Tutorial 1

Reconstruction of a gene or genomic segment using a nucleotide or protein sequence seed

In this tutorial, we describe how to run GenSeed to reconstruct a genomic segment containing several genes. For this purpose, we will use either a nucleotide or a protein sequence seed of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) of *Campylobacter jejuni*.

1. Seed sequence

The seed directory contains four files:

- `GAPDH_nt_seed.fasta` – a 140-bp nucleotide sequence of the GAPDH gene (accession code AL111168, region 1337856-1337995).
- `GAPDH_prot_seed.fasta` – a 25-aa protein sequence fragment of the GAPDH.

2. Database

We will use a database of shotgun reads of *Campylobacter jejuni* NCTC 11168 genome. For this tutorial, we are providing a copy of the database at `/tutorial_1/db/`. You can also download the original file from ftp://ftp.sanger.ac.uk/pub/pathogens/cj/CJ_shotgun.dbs.gz (09-02-2000).

3. 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.

To run GenSeed, first go to the `/tutorial_1/test` directory, and then type the following command:

```
genseed.pl -s ../seed/GAPDH_nt_seed.fasta -d ../db/CJ_shotgun.dbs -o
output_nt -r 5
```

4. 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/GAPDH_nt.fasta` – seed sequence file with path
- `-d ../db/CJ_shotgun.dbs` - database file with path

- `–o output_nt` - `output_nt` as the output directory name
- `–r 5` – use five walking rounds

Please notice that in this example we are using a nucleotide sequence seed. At the end of this tutorial we will perform a second reconstruction using a protein seed (`GAPDH_prot.fasta` file).

5. Understanding GenSeed output files:

If everything goes well, a new subdirectory will be created:

`output_nt`

Also, a log file (`genseed.log`) will be created in the `/tutorial_1/test/output_nt` directory. This is a text file that displays all steps followed by GenSeed. In case of an error, you should expect to find the corresponding error message in this file. Inspect the file using any text editor/viewer. It should present the following content:

```
Fri Nov 16 16:11:25 BRST 2007
genseed.pl –s ../seed/GAPDH_nt_seed.fasta –d ../db/CJ_shotgun.dbs  –o
output_nt –r 5

seed type: DNA

####   Round 1   ####
Total # of reads for CAP3: 4
Length of the seed-contig C_jenuni: 755
Accumulative number of reads: 4

####   Round 2   ####
Total # of reads for CAP3: 15
Length of the seed-contig C_jenuni: 1739
Accumulative number of reads: 19

####   Round 3   ####
Total # of reads for CAP3: 13
Length of the seed-contig C_jenuni: 2813
Accumulative number of reads: 32

####   Round 4   ####
Total # of reads for CAP3: 22
Length of the seed-contig C_jenuni: 3756
Accumulative number of reads: 54

####   Round 5   ####
Total # of reads for CAP3: 13
Length of the seed-contig C_jenuni: 4928
Accumulative number of reads: 67
####   Last Round   ####
Length of the final seed-positive contig C_jejuni: 4928
```

According to this file, GenSeed ran four rounds, as chosen by the parameter –r 5. In total, 67 reads were used for the final assembly step. No errors were reported. To see the results, go to the `output_nt` directory, where you will find the following directories…

