Omnigraph format specificaion, version 0.1

Mikhail Dvorkin, Max Alekseyev

May 27, 2011

Omnigraph format (OMG) serves for storing graphs of certain type that arise in the genome assembly software.

The considered graphs basically store lengths of edges (and possibly vertices), and the information on distances between some pairs of edges. (In fact, they can store any other information, but it would be supplementary to OMG format.) DNA sequences (reads) correspond to paths in such graphs.

Omnigraph format is a subset of Sequence Graph format (SQG).

A reference C++ implementation is coming soon. An interface will be provided to create, input and output Omnigraphs.

1 Example Omnigraph

Consider the following sample genome, with a corresponding (say, A Bruijn) graph that has five vertices a, b, c, d and e, and three edges (a, b), (b, c) and (d, e); also the distance between (b, c) and (d, e) is known to be 100 ± 5. Note that c and d come from the other DNA strand.

```
GATTACATCTACATTG.............TAGGATAG
-a->      -b->     100 ± 5         -e->
            <-c-              <-d-
```

Here's a OMG file (proposal) that represents this graph.

```
SQG 0 1
HISTORY 2011/04/18_18:05:00 made by hand
PROPERTY Vertex flag Segment corresponds to a vertex in original graph
PROPERTY Edge flag Segment corresponds to an edge in original graph
```

```
PROPERTY Incidence flag Connection between an edge and its end vertex
PROPERTY Error int Acceptable absolute difference from actual distance
SEGMENT a 4 Vertex
SEGMENT b 4 Vertex
SEGMENT c 4 Vertex
SEGMENT d 4 Vertex
SEGMENT e 4 Vertex
SEGMENT a_b 14 Edge
SEGMENT b_c 6 Edge
SEGMENT d_e 8 Edge
CONNECTION a >> a_b -4 Incidence
CONNECTION a_b >> b -4 Incidence
CONNECTION b >> b_c -4 Incidence
CONNECTION b_c >< c -4 Incidence
CONNECTION d <> d_e -4 Incidence
CONNECTION d_e >> e -4 Incidence
CONNECTION b_c >> d_e 100 Error=5
```

Records corresponding to vertices and edges of the graph may optionally specify the nucleotide sequence.

2 Discussion

- Each original vertex is represented as 1 segment of type "Vertex".
- Each edge is represented as 1 segment of type "Edge" and 2 connections of type "Incidence".
- Each distance-info is represented as 1 connection.

Therefore the OMG file size is linear of the size of original graph.

2.1 Support for Reverse-complementaries

Consider two vertices corresponding a and a^{RC}. If the original graph has the knowledge of their reverse-complementarness, it would be a good practice to put only 1 of them into the OMG file.

Note that the SQG file provides enough support for storing all the information about vertex a^{RC} in the case only a is contained as a segment.

3 References

AMOS seem to have their own file format. To do: comparison.
http://sourceforge.net/apps/mediawiki/amos/index.php?title=AMOS