JUST - EVOTEC BIOLOGICS, INC. WHITE PAPER March 2020

CoreAb Sequence Classification

J. Alex Taylor

Keywords: antibody; antibody numbering; structural numbering; antibody engineering; antibody analysis

INTRODUCTION

A methodology has been developed for the structural alignment and classification of full sequences from antibodies and antibody-like structures using the Antibody Structural Numbering system (ASN). For a description of the Antibody Structural Numbering system (ASN), please see the accompanying white paper. The classification process generates a series of ASN-aligned regions which can be used to uniquely describe residue locations in common across different molecules.

CLASSIFICATION PROCESS

The CoreAb Java library (developed at Just - Evotec Biologics) contains algorithms for the classification and alignment of antibodies and antibody-like sequences. A high-level summary of the classification process is presented in Figure 1. The first step in the classification process is the detection of antibody variable and constant regions specified in the detection settings. The default regions for detection are kappa variable, lambda variable, heavy variable, light constant, heavy constant Ig (CH1), heavy constant Fc-N (CH2), and heavy constant Fc-C (CH3). A Position-Specific Sequence Matrix (PSSM) has been pre-built for each type and is used as a low threshold first pass filter for region detection using the Smith-Waterman algorithm to find local alignments. Each local alignment is then refined by a more careful alignment comparison to the germline gene segments from species specified in the detection settings. If germline data for the query's species of origin does not exist or is incomplete in the resource files contained in CoreAb, other, more complete, germline gene data sets from other species can be used to identify homologous regions. The germline sequences are stored as ASN-aligned so that the resulting region alignments are also ASN-aligned.

To generate an alignment for variable regions, the PSSM-matched sub-sequence is aligned to both germline V-segments and J-segments and these results are combined to synthesize an alignment for the entire variable region. For heavy variable regions, the germine D-segments are aligned to the residues between the V-segment match and the J-segment match. As a final step in the variable region alignment refinement process, CDR regions are center-gapped to match the AHo/ASN numbering system.

Fig. 1
High-level antibody classification pseudocode

1. Identify antibody variable and constant regions (domains)
    a. Loop over the region types that were specified in settings
        i. Use a PSSM for the region type to find local alignments in the query
        ii. Loop over each local alignment from the query
            1. Loop over the germline sets that were specified in settings
                a. Generate a refined region alignment for the PSSM alignment
                    i. Only keep alignments that meet the minimum number of identities and region percent identity specified in settings
                    ii. Assign ASN numbering
                    iii. If variable region, refine the alignment and adjust CDR gapping
                    iv. If constant region and alignment is < 10 aa, toss it unless it is at the start of the region
2. Identify potential leader region matches (can use SeqParts)
3. Resolve overlapping regions giving priority to the higher scoring region
4. Assign gaps between identified regions (can use SeqParts)
5. Cleanup constant regions
6. Assign chain and structure format (based on the arrangement of regions)

Resulting germline-aligned regions are subjected to minimum percent identity thresholds which can be specified in the detection settings. The default threshold is 80% identity for constant regions and 60% identity for variable region frameworks. Constant region results of less than 10 residues are removed unless they occur at the start of a region. Regions that meet these thresholds are then compared to the other results for the same region and, if overlaps are found, the lower scoring region is removed.

Step 2 of the classification process is the detection of a leader sequence. If, after the variable and constant regions have been detected, there remains an N-terminal portion of the query sequence that is unmatched, the N-terminal portion is aligned to germline leaders from the specified germline gene sets and also to user-specified SeqPart sequences which have been provided to the detector. Resulting leader regions are subjected to a minimum percentage identity threshold which can be specified in the detection settings. The default threshold is 80% identity for leader regions. The highest scoring region result that meets this threshold is retained as the leader region.

In step 3, remaining regions are sorted by their score and then overlaps are resolved by giving preference to the higher scoring region except in cases where the overlapping residues are identities in the lower scoring region and are not identities in the higher scoring region. This step may result in the removal of the lower scoring region.

