BIO508: Lab Session 6

Announcements

- Homework 5 is due at 11:55pm on Monday, 3/10.
- In homework 5, don't use the asterisks in your file names: they are only there for emphasis.
- For Windows users: Cygwin is available at `http://cygwin.com/install.html`.
- Homework 4 grades will be posted on Friday, 3/7.
- Bioinformatics Bruch tomorrow 10-11am in 4th floor Breezeway (between buildings 1 and 2).

Exercise

We will use the practice problem from the homework to practice the class concepts in lab today.

1. It would be great if we could run a whole genome assembly from these reads, but A) I haven't provided you with enough coverage to get a half-decent assembly because B) there's no good assembly software that runs on Windows. But fret not! I (or really, the Human Microbiome Project) have conveniently provided a draft assembly of your bug (probably run using SOAPdenovo, for those who are curious). Rather than mapping individual reads, let's align the entire set of resulting contigs against the *Bli*15697 reference to see what synteny and conservation looks like (we'll discuss comparative genomics in more detail later on in the class).

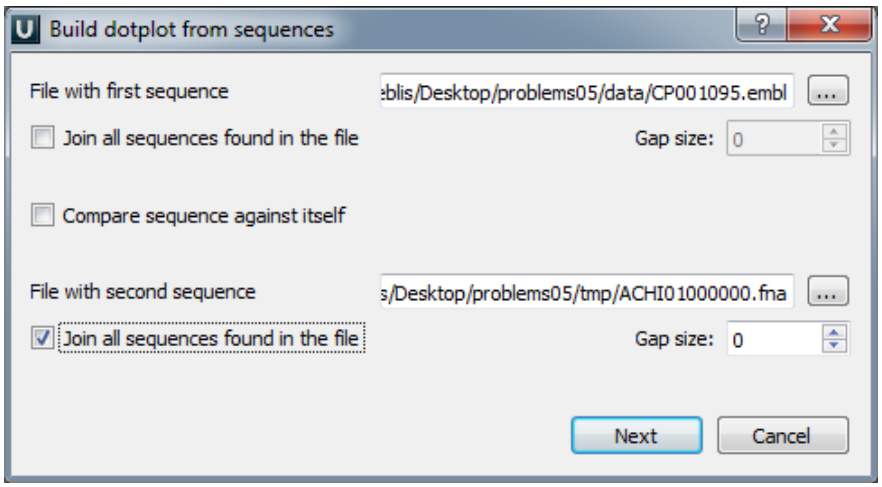
   (a) It turns out that your bug is *Bifidobacterium longum infantis* ATCC 55813; from now on, for clarity's sake, I'll refer to it as *Bli*55813 (not to be confused with the reference strain, *Bli*15697). You can download its assembled contigs from: `ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria_DRAFT/Bifidobacterium_longum_infantis_ATCC_55813_uid31437/`. Note that you need only the assembled contigs as plain FASTA files, `ACHI00000000.contig.fna.tgz`, and you can ignore all the other stuff in the directory (although it is interesting, if you want to poke around!) Also note that the GenBank FTP site is occasionally mysteriously unavailable - your tax dollars at work. In such a case, you can alternatively retrieve a near-identical genomic FASTA file from: `http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=ACHI01` by scrolling down and clicking on `ACHI01.fasta.gz`. This lets you skip the next step, too, by the way.

   (b) Untar this package of 140 contigs into files `ACHI01000001.fna` through `ACHI01000140.fna`. Using `cat` or a text editor, combine them into a single FASTA file:

