Tutorial 3

Reconstruction of cDNAs using proteomic data

1. Introduction

In Tutorial 2, we covered cDNAs reconstruction from EST datasets using either DNA or protein seeds. At that time, we used protein seeds of 33 residues, which is a relatively long size. Proteomic data is becoming increasingly abundant, and the correct identification of the corresponding proteins and/or coding genes is a crucial step in such studies. However, proteome-derived sequences are most of times very short, sometimes in the range of 7-8 amino acid residues.

GenSeed can perform a really good job in cDNA reconstruction, even when using the typically short sequences produced by proteomics projects. In this tutorial, we will use real-life proteomic data of *Toxoplasma gondii*, a protozoan pathogen of medical and veterinary relevance.

2. Seed sequences

We will use several seed sequences that represent part of the rhoptry organelles proteome of *T. gondii*, as reported by Bradley *et al.* (Proteomic analysis of rhoptry organelles reveals many novel constituents for host-parasite interactions in *Toxoplasma gondii. J. Biol. Chem.* **280**: 34245-34258, 2005).

These authors have determined the sequence of two peptides for each excised band of a rhoptry-purified lysate separated by 1D gel electrophoresis. First, we will use these peptides as separate seeds, showing that they reconstruct the same gene. Next, we will use them simultaneously, through GenSeed's ability to use multiple seeds.

The `/tutorial_3/seed` directory contains nine protein seed files derived from proteomic data described by Bradley *et al.* (2005):

- `0176AB.fasta` – multiple sequence FASTA file containing peptide sequences #0176A and #0176B
- `0176A.fasta` – FASTA file containing peptide sequence #0176A
- `0176B.fasta` – FASTA file containing peptide sequence #0176B
- `1180AB.fasta` – multiple sequence FASTA file containing peptide sequences #1180A and #1180B
- `1180A.fasta` – FASTA file containing peptide sequence #1180A
- `1180B.fasta` – FASTA file containing peptide sequence #1180B
- `1762AB.fasta` – multiple sequence FASTA file containing peptide sequences #1762A and #1762B
- `1762A.fasta` – FASTA file containing peptide sequence #1762A
- `1762B.fasta` – FASTA file containing peptide sequence #1762B

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_3/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 the "GenSeed - Quick Guide" document. Please refer to it if you need more information.

Let's start this tutorial by reconstructing three genes using peptide data from *T. gondii* rhoptry organelles. Each pair of peptide sequences (A and B) was determined from a particular band excised from a gel, so they are in principle expected to be part of the same protein.

Go to the `/tutorial_3/test` directory and type the commands below:

```
genseed.pl  -s  ../seed/0176A.fasta  -d  ../db/T_gondii_EST.fasta  -o
output_0176A -b "-e 1000 -b 500 -F F"

genseed.pl  -s  ../seed/0176B.fasta  -d  ../db/T_gondii_EST.fasta  -o
output_0176B -b "-e 1000 -b 500 -F F"

genseed.pl  -s  ../seed/1180A.fasta  -d  ../db/T_gondii_EST.fasta  -o
output_1180A -b "-e 1000 -b 500 -F F"

genseed.pl  -s  ../seed/1180B.fasta  -d  ../db/T_gondii_EST.fasta  -o
output_1180B -b "-e 1000 -b 500 -F F"

genseed.pl  -s  ../seed/1762A.fasta  -d  ../db/T_gondii_EST.fasta  -o
output_1762A -b "-e 1000 -b 500 -F F"

genseed.pl  -s  ../seed/1762B.fasta  -d  ../db/T_gondii_EST.fasta  -o
output_1762B -b "-e 1000 -b 500 -F F"
```

These commands will invoke GenSeed to reconstruct the corresponding cDNAs of the three genes using either the peptides "A" or "B".

If everything goes well, six new subdirectories will be created:

```
output_0176A
output_0176B
output_1180A
output_1180B
output_1762A
output_1762B
```

