NAME

SRAssembler – Selective and Recursive local Assembler

SYNOPSIS

SRAssembler [**-q** *query_file*] [**-s** *species*] [**-p** *parameter_file*] [**-l** *library_file*] [**-1** *left_end_file*] [**-2** *right_end_file*] [**-t** *query_type*] [**-o** *output_directory*] [**-r** *preprocessed_directory*] [**-m** *min_contig_length*] [**-M** *max_contig_length*] [**-e** *min_score*] [**-c** *min_coverage*] [**-k** *kmer*] [**-z** *insert_size*] [**-n** *num_rounds*] [**-x** *num_reads_per_file*] [**-A** *assembler_program*] [**-S** *spliced_alignment_program*] [**-G** *gene_finding_program*] [**-P**] [**-y**] [**-w**] [**-a** *assemble_round*] [**-b** *clean_round*] [**-v**] [**-h**]

DESCRIPTION

SRAssembler is a pipeline program that can assemble genomic DNA reads using homologous sequences as input. **SRAssembler** first aligns the reads that can locally be mapped onto those query sequences. These mapped reads are then assembled as contigs, which are further used to find other reads partially mapped to them. This *in-silico* chromosome walking strategy can recursively gather reads that are associated with regions of interest. The gene structure of the final contigs is predicted by spliced alignment and *ab initio* gene finding programs.

The **SRAssembler** program was developed in the group of Prof. Volker Brendel. Information on program availability may be obtained at http://brendelgroup.org/bioinformatics2go/bioinformatics2go.php.

Correspondence relating to **SRAssembler** should be addressed to

Volker Brendel

School of Informatics and Computing

Indiana University, Bloomington, Indiana 47405

Tel: (515) 294-9884, Fax: (515) 294-6755

Email: vbrendel@indiana.edu

OPTIONS

-q *query_file*

Required. The FASTA format query file.

-s *species*

Required. Set species to select the most appropriate splice site models. Options: “human”, “mouse”, “rat”, “chicken”, “drosophila”, “nematode”, “fission_yeast”, “aspergillus”, “arabidopsis”, “maize”, “rice”, “medicago”.

-p *parameter_file*

Required. Parameter configuration file. **SRAssembler** allows users to specify the parameters of the programs used in **SRAssembler** such as **Vmatch**, **GenomeThreader**, **GeneSeqer**, **Exonerate** and **Snap**. The parameters for each program are grouped by **[program_name]**. For example:

```
[Vmatch_protein_init]
e=1
l=11

[Vmatch_extend_contig]
e=0
l=30

[GenomeThreader]
gcmincoverage=10
prminmatchlen=15

[GeneSeqer]
x=14
y=14
z=25

[Exonerate]
percent=20

[Snap]
snaphmm=A.thaliana.hmm
```

For aligners, the parameters for initial round can be specified by the section with the suffix of query type and "init". For example, the initial round parameter section for **Vmatch** is **[Vmatch_protein_init]** or **[Vmatch_cdna_init]**. **-l** option in **[Vmatch_protein_init]** or **[Vmatch_cdna_init]** is the match length in the first round. Since the first round is the alignment of homologous genes, if you set it too high, you may get very few hits. The default value for protein query sequences is 10; 30 for cDNA query sequences. If your assembly has very poor spliced alignment results, you can decrease this value to gather more reads to improve your assembly results. **-e** option in **[Vmatch_protein_init]** or **[Vmatch_cdna_init]** is the number of mismatches allowed in the first round. The default value is 1. If this value is too low, you may get very few matched reads if your query sequences are not very well conserved. On the other hand, if this value is too high, some false positive reads may be fetched.

For the recursive rounds, the parameters are set in the section with the suffix of "extend_contig". For **Vmatch**, the section name is **[Vmatch_extend_contig]**. Similar to the initial round, the **-l** option is the match length in recursive rounds. The default value is 30. This value controls the speed of chromosome walking. If your dataset is very deep, you can specify higher value and save the running time. **-e** option **in [Vmatch_extend_contig]** is the mismatches allowed in recursive round. The default value is 0.

For snap, please set the environment variable **ZOE** to the path of HMM files.

-l *library_file*

Library definition file. Each library is specified by **[LIBRARY]** section. Each section includes the following items:

insert_size: the insert size of the library. This value is used in paired-end reads. Default: 300.

direction: the sequencing direction of paired-end reads. 0 : forward-reverse; 1: reverse-forward. Default: 0.

r1: left-end reads file or single-end reads file.

r2: right-end reads file. Do not specify if your library is single-end.

format : “fastq” or “fasta”. Default: “fastq”.