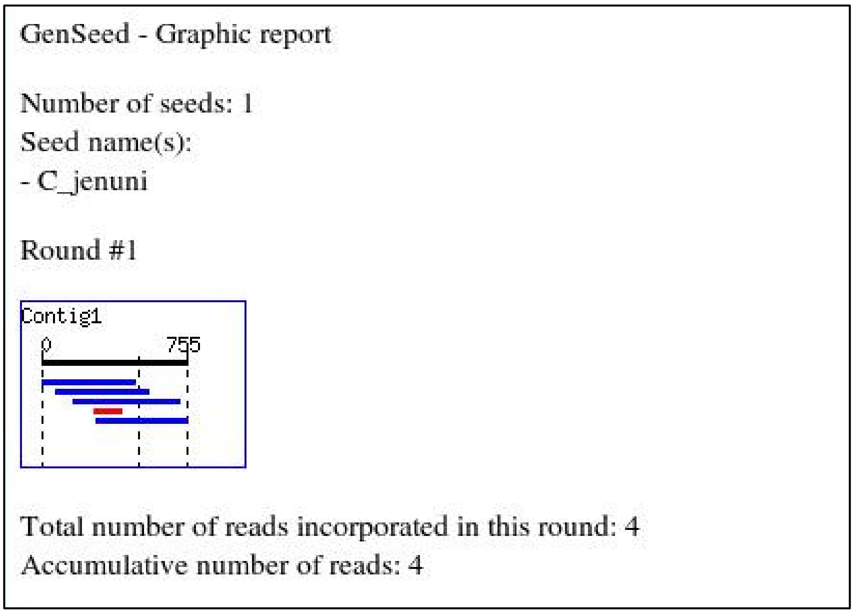
```
BLAST_dir
CAP3_dir
fasta_dir
image_dir
```

…and files:

```
final_contigs.fasta
list_of_reads.html
report.html
seed_contigs.html
```

Let's know understand all output files and directories created by GenSeed:

- `genseed.log` – This file is created by GenSeed on every run. If a file already exists, GenSeed appends the information of the new run.
- `output_nt` – This is the directory that contains all output files stored by GenSeed. It contains several files and subdirectories as follows:
  - `CAP3_dir` – This directory contains CAP3 input and output files of all rounds, plus the final assembly files.
  - `BLAST_dir` – This directory contains BLAST output files produced in each round. Files typically are named as `blast_1.out`, `blast_2.out`, etc.
  - `fasta_dir` – This directory contains all seed sequences used as inputs for BLAST searches in all rounds. At the first round the user provides the seed sequence(s) to be used. After the first round, GenSeed automatically extracts the ends from the assembled sequence and then use them as seeds in the subsequent round. Seed sequences are named `seed_1.fasta`, `seed_2.fasta`, etc. A second set of files is composed by the assembled consensus sequences of each round (`consensus_1.fasta`, `consensus_2.fasta`, etc.). The consensus sequences correspond to contig sequences assembled by CAP3 in each round. The final assembled sequence (`final_reads.fasta`) contains all the reads selected in all rounds.
  - `report.html` – This file contains the information on the size of the contig(s), number of reads included in the last assembly step, and graphical outputs that display the assembly of each walking cycle. The bitmap *.png files are stored in the `image_dir` directory. The series of screenshots below show the graphical outputs obtained for the dataset of the current tutorial.



GenSeed - Graphic report

Number of seeds: 1

Seed name(s):