5. Understanding GenSeed parameters

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

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

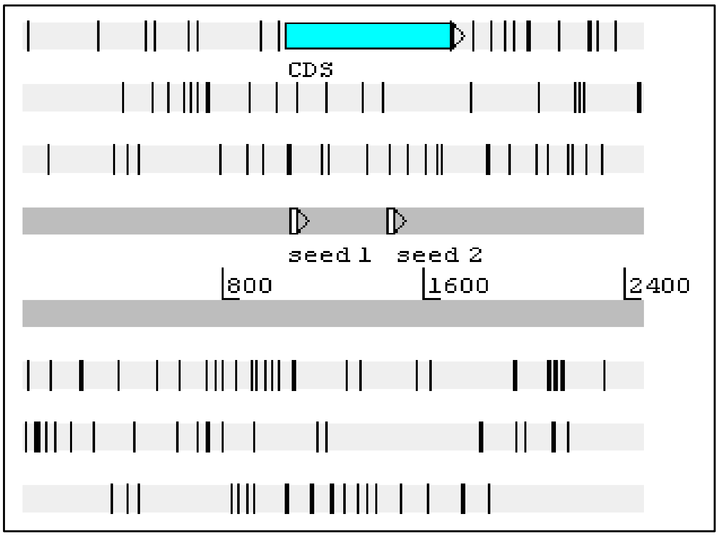
- `-s ../seed/name.fasta` – seed sequence file with path
- `-d ../db/name.fasta` - database file with path
- `-o output_name` - as the output directory name
- `-b "-e 1000 -b 500 -F F"` – set of parameters for BLAST runs

6. Inspecting GenSeed's output files

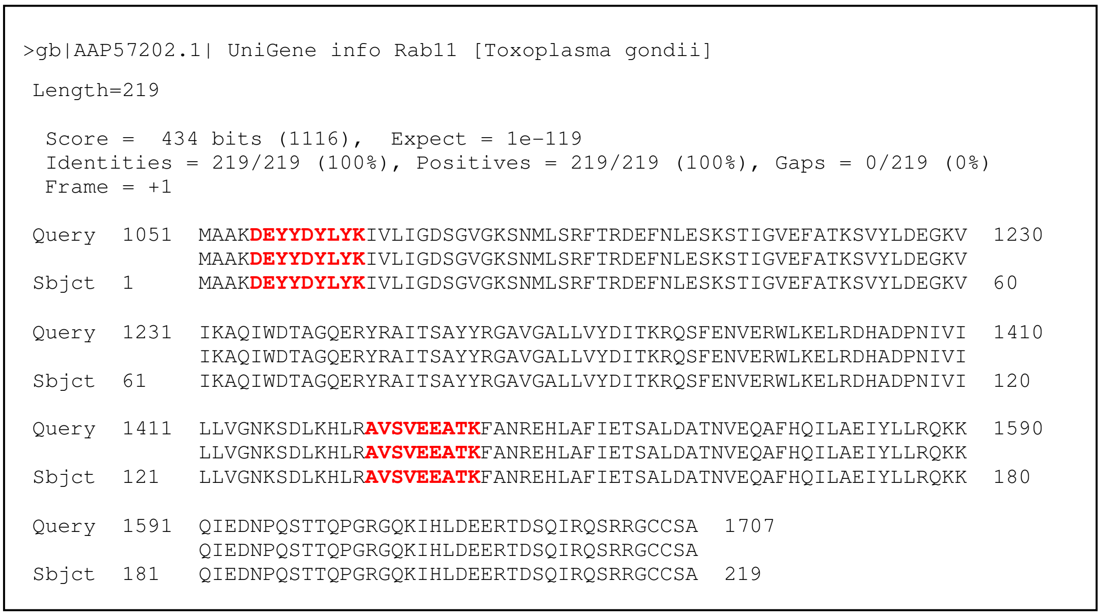
Seed	cDNA size (bp)	Protein
0176A	2470	Rab11 (*Toxoplasma gondii*)
0176B	2470	Rab11 (*Toxoplasma gondii*)
0176A + 0176B	2469	Rab11 (*Toxoplasma gondii*)
1180A	1424	Toxofilin (*Toxoplasma gondii*)
1180B	1424	Toxofilin (*Toxoplasma gondii*)
1180A + 1180B	1424	Toxofilin (*Toxoplasma gondii*)
1762A	3905	Nucleoside triphosphate hydrolase 2 (*Toxoplasma gondii*)
1762B	3905	Nucleoside triphosphate hydrolase 2 (*Toxoplasma gondii*)
1762A + 1762B	3905	Nucleoside triphosphate hydrolase 2 (*Toxoplasma gondii*)

Now inspect the `final_contigs.fasta` file of each output directory. You will notice that sequence reconstructions for peptides 1180A and 1180B, 0176A and 0176B, and 1762A and 1762B were practically identical.

