Comparison of VDJ Alignment Representations: BTOP vs CIGAR

Nishanth M

24th October, 2016

Introduction to BTOP & CIGAR formats

CIGAR (from SAMv1 document)

CIGAR string 30M1I69M means 30 bases aligning to the reference (30M), 1 base insert (1I), and 69 bases aligning (69M).

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

- H can only be present as the first and/or last operation.
- S may only have H operations between them and the ends of the CIGAR string.
- For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

CIGAR dependency on another entry in the SAM formatted data :

POS: 1-based leftmost mapping POSition of the first matching base. The first base in a reference sequence has coordinate 1. POS is set as 0 for an unmapped read without coordinate. If POS is 0, no assumptions can be made about RNAME and CIGAR.

Blast Trace-back OPerations (BTOP)

- BTOP, in comparison to CIGAR, lists not only the aligned region but also matches and mismatches.
- BTOP operations consist of
  1. a number with a count of matching letters,
  2. two letters showing a mismatch (e.g., “AG” means A was replaced by G),
  3. a dash (“-“) and a letter showing a gap. *Note also that BTOP always shows the query on the plus strand, whereas the CIGAR string always has the subject on the plus strand.*

Examples:

7AG39

7A-39

6-G-A41

BTOP equivalent

7AG39

7A-39

6-G-A41

```
Query= query1
Length=47

Subject=
Length=142


 Score = 82.4 bits (44),  Expect = 9e-22
 Identities = 46/47 (97%), Gaps = 0/47 (0%)
 Strand=Plus/Plus


Query  1   ACGTCCGAGACGCGAGCAGCGAGCAGCAGAGCGACGAGCAGCGACGA  47
           ||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  47  ACGTCCGGGACGCGAGCAGCGAGCAGCAGAGCGACGAGCAGCGACGA  93



 Score = 80.5 bits (43),  Expect = 3e-21
 Identities = 46/47 (97%), Gaps = 1/47 (2%)
 Strand=Plus/Plus


Query  1   ACGTCCGAGACGCGAGCAGCGAGCAGCAGAGCGACGAGCAGCGACGA  47
           ||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  1   ACGTCCG-GACGCGAGCAGCGAGCAGCAGAGCGACGAGCAGCGACGA  46



 Score = 78.7 bits (42),  Expect = 1e-20
 Identities = 47/49 (95%), Gaps = 2/49 (4%)
 Strand=Plus/Plus


Query  1    ACGTCC--GAGACGCGAGCAGCGAGCAGCAGAGCGACGAGCAGCGACGA  47
            ||||||  ||||||||||||||||||||||||||||||||||||||||||
Sbjct  94   ACGTCCGAGAGACGCGAGCAGCGAGCAGCAGAGCGACGAGCAGCGACGA  142
```

An Example (sort of a dry run): Representing annotations of an IGH sequence using BTOP & CIGAR formats

IGH sequence queried using IgBlast

```
>IGH sequence with the primer sequence
1    aagcagtggtatcaacgcaaagtacatgggatggaatggctgtggaacatgctagttcncatggcagaagcacaaaatatgcaatcaCA  89
90   GATCCAGTTGGTGCAGTCTGGACCTCAGCTGAAGAAGCCTGGAGAGACAGTCAAGATCTCCTGCAAGGCTACTGGTTATACCTTCACAA  179
180  CTGCTGGAATGCAGTGCGTGCAAAAGATGCCAGGAAAGGGTTTTAAGTGGATTGGCTGGATAGACACCCACTCTGGAGAGTCTAAATAT  269
270  GCAGAAGACTTCAAGGGACGGTTTGCCTTCTCTTTGGAAACCTCTGCCAACACTGCCTTTTTACAGATAACCAACCTCAAAAATGAGGA  359
360  CACGGCTACGTATTTCTGTGCAACTAGGGGGAATGCTATGGACTACTGGGGTCAAGGAACCTCAGTCACCGTCTCCTCAGCCAAAACAA  449
450  CACCCCCATCA                                                                                460
```