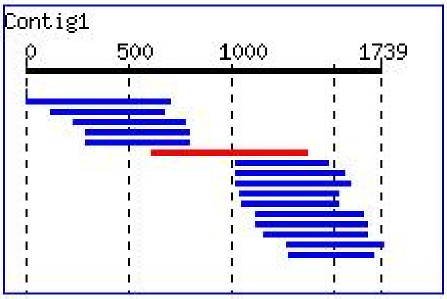
- C_jenuni

Round #1

Total number of reads incorporated in this round: 4

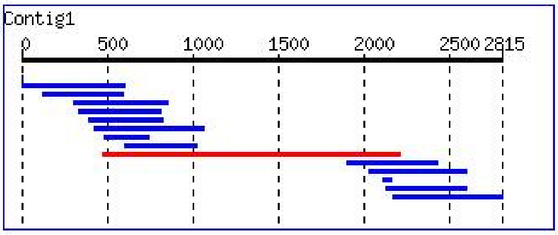
Accumulative number of reads: 4

Round #2



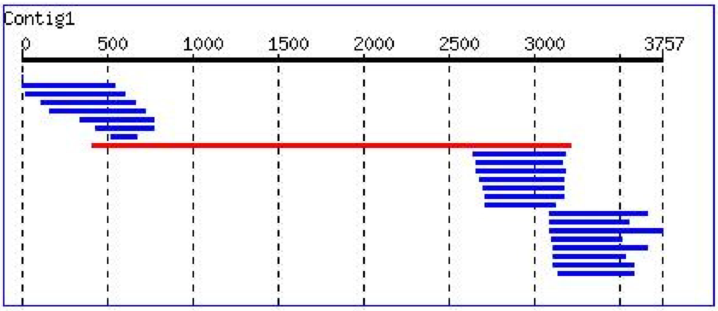
Total number of reads incorporated in this round: 15
Accumulative number of reads: 19

Round #3

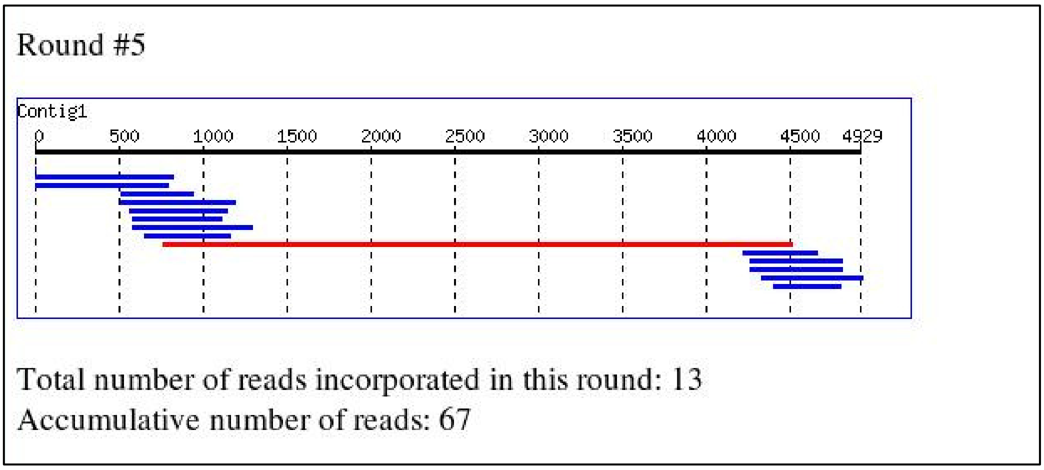


Total number of reads incorporated in this round: 13
Accumulative number of reads: 32