Sequence #0176



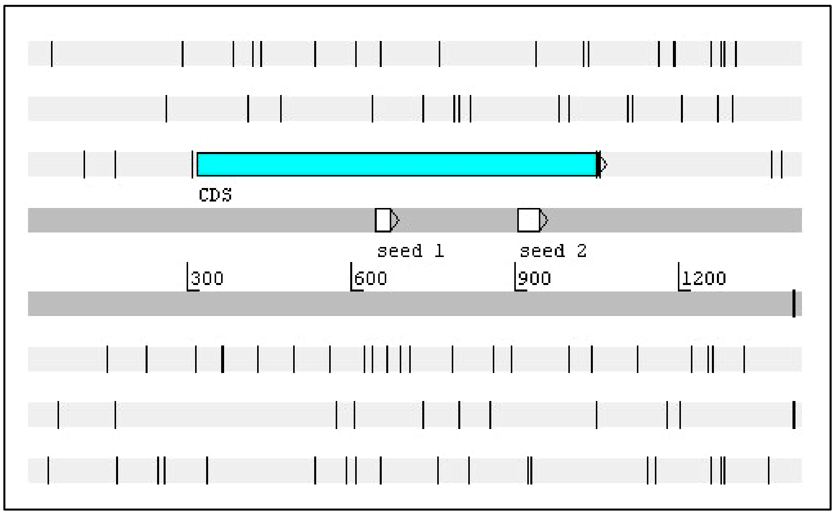
Screenshot of Artemis annotation editor, showing the reconstructed #0176 sequence using seeds 0176A (1) and 0176B (2). The blue horizontal box represents the coding sequence and the white boxes the seed sequences.



```
>gb|AAP57202.1| UniGene info Rab11 [Toxoplasma gondii]
 Length=219

  Score =  434 bits (1116),  Expect = 1e-119
  Identities = 219/219 (100%), Positives = 219/219 (100%), Gaps = 0/219 (0%)
  Frame = +1

 Query  1051  MAAKDEYYDYLYKIVLIGDSGVGKSNMLSRFTRDEFNLESKSTIGVEFATKSVYLDEGKV  1230
               MAAKDEYYDYLYKIVLIGDSGVGKSNMLSRFTRDEFNLESKSTIGVEFATKSVYLDEGKV
 Sbjct  1     MAAKDEYYDYLYKIVLIGDSGVGKSNMLSRFTRDEFNLESKSTIGVEFATKSVYLDEGKV  60

 Query  1231  IKAQIWDTAGQERYRAITSAYYRGAVGALLVYDITKRQSFENVERWLKELRDHADPNIVI  1410
               IKAQIWDTAGQERYRAITSAYYRGAVGALLVYDITKRQSFENVERWLKELRDHADPNIVI
 Sbjct  61    IKAQIWDTAGQERYRAITSAYYRGAVGALLVYDITKRQSFENVERWLKELRDHADPNIVI  120

 Query  1411  LLVGNKSDLKHLRAVSVEEATKFANREHLAFIETSALDATNVEQAFHQILAEIYLLRQKK  1590
               LLVGNKSDLKHLRAVSVEEATKFANREHLAFIETSALDATNVEQAFHQILAEIYLLRQKK
 Sbjct  121   LLVGNKSDLKHLRAVSVEEATKFANREHLAFIETSALDATNVEQAFHQILAEIYLLRQKK  180

 Query  1591  QIEDNPQSTTQPGRGQKIHLDEERTDSQIRQSRRGCCSA  1707
               QIEDNPQSTTQPGRGQKIHLDEERTDSQIRQSRRGCCSA
 Sbjct  181   QIEDNPQSTTQPGRGQKIHLDEERTDSQIRQSRRGCCSA  219
```

Sequence alignment of the best hit of sequence #0176 against nr database using BLASTX. Regions labeled in red correspond to seed sequences.

