Tutorial 2

Reconstruction of cDNAs using nucleic acid or protein seeds

In this tutorial we will use both DNA or protein seeds to reconstruct full-length cDNAs of *Toxoplasma gondii*. Three genes were chosen as examples: microneme protein 7 (mic7), surface antigen glycoprotein 3 (SAG3) and pyruvate kinase (pk).

1. Seed sequences

The `/tutorial_2/seed` directory contains three DNA seed files…

- `mic7_nt_seed.fasta` - microneme protein 7 (AF357911.1, DNA region 501-600)
- `sag3_nt_seed.fasta` – surface antigen glycoprotein 3 gene (SAG3) (accession code AY187280.1, region 531-630).
- `pk_nt_seed.fasta` – pyruvate kinase (accession code AB050726.1, region 1051-1150)

…and the corresponding protein seed files:

- `mic7_prot_seed.fasta` - microneme protein 7 (AF357911.1, protein region 502-600)
- `pk_prot_seed.fasta` – pyruvate kinase (accession code AB050726.1, region 1053-1151)
- `sag3_prot_seed.fasta` – surface antigen glycoprotein 3 gene (SAG3) (accession code AY187280.1, region 532-630).

2. Original sequences

The `/tutorial_2/original_sequences` directory contains the original cDNA and proteins sequences of microneme protein 7, SAG3 and pyruvate kinase. We are providing these sequences as a mean to compare the final reconstructed sequences and thus evaluate GenSeed performance.

- `mic7_nt.fasta` - microneme protein 7 (AF357911.1) - nucleotide sequence
- `mic7_prot.fasta` - microneme protein 7 (AAK35070.1) - protein sequence
- `pk_nt.fasta` – pyruvate kinase (accession code AB050726.1) - nucleotide sequence
- `pk_prot.fasta` – pyruvate kinase (accession code BAB47171.1) – protein sequence
- `sag3_nt.fasta` – surface antigen glycoprotein 3 gene (SAG3) (accession code AY187280.1) - nucleotide sequence
- `sag3_prot_seed.fasta` – surface antigen glycoprotein 3 gene (SAG3) (accession code AAO72428.1) – protein sequence

3. Database

We will use a database of *Toxoplasma gondii* containing 129,421 unclustered ESTs. You can download such database from the NCBI (http://www.ncbi.nlm.nih.gov) choosing EST database (dbEST) and `txid5810[Organism:exp]` as query. The total number of sequences may be changed by the time you download it. The database file should be saved in the `/tutorial_2/db` directory. We assume from now on that the database file is named `T_gondii_EST.fasta`.

4. Running GenSeed

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

First, we will run GenSeed for each one of the DNA seeds, as follows:

```
genseed.pl -s ../seed/mic7_nt_seed.fasta -d ../db/T_gondii_EST.fasta
-o output_mic7_nt
```

and…

```
genseed.pl -s ../seed/pk_nt_seed.fasta -d ../db/T_gondii_EST.fasta -
o output_pk_nt
```

and…

```
genseed.pl -s ../seed/sag3_nt_seed.fasta -d ../db/T_gondii_EST.fasta
-o output_sag3_nt
```

If everything works well, then three new subdirectories will be created:

```
output_mic7_nt
output_pk_nt
output_sag3_nt
```

Now let's run GenSeed for the reconstruction of the same three genes using protein seeds. Type the following commands:

```
genseed.pl -s ../seed/mic7_prot_seed.fasta -d ../db/T_gondii_EST.fasta -o
output_mic7_prot
```

and…

```
genseed.pl -s ../seed/pk_prot_seed.fasta -d ../db/T_gondii_EST.fasta -o
output_pk_prot
```

and…

```
genseed.pl -s ../seed/sag3_prot_seed.fasta -d ../db/T_gondii_EST.fasta -o
output_sag3_prot
```

If everything works well, then three more subdirectories will be created:

```
output_mic7_prot
output_pk_prot
output_sag3_prot
```

5. Understanding GenSeed parameters

A comprehensive explanation on 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/name.fasta` – seed sequence file with path
- `-d ../db/name.fasta` - database file with path
- `-o output_name` - as the output directory name

6. Inspecting GenSeed output files

We will not cover in this tutorial the content of each output file. For a better explanation, please consult Tutorial 1.

Now please do the following steps:

1. Check the size of the sequences present in the `final_contigs.fasta` files of each of the output directories. Is there any difference between the sequences reconstructed with DNA and protein seeds?

Run a BLAST 2 sequences between the contigs of both assemblies to check if they are identical. You can use NCBI's BLAST server at http://www.ncbi.nlm.nih.gov/blast/. *Most of times, the sequences will be identical or very similar. However, some small differences may occur.*

2. Do the reconstructed sequences contain the genes which the seeds were derived from?

Run a BLASTX similarity search of the `final_contigs.fasta` *sequence against nr database. You can use NCBI's BLAST server at* http://www.ncbi.nlm.nih.gov/blast/.

3. If you run a BLASTX search of the reconstructed pyruvate kinase gene, you will notice that the 3' end of the gene did not match exactly the original gene. Would you have any explanation?

Sequence similarity extends until position 1719. From this coordinate, the sequence does not align well. Using Consed, open the assembly file located at:

`/output_pk_nt/CAP3_dir/final_reads.fasta.cap.ace` *or...*

`/output_pk_prot/CAP3_dir/final_reads.fasta.cap.ace`

Please be reminded that Consed should be run with the `-nophd` *flag, since no PHDs (Phred's quality files) are available.*

You will notice that, starting exactly at position 1720, there is a sequence discrepancy among some reads. CAP3 chose the sequence of one of these reads to compose the consensus. However, it looks like this sequence is chimeric. This is unfortunately a not very rare occurrence. For our luck, it only avoided a perfect reconstruction of the last 40 residues (from a total of 531 amino acid residues).

