Tutorial 5

Reconstruction of a complete extrachromosomal genome using DNA or protein seeds

1. Introduction

Extrachromosomal genomes such as mitochondrial and plastid episomes can also be reconstructed using GenSeed. In this tutorial, we will use real-life data to show how feasible and easy is this application. For this purpose, we will use shotgun genomic reads from *Eimeria tenella*, a coccidian protozoon that infects chickens. *E. tenella*, as several other apicomplexan parasites, presents an mtDNA composed by linear molecules containing a variable number of tandemly repeated 6-kb units (Dunn *et al.* - *Eimeria tenella*: two species of extrachromosomal DNA revealed by pulsed-field gel electrophoresis. *Parasitol Res.* **84**: 272-275). In our laboratory, we have determined a short stretch of sequence from the cytochrome b gene. Since cytochrome b gene is a reliable mitochondrial marker, a 100-bp DNA region, and the corresponding 33-aa protein sequence, will be used as seeds.

2. Seed sequences

The seed directory contains four files:

- `cytb_nt_seed.fasta` – a 100-bp nucleotide sequence of the cytochrome b gene of *Eimeria tenella* (determined in our laboratory).
- `cytb_prot_seed.fasta` – a 33-aa protein sequence fragment deduced from the cytochrome b gene.

3. Database

We will use a database comprising 936,122 shotgun reads of the coccidian parasite *Eimeria tenella*. This database was available in November 2007 and must be downloaded from (ftp://ftp.sanger.ac.uk/pub/pathogens/Eimeria/tenella/genome/reads/eimeriareads. 050808.gz).

4. Running GenSeed with DNA and protein seeds

Let's first reconstruct the mitochondrial genome using a DNA seed. Go to the `/tutorial_5/test` directory and type the following command:

```
genseed.pl -s ../seed/cytb_nt_seed.fasta -d ../db/eimeriareads.050808
-o output_nt -v ../vector/UniVec -b "-e 1e-06 -b 100 -F F" -g no
```

Now let's run GenSeed for the reconstruction of the same mtDNA using a protein seed. Type the following command:

```
genseed.pl -s ../seed/cytb_prot_seed.fasta -d
../db/eimeriareads.050808 -o output_prot -v ../vector/UniVec -b "-e
1e-06 -b 100 -F F" -g no
```

5. Understanding GenSeed parameters:

A comprehensive explanation of all *GenSeed* parameters is depicted in "GenSeed - Quick Guide" document. Please refer to it if you need more information.

Shortly, the command line used above specifies the following parameters:

- `-s ../seed/filename.fasta` – seed sequence file with path
- `-d ../db/eimeriareads.050808` - database file with path
- `-o output_nt` - `output_nt` as the output directory name
- `-b "-e 1e-06 -F F -b 100"` – these are parameters for BLAST. We use parameters `-e 1e-06` and `-F F` as default. The parameter we have changed here is `-b 100`, implying that only the top 100 positive hits will be selected on each round for the assembly process. Because the database is very big and coverage is highly redundant, we decided to limit the number of selected reads to improve performance, as CAP3 would only have to deal with a relatively small dataset in each walking round.
- `–v ../vector/UniVec` – vector database file with path. This database will be used to screen and mask the selected reads before the assembly steps. You can download this file from ftp://ftp.ncbi.nih.gov/pub/UniVec/.
- `-g no` – do not use the last assembly consensus sequence as a template in the final assembly step. We turned this option off also to reduce computational load. You can try re-running the reconstruction with `-g yes` option and compare the processing time.

6. Understanding GenSeed output files:

If everything works well, two new subdirectories will be created: `output_nt` and `output_prot`. Also, log files (`genseed.log`) will be created in the `/tutorial_5/test/output_nt` and `/tutorial_5/test/output_prot` directories. Let's inspect the content of the first one.

According to this file, in the first run GenSeed processed 14 rounds to complete the reconstruction of the mtDNA, recruiting in total 1,300 reads. If we had chosen no limit for the number of reads (`-b` parameter, see above), GenSeed would probably require less rounds to complete the assembly, as more reads would be incorporated, with some of them spanning longer regions in each round. However, the performance gain we obtain by limiting the number of selected reads may be worth. Please notice that we used BLAST's parameter `-b 100` and hence only 100 reads were selected in each round (see below). If we inspect the final part of the file, containing data about the reconstruction using the protein seed, you will see that the results were almost identical.

