Nucleics

UniFinish

Users manual

Version 1.1

August 2002

Table of contents

Overview

UniFinish is intended for use with the UniSelect software. UniFinish analyses PHRAP generated assembly files (ACE formate) and identifies those regions in the assembly that require additional UniSeq walking experiments to either close contig gaps or resolve internal regions of low quality within existing contigs.

The user can select to either "gap closure" or "low-quality regions" UniSeq reactions. UniFinish reads the PHRAP ACE assembly file and extracts information concerning the relationships between contigs. A series of PHD format files are produced for use with the UniSelect program. The PHD files are organised into directories representing 96 well trays. Each tray directory has two subdirectories "fwd" and "rev" within which the forward and reverse walks relative to a particular insert are written. The names of these PHD files are produced by the concatenation of the new location and the old template name separated by a hyphen (Example 1).

Example 1. UniFinish file format

```
u01a01-001c5.b.phd
```

u01a01 is the new tray/well location and

001c5 is the original tray/well location

UniFinish also produces a text file (`templates.txt`) of all templates selected. Each line contains the new re-arrayed location, the template name, and the two contigs that the insert is believed to connect. It is expected that a user will re-array the selected templates into new trays corresponding to the location given in this file (example 2).

Example 2. template.txt file

```
u01a01    001c5     Contig641 Contig714
u01a02    001d8     Contig728 Contig778
u01a03    001e3     Contig700 Contig643
```

If the "gap closure" mode is selected there will be a forward and a reverse PHD file for each template, however, "low quality regions" mode may not always generate two files. UniFinish attempts to select an insert that spans the low quality region and walk from both directions. If there is no spanning insert, UniFinish will attempt to select an insert that will likely cover the region from one direction. This selection process depends on the user defined insert length.

During the run UniFinish performs a number of statistical calculations and outputs the result to the shell (example 3). The calculated expected insert and read length are not used in subsequent analysis, but these values can be used on subsequent UniFinish runs. Invalid inserts are those that have both the forward and reverse reads aligning in the same direction on the same contig. Paired inserts are those that have both forward and reverse reads, unpaired have only one read. Internal inserts are those existing entirely within one contig, and gap spanning as the name suggests are those inserts whose ends belong to different contigs.

Example 3.

```
Statistics:
===========
   14222 valid reads in assembly
   12 invalid reads in assembly (0.08%)
   8634 valid inserts in assembly
```

Command line switches and parameters

UniFinish has the following command line parameters and options:

1.1. Help `-h`

Lists the help for UniFinish

1.2. version `-V`

Lists the version of UniFinish

1.3. input file `-if [string]`

Specify the input assembly file for processing. This must be a PHRAP ACE assembly file.

Example 4. ACE input file

```
-if my_assembly.ace
```

1.4. output directory `-od [string]`

Specify the output directory. By default UniFinish will write its output to the current working directory. A user can change this to any alternative directory using this command line parameter.

1.5. Gap closure `-lq`

UniFinish by default will process an assembly file to close gaps. Setting this switch, UniFinish will instead search the assembly file for internal low quality regions within contigs.

1.6. Reject templates `--rejects [string]`

A user can exclude templates from consideration as targets to use for walks. The rejects command line parameter specifies a text file that contains the list of templates to exclude, one template per line.

Example 5. Template exclusion file

```
001a10

012c05

034d12
```

1.7. Tray naming --start-tray (integer >= 1)

UniFinish produces a number of PHD files for selecting UniSeq primers. The PHD files are stored in directories that symbolically represent trays. By default these symbolic trays and their correspondingly named PHD files will begin at u01. If a user so chooses they can begin numbering trays with another integer above 1.

1.8. Contig read number `--contig-read [integer >=1]`

This parameter can be used to restrict processing to only those contigs with formed from a minimum number of reads. The allows single read contigs to be excluded from the analysis.

1.9. Consensus quality `--quality-limit [integer >= 0]`

This parameter allows the threshold quality score below which consensus sequence is considered to be unreliable. This is used to determine sequence on contigs ends.

1.10. Trace read length `--read-length [integer >= 1]`

This parameter allows the expected read length to be set. This is important for re-sequencing internal low-quality regions.

1.11. Insert length `--insert-length [integer >= 1]`

This parameter allows the expect insert length to be set. This is important for re-sequencing internal low-quality regions.

1.12. Low quality `--low-quality [integer >= 0]`

When processing low-quality regions, the threshold quality score below which sequence is considered unacceptable can be set. This threshold is then used to determine regions of low quality suitable for primer walks.

Setting this number higher results in a greater amount of sequence being considered unacceptable and consequently produces a greater number of walk jobs. Lower numbers will result in less sequence being considered unacceptable and will result in fewer walk jobs.

1.13. Minimum low quality length `--min-length [integer >= 1]`

This parameter sets the minimum length of a low-quality region to be considered as requiring a walk. The can be used to avoid unnecessary walking reactions.