Terminator Module

Knut Reinert

1. Overview

The terminator module is at the end of the assembler pipeline. It gathers and transforms data to deliver to the post-assembly team. There are a number of new messages created by the assembler. However, external accession numbers (UIDs) need to be provided for objects computed during a run of the assembler. This numbers are taken from Celeras UID server (see /cm/cvs/SYS/UID/doc). Since an UID assigned by this server cannot be given back, the terminator by default assigns a dummy UID for testing purposes.
If real UIDs are needed this must be triggered with the –u flag. The dummy UIDs start with number 4711 and are only unique during one run of the terminator.

2. Interface

Command line interface for AS_TER module:

```
terminator [-Pu] <filename prefix>

      -P      outputs ProtoIO format instead of binary.
      -u      queries the UID server
```

3. Design

The TERMINATOR module reads several messages from the assembler pipeline in either binary or ASCII format. It then modifies some of these messages or generates new messages using the information in the input messages. Finally it outputs a number of messages that form a *genome snapshot* (see also Chapter 8 ..\BigPicture\ProtoSpec.rtf - Genome_Snapshot).
Information is taken out of three locations:

- the frag store. It contains the IID to UID mapping as well as the clear range and a list of screen matches for each fragment.
- The dist store. It contains the IID to UID mapping for all distance messages.
- the .cns file generated by the final call to consensus.

The terminator first reads the two stores and records all IID to UID mappings in variable arrays (VAs). Then it opens the .cns file and streams in the messages in it. It records the IID to UID mapping of all ISN (Internal ScreeN) messages, and for each internal message type that should be converted into a genome snapshot message type it converts the internal into the external message type. Doing so it has to spot all IID occurences in the specific type and either assign them their known UIDs, or it has to assign to a message a totally new UID. A new UID is also recorded for later reference. The terminator also checks, whether IIDs occur twice, or whether the UID of a message is needed that has not yet occurred on the stream (reference before definition). In both cases it exists with an error message.

In the following we shortly describe the messages and what is changed in them.

1. **Augmented Fragment Messages (AFG)** The augmented fragment messages need to lookup their UIDs in a variable array. Then they have to look up their associated internal screen match list and replace in each internal screen match the IID by the corresponding UID (obtained by looking up the VA containing the IID to UID mapping for the internal screen items).
2. **Mate Distance Messages (MDI)** Here only the IID has to be replaced by the UID.
3. **Unitig Messages (UTG)** The unitig messages have to lookup their UID in a variable array. In addition the internal fragment Ids in the internal multipos messages need to be replaced by the corresponding UIDs.
4. **Unitig Link Messages (ULK)** The unitig link messages have to replace two internal unitig IDs by the corresponding UIDs. Also the internal fragment IDs in their jump lists need to be replaced by UIDS.
5. **ConConMessages (CCO)** The contig messages have to lookup their UID in a variable array. In addition the internal contig IDs in the internal multipos and elementpos messages need to be replaced by the corresponding UIDs.
6. **Contig Link Messages (CLK)** The contig link messages have to replace two internal contig messages by the corresponding UIDs. Also the internal fragment IDs in their jump lists need to be replaced by UIDS.
7. **Scaffold Messages (SFG)** Each internal scaffold message has its internal ID replaced by a new UID. Also the list of ContigPairs has to be traversed and the internal contig Ids are replaced.

In addition to changing all IIDs to UIDs the terminator also outputs this mapping to a file with the extension .map for inspection reasons.