**IMGT Rules**

The best way to get the most accurate linguistic annotation is to use **ICK-Data** (based on Kabat R, Chothia R and IMGT R) so check and study ***“ICK-Data about.doc”*** and ***“ICK-Data.xls”***.

FR1 – **ends *nearly* after Cys23**, length = 25-26, 1 gap possible (at 10), (1-26)

CDR1 – **begins *nearly* after Cys23 and ends *nearly* before Trp41**, length <= 12, (27-38)

FR2 – **begins *nearly* before Trp41**, length = 17 + 2, (39-55) ///// 46A, B

CDR2 – length <= 10, (56-65)

FR3 – **ends with Cys104**, 89 – conserved hydrophobic, length = 36-39 + 1, 3 gaps possible (at 73, 81 and 82), (66-104) ///// 84A

CDR3 – **begins after Cys104 and is followed by Phe/Trp 118**, length >= / <= 13, (105-117)

FR4 – **begins with Phe/Trp 118**, F/W-G-X-G at positions 118–121, length = 10-11 (gap at the last position 128), (118-128)

***Gaps and Additional Positions:*** For CDR loops shorter than 12, 10, 13 amino acids, respectively, gaps are created at the apex. The gaps are placed at the apex of the loop with an equal number of amino acids on both sides if the loop length is an even number, or with one more amino acid in the left part if it is an odd number.

For example, for FG loops <13 amino acids, gaps are created from the apex of the loop, in the following order: 111, 112, 110, 113, 109, etc. For FG loops >13 amino acids, additional positions are created, between positions 111 and 112 at the top of the FG loop, in the following order: 112.1, 111.1, 112.2, 111.2, 112.3, etc.

***Numbering:*** 1 … 46 46A 46B 47 … 84 84A 85 … 111 <CDR3 additional positions> 112 … 128

(… - numbering equals to the sequence of natural numbers)

(add-positions marked with grey color – probably, aren’t present in the V-domain (95% sure, but be careful))

<CDR3 additional positions>:

Even number of add-positions (=2\*n): 111.1 111.2 … 111.n 112.n 112.(n-1) … 112.1

Odd number of add-positions (= 2\*n-1): 111.1 111.2 … 111.(n-1) 112.n 112.(n-1) … 112.1

**Afterword:** region – **information about (mostly) invariable key residues that helps to find the desired region (if they exist),** extra information that may help too (if exists)**,** length of the region (that help to move from the key residues and find the regions)**,** **(x-y)** – region’s boundaries corresponding to the IMGT numbering (can help if the numbering already exists or to check the already annotated sequence) **/////** addition positions in the numbering, that are located in this region (can help with making the correct numbering)

*<< 29-32, IMGT-Booklet >> describing concept, definitions; some of key residues*

*<< 46-47, IMGT-Booklet >> regions’ boundaries, lengths, conserved (invariant) residues (key positions), gaps and additional positions*

**Extra R (IMGT\*)**

**invariant residues: 23 C, 41 W, 44 Q, 98 D, 102 Y, 104 C, 119 G, 121 G, 122 T**

**closely related residues: 4 LM, 6 QE, 16 GS, 21 VLIM, 22 ST, 53 VLIM, 75 RK, 77 ST, 79 ST, 89 LIMF, 91 LIM, 94 VLM, 100 GA, 103 YF, 118 FW, 124 VL, 126 VLI**

***bn* sites (C, V ,L, I, M, F, W + P, H ,Y ,G, A, S, T) = 11 (+RKE), 12, 13 (+RKE), 15 (P), 19 (VLIM), 25, 39, 42, 50, 52, 54 (I)/(GAS), 76 (*b*), 78 (G)/(M,I), 87 (A if (lambda), F if (k))/, 99 (+E), 101 (+D), 117**