Round #4

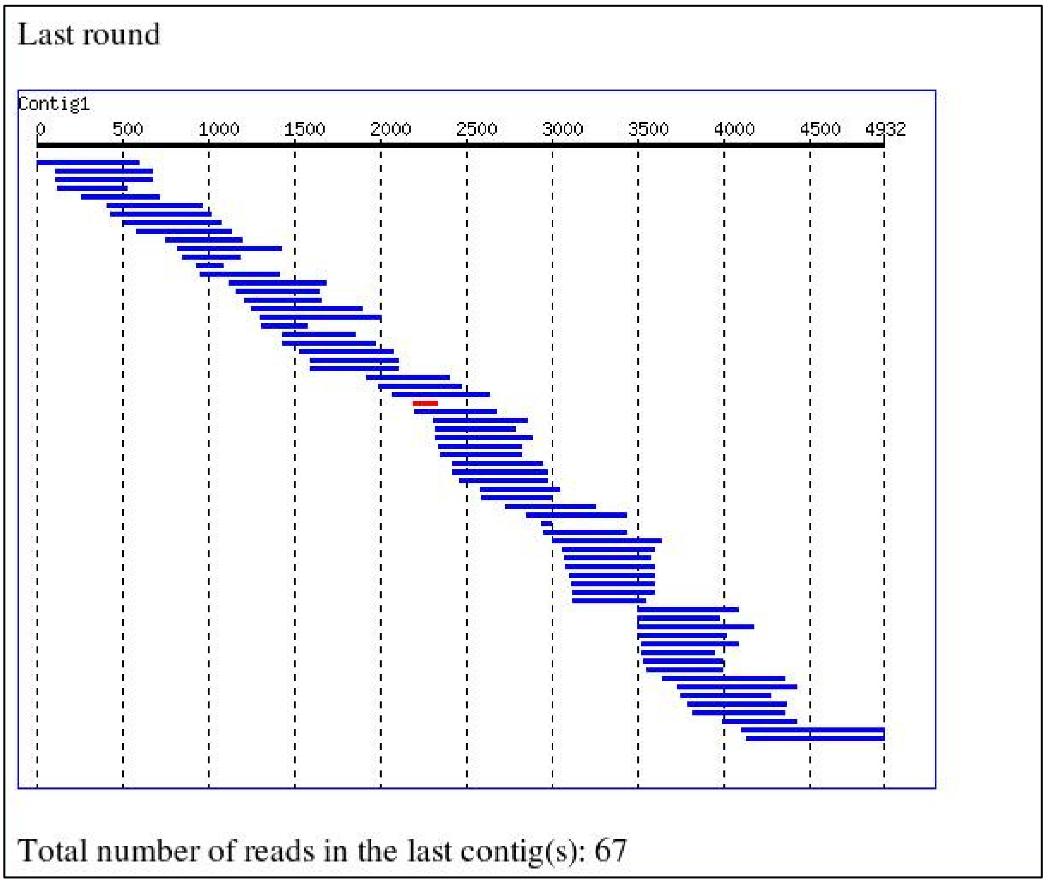


Total number of reads incorporated in this round: 22
Accumulative number of reads: 54



Round #5

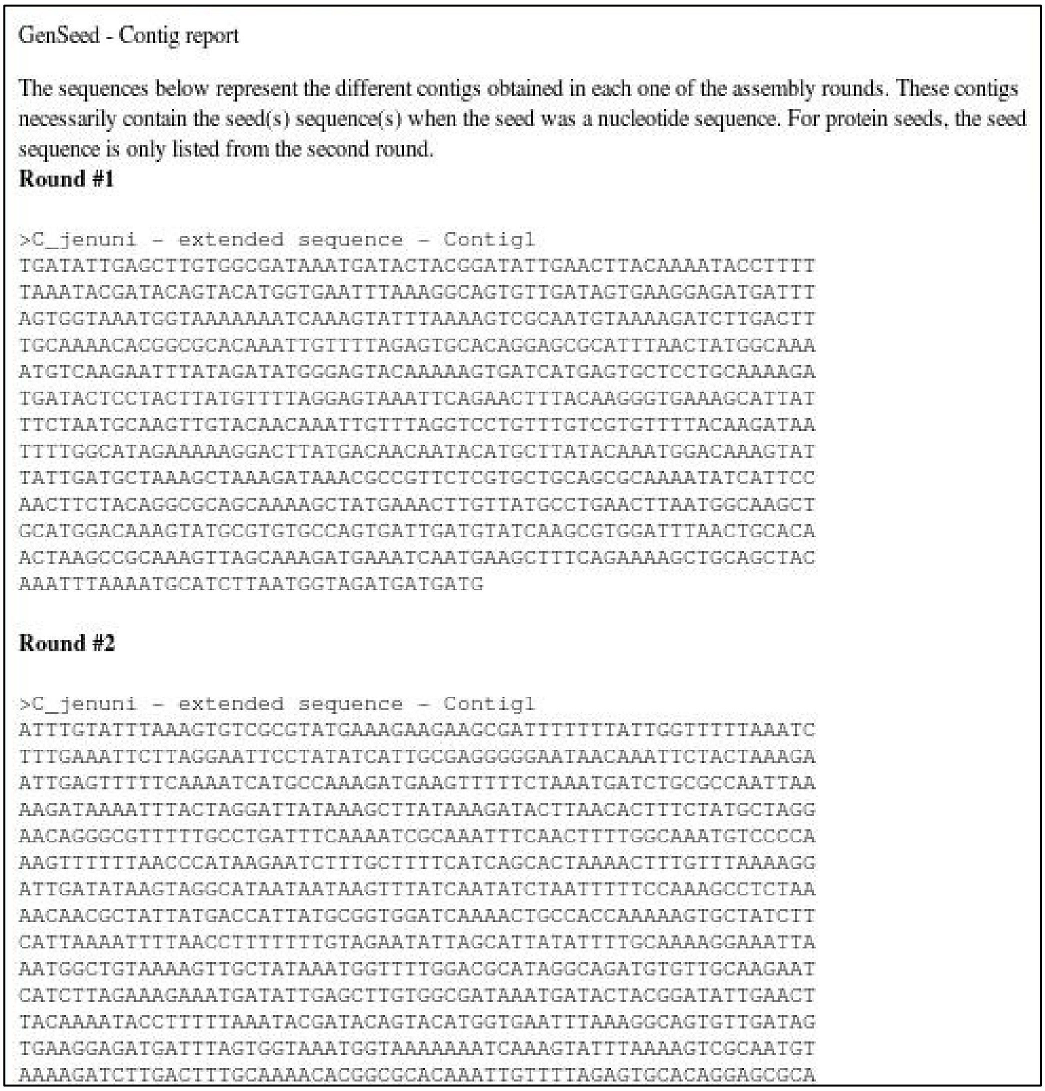
Total number of reads incorporated in this round: 13
Accumulative number of reads: 67