Step 4 assigns regions to any portions of the query which fall before, between, or after the remaining identified regions. If such regions fall after a constant Ig region or constant Fc-C region, germline hinge or post-constant regions from the germline gene matching the preceding region are respectively aligned to the query subsequence. If the resulting alignment percent identity meets the constant region threshold, the regions are added. Remaining unmatched portions of the query are then compared to

SeqParts if a SeqParts resource has been provided to the detector and resulting regions with a percent identity of greater than or equal to 80% are retained. Any portions of the query that still remain unassigned are assigned to unrecognized regions.

In step 5, the assigned constant region germline genes are harmonized if necessary. In many cases a region may have the same sequence for different alleles. In this step, the overall best scoring germline gene is determined and then any regions that are assigned to another germline gene are checked to determine if the overall best scoring germline has an equivalent score. If so, then the assignment for the region is changed to the overall best scoring germline.

The final step in the sequence classification process is to assign a chain format. If an AbFormatCache is provided to the detector, it is used to match the pattern of regions to a reference pattern associated with a particular chain format. Figure 2 shows a portion of the default AbFormatCache contained in the CoreAb library.

After all sequence chains have been classified they can be grouped into structures, often based on a common base name. An AbFormatCache can then be used to assign a structure format such as IgG1 Antibody or IgG1 Fc-Fusion to the structure by matching the chain formats present in the structure to structure format definitions that are made up of possible combinations of chain formats.

Fig. 2
Snippet of the default AbFormatCache from CoreAb. Three chain format definitions and three structure format definitions are shown. Regions in curly braces are optional.

```
<AbFormatCache version='2'>
   <ChainFormats>

    ...

    <ChainFormat id='55' name='DVD-IgG1 Heavy Chain' abbrev='DVD-IgG HC' description='DVD-IgG1 Heavy Chain'>
            {Ldr} ; HV ; Lnk ; HV ; IgG1:HCnst-Ig ; IgG1:Hinge ; IgG1:Fc-N ; IgG1:Fc-C ; IgG1:HCnst-Po
    </ChainFormat>
    <ChainFormat id='56' name='IgG1 Fc-Fusion' abbrev='IgG1 Fc-Fusion' description='IgG1 Fc-Fusion'>
            {Ldr} ; Unk ; {Lnk}; IgG1:Hinge ; IgG1:Fc-N ; IgG1:Fc-C ; IgG1:HCnst-Po
    </ChainFormat>
    <ChainFormat id='57' name='IgG1 Heavy Chain F(ab&apos;)2' abbrev='IgG1 HC F(ab&apos;)2' description='IgG1 Heavy Chain F(ab&apos;)2'>
            {Ldr} ; HV ; IgG1:HCnst-Ig ; IgG1:Hinge&lt;C113,C116>
    </ChainFormat>

    ...

   </ChainFormats>

   <StructureFormats>

      ...

      <StructureFormat id='28' name='DVD-IgG1' abbrev='DVD-IgG1' description='Dual Variable Domain (DVD) Bispecific IgG1 Antibody'>
          <ChainCombination>
              <ChainFormat id='53' stoichiometry='2' />
              <ChainFormat id='55' stoichiometry='2' />
          </ChainCombination>
          <ChainCombination>
              <ChainFormat id='54' stoichiometry='2' />
              <ChainFormat id='55' stoichiometry='2' />
          </ChainCombination>
      </StructureFormat>
```

```
      <StructureFormat id='29' name='IgG1 Fc-Fusion' abbrev='IgG1 Fc-Fusion' description='IgG1 Fc-Fusion'>
          <ChainCombination>
              <ChainFormat id='56' stoichiometry='2' />
          </ChainCombination>
      </StructureFormat>
      <StructureFormat id='30' name='IgG1 F(ab&apos;)2' abbrev='IgG1 F(ab&apos;)2' description='IgG1 F(ab&apos;)2 Fragment'>
          <ChainCombination>
              <ChainFormat id='5' stoichiometry='2' />
              <ChainFormat id='57' stoichiometry='2' />
          </ChainCombination>
          <ChainCombination>
              <ChainFormat id='8' stoichiometry='2' />
              <ChainFormat id='57' stoichiometry='2' />
          </ChainCombination>
      </StructureFormat>

      ...

   </StructureFormats>

</AbFormatsCache>
```