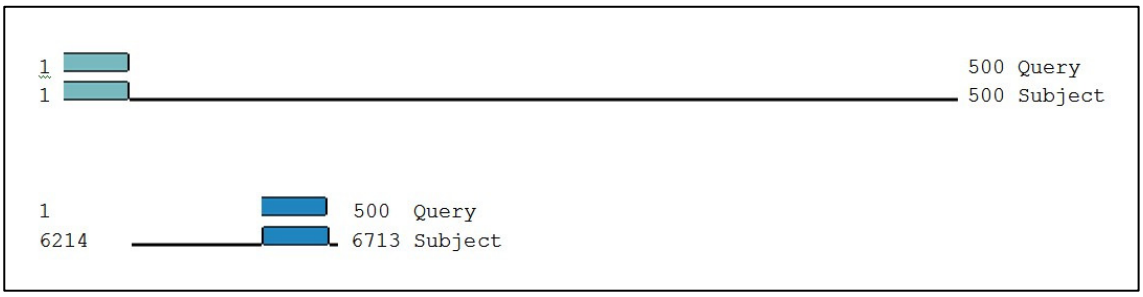
```
   Mon Nov 19 20:37:11 BRST 2007
   genseed.pl -s ../seed/cytb_nt_seed.fasta -d ../db/eimeriareads.050808 -o output_nt
-v ../vector/UniVec -b "-e 1e-06 -b 100 -F F" -g no

   seed type: DNA

   ####   Round 1   ####
   Total # of reads for CAP3: 100
   Length of the seed-contig E_tenella: 1267
   Accumulative number of reads: 99
   ####   Round 2   ####
   Total # of reads for CAP3: 100
   Length of the seed-contig E_tenella: 1704
   Accumulative number of reads: 199
   ####   Round 3   ####
   Total # of reads for CAP3: 100
   Length of the seed-contig E_tenella: 2355
   Accumulative number of reads: 299
   ####   Round 4   ####
   Total # of reads for CAP3: 100
   Length of the seed-contig E_tenella: 2958
   Accumulative number of reads: 399
   ####   Round 5   ####
   Total # of reads for CAP3: 100
   Length of the seed-contig E_tenella: 3390
   Accumulative number of reads: 499
   ####   Round 6   ####
   Total # of reads for CAP3: 100
   Length of the seed-contig E_tenella: 3685
   Accumulative number of reads: 599
   ####   Round 7   ####
   Total # of reads for CAP3: 100
   Length of the seed-contig E_tenella: 4164
   Accumulative number of reads: 699
   ####   Round 8   ####
   Total # of reads for CAP3: 100
   Length of the seed-contig E_tenella: 4677
   Accumulative number of reads: 799
   ####   Round 9   ####
   Total # of reads for CAP3: 100
   Length of the seed-contig E_tenella: 4985
   Accumulative number of reads: 899
   ####   Round 10   ####
   Total # of reads for CAP3: 100
   Length of the seed-contig E_tenella: 5515
   Accumulative number of reads: 999
   ####   Round 11   ####
   Total # of reads for CAP3: 100
   Length of the seed-contig E_tenella: 5896
   Accumulative number of reads: 1099
   ####   Round 12   ####
   Total # of reads for CAP3: 100
   Length of the seed-contig E_tenella: 6405
   Accumulative number of reads: 1199
   ####   Round 13   ####
   Total # of reads for CAP3: 100
   Length of the seed-contig E_tenella: 6860
   Accumulative number of reads: 1299
   ####   Round 14   ####
   Total # of reads for CAP3: 1
   Length of the seed-contig E_tenella: 6860
   Accumulative number of reads: 1300
   The end! Contigs have the same size
   ####   Last Round   ####
   Length of the final seed-positive contig E_tenella: 6807
```

Analyzing the `final_contigs.fasta` file, we will find one single contig of 6,807 bp. As we have previously commented in the introduction, the mitochondrial genome of *Eimeria* spp. is composed by tandemly repeated units of approximately 6 kb, arrayed in linear concatamers. Due to this feature, GenSeed may have constructed a sequence which still presents some redundancy in the ends. To precisely define the limits of the

mtDNA, use Blast 2 sequences and perform a BLASTN similarity search between the first 500 bp of `final_contigs.fasta` sequence versus the whole sequence itself. You will find two alignment blocks, as shown in the figure below: (1) a block starting from position 1 to 500 for both sequences and a (2) a block starting from position 1 for the query sequence (the short one) and position 6,214 for the subject sequence (the whole sequence).



This result means that the sequence at the 5' end is repeated exactly after position 6,214, which lead us to conclude that the mtDNA unit is comprised within positions 1 to 6,213 bp. This is in fact the exact length of the mitochondrial genome, as determined in our laboratory by conventional DNA sequencing of PCR-amplified products. If you now run BLASTX searches of this reconstructed sequence against the non-redundant database (nr), you should find three mitochondrial genes: cytochrome b, cytochrome oxidase c subunit 1 and cytochrome oxidase c subunit 3.