practical you will use a set of tools to annotate and identify three microbial genomes. Of these genomes, two were sequenced as part of the Human Microbiome Project and one was sequenced by the Broad Institute. One of the microbes was isolated from the gut, one from the mouth, and one from the airways of a person with respiratory disease. You will use genomic sequence data to identify each microbe taxonomically, look for specific genome features such as functional RNAs, and carry out a whole genome annotation to determine important metabolic functions. Based on these analyses, you should be able to identify which microbe came from which source.

Note: Some of the web services you'll be using may take several minutes to complete, and are marked appropriately. You may wish to analyse the genomes in parallel (i.e., run Genome1, 2 and 3 through a single step at the same time). If you decide to do the optional exercises, be aware that the KAAS annotation can be slow.

Files

For this week's practical you will need this guide, and genomic (.fasta) files for each of Genome1, Genome2 and Genome3, and the SampleContig.fasta file. We've also provided proteomic (.pfasta) files, but you will not need them except for optional material.

Nucleotide and amino acid sequence data

Nucleotide sequences are used for extracting features such as rRNAs and tRNAs, and can be used for similarity searches, however amino acid data is typically used for searches against protein databases. The nucleotide sequences that you have are contigs formed by assembling raw sequence reads into larger pieces, which is why in some files there are multiple sequences instead of one sequence representing the complete genome. This is the type of data you will usually get in the first attempts to sequence a complete genome. The amino acid sequences are determined by translating the nucleotide sequence in the correct reading frame and looking for open reading frames.

Analyzing GC Content

Microbial genomes vary in the percentage of GC nucleotides (as opposed to AT) they contain; this property is referred to as GC content. Whole genomes can sometimes be distinguished by their GC content. In addition, it is sometimes possible to identify genomic regions that have been newly introduced, for example via lateral gene transfer, as they may have a different GC percentage to the rest of the genome.

1) Calculate the GC content for your nucleotide data of each genome independently using the EMBOSS tool geecee available at
   http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::geecee

Download the results for each genome and calculate the mean GC content for all genomes as well as the standard deviation for the contigs. For normally distributed data, two standard deviations are said to capture 95% of the data, and values outside this interval may be considered unusual observations. Calculate the GC percentages that define the lower and upper bounds for this 95% interval for each genome and report these in a table with the mean and standard deviation.

Investigate whether we can do this with Python.

Taxonomic identification using 16S rRNA

Before we do the whole genome annotation, we will look at specialised databases to identify genes and genomic features of interest, starting with the 16S rRNA. Prokaryotic ribosomes have a large and small subunit, which contain individual ribosomal RNA molecules (rRNAs) that can be used to distinguish microbial taxa. The 16S rRNA in the small subunit is a universal marker gene for Bacteria and Archaea. These genes are easy to detect in sequence data because they contain highly conserved regions. Variable regions in the 16S rRNA differ between microbial species and allow for the identification Bacteria and Archaea to the species level.

2) Find rRNA genes in each genome using the RNAmmer web server at http://www.cbs.dtu.dk/services/RNAmmer/

   Use the nucleotide sequence files (.fasta) for this exercise, making sure to select the “Bacteria” option. You may wish to enter your email address and have the server email you when the analysis is completed. Once you have your results download the fasta file containing the rRNA sequences.
   [Run time: ~5 minutes for each genome]

   Go to the Greengenes website at http://greengenes.lbl.gov Select the compare tool and the choose the BLAST option. Make sure that you select a *single* 16S rRNA result from your fasta RNAmmer results file, and paste it into the box. Run BLAST using the default settings.

   For each of your genomes, give the full taxonomic listing for the closest BLAST hit (i.e., from the Kingdom level down to the Genus/Species).

Translation and ORF finding

In order to annotate a genome, you must first determine the open reading frames (ORFs), i.e. genomic regions encoding protein sequences. Nucleotide sequences can be translated in 6 reading frames: 3 on the forward strand and 3 on the reverse strand. Translation in the proper frame is essential to determining the location of ORFs, and the correct frame is generally the one with the longest sequence between valid start and stop codons.

3) Use the nucleotide sequence in the file SampleContig.fasta for this exercise. Translate the sequence in all 6 reading frames using Transeq http://www.ebi.ac.uk/Tools/st/emboss_transeq/ . Which reading frame do you think is the correct translation? Why? From the correct translation, select the actual protein sequence (from start to stop) and include it with your answer.
[Run time: 1-2 minutes]
Investigate whether we can do this with Python (with translate.py).