Sequence #1180

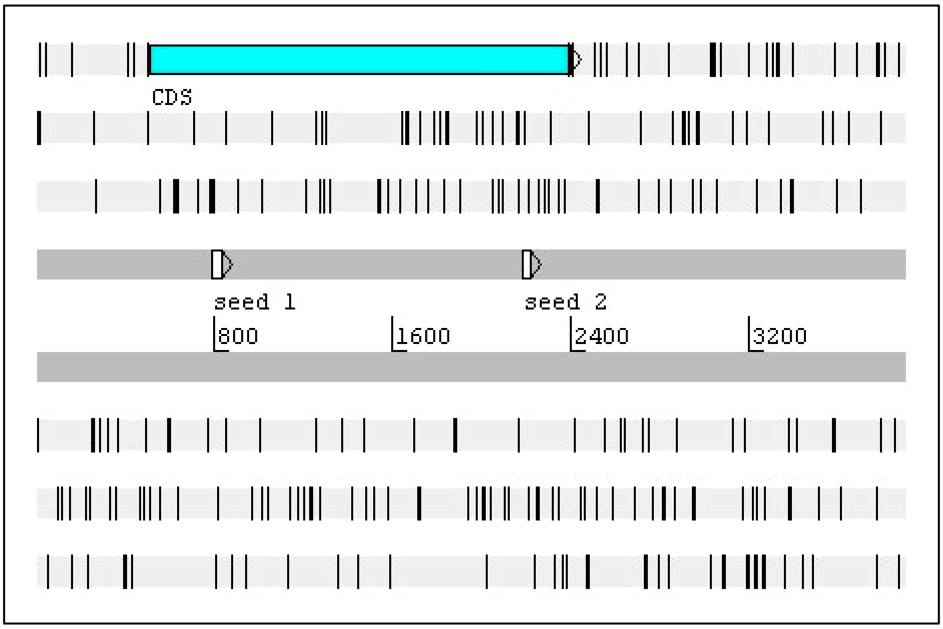


Screenshot of Artemis annotation editor, showing the reconstructed #1180 sequence using seeds 1180A (1) and 1180B (2). The blue horizontal box represents the coding sequence and the white boxes the seed sequences.

```
>emb|CAB72264.2| UniGene info toxofilin [Toxoplasma gondii]
Length=245

 Score =  434 bits (1115),  Expect = 9e-120
 Identities = 225/243 (92%), Positives = 237/243 (97%), Gaps = 0/243 (0%)
 Frame = +3

Query  318   MAQYKSRPLAAVLLLITVGSLLTASESVQLSEGMKRLSMRGRSPSPKTGRFESGDEGTST  497
             MAQYKSRPLAA LLLITVGSLLTASESVQLSEGMKRLSMRGRSPSPK GRFESGDEGTST
Sbjct  1     MAQYKSRPLAAFLLLITVGSLLTASESVQLSEGMKRLSMRGRSPSPKRGRFESGDEGTST  60

Query  498   MSPSVAARQQELGLLRPEERLIAGQAKAAALQTVHQLGAVVLTPEQAKAALLDEILRATQ  677
             MSPSVAARQQELGLLRPEERLIAGQAKAAALQTVHQLGAV LTPEQAKAALLDEILRATQ
Sbjct  61    MSPSVAARQQELGLLRPEERLIAGQAKAAALQTVHQLGAVALTPEQAKAALLDEILRATQ  120

Query  678   NLDLKKYENLNTEQQKAYEQVQKDLSLLSPETKALLIENHRKEKSLLEQAKRLFRKRHYH  857
             NLDL+KYENLNTEQQKAYEQVQ+DLS LSPETKALLIEN RKEK+LLE+A++LF++RHYH
Sbjct  121   NLDLRKYENLNTEQQKAYEQVQRDLSQLSPETKALLIENQRKEKTLLEKARKLFQRRHYH  180

Query  858   VTRQAALAGQILNEQRDASGALQSGAVKAAIRKANEQYNVAEEDKNFNEEQHAAQLKKVG  1037
             VT+QAALAGQILNEQRDASGALQSGAVK AI++ANEQYNVAEEDKNFNEEQHA+QLKKVG
Sbjct  181   VTKQAALAGQILNEQRDASGALQSGAVKTAIQRANEQYNVAEEDKNFNEEQHASQLKKVG  240

Query  1038  AMP  1046
             AMP
Sbjct  241   AMP  243
```

Sequence alignment of the best hit of sequence #1180 against nr database using BLASTX. Regions labeled in red correspond to seed sequences.

Sequence #1762

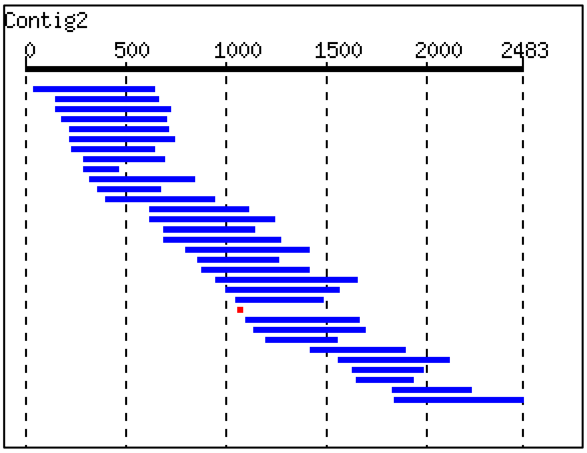


Screenshot of Artemis annotation editor, showing the reconstructed #1762 sequence using seeds 1762A (1) and 1762B (2). The blue horizontal box represents the coding sequence and the white boxes the seed sequences.