Lower case	– Primer and leader sequence
Red font	– Poor quality bases at 5’ end of V-gene
Green font	– Constant region

IgBlast result showing the top most hits for V, D and J genes
(Alignments in CIGAR format is given from the perspective of the full IGH query sequence/read)

```
Alignments

                                        <-----------------------FR1-IMGT-----------------------><-------CDR1-IMGT------><---------
                                           G  P  Q  L  K  K  P  G  E  T  V  K  I  S  C  K  A  T  G  Y  T  F  T  T  A  G  M  Q  C
                    Query_1          107   CTGGACCTCAGCTGAAGAAGCCTGGAGAGACAGTCAAGATCTCCTGCAAGGCTACTGGTTATACCTTCACAACTGCTGGAATGCAGTGCG  196
V  96.3% (261/271)  IGHV9-4*01        20   ........G.............................................T....G.............................G.  109

                                        -----------FR2-IMGT--------------------><-------CDR2-IMGT------><-----------------------
                                        V  Q  K  M  P  G  K  G  F  K  W  I  G  W  I  D  T  H  S  G  E  S  K  Y  A  E  D  F  K  G
                    Query_1          197   TGCAAAAGATGCCAGGAAAGGGTTTTAAGTGGATTGGCTGGATAGACACCCACTCTGGAGAGTCTAAATATGCAGAAGACTTCAAGGGAC  286
V  96.3% (261/271)  IGHV9-4*01        110   ............................................A.................C.A.........................  199

                                        ---------------------------FR3-IMGT-------------------------------------------------------><
                                        R  F  A  F  S  L  E  T  S  A  N  T  A  F  L  Q  I  T  N  L  K  N  E  D  T  A  T  Y  F  C
                    Query_1          287   GGTTTGCCTTCTCTTTGGAAACCTCTGCCAACACTGCCTTTTTACAGATAACCAACCTCAAAAATGAGGACACGGCTACGTATTTCTGTG  376
V  96.3% (261/271)  IGHV9-4*01        200   ..............................G........A...........G.....................................  289

                                        --------CDR3-IMGT-------->
                                        A  T  R  G  N  A  M  D  Y  W  G  Q  G  T  S  V  T  V  S  S
                    Query_1          377   CAACTAGGGGGAATGCTATGGACTACTGGGGTCAAGGAACCTCAGTCACCGTCTCCTCAG  436
V  96.3% (261/271)  IGHV9-4*01        290   .------------------------------------------------------------  290
D  100.0% (6/6)     IGHD3-3*01        8     ---......-----------------------------------------------------   13
J  100.0% (47/47)   IGHJ4*01          7     -----------.............................................   54
```

BTOP (from IgBlast)