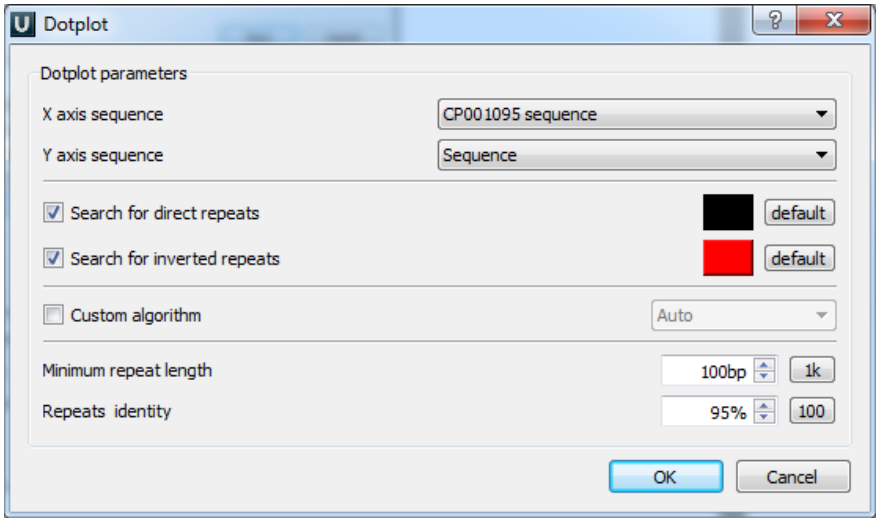
   ```
   cat ACHI01000*.fna > ACHI01000000.fna
   ```

   (c) Let's get a fully-annotated version of the reference strain *Bli*15697 from Ensembl, which has its own GenBank-like format. Check out the list of Ensembl bacteria at: `http://www.ebi.ac.uk/genomes/bacteria.html`. Find "Bifidobacterium longum subsp. infantis ATCC 15697" and click on it, taking you to: `http://www.ebi.ac.uk/ena/data/view/Taxon:391904` Click the + if necessdary to expand "Assembled & Annotated Sequences (EMBL-Bank)", then click the first link under "Taxon & its descendants" to go to the genome page. Click on "GCA_000020425.1" to take you to the correct assembly. Finally, right click on the "TEXT" link by "Download:" in the lower right, choose "Save Link As", and save it as the file `CP001095.embl` (instead of just `.txt`, so as not to confuse some programs later).

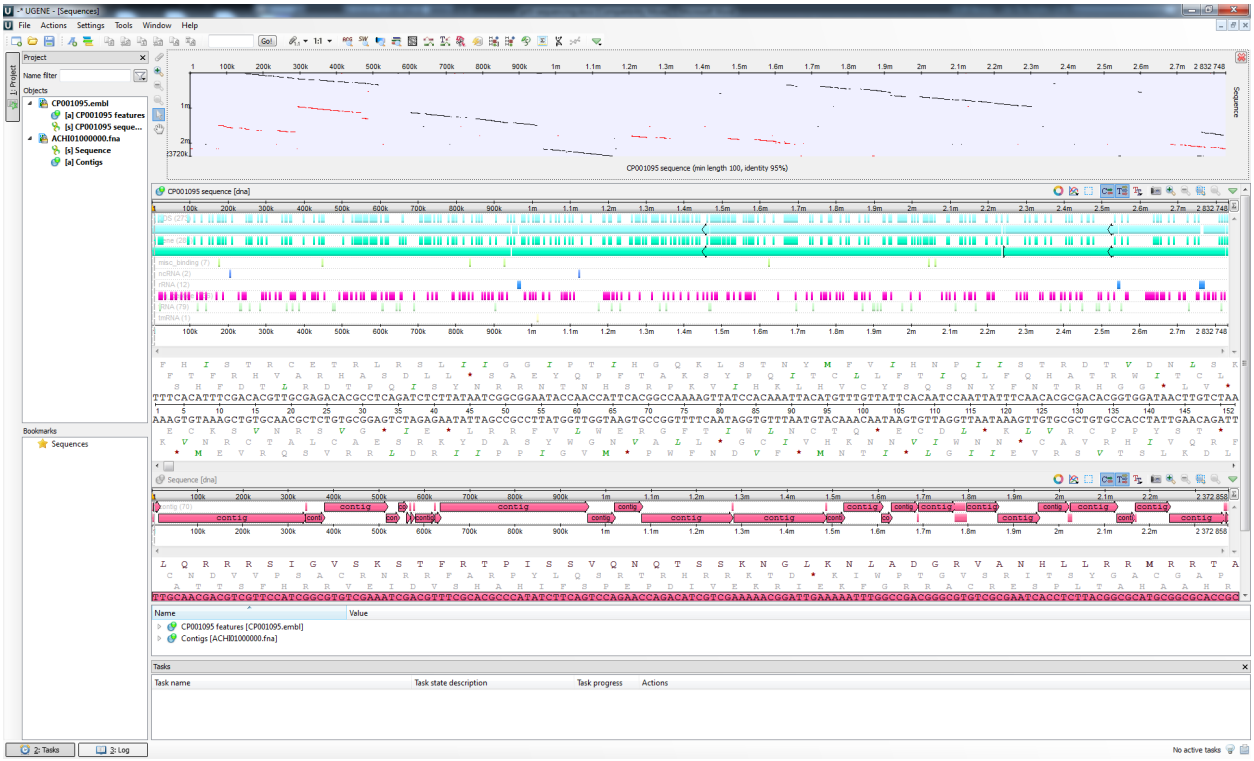
(d) Grab and install UGENE from: `http://ugene.unipro.ru/download.html` Fire it up, create a new project wherever you'd like when prompted, and then go to Tools/Build dotplot. Click on the "..." button for the "File with first sequence", browse to your reference genome `CP001095.embl`, and select it. Then click on "..." for the "File with second sequence", browse to your concatenated `ACHI01000000.fna` file, and select it. Check "Join all sequences found in file" to end up with a dialog something like this (then hit Next):