REFERENCE GERMLINE DATA EXTRACTION PROCESS

The extraction and compilation of antibody germline gene data can be a difficult and time consuming process. In cases where gene annotation is provided by the NCBI, a CoreAb tool is used to extract and align the gene information. Incomplete or unannotated genomes require a more *de novo* approach. CoreAb also contains a tool that can scan for potential V-segments, J-segments, and D-segments using PSSMs designed to locate the Recombination Signal Sequence (RSS) sequences used to join the variable region segments. Manual curation is still required to filter and adjust the results but this automation can alleviate most of the tedious work. When possible, names for extracted genes are set to those from IMGT since that is the source of official naming. Figure 3 displays a section of an XML-formatted germline data resource file. Default XML-formatted germline data is included in CoreAb and loaded at runtime. Additional or alternate germline data can be provided by the user. Full or partial antibody gene data is currently included in CoreAb for the following organisms: *Bos taurus*, *Camelus bactrianus*, *Camelus dromedarius*, *Canis familiaris*, *Cavia porcellus*, *Gallus gallus*, *Homo sapiens*, *Macaca mulatta*, *Mus musculus*, *Oryctolagus cuniculus*, *Ovis aries*, *Protopterus dolloi*, *Rattus norvegicus*, *Struthio camelus*, and *Vicugna pacos*.

Fig. 3
Snippet of the Bos taurus HV.xml germline data file of heavy variable genes extracted from genomic sequences.

```
<?xml version='1.0' encoding='UTF-8' standalone='yes' ?>
<GermlineGeneSetRsrc igGeneGroup='IGHV' taxonId='9913'>

...

  <GermlineGene id='IGHV1-20*01'>
    <Note>Num RSS mismatches: 0</Note>
    <GenomicLoc build='ARS-UCD1.2' chromosome='21' contig='NW_020190105.1' contigLength='69862954'
strand='Forward'>join(278220..278274,278353..278658)</GenomicLoc>
    <DB_Xref db='Just-TempId'>IGHV1-4S*01</DB_Xref>
    <DB_Xref db='IMGT/GENE-DB'>IGHV1-20*01</DB_Xref>
    <Intron donorSite='a gtgtc' acceptorSite='acag g' seqLength='78'>
      <GenomicLoc build='ARS-UCD1.2' chromosome='21' contig='NW_020190105.1' contigLength='69862954'
strand='Forward'>278275..278352</GenomicLoc>
      <DNA>
         gtgtctctgggtcagacatgggcacgtggggaagctgcctctgagcccacgggtcaccgtgcttctctctctccacag
```

```
      </DNA>
    </Intron>
    <AlignedRegions>
      <AlignedRegion region='HLdr'>
        <Protein>
          MNPLWEPPLVLSSPQSGVRVLS
        </Protein>
        <DNA>
          atgaacccactgtgggaacctcctcttgtgctctcaagcccccagagcggagtgagggtcctgtcc
        </DNA>
      </AlignedRegion>
      <AlignedRegion region='HV'>
        <Protein>

QVQLRES-GPSLVKPSQTLSLTCTVSG-FSLSS-----YAVGWVRQAPGKALEWLGGISS----GGSTYYNPALKSRLSITKDNSKSQVSLSVSSVTPEDTATYYCAK--------------------------------
----------
        </Protein>
        <DNA RSS_3prime='cacagtg aggggaaatcagtgtgagcccag acaaaaacc' splitCodon3prime='ga'>
          caggtgcagctgcgggagtcgggccccagcctggtgaagccctcacagaccctctccctcacctgcacggtctctggattctcactgagcagctatgctg
          taggctgggtccgccaggctccagggaaggcgctggagtggctcggtggtataagcagtggtggaagcacatactataacccagccctgaaatcccggct
          cagcatcaccaaggacaactccaagagccaagtctctctgtcagtgagcagcgtgacacctgaggacacggccacatactactgtgcgaagga
        </DNA>
      </AlignedRegion>
    </AlignedRegions>
  </GermlineGene>

...

</GermlineGeneSetRsrc>
```