V	8CG44AT4TG29CG45GA17TC1TA55AG8TA11CG39
D	6
J	48

CIGAR string from the perspective of the <u>query</u> IGH sequence

V	106S8=1X44=1X4=1X29=1X45=1X17=1X1=1X55=1X8=1X11=1X39=59N
D	273N6=51N
J	282N48=

Full germline sequence for the top hit genes and the sequences removed at junctions are shown (Alignments in CIGAR format is given from the perspective of each germline gene sequence)

```
Alignments

                                        <-----------------------FR1-IMGT-----------------------><-------CDR1-IMGT------><---------
                                          G  P  Q  L  K  K  P  G  E  T  V  K  I  S  C  K  A  T  G  Y  T  F  T  T  A  G  M  Q  C
                    Query_1        107  CTGGACCTCAGCTGAAGAAGCCTGGAGAGACAGTCAAGATCTCCTGCAAGGCTACTGGTTATACCTTCACAACTGCTGGAATGCAGTGCG  196
V  96.3% (261/271)  IGHV9-4*01      20  ........G.........................................T....G.........................G.  109

                                        -----------FR2-IMGT--------------------><-------CDR2-IMGT------><-----------------------
                                        V  Q  K  M  P  G  K  G  F  K  W  I  G  W  I  D  T  H  S  G  E  S  K  Y  A  E  D  F  K  G
                    Query_1        197  TGCAAAAGATGCCAGGAAAGGGTTTTAAGTGGATTGGCTGGATAGACACCCACTCTGGAGAGTCTAAATATGCAGAAGACTTCAAGGGAC  286
V  96.3% (261/271)  IGHV9-4*01     110  ..........................................A.................C.A.......................  199

                                        --------------------------FR3-IMGT-------------------------------------------------><
                                        R  F  A  F  S  L  E  T  S  A  N  T  A  F  L  Q  I  T  N  L  K  N  E  D  T  A  T  Y  F  C
                    Query_1        287  GGTTTGCCTTCTCTTTGGAAACCTCTGCCAACACTGCCTTTTTACAGATAACCAACCTCAAAAATGAGGACACGGCTACGTATTTCTGTG  376
V  96.3% (261/271)  IGHV9-4*01     200  ..............................G........A...........G...................................  289

                                        --------CDR3-IMGT-------->
                                        A  T  R  G  N  A  M  D  Y  W  G  Q  G  T  S  V  T  V  S  S
                    Query_1        377  CAACTAGGGGGAATGCTATGGACTACTGGGGTCAAGGAACCTCAGTCACCGTCTCCTCAG  436
V  96.3% (261/271)  IGHV9-4*01     290  .----------------------------------------------------------  290
D  100.0% (6/6)     IGHD3-3*01      8   ---......-------------------------------------------------   13
J  100.0% (47/47)   IGHJ4*01        7   ----------------.........................................   54
```

lower case	– Sequences trimmed due to low quality
Red font	– Sequences removed during imprecise joining at junctions

```
>AC073590|IGHV9-4*01|Mus musculus_C57BL/6|F|V-REGION|65900..66193|294 nt|1| | || |294+0=294| | |
cagatccagttggtgcagtCTGGACCTGAGCTGAAGAAGCCTGGAGAGACAGTCAAGATCTCCTGCAAGGCTTCTGGGTATACCTTCACAACTGCTGG
AATGCAGTGGGTGCAAAAGATGCCAGGAAAGGGTTTTAAGTGGATTGGCTGGATAAACACCCACTCTGGAGAGCCAAAATATGCAGAAGACTTCAAGG
GACGGTTTGCCTTCTCTTTGGAAACCTCTGCCAGCACTGCCTATTTACAGATAAGCAACCTCAAAAATGAGGACACGGCTACGTATTTCTGTGCGAGA
```

```
>AJ851868|IGHD3-3*01|Mus musculus_129/Sv|F|D-REGION| |17 nt|1| | | | |17+0=17| | |
GGGACAGCTAGGGCTGT
```

```
>V00770|IGHJ4*01|Mus musculus|F|J-REGION|1333..1386|54 nt|3| | | | |54+0=54| | |
ATTACTATGCTATGGACTACTGGGGTCAAGGAACCTCAGTCACCGTCTCCTCAG
```

BTOP (from IgBlast)

V	8CG44AT4TG29CG45GA17TC1TA55AG8TA11CG39
D	6
J	48

CIGAR string from the perspective of the germline sequence

V	19N8=1X44=1X4=1X29=1X45=1X17=1X1=1X55=1X8=1X11=1X39=4D
D	7D6=4D
J	6D48=

Few observations and questions:

- The CIGAR string is slightly longer than the BTOP representation
- BTOP format allows the reconstruction of the germline sequence (at least the aligned part) from the BTOP string.
- CIGAR format would require the germline sequences used by the annotation tool to be provided separately
- CIGAR format allows for representation of the trimming from the original reads
- CIGAR format can be represented from the perspective of the query sequence or the individual germline sequence. From which perspective should the VDJ sequence alignment be represented?