INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 MPEG2017/M42063
January 2018, Gwangju, KR

Source: **DMAG-UPC**
Status: **Proposal**
Title: **Minor improvements to MPEG-G Part 2**
Authors: **Daniel Naro, Jaime Delgado, Omair Iqbal (Distributed Multimedia Applications Group – Universitat Politècnica de Catalunya)**

Table of Contents

1 Improvement to the definition of `rcomp` and `pair` genomic descriptors

As explained in document [1], the wording used to describe the descriptors `rcomp` and `pair` could be misleading.

1.1 Problem

In Part 2 of the standard [2], the description of both `rcomp` and `pair` streams state that both streams encode the strandedness. However, `pair` encodes only information on the distance between the first and second mate, and which segment is the first or second mate.

Additionally, in clause 11.2 *Paired reads* of [2], the text needs to be modified, at least for Illumina's devices, as it links read 1 to the forward strand and read 2 to the reverse strand. The current text is:

"The first read (or read 1) in a pair is the one sequenced from the forward strand of the sequenced DNA, while the second read (or read 2) is the one sequenced from the reverse strand."

In order to verify the correctness of our proposal, we have used the low coverage BAM file listed as input 05 in [3], to provide some statistics concerning this point. We have filtered the data to consider only the data perfectly aligned (i.e. we have discarded reads with unknown bases, or whose alignment requires a mutation, insertion, deletion or clipping operation), so a spurious alignment is therefore unlikely. Iterating over the file, and for each pair of reads, we have updated a ledger to reflect to which of the four cases the pair belongs:

	Read 1 on	Read 2 on
Case 0	Reverse strand	Reverse strand
Case 1	Reverse strand	Forward strand
Case 2	Forward strand	Reverse strand
Case 3	Forward strand	Forward strand

In our table, read 1 and read 2 are not determined by the order of appearance in the file, but by the information provided by the aligner (i.e., read 1 is the one with the SAM flag 0x40 set to true, and read 2 the one with the flag 0x80).

In the case of the selected input file, case 1 and case 2 represent each of them almost 50% of the cases. There are 7296291 instances of case 1 (where read 1 is on the reverse strand), and 7296783 instances of case 2 (where read 2 is on the reverse). As these are perfect alignments, it cannot be attributed to only errors in the alignment, thus proving that both read 1 and read 2 might be either on the forward or reverse strand.

Case 0 and 4 represent together only 0.4% of the reads. As the number is quite negligible, it might be within the error margin of the alignment (although these are perfect alignments). Thus, this does not provide evidence against the fact that both segments are mapped from a different strand.

1.2 Proposed solutions

Just by removing the last sentence of 11.4.9:

"This implies that the information about reads *strandedness* is encoded in the sign of the pairing distance descriptor."

the problem concerning streams descriptions should be solved.

In addition, to solve the problem of read 1 and read 2 definitions, the following sentence of clause 11.2:

"The first read (or read 1) in a pair is the one sequenced from the forward strand of the sequenced DNA, while the second read (or read 2) is the one sequenced from the reverse strand."

may be reformulated to something such as:

"The two reads are not sequenced from the same strand, but might be aligned to the same strand. The sequencing hardware determines which read in the pair is marked as read 1, while the other one will be read 2 ."

2 Improvement to the definition of Extended Cigar

2.1 Problem

In the definition of the Extended Cigar in clause 5.15, the text says that the E-Cigar operation C is similar but not equivalent to X. This is a typo, the not equivalent statement is meant for the M operation as in the previous row.

2.2 Proposed solutions

Concerning the definition of E-Cigar, it would just be required to move the text in parenthesis to the end of the sentence concerning the M operation in Cigar.

3 References

[1] Daniel Naro, Jaime Delgado, Silvia Llorente, "M41070 Comments and issues in current MPEG-G Working Drafts (Parts 1 and 2)" Torino, July 2017.

[2] ISO/IEC JTC 1/SC 29/WG 11, N17140 "Text of ISO/IEC 23092-2 CD, Coding of Genomic Information" Macao, October 2017.

[3] ISO/IEC JTC 1/SC 29/WG 11, N16145 "Database for Evaluation of Genomic Information Compression and Storage" San Diego, February 2016.