Last round

Total number of reads in the last contig(s): 67

- `seed_contigs.html` – This file contains the consensus sequence(s) obtained in each round.



GenSeed - Contig report

The sequences below represent the different contigs obtained in each one of the assembly rounds. These contigs necessarily contain the seed(s) sequence(s) when the seed was a nucleotide sequence. For protein seeds, the seed sequence is only listed from the second round.

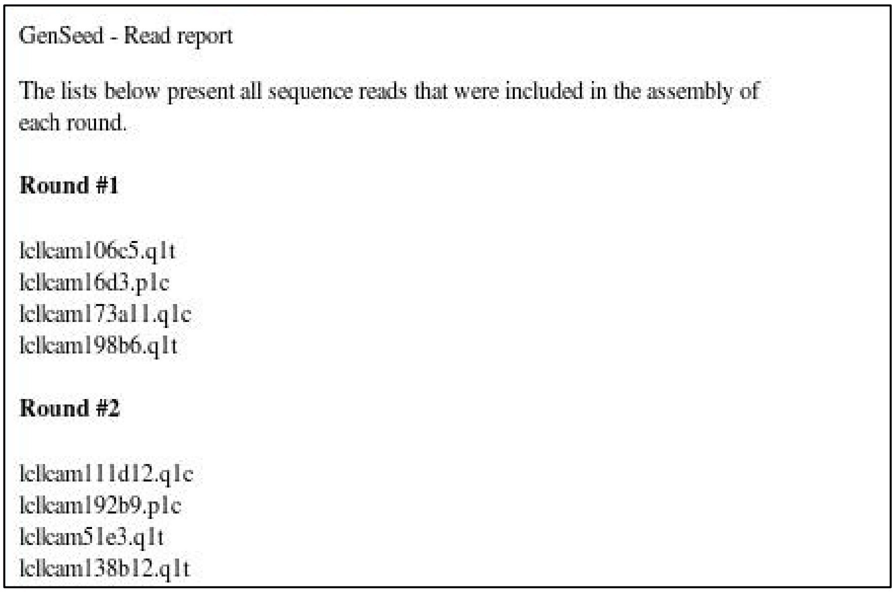
Round #1

```
>C_jenuni - extended sequence - Contig1
TGATATTGAGCTTGTGGCGATAAATGATACTACGGATATTGAACTTACAAAATACCTTTT
TAAATACGATACAGTACATGGTGAATTTAAAGGCAGTGTTGATAGTGAAGGAGATGATTT
AGTGGTAAATGGTAAAAAAATCAAAGTATTTAAAAGTCGCAATGTAAAAGATCTTGACTT
TGCAAAACACGGCGCACAAATTGTTTTAGAGTGCACAGGAGCGCATTTAACTATGGCAAA
ATGTCAAGAATTTATAGATATGGGAGTACAAAAAGTGATCATGAGTGCTCCTGCAAAAGA
TGATACTCCTACTTATGTTTTAGGAGTAAATTCAGAACTTTACAAGGGTGAAAGCATTAT
TTCTAATGCAAGTTGTACAACAAATTGTTTAGGTCCTGTTTGTCGTGTTTTACAAGATAA
TTTTGGCATAGAAAAAGGACTTATGACAACAATACATGCTTATACAAATGGACAAAGTAT
TATTGATGCTAAAGCTAAAGATAAACGCCGTTCTCGTGCTGCAGCGCAAAATATCATTCC
AACTTCTACAGGCGCAGCAAAAGCTATGAAACTTGTTATGCCTGAACTTAATGGCAAGCT
GCATGGACAAAGTATGCGTGTGCCAGTGATTGATGTATCAAGCGTGGATTTAACTGCACA
ACTAAGCCGCAAAGTTAGCAAAGATGAAATCAATGAAGCTTTCAGAAAAGCTGCAGCTAC
AAATTTAAAATGCATCTTAATGGTAGATGATGATG
```