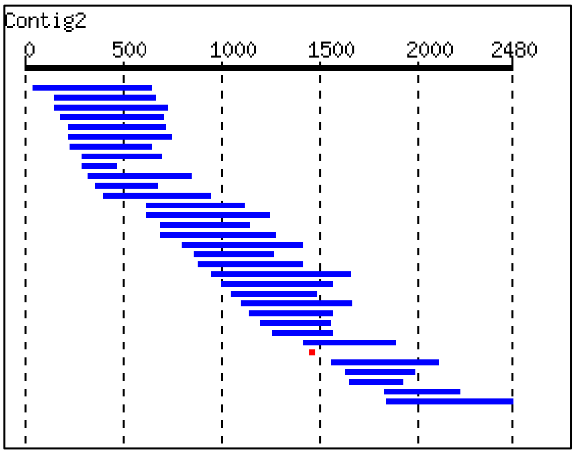
```
> sp|Q27895|NTP2_TOXGO  Nucleoside-triphosphatase 2 precursor (Nucleoside-
triphosphatase
II) (NTPase-II) (Nucleoside triphosphate hydrolase 2)
 gb|AAC41570.1|   adenosinetriphosphatase
 gb|AAC80187.1|  nucleoside triphosphate hydrolase 1 [Toxoplasma gondii]
Length=628

 Score = 1250 bits (3235),  Expect = 0.0
 Identities = 623/628 (99%), Positives = 624/628 (99%), Gaps = 0/628 (0%)
 Frame = +1

Query  508   MWLPVYVPLLLVFGVSLSLPHGSLGTDSSSLRGVDADTEKRINVGKKHLQTLRNLETRCH  687
             MWLPVYVPLLLVFGVSLSLPHGSLGTDSSSLRGVDADTEKRINVGK HLQTLRNLETRCH
Sbjct  1     MWLPVYVPLLLVFGVSLSLPHGSLGTDSSSLRGVDADTEKRINVGKTHLQTLRNLETRCH  60

Query  688   DSLQALVVIDAGSSSTRTNVFLAKTRSCPNKGRSIDPDSIQLIGAGKRFAGLRVVLEEWL  867
             DSLQALVVIDAGSSSTRTNVFLAKTRSCPNKGRSIDPDSIQLI  GKRF GLRVVLEEWL
Sbjct  61    DSLQALVVIDAGSSSTRTNVFLAKTRSCPNKGRSIDPDSIQLIREGKRFTGLRVVLEEWL  120

Query  868   DTYAGKDWESRPVDARLLFQYVPQMHEGAKKLMQLLEEDTVAILDSQLNEKQKVQVKALG  1047
             DTYAGKDWESRPVDARLLFQYVPQMHEGAKKLMQLLEEDTVAILDSQLNE+QKVQVKALG
Sbjct  121   DTYAGKDWESRPVDARLLFQYVPQMHEGAKKLMQLLEEDTVAILDSQLNEEQKVQVKALG  180

Query  1048  IPVMLCSTAGVRDFHEWYRDALFVLLRHLINNPSPAHGYKFFTNPFWTRPITGAEEGLFA  1227
             IPVMLCSTAGVRDFHEWYRDALFVLLRHLINNPSPAHGYKFFTNPFWTRPITGAEEGLFA
Sbjct  181   IPVMLCSTAGVRDFHEWYRDALFVLLRHLINNPSPAHGYKFFTNPFWTRPITGAEEGLFA  240

Query  1228  FITLNHLSRRLGEDPARCMIDEYGVKHCRNDLAGVVEVGGASAQIVFPLQEGTVLPSSVR  1407
             FITLNHLSRRLGEDPARCMIDEYGVKHCRNDLAGVVEVGGASAQIVFPLQEGTVLPSSVR
Sbjct  241   FITLNHLSRRLGEDPARCMIDEYGVKHCRNDLAGVVEVGGASAQIVFPLQEGTVLPSSVR  300

Query  1408  AVNLQRERLLPERYPSADVVSVSFMQLGMASSAGLFLKELCSNDEFLQGGICSNPCLFKG  1587
             AVNLQRERLLPERYPSADVVSVSFMQLGMASSAGLFLKELCSNDEFLQGGICSNPCLFKG
Sbjct  301   AVNLQRERLLPERYPSADVVSVSFMQLGMASSAGLFLKELCSNDEFLQGGICSNPCLFKG  360

Query  1588  FQQSCSAGEVEVRPDGSASVNEDVRKNRLKPLATYCSVHNPEISFKVTNEMQCRENSIDP  1767
             FQQSCSAGEVEVRPDGSASVNEDVRKNRLKPLATYCSVHNPEISFKVTNEMQCRENSIDP
Sbjct  361   FQQSCSAGEVEVRPDGSASVNEDVRKNRLKPLATYCSVHNPEISFKVTNEMQCRENSIDP  420

Query  1768  TKPLAERMKIENCSIIEGTGNFDKCVSQVESILVAPKLPLPANIEAASSGFESVDQVFRF  1947
             TKPLAERMKIENCSIIEGTGNFDKCVSQVESILVAPKLPLPANIEAASSGFESVDQVFRF
Sbjct  421   TKPLAERMKIENCSIIEGTGNFDKCVSQVESILVAPKLPLPANIEAASSGFESVDQVFRF  480

Query  1948  ASSTAPMFITGREMLASIDTLKDHRLLRSDFSGDVEELAEAAREFCSSEVIIRTDGPVIQ  2127
             ASSTAPMFITGREMLASIDTLKDHRLLRSDFSGDVEELAEAAREFCSSEVIIRTDGPVIQ
Sbjct  481   ASSTAPMFITGREMLASIDTLKDHRLLRSDFSGDVEELAEAAREFCSSEVIIRTDGPVIQ  540

Query  2128  LPNARGEQKLNSLNFDLCKTMALTVSLLRHMAAGENQPSFIKWEKSIAGPDGKPLADLGW  2307
             LPNARGEQKLNSLNFDLCKTMALTVSLLRHMAAGENQPSFIKWEKSIAGPDGKPLADLGW
Sbjct  541   LPNARGEQKLNSLNFDLCKTMALTVSLLRHMAAGENQPSFIKWEKSIAGPDGKPLADLGW  600

Query  2308  QVGVILHHVLFTEEWGRTAYEAGYSHNL  2391
             QVGVILHHVLFTEEWGRTAYEAGYSHNL
Sbjct  601   QVGVILHHVLFTEEWGRTAYEAGYSHNL  628
```