3. Inspect the graphical output files of each assembly. You will find them at:

`/output_directory/report.html`

These files will show you the relative position of all assembled reads, as well as the position of the seed sequences. In the intermediate rounds, the red bar corresponds to the consensus sequence of the previous round. In the first and last (final) rounds, the red bar corresponds to the original seed sequence.

4. What is the size of the reconstructed cDNAs? Why haven't we used any limit to the walking process?

You may have noticed that, differently from what was done in Tutorial 1, we have not established any limit for the walking process in terms of number of rounds (parameter `-r`*) or size (parameter* `-f`*). If you inspect* the reconstructed sequences, you should find that the following sizes were obtained:*

a. *Microneme protein 7 – 1,905 bp*
b. *Pyruvate kinase – 1,847 bp*
c. *SAG3 – 1,887 bp*

When we use shotgun reads from a complete genome as a database, the size limit of the reconstruction process is, in theory, the own genome size. This is the reason why we limited the walking process in Tutorial 1 by the number of rounds or, alternatively, by the size of the reconstructed sequence. This time, because we are reconstructing cDNAs, the process is self-limited by the size of the original transcript. Thus, provided that the database contains reads covering the whole transcript sequence, the process will automatically end up when GenSeed detects no additional read increment in the walking process. This will happen when the transcript ends have been reached. Conversely, the walking process may finish prematurely if the transcript is not fully covered by the reads of the database.

**You can use any sequence editor/viewer such as Artemis or BioEdit to determine the size of the reconstructed sequences.*

6. Reconstructing orthologous cDNAs using heterologous seeds

As we have seen in the former steps of this tutorial, GenSeed can reconstruct cDNAs starting from either a DNA or protein sequence seed. So far, we have used only homologous seeds in our tutorials. Since orthologous sequences often contain highly conserved domains, especially at the protein level, we can use protein seeds to reconstruct orthologs.

This application will be demonstrated here using the same database of *T. gondii* ESTs that used in this tutorial. We will reconstruct a *T. gondii* sugar transporter cDNA using a 22-aa seed from the orthologous protein of *Eimeria tenella*, another coccidian parasite. Because the seed is heterologous in regard to the database, we will use less stringent conditions.

Go to `tutorial_2/test` directory and type the command below:

```
genseed.pl -s ../seed/ortholog_seed.fasta -d ../db/T_gondii_EST.fasta -o
output_ortholog -b "-e 1e-02 -b 100 -F F"
```

Note that we have changed the default BLAST parameter, so now we have a much higher expectation value (1e-02 instead of default 1e-06), meaning that the cutoff for selecting positive hits is now less stringent.

You will see that reconstruction will be complete after 5 rounds:

```
seed type: Protein

####   Round 1   ####
Total # of reads for CAP3: 3
Length of the seed-contig E.: 899
Accumulative number of reads: 3

####   Round 2   ####
Total # of reads for CAP3: 27
Length of the seed-contig E.: 1843
Accumulative number of reads: 30

####   Round 3   ####
Total # of reads for CAP3: 23
Length of the seed-contig E.: 2773
Accumulative number of reads: 53

####   Round 4   ####
Total # of reads for CAP3: 8
Length of the seed-contig E.: 2906
Accumulative number of reads: 61

####   Round 5   ####
Total # of reads for CAP3: 112
Fasta seed is not present in any contig!
Finishing...

####   Last Round   ####
Length of the final seed-positive contig E_tenella: 2906
```

The final reconstructed cDNA comprises 2,906 bp. Use this sequence to perform a similarity search with BLASTX against nr database. You will find that the best hit is a facilitative glucose transporter of *Toxoplasma gondii* (accession code AAM69350). The alignment block covers the whole 568-aa protein with 99% similarity and 98% of identities. The CDS is present at coordinates 161-1864, frame +2.

The seed sequence in this alignment block is presented below:

E. tenella seed	1	IALIANLVDRHNTGVQWLTVAC	22
		IA +ANLVD+ NT VQW+TVAC	
T. gondii protein	388	IAFVANLVDQSNTAVQWVTVAC	409

Important notes:

Be reminded that ortholog reconstruction can only succeed if the following conditions are fulfilled:

a. Seed covers a sequence that is relatively well conserved in the orthologous protein whose gene is going to be reconstructed;

b. Database reads covers the gene to be reconstructed;

c. Low stringency conditions are used. Use low `-b` parameter with high BLAST's `-e` values (e.g. `-b "-e 1-e02 -b 100 -F F"`).

Also, consider that the longer the seed sequence, the best the chances to find significant matches that will permit cDNA reconstruction. Another important aspect is the fact that because we are using an orthologous seed sequence, and low stringency parameters to select the reads that will be used for the assembly steps, two scenarios may arise:

a. We find significant hits in the database and successfully reconstruct the ortholog;

b. Database (and/or the source organism) does not present sequences of the ortholog. In this case, GenSeed will try to reconstruct the sequence derived from the top hits. Because these hit may in fact be spurious due to the low stringency, a non-related sequence may be reconstructed. To avoid such occurrence, we suggest using relatively high `-i` and `-j` parameters (see "GenSeed - Quick Guide" for more details). Using, for instance, `-i 90` and `-j 90`, GenSeed will only reconstruct a sequence if the seed has found a match in the database presenting an alignment block with a length equal or higher than 90% of its length (parameter `-i`). Also, as a second parameter, the read will only be selected if the alignment block presents a similarity higher than 90% (parameter `-j`). If the read does not fulfill these criteria, than GenSeed will return a message informing that no hit has been found in the database and suggesting to use another seed.