Note that if your library contains both paired-end and single-end reads, please treat them as two libraries.

Here is an example of two libraries:

```
[LIBRARY]
insert_size=200
direction=0
r1=reads1_200.fq
r2=reads2_200.fq
format=fastq

[LIBRARY]
insert_size=1000
direction=0
r1=reads1_1000.fq
r2=reads2_1000.fq
format=fastq
```

-1 *left_end_file*

Required if the -l option is not used; use this option to specify the single-end reads file or the left-end reads file for paired-end reads.

-2 *right_end_file*

Right-end reads file for paired-end reads.

-z *insert_size*

The insert size of the paired-end reads [default: 300]

-t *query_type*

Query file type: Options: “protein”, “cdna” [default: protein].

-o *output_directory*

Output directory [default: current directory]

-r *preprocessed_directory*

Directory in which to store or from which to retrieve the pre-processed reads [default: output directory/reads_data]. For each **SRAssembler** run, reads data will be split into several parts. This preprocessing step is executed in the first time only. When **SRAssembler** is rerun, the preprocessed reads will be reused. If users have several runs which share the same dataset, this option can be used to avoid the preprocessing step.

-m *min_contig_length*

Minimum contig length to be reported [default: 200]

-M *max_contig_length*

Maximum contig length to be reported [default: 10000]. If the contig size is larger than this value, **SRAssembler** will stop assembling such contigs. Because they are long enough, we do not want to waste our time to keep assembling them again. We also remove the reads associated with these contigs, therefore improving the running time. This is done by the following steps:

1. In each round, we test if the contig length is larger than the maximum contig size. We trim the head and tail of the contigs and make their size be equal to the maximum contig size, and then copy these contigs to the candidate long contig file. Note that we do not remove them immediately, because we want to do the double check if these long contigs are correctly assembled. If such contigs are assembled again, we can confirm they are our final contigs.

2. In the next round, we align the candidate long contigs to the current assembled contig file (done by **Vmatch**). If matched, we move the contigs to the permanent long contig file.

3. We align current matched reads to the long contigs (done by **Bowtie**). If matched, those reads are removed from the reads pool.

4. Long contigs are removed from the query file of the next round.

-e *min_score*

Minimum score to determine the query genes have been assembled. [default: 0.5]. In each round, **SRAssembler** will align the query sequences to the assembled

contigs. If the alignment score and coverage (see **-c** option) are above the specified threshold, the assembling of these genes is considered finished. These query sequences will be removed and **SRAssembler** will not assemble them anymore.

-c *min_coverage*

Minimum coverage to determine the query genes are assembled. [default: 0.5]. See the explanation of option **-e**.

-k *kmer*

The k-mers of assembler. **SRAssembler** can test multiple k-mers at the same time. The format is: **start_k:interval:end_k**. The start_k and end_k must be an odd value, and the interval must be an even value. For example, **15:10:45** means k-mer value 15, 25, 35, 45 will be tested. [default: 15:10:45]

-n *num_rounds*

Number of maximum rounds [default: 10]

-x *num_reads_per_file*

Number of reads in the split files [default: 500000]

-A *assembler_program*

The internal assembler program used in **SRAssembler**. Options: **0**=>**SOAPdenovo 1**=>**ABySS** [default: 0]

-S *spliced_alignment_program*

The spliced alignment program used in **SRAssembler**. Options: **0**=>**GenomeThreader**, **1**=>**GeneSeqer**, **2**=>**Exonerate** [default: 0]

-G *gene_finding_program*

The gene finding program used in **SRAssembler**. Options: **0**=>none, **1**=>Snap [default: 0]

-P Run the read pre-processing step only, then terminate SRAssembler.

-y Disable SRAssembler resumption from previous checkpoint (will overwrite existing output). By default, **SRAssembler** will continue the previous run if the **-n** option is larger than previous one. For example, if the previous run stops at round 6. You can continue previous run starting from 7 if **-n** is larger than 6. If you specify this **-y** option, **SRAssembler** will start the assembly from first round.