Sequence alignment of the best hit of sequence #1762 against nr database using BLASTX. Regions labeled in red correspond to seed sequences.

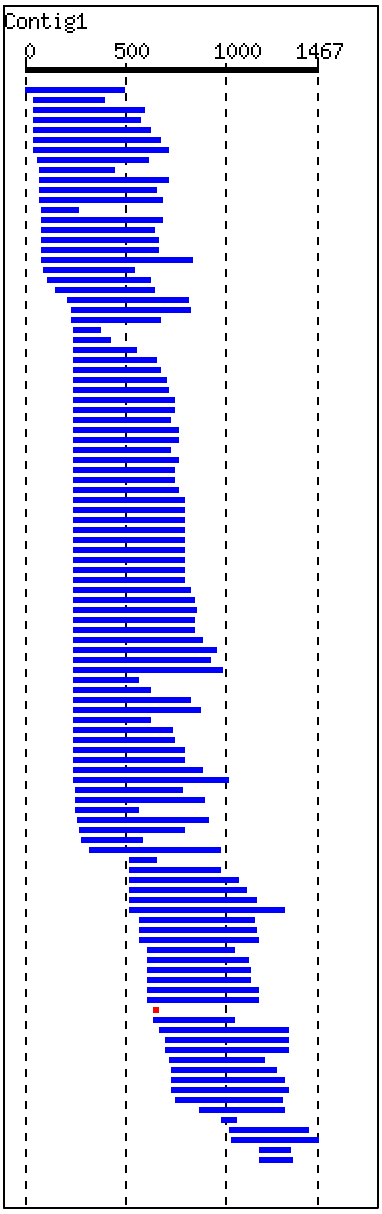
Let’s now visualize the assembly of each gene by opening the `report.html` files of each directory. The figures below represent the assemblies, with each horizontal blue bar representing a read. The arrows, when present, point to the seed (red bars).