***n* sites (P, H ,Y ,G, A, S, T) = 46 (P,S), 88 (S,T,A)/(Y,F)**

***sn* sites (R, K, E, D, Q, N + P, H ,Y ,G, A, S, T) = 5, 14, 17, 18, 20, 24, 40, 45, 47 (G), 48, 49 (*n*), 74, 80, 80+1 (S)/(D), 80+2 , 80+3, 86, 90, 92, 93, 95, 96, 97, 105, 106, 120, 123, 125, 127**

***s* sites (R, K, E, D, Q, N) = 43 (RKQ), 51**

**(s)-(n)-(b) groups:**

**(s) surface group: R, K, E, D, Q, N (high probability to be on the surface)**

**(n) neutral group: P, H ,Y ,G, A, S, T (equal probabilities)**

**(b) buried group: C, V ,L, I, M, F, W (high probability to be buried)**

**Afterword: m – residue’s number** in the IMGT numbering ///// y(X) means that X (or X-group) is present at y residue rather than other possible // (Y)/(Z) – (Y) for the light chain, and (Z) for the heavy chain

**\*numbering considered in this article is different from the IMGT one, so the obtained info was accurately translated into the IMGT annotation (that wasn’t really easy to do because of the IMGT gaps, but no matter what it was successfully done)**

*<< 5, 1998All >> s, n, b groups of amino acids*

*<< 11, 1998All >> invariant and closely related residues in VH, VL*

*<< 11, 1998All >> positions of bn, n, sn, s sites*

*<< 10, 12, 15, 17, 19, 21, 1998All >> more info about some positions, i.e.*

*Rather than amino acids of their groups:*

*10 - 40/41 P, 43/44 n, 62/67 b*

*12 - 67Light S, 72Light STA, 72Heavy D, 79Heavy YF*

*15 - 19/18 VLIM, 37/38 s (usually RKQ)*

*17 - 71/78 A in V(lambda) and FY in V(k)*

*19 - 15/14 P, 16/15 GS, 40/41 PS, 41/42 G*

*21 - 48Light I, 64Light G; 49Heavy GAS, 69Heavy MI*

***//add this info to other positions later << 6-9, 1998All >>***

**Extra RR (IMGT specified)\***

**L1** – **G**, DEQ, RKQTS, VA, TS, ILM, STN, **C (23)**,RKST, ASG, **S**T = … YNFAW = LVMIA, ANH, **W (41)**, **Y**FLV, QL, **Q**EH, **K**R, **P**SQ, **G**DH, QKGT

**L2** – **G**DH, QKGT, SAPT, **P**FY, KRQT, LRGTV, **L**WV, **I**VM, **Y**KG (55) = YKW, ATV … LRS, AF, STDP, **G (70)**, **V**I, **P (72)**,! DSAV (74), **R**, **F**, **S**T, **G**

**L3** – **E**D, **D**, LFEIAV, AG, TVDI, **Y**, YF, **C (104)** = QFLAS, QHL, YGSWH … **P**LH, YLPRWF, **T**V = **F (118)**, **G**, GQAST, **G**, **T**, **K**R, **L**V, **E**TD

**H1** – **G**S, GAQ, **S**T, LVM, KRS, LIMV, **S**T, **C (23)**, KATS, AVT, **S**T = … YWGA = MIWV, HNSG, **W (41)**, **V**IF, RK, **Q**K, ARPFST, **P**H, **G**E, KQNH

**H2** – **G**E, KQNH, GRKEA, **L**R, **E**K, **W**YG, IVML, GA, YW (55) = **I**V, SNYD … SNYDT, TIPSKA = YN, **Y**F (67), NASVG, DPEQA, SKADT, FVL, **K**QR (72),! GSD (74), RK (75)

**H3** – **E**DA, **D**, TS, **A**G, VTIML, **Y**, **Y**F, **C (104)** = ATV, RASN, GW … FMGLY, DAGV, YV = **W (118)**, **G**, **Q**AEKHP, **G**, **T**, TLSQ, **V**L, **T**

