



SSPACE Basic Tutorial



Tutorial

Date: 08 November 2011
Software Version: 2.0
Developed by: Marten Boetzer and Walter Pirovano



SSPACE Basic Tutorial

getting started

=====

Download and obtaining data

Download the E.coli 200bp paired-end read data from
<http://www.ebi.ac.uk/ena/data/view/SRR001665>

Download both .fastq files and store them on disk. For this example, we place the two datasets in the 'example' folder which is present in the SSPACE .zip file.

In the 'example' folder, a contig datasets is present. This file is generated by assembling the E.coli 200bp paired-end dataset with ABYSS, and contains only contigs larger or equal to 100 basepairs.

In the same folder, a library file is available. This library file describes the datasets and distances between the reads for each library. Multiple libraries are allowed, here only is described. Each column per library should be separated by a space. In the file 'library.txt' library is described for the E.coli 200bp paired-end dataset. Each column is described below;

- First column contains a desired name, here we take 'ecoli'.
- Second and third columns are the datasets.
- Fourth column is the mean distance between the paired reads, which is 200.
- Fifth column is a deviation of the mean distance that is allowed, here we take 0.75. This means that any pair having a distance between 150 and 250 is allowed.
- The final column is a value to indicate whether to take reverse complement of the reads (usually the case for mate pair data). We set this to 0, so no reverse complement;

Before running SSPACE, set your current working directory to the 'example' directory(`cd (path_to_SSPACE)/example`) were both the library file and the contig sequences are stored.

Running SSPACE without extension

SSPACE Basic Tutorial

We perform scaffolding on the ABYSS contigs, without extending them. To run SSPACE without contig extension, -x should be set to 0. Here we run SSPACE with a minimal number of five links (-k 5) and a ratio less than 0.7 (-a 0.7). If a contig pair satisfy these two parameters, they are considered for scaffolding.

```
perl (path_to_SSPACE)/SSPACE_v2-0.pl -l libraries.txt -s  
contigs_abyss.fa -k 5 -a 0.7 -x 0 -b ecoli_scaffolds_no_extension
```

A number of files and folders are generated.

The main file is the 'ecoli_scaffolds_no_extension_summaryfile.txt'. This file contains information of the final scaffolds, e.g. number of scaffolds, N50 and scaffold statistics. Also, you can see how many paired reads failed to map to contigs, and how many paired reads are used for scaffolding.

The final scaffolds in .fasta format and an .evidence file are outputted in the current working directory. The evidence file shows all the contigs present on each scaffold.

Running SSPACE with extension

To run SSPACE with contig extension, -x should be set to 1. Here, the minimal overlap of the reads with the contigs should at least be 30 bases (-m 30). The minimal number of reads needed to extend a contig is set to 20 (-o 20).

```
perl (path_to_SSPACE)/SSPACE_v2-0.pl -l libraries.txt -s  
contigs_abyss.fa -k 5 -a 0.7 -x 1 -m 30 -o 20 -b  
ecoli_scaffolds_extension
```

If there are contigs extended, you can view the information in the 'ecoli_scaffolds_extension_summaryfile.txt' file. Information is provided about the number of contigs that are extended and the number of reads that are used.

The sequences of the extended contigs are stored in the 'intermediate_results' folder. For each contig it is indicated if it has been extended or not.



SSPACE Basic Tutorial

Visualise scaffolds

To make a .dot file for visualisation, the -p option should be set to 1.

```
perl (path_to_SSPACE)/SSPACE_v2-0.pl -l libraries.txt -s  
contigs_abyss.fa -k 5 -a 0.7 -x 0 -p 1 -b ecoli_scaffolds_no_extension
```

A .dot file is generated in the 'dotfiles' folder. To convert the file to any desired format, GraphViz should be downloaded at www.graphviz.org. Once installed, the .dot file can be converted to a .ps file with;

```
dot -Tps2 (dotfile.dot) -o visualisation_scaffolds.ps
```

where 'dotfile.dot' stands for the produced .dot file. The output is a .ps file named 'visualisation_scaffolds.ps', containing the scaffolds and its contigs. For more information see the help file of dot.