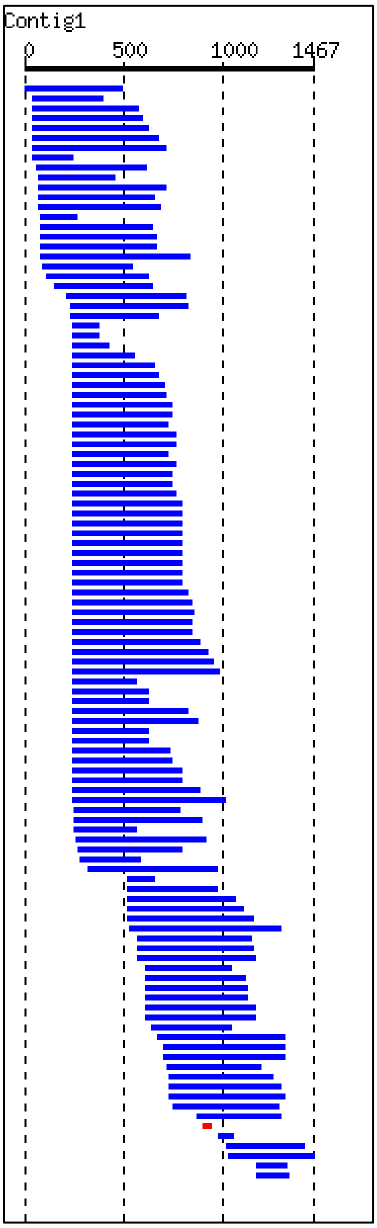
0176A



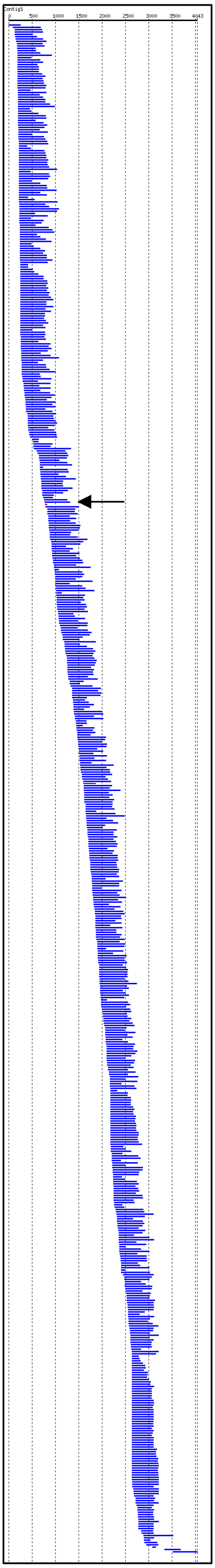
0176B



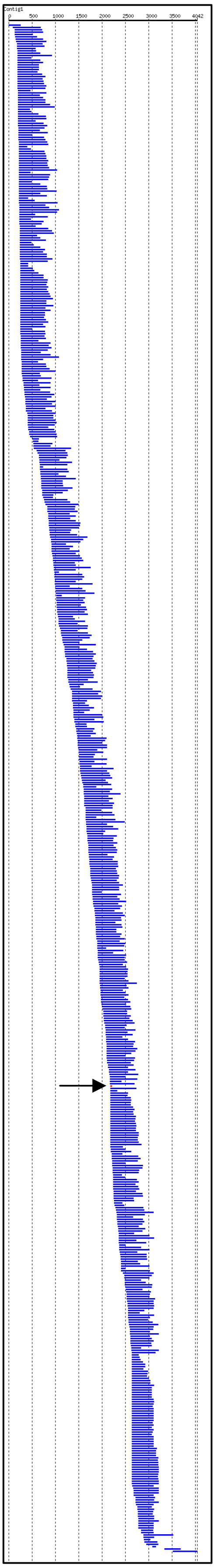
1180A



1180B


Contig1
0
500
1000
1500
2000
2500
3000
3500
4043

1762A


Contig1
0
500
1000
1500
2000
2500
3000
3500
4042

1762B

As we can see, the assemblies were the same for peptides 0176A and 0176B, 1180A and 1180B, and 1762A and 1762B.

7. Using multiple seeds

In the examples above, we commented that two peptides were generated from each excised band of the proteomics study. In order to facilitate gene reconstruction, as well as to improve performance, GenSeed may accept multiple sequence files as seeds. To test this feature, we will re-run the same reconstructions, but this time using files that contain each both peptides A and B.

Go to the `/tutorial_3/test` directory and type the commands below:

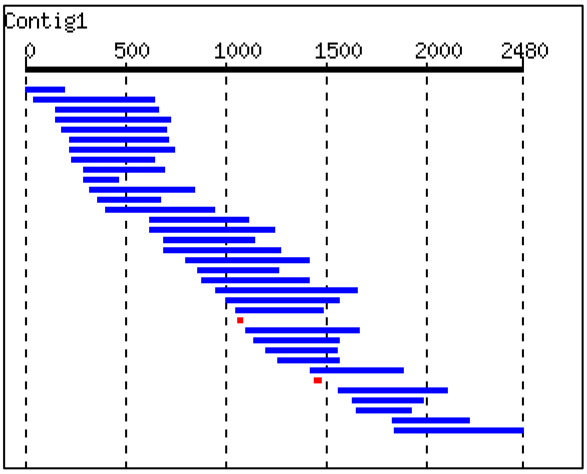
```
genseed.pl -s ../seed/0176AB.fasta -d ../db/T_gondii_EST.fasta -o
output_0176AB -b "-e 1000 -b 500 -F F"

genseed.pl -s ../seed/1180AB.fasta -d ../db/T_gondii_EST.fasta -o
output_1180AB -b "-e 1000 -b 500 -F F"

genseed.pl -s ../seed/1762AB.fasta -d ../db/T_gondii_EST.fasta -o
output_1762AB -b "-e 1000 -b 500 -F F"
```