-w Forgo spliced alignment check after intermediate assembly rounds (SRAssembler will continue for the -n specified number of rounds). By default, **SRAssembler** will align the query sequences to the assembled contigs. If the alignment score (see **-e** option) and coverage (see **-c** option) are above the specified threshold, the assembling of these genes is considered finished. If you specify this option, **SRAssembler** will NOT check if genes are assembled. This is particular useful if you want to assemble longer contigs and do not want to stop the assembling when the coding regions are covered. For example, if you want to assemble the UTR or promoter regions, you can turn this option on to get longer contigs.

-a *assemble_round*

The number of the round in which to start read assembly [default: 1].This option indicates which round **SRAssembler** starts to do the assembly. For the deeper dataset, we can start the assembly from round 1. But for some datasets with lower coverage, assembly in the early rounds may cause the wrong contigs, which will further affect your final results.

-b *clean_round*

The number of the round in which to periodically remove unrelated contigs and reads. For example, "-b 3" specifies that **SRAssembler** will remove unrelated contigs and reads after assembly rounds 3, 6, 9, ... [Default: 3]. **SRAssembler** may assemble many contigs which are unrelated to the query sequences. So in the cleaning rounds, **SRAssembler** align the query sequences to the contigs. If the contigs have 0 hits, they will be deleted. The reads which fail to map to the survival contigs will be removed as well.

-v Debug mode. The detailed information will be printed.

-h Print this usage synopsis.

OUTPUT

- **output_dir/output/summary.html:** The summary html file.
- **output_dir/output/all_contigs.fasta:** All assembled contigs.
- **output_dir/output/hit_contigs.fasta:** The contigs identified by spliced alignment program.
- **output_dir/output/output.aln:** The spliced alignment report.
- **output_dir/output/output.ano:** The *ab initio* gene prediction report.
- **output_dir/output/msg.log:** The detailed log file.
- **output_dir/output/intermediates directory:** All intermediate contigs are saved in this directory. For example, the contigs assembled in round 3 is contigs-3.fasta.
- **output_dir/reads_data directory** : The original reads data files will be split into smaller parts in FASTA format. We save these preprocessed files in this directory.

If you want to rerun this dataset, **SRAssembler** will read the preprocessed data directly. This directory can also be specified by **-R** option.

- **output_dir/tmp directory** : This directory is used to keep all the temporary files. They serve as the checkpoint files so that SRAssembler can continue previous assembling. If this directory is removed, SRAssembler must start from scratch in the next run.

TEST FILES

We used SAMTools' wgsim program to simulate 2 libraries from region 300,000 ~ 400,000 of chromosome 1 of Arabidopsis with insert size 200bp and 1kb. Each library has 50,000 paired reads. The base error rate is 0.002 and the fraction of indels is 0.001. In data directory, several files are provided:

- input/LOC_Os06g04560.pep: the protein sequence of rice gene Os06g04560.1. FASTA format.
- input/reads1_200.fq: the left-end of simulated reads with insert size 200bp. The length is 70bp, FASTQ format.
- input/reads2_200.fq: the right-end of simulated reads with insert size 200bp. The length is 70bp, FASTQ format.
- input/reads1_1000.fq: the left-end of simulated reads with insert size 1000bp. The length is 70bp, FASTQ format.
- input/reads2_1000.fq: the right-end of simulated reads with insert size 1000bp. The length is 70bp, FASTQ format.
- input/reads_2.fastq: the right-end of simulated reads. The length is 70bp, FASTQ format
- libraries_200bp.config: library definition file for 200bp library.
- libraries_200bp_1kb.config: library definition file for 200bp and 1kb libraries.

RUNNING TESTS

In “data” folder, a simple script is provided: xtest. This script will run 5 tests. Type:

```
cd data
./xtest-all
```

- **Example 1 : (use -1, -2 options to assign reads files)**
  In this test, we use one libraries with insert size 200bp using **-1** and **-2** options. This script executes the following command:

```
../bin/SRAssembler -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -1 input/reads1_200.fq -2 input/reads2_200.fq -z 200 -r ./reads_data -x 15000 -o Stestout1 -A 0 -S 0 -s arabidopsis -n 10
```

- **Example 2 : (use library definition file)**
  In this test, we use one libraries with insert size 200bp using a library definition file. This script executes the following command:

```
../bin/SRAssembler -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp.conf -r ./reads_data -x 15000 -o Stestout2 -A 0 -S 0 -s arabidopsis -n 10
```

- **Example 3 : (use two libraries)**
  In this test, we use two libraries with insert size 200bp and 1kb. You will see the assembly is more efficient than the one library case.

```
../bin/SRAssembler  -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp_1kb.conf -r ./reads_data -x 15000 -o Stestout3 -A 0 -S 0 -s arabidopsis -n 3
```