4) Identify this protein by using BLAST at NCBI
http://blast.ncbi.nlm.nih.gov/blast/Blast.cgi

Which BLAST should you use?

From the BLAST results, what is the identity of this protein? What organism is it likely to be from?
[Run time: less than 1 minute]
Investigate whether we can do this with Python (with runBLAST).

There are bioinformatic tools (such as Prodigal) which can select all of the likely ORFs from complete genomes (or contig sequences) at once. Such tools can produce protein sequences for possible ORFs, which can then be used for whole genome annotation.

Whole genome annotation

Genome annotation servers can be used to compare the full set of translated ORFs from a genome to protein databases in an automated fashion. These servers generally also provide information on how the annotated genes fit into metabolic pathways and functional networks. We can use the KAAS server which performs BLAST-based comparisons to annotated genomes to characterise ORFs and also provides graphical representations of metabolic pathways (KEGG maps).

5) Because the KAAS annotation takes time, we've provided you with the top 10 pathways for each of the provided genomes. Looking at the table below, answer the following questions:
   1. Which pathways which make the top 10 for all 3 genomes?
   2. Why might these pathways be important and highly conserved in all 3 genomes?

Genome1	**Count**
03010 Ribosome	53
00230 Purine metabolism	51
00190 Oxidative phosphorylation	43
02020 Two-component system	40
00240 Pyrimidine metabolism	37
00910 Nitrogen metabolism	34
02010 ABC transporters	33
00860 Porphyrin and chlorophyll metabolism	30
00720 Carbon fixation pathways in prokaryotes	27

00330 Arginine and proline metabolism	27

Genome2	**Count**
00680 Methane metabolism	86
03010 Ribosome	59
00230 Purine metabolism	36
00240 Pyrimidine metabolism	35
00860 Porphyrin and chlorophyll metabolism	24
00970 Aminoacyl-tRNA biosynthesis	24
00720 Carbon fixation pathways in prokaryotes	19
00400 Phenylalanine, tyrosine and tryptophan biosynthesis	18
02010 ABC transporters	18
00250 Alanine, aspartate and glutamate metabolism	16

Genome3	**Count**
03010 Ribosome	49
00230 Purine metabolism	47
02010 ABC transporters	40
00240 Pyrimidine metabolism	38
00860 Porphyrin and chlorophyll metabolism	32
00970 Aminoacyl-tRNA biosynthesis	25
00270 Cysteine and methionine metabolism	24
00720 Carbon fixation pathways in prokaryotes	21
00910 Nitrogen metabolism	21
00400 Phenylalanine, tyrosine and tryptophan biosynthesis	19

6) The set of organisms you're identifying consists of a methanotrophic archaeon, a Mycobacterium and Veilonella. You should already have a good idea which is which from the rRNA analysis earlier, but the pathways will provide supporting evidence.

Look up each organism, noting possible interesting differences. Consider the pathways that are *different* between the organisms; which pathways tell you something about which genome belongs to which organism, and why?

Optional exercises

7) You can run the KAAS jobs yourself to get a full list of pathways annotated for each organism, and links to details of the pathway coverage. We've provided you with `pfasta` files – one for each genome – that you can use for this next purpose.

   On the KAAS homepage
   http://www.genome.jp/tools/kaas/
   select “KAAS job request (SBH method)”, which will re-direct you to the input form. Use the provided *amino acid* sequence files for these analyses with

the file upload option (do not check nucleotide). Select the standard Prokaryotic set for comparison to your input genome (click “for Prokaryotes”) which includes 28 organisms. Submit your genome by pressing Compute. **[Run time: 5-15 minutes for each genome]**

After receiving the email that the annotation has finished, go to the html results for each genome.

a. Based on the KEGG maps for Methane metabolism, which organism is most likely producing methane and why? Do you think this organism has cytochromes? Why or why not?

b. Using the KEGG maps for Oxidative Phosphorylation only, tell how you can identify which organism is an Archaeon, which is an anaerobic Bacterium, and which is an aerobic Bacterium. This will require you to do additional research into the biology of these organisms based on the differences you see in the KEGG maps to justify your answer.

Assessment

Answer the following questions/complete the following tasks in a PDF or Word document and submit it to Blackboard as an attached file to the assessment by the due date.

- Provide the table from Exercise 1.
- Provide the taxonomies from Exercise 2.
- Provide the protein sequence and justification for the translation you selected in Exercise 3.
- Provide the protein identity and organism from Exercise 4.
- Provide the list of common shared pathways and an answer to the question in Exercise 5.
- State your organism identifications and justification, including pathway information, from Exercise 6.