Round #2

```
>C_jenuni - extended sequence - Contig1
ATTTGTATTTAAAGTGTCGCGTATGAAAGAAGAAGCGATTTTTTTATTGGTTTTTAAATC
TTTGAAATTCTTAGGAATTCCTATATCATTGCGAGGGGGAATAACAAATTCTACTAAAGA
ATTGAGTTTTTCAAAATCATGCCAAAGATGAAGTTTTTCTAAATGATCTGCGCCAATTAA
AAGATAAAATTTACTAGGATTATAAAGCTTATAAAGATACTTAACACTTTCTATGCTAGG
AACAGGGCGTTTTTGCCTGATTTCAAAATCGCAAATTTCAACTTTTGGCAAATGTCCCCA
AAGTTTTTTAACCCATAAGAATCTTTGCTTTTCATCAGCACTAAAACTTTGTTTAAAAGG
ATTGATATAAGTAGGCATAATAATAAGTTTATCAATATCTAATTTTTCCAAAGCCTCTAA
AACAACGCTATTATGACCATTATGCGGTGGATCAAAACTGCCACCAAAAAGTGCTATCTT
CATTAAAATTTTAACCTTTTTTTGTAGAATATTAGCATTATATTTTGCAAAAGGAAATTA
AATGGCTGTAAAAGTTGCTATAAATGGTTTTGGACGCATAGGCAGATGTGTTGCAAGAAT
CATCTTAGAAAGAAATGATATTGAGCTTGTGGCGATAAATGATACTACGGATATTGAACT
TACAAAATACCTTTTTAAATACGATACAGTACATGGTGAATTTAAAGGCAGTGTTGATAG
TGAAGGAGATGATTTAGTGGTAAATGGTAAAAAAATCAAAGTATTTAAAAGTCGCAATGT
AAAAGATCTTGACTTTGCAAAACACGGCGCACAAATTGTTTTAGAGTGCACAGGAGCGCA
```

- `list_of_reads.html` – This file contains a list of names of the reads incorporated in each round.



GenSeed - Read report

The lists below present all sequence reads that were included in the assembly of each round.