Have a look at `genseed.log` file:

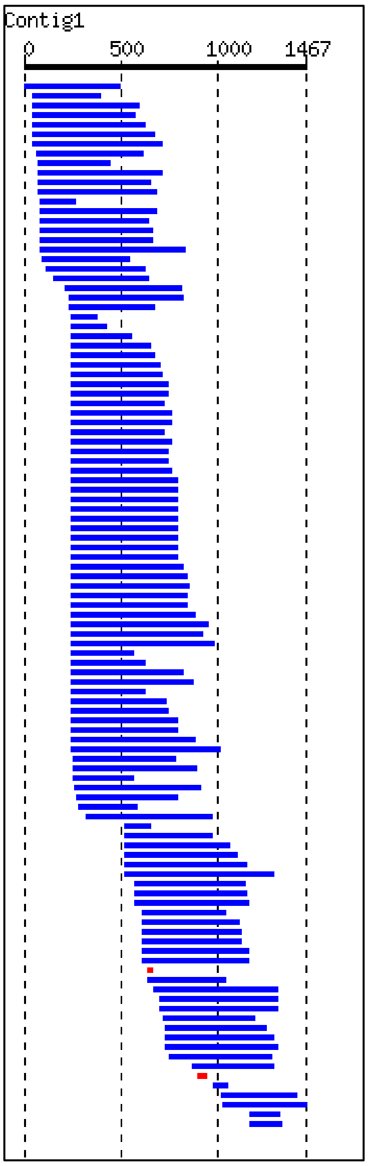
You will see that when the selected reads cover both seeds, they are all merged and assembled together. In the example below, both seeds generated distinct contigs in the first round, but at the second round, they were assembled together (`contig 1762A_1762B`).

```
seed type: Protein
####   Round 1   ####
Total # of reads for CAP3: 237
Length of the seed-contig 1762A: 1259
Length of the seed-contig 1762B: 1233
Accumulative number of reads: 215
####   Round 2   ####
Total # of reads for CAP3: 491
Length of the seed-contig 1762A_1762B: 3088
Accumulative number of reads: 646
####   Round 3   ####
Total # of reads for CAP3: 436
Length of the seed-contig 1762A_1762B: 3387
Accumulative number of reads: 652
####   Round 4   ####
Total # of reads for CAP3: 23
Length of the seed-contig 1762A_1762B: 3544
Accumulative number of reads: 653
####   Round 5   ####
Total # of reads for CAP3: 58
Length of the seed-contig 1762A_1762B: 3905
Accumulative number of reads: 654
####   Round 6   ####
Total # of reads for CAP3: 371
```

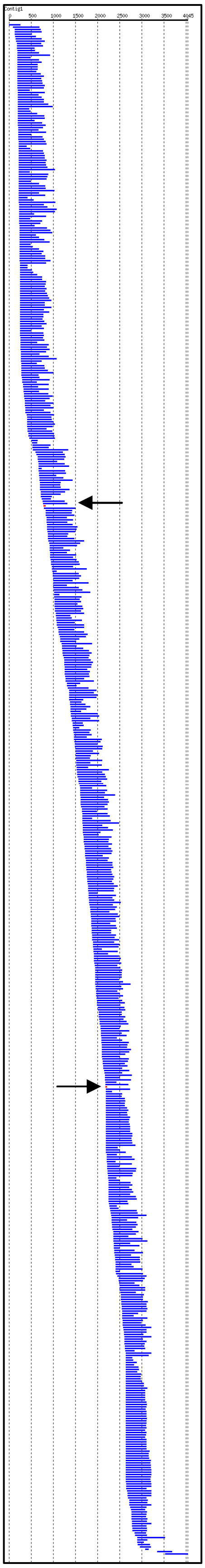
Also, inspect the `report.html` files of each assembly. The assemblies now contain both peptides used as seeds (red horizontal bars).



0176A + 0176B



1180A + 1180B



1762A +1762B