(e) When UGENE asks you to configure the dotplot, let's enable inverted as well as direct repeats (check "Search for inverted repeats") and make them red. Click on the black box beside inverted repeats, select a nice red, and click OK. Also crank down the stringency to only 95% identity. You should end up with something like this (then click OK):

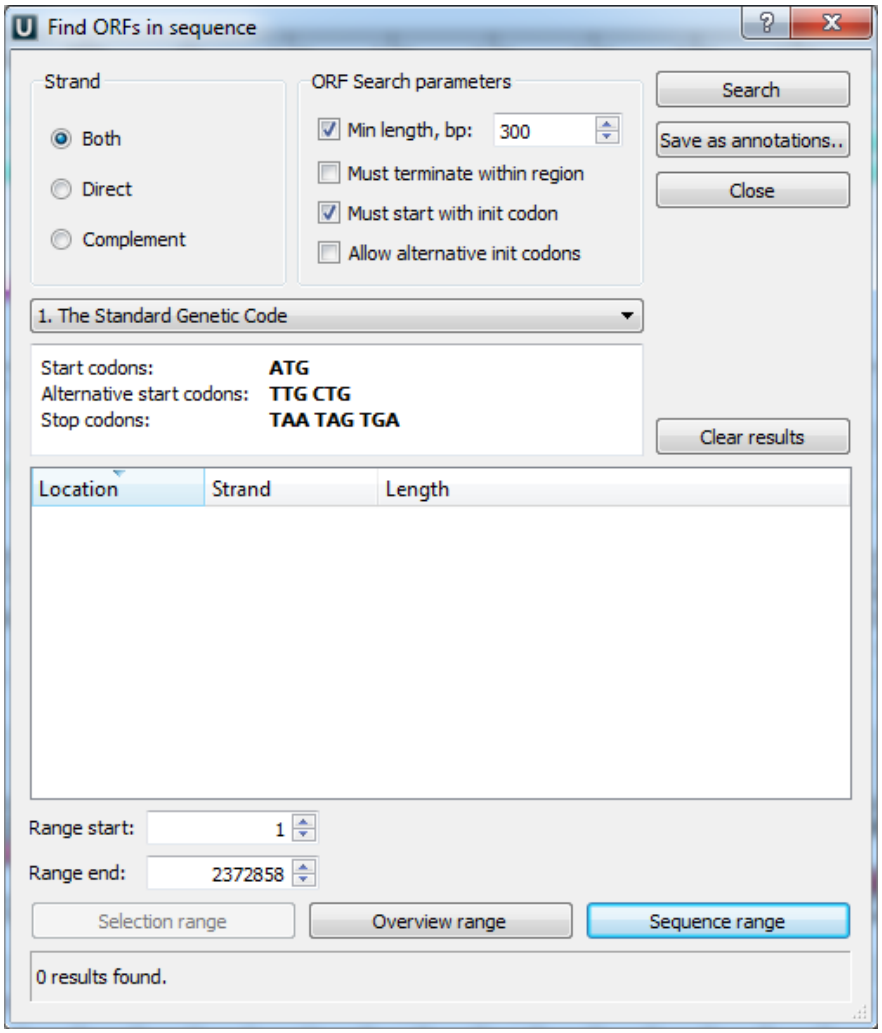


(f) Whoah, colors!