Round #1

lcl|cam106c5.q1t
lcl|cam16d3.p1c
lcl|cam173a11.q1c
lcl|cam198b6.q1t

Round #2

lcl|cam111d12.q1c
lcl|cam192b9.p1c
lcl|cam51e3.q1t
lcl|cam138b12.q1t

- `final_contigs.fasta` – This file contains the final consensus sequence assembled sequence(s) obtained by GenSeed. In this dataset we should obtain a sequence containing 4,928 bp. If you use this sequence to perform BLAST searches against the annotated genome of *C. jejuni*, you will find that the reconstructed region covered in total six full-length protein-coding genes (accession codes NP_282542 to NP_282547), including the GAPDH.

6. Making a unidirectional progressive assembly

GenSeed can perform progressive assemblies in both directions of the seed, as we have done so far. However, the program also permits the user to choose a 5' or 3' direction, such that the assembly process proceeds unidirectionally. To test this feature, please go to the `/tutorial_1/test` directory, and then type the following command:

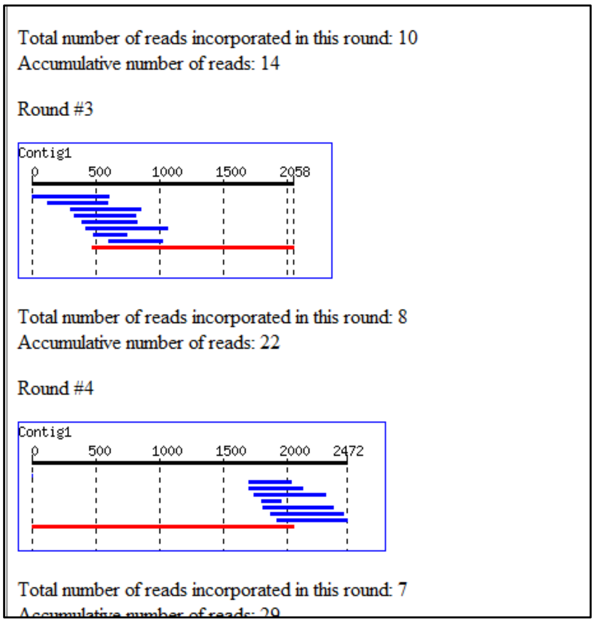
```
genseed.pl -s ../seed/GAPDH_nt_seed.fasta -d ../db/CJ_shotgun.dbs -o
output_nt_l -r 5 -e l
```

The `-e` (expansion) parameter defines the walking direction, which in this case is 5' or left (`l`).

According to `genseed.log` file, the sequence was grown in an average of 600 bp per cycle (3,017 bp divided by 5 cycles). Compare this growing performance with the one obtained in the previous step, using a bidirectional expansion. In fact, the bidirectional walking process resulted in a growing speed of almost 1 kb per cycle.

```
seed type: DNA

####   Round 1   ####
Total # of reads for CAP3: 4
Length of the seed-contig C_jenuni: 755
Accumulative number of reads: 4

####   Round 2   ####
Total # of reads for CAP3: 10
Length of the seed-contig C_jenuni: 1583
Accumulative number of reads: 14

####   Round 3   ####
Total # of reads for CAP3: 8
Length of the seed-contig C_jenuni: 2057
Accumulative number of reads: 22

####   Round 4   ####
Total # of reads for CAP3: 7
Length of the seed-contig C_jenuni: 2471
Accumulative number of reads: 29

####   Round 5   ####
Total # of reads for CAP3: 7
Length of the seed-contig C_jenuni: 3017
Accumulative number of reads: 34
####   Last Round   ####
Length of the final seed-positive contig C_jejuni: 2883
```

Now open the `report.html` file and compare the walking process with the bidirectional process executed in the previous step. Run a BLAST search of the final reconstructed sequence against the annotated genome of *C. jejuni*. Now inspect which genes were reconstructed compared to the bidirectional assembly.