- **Example 4 : (test 200bp library with MPI feature)**
  In this test, we use 4 cores to run the **SRAssembler** simultaneously. You can notice the improvement of running time. (You may get errors if you do not install MPI environment properly)

```
mpirun -n 4 ../bin/SRAssembler_MPI -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp_1kb.conf -r ./reads_data -x 15000 -o Stestout4 -A 0 -S 0 -s arabidopsis -n 3
```

- **Example 5 : (The checkpoint feature)**
  In this test, we first test 3 rounds. Then the **-n** option is changed to 7. The **SRAssembler** will starts from 4th round. Finally, we use **-y** option to force SRAssembler to start the assembling from scratch:

```
../bin/SRAssembler -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp.conf -r ./reads_data -x 15000 -o Stestout5 -A 0 -S 0 -s arabidopsis -n 4

../bin/SRAssembler -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp.conf -r ./reads_data -x 15000 -o Stestout5 -A 0 -S 0 -s arabidopsis -n 7

../bin/SRAssembler -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp.conf -r ./reads_data -x 15000 -o Stestout5 -A 0 -S 0 -s arabidopsis -n 7 -y
```

- **Example 6: Looking for potentially better alignments ... (Use higher score (-e 0.6) and coverage (-c 0.9) values to declare spliced alignment hits. Only report contigs of at least 7000 nucleotides.)**
  In this test, we use higher score and coverage to check if the genes are assembled:

```
mpirun -n 8 ../bin/SRAssembler_MPI -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp_1kb.conf -r ./reads_data -x 15000 -o Stestout6 -A 0 -S 0 -s arabidopsis -n 20 -e 0.6 -c 0.9 -m 7000
```

- **Example 7 : (use GeneSeqer as spliced alignment program (-S 1))**
  In this test, we use GeneSeqer as the spliced alignment program. You might get errors if GeneSeqer is not properly installed:

```
../bin/SRAssembler -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp.conf -r ./reads_data -x 15000 -o Stestout7 -A 0 -S 1 -s arabidopsis -n 10
```

- **Example 8 : (use Exonerate as spliced alignment program (-S 2))**
  In this test, we use Exonerate as the spliced alignment program. You might get errors if Exonerate is not properly installed. Note the coverage here refers to the total nucleotides covered by query sequences:

```
../bin/SRAssembler -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp.conf -r ./reads_data -x 15000 -o Stestout8 -A 0 -S 2 -s arabidopsis -n 10
```

- **Example 9 : (use Snap as *ab initio* gene prediction program (-G 1))**
  In this test, we use Snap as *ab initio* gene prediction program. You might get errors if Snap is not properly installed:

```
../bin/SRAssembler -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp.conf -r ./reads_data -x 15000 -o Stestout9 -A 0 -G 1 -s arabidopsis -n 10
```

- **Example 10 : (Pre-process reads only (-P))**
  In this test, we do the preprocessing only. SRAssembler will stop right after the preprocessing is complete:

```
../bin/SRAssembler -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp.conf -r ./reads_data -x 15000 -o Stestout10 -P
```

- **Example 11 : (Remove unrelated contigs and reads every 2 rounds (-b 2))**
  In this test, SRAssembler will clean unrelated reads and contigs every 2 rounds (round 2, 4, 6 …):

```
../bin/SRAssembler -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp.conf -r ./reads_data -x 15000 -o Stestout11 -A 0 -S 0 -s arabidopsis -n 10 -b 2
```

- **Example 12 : (Assemble contigs only from round 3 onward (-a 3))**
  In this test, SRAssembler will start assembling from 3rd round. SRAssembler will not assemble the first round and all the hit reads in the first round will be the "seed" reads and will be used to find adjacent reads associated with them:

```
../bin/SRAssembler -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp.conf -r ./reads_data -x 15000 -o Stestout12 -A 0 -S 0 -s arabidopsis -n 10 -a 3
```

- **Example 13 : (Do not check if the query genes are assembled until the last round (-w))**
  In this test, SRAssembler will not check if the query genes are assembled until the last round:

```
../bin/SRAssembler -q input/LOC_Os06g04560.pep -t protein -p SRAssembler.conf -l libraries_200bp.conf -r ./reads_data -x 15000 -o Stestout13 -A 0 -S 0 -s arabidopsis -n 10 -w
```

AUTHOR

Hsien-chao Chou <hsienchao.chou@gmail.com>