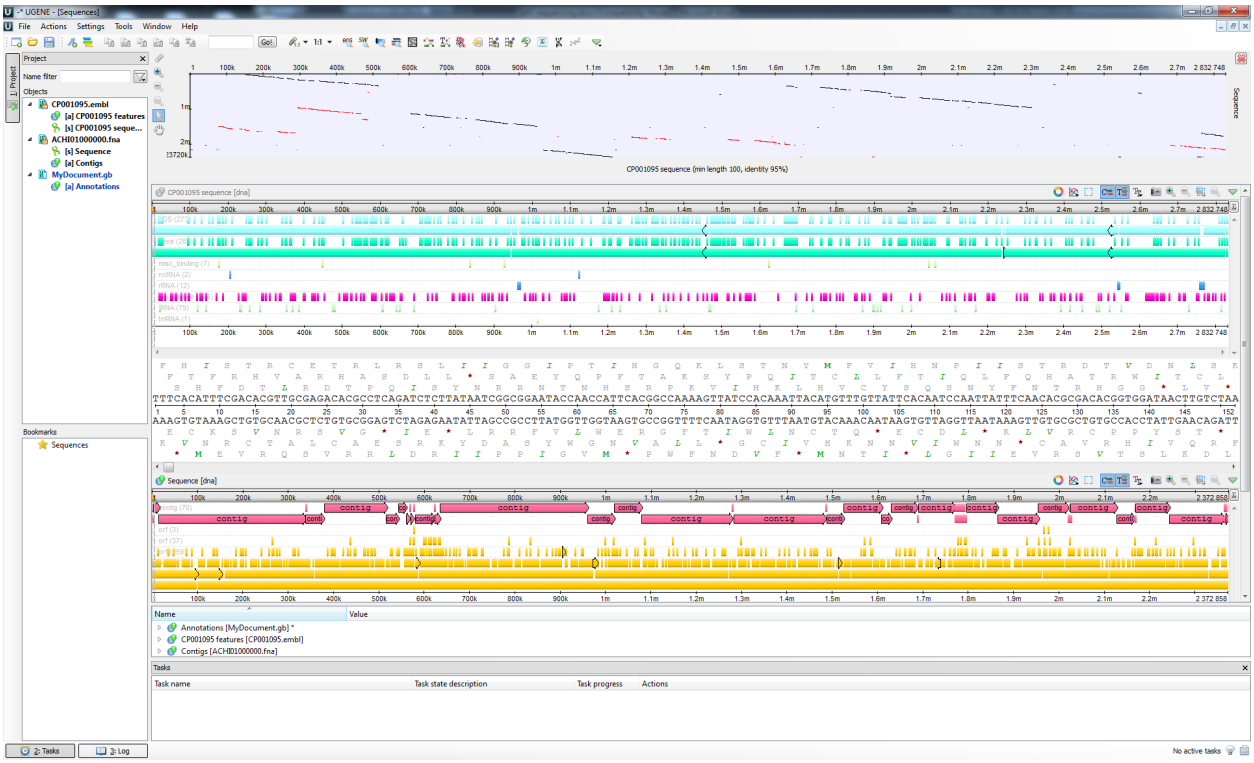


UGENE is showing you a variety of important things: the reference vs. query dotplot on top, the reference strain *Bli*15697 with annotated features on top, and the query strain *Bli*55813 contigs below. It would take a lot of typing to explain what all of these views are, but feel free to poke around, zoom in and out, and most importantly read the UGENE documentation liberally as needed: `http://ugene.unipro.ru/documentation.html` If you'd like us to check this, provide a screenshot file ***screenshot_dotplot.png*** when you get this far.