Total number of reads incorporated in this round: 10
Accumulative number of reads: 14

Round #3

Total number of reads incorporated in this round: 8
Accumulative number of reads: 22

Round #4

Total number of reads incorporated in this round: 7
Accumulative number of reads: 29

7. Limiting the walking process by size

In the former steps we have reconstructed a 4,298-bp segment of *C. jejuni* genome by limiting the reconstruction process to five walking cycles (`-r` parameter). We will now re-run the program with the same seed and database, but this time limiting the reconstruction according to the length of the reconstructed fragment. We will use the parameter `-f 5000` to limit the walking process to 5 kb.

Go to the `/tutorial_1/test` directory, and then type the following command:

```
genseed.pl -s ../seed/GAPDH_nt_seed.fasta -d ../db/CJ_shotgun.dbs -o
output_nt_5kb -f 5000
```

According to `genseed.log` file, six rounds were performed, until contig of round #6 reached 5,733 bp. At this moment, the program stopped the walking cycling and ran the final assembly phase, resulting in a final reconstructed sequence of 5,732 bp (see `final_contigs.fasta` file). Please notice that 5 kb was just an upper limit for the program to know when to stop the walking process. The final contig length is slightly longer than this value.

```
seed type: Protein

####   Round 1   ####
Total # of reads for CAP3: 4
Length of the seed-contig C_jejuni: 755
Accumulative number of reads: 4

####   Round 2   ####
Total # of reads for CAP3: 15
Length of the seed-contig C_jejuni: 1739
Accumulative number of reads: 19

####   Round 3   ####
Total # of reads for CAP3: 13
Length of the seed-contig C_jejuni: 2813
Accumulative number of reads: 32

####   Round 4   ####
Total # of reads for CAP3: 22
Length of the seed-contig C_jejuni: 3756
Accumulative number of reads: 54

####   Round 5   ####
Total # of reads for CAP3: 13
Length of the seed-contig C_jejuni: 4928
Accumulative number of reads: 67

####   Round 6   ####
Total # of reads for CAP3: 21
Length of the seed-contig C_jejuni: 5733
Accumulative number of reads: 88
A contig reached the maximum number of bases!
####   Last Round   ####
Length of the final seed-positive contig C_jejuni: 5732
```

8. Running GenSeed with a protein sequence seed

This time we will run GenSeed against the same database, but using a protein sequence seed of the same GAPDH gene. To run GenSeed, first go to the `/tutorial_1/test` directory, and then type the following command:

```
genseed.pl -s ../seed/GAPDH_prot_seed.fasta -d ../db/CJ_shotgun.dbs
-o output_prot -r 5
```

You should now find a newly created `output_prot` subdirectory.

Start inspecting the log file (`genseed.log`). You will see that GenSeed has appended new information on the previously existing file. This new information should like as follows:

```
Fri Nov 16 16:51:10 BRST 2007
genseed.pl -s ../seed/GAPDH_prot_seed.fasta -d  ../db/CJ_shotgun.dbs
-o output_prot -r 5

seed type: Protein

####   Round 1   ####
Total # of reads for CAP3: 4
Length of the seed-contig C_jenuni: 755
Accumulative number of reads: 4
####   Round 2   ####
Total # of reads for CAP3: 15
Length of the seed-contig C_jenuni: 1739
Accumulative number of reads: 19
####   Round 3   ####
Total # of reads for CAP3: 13
Length of the seed-contig C_jenuni: 2813
Accumulative number of reads: 32
####   Round 4   ####
Total # of reads for CAP3: 22
Length of the seed-contig C_jenuni: 3756
Accumulative number of reads: 54
####   Round 5   ####
Total # of reads for CAP3: 13
Length of the seed-contig C_jenuni: 4928
Accumulative number of reads: 67
####   Last Round   ####
Length of the final seed-positive contig C_jenuni: 4928
```

Compare this information with the reconstruction performed with the DNA seed. You will see that both reconstructions were almost identical in respect to the size of the contig and the number of composing reads. Check the report files and compare the results.