(g) Note that there are a zillion features annotated by Ensembl in the reference genome, but nothing but the contig boundaries in your query. There are sophisticated algorithms for calling Open Reading Frames (ORFs) based on machine learning, but it's easy to call them just by looking for appropriately separated start and stop codons on the same strand. Let's do that: click on the query sequence panel (on the bottom, where it says "Sequence [dna]"), then click on the "Find ORFs" button in the toolbar (a magnifying glass with a little horizontal blue bar). Make sure to search both strands using the standard genetic code, with a minimum length of 300bp and using the whole "Sequence range". When you're done, it should look like this (and hit Search):

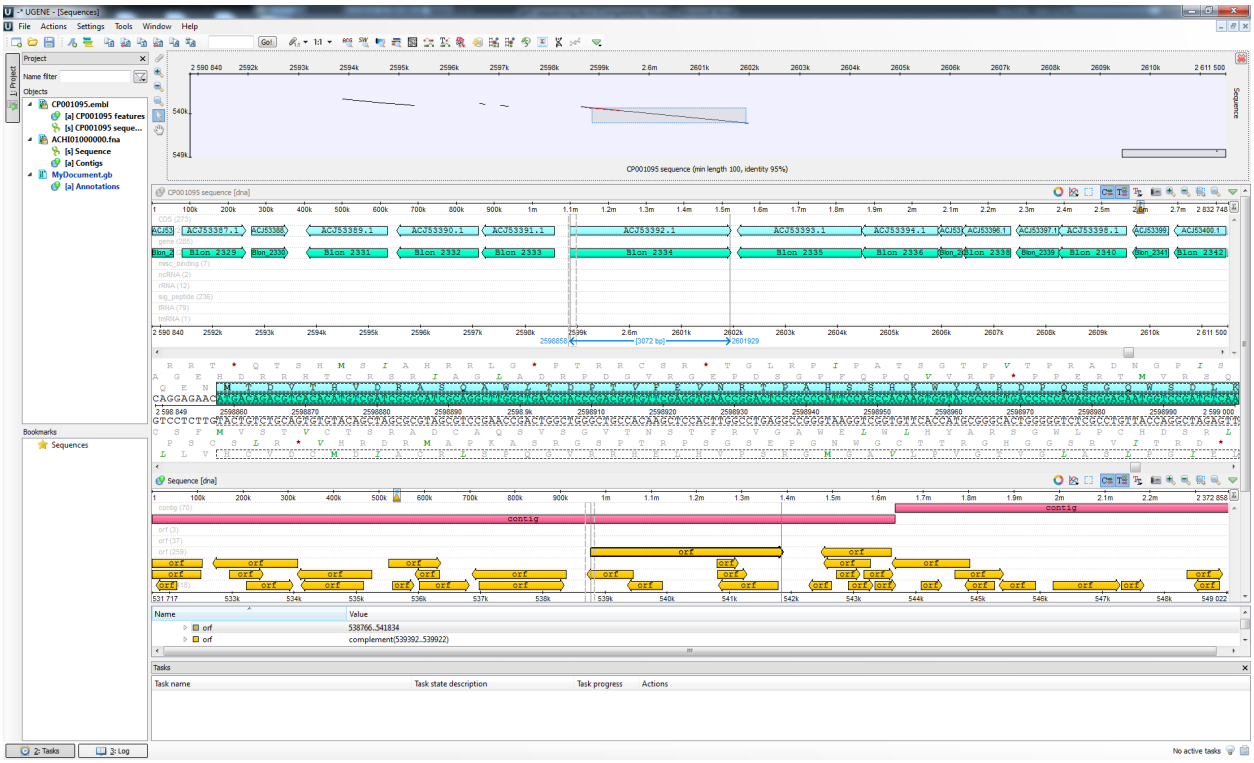


(h) Click “Save as annotations”, create a new table file wherever you’d like, and click Create to get these new annotations to show up in UGENE. You should now have a bright yellow ORF track on your query genome listing every possible putative ORF (including a bunch of false positives):



Again, if you’d like us to check it, submit a screenshot called ***screenshot_orfs.png***

(i) If you read the *Bli*15697 genome paper, you might have noticed they were particularly excited about the β-galactosidase operon anchored around reference ORF Blon_2334. What does that operon look like in our *Bli*55813 genome? Hint: if you search around, you should end up with a view something like this:



(j) Is Blon_2334 conserved?

(k) Is the β-gal operon conserved? Syntenically?

(l) What's unfortunate about our contigs? What might be some solutions to remedy the situation?