**Afterword: X** – CDR region and its neighborhood - each residue position is separated with other by “,” ///// for each position, the most common amino acids are listed, **but!** the variability of the amino acids on the different positions don’t correlate with each other

///// there are three types of marking here: a) **black and just bolded** – the possibility of this residue at this position is high b) **purple and bolded** - the possibility of this residue at this position is super high (mostly invariant) c) **purple, bolded and bigger** - the possibility of this residue at this position is also super high, but it’s almost noticed as invariant referring to the numberings (Kabat, Chothia and IMGT)

///// **the region marked with a red color corresponds to the CDR**

///// some positions are numbered (in the IMGT numbering**\***) to help with finding these residues in the sequences (other positions can be numbered just with using the sequence of natural numbers (but without crossing the ellipsis bound)**\*\***)

///// grey color means that the marked residue’s number is not surely correct (because of the Kabat-to-IMGT translation) *// but these numbers are probably correct, since they were tested on several examples*

**\*numbering considered in this article are different from IMGT one, so the obtained info was accurately translated into the IMGT annotation (that wasn’t really easy to do because of IMGT gaps, but no matter what it was successfully done)**

**\*\*the sequence of natural numbers interrupts between some positions (because of IMGT gaps (10, 73, 81, 82)), so be careful with them (this interrupts are marked with ! to help with this problem) // and also don’t forget about gaps in the CDRs – something can be changed in “the red-colored regions”**

**This information can be only an addition to the main rules!**

*<< 3, North >> Extra RR*

*<< 26-27, North >> Extra RR additions (i.e. more possible amino acids for some of considered positions)*

**Extra RRR (IMGT specified)\***

**L1** – length 5-12 (mostly 6 or 11)

**///// 24-40 ([24-26] FR1 + CDR1 + [39-40] FR2)**

L1-11-1 – R**ASQ**DISNYLA (76, k, HM) ///// F71

L1-16-1 – **RSS**QSLVHSN**G**N**TYL**E (68, k, HM)

L1-11-2 – R**AS**QD**I**SNY**L**N (55, k, M) ///// T/G 71

L1-17-1 – **KSSQSL**LN**S**RTRK**NYLA** (21, k, HM)

L1-10-1 – S**A**S**SSV**S**Y**MH (20, k, M)

L1-14-1 – S**A**S**SSV**S**Y**MH (14, lambda, H)

L1-15-1 – R**AS**E**SVD**SY**G**N**S**F**M**N (11, k, HM)

L1-13-1 – **SG**SS**SNIG**N**N**Y**V**S (7, lambda, H)

L1-12-1 – R**AS**S**S**V**SS**SYLH (5, k, M)

L1-12-2 – R**AS**Q**S**VSSNYL**A** (5, k, HM)

L1-11-3 – SGNNLGS-SVH (5, lambda, H)

L1-13-2 – **TRSSG**N**I**AS**NYV**Q (4, lambda, H)

L1-14-2 – S**A**S**SSV**S**Y**MH (4, lambda, M)

L1-10-2 – **SASSSVSY**MY (2, k, M)

L1-12-3 – **TLS**S**QHSTYTIE** (2, lambda, HM)

L1-15-2 – **RASKSVSTSGY**N**YMH** (2, k, M)

*<< 7, North >> L1’s clusters*

*<< 8, 12-13, North >> commentaries for some L1 clusters*

***//add more detailed commentaries later***

**L2** – length only 3 or 7

**///// 55-69 ([55] FR2 + CDR2 + [66-69] FR3)**

L2-8-1 (mostly this) – **Y**-ASNLAS (290, k, HM)

L2-8-2 – YAASNLDS (9, k, HM)

L2-8-3 – SEG**N**TLR**P** (3, k/lambda, M)

L2-8-4 – G**G**TN**NR**VP (2, k/lambda, M)

L2-8-5 – **Y**SA**S**Y**R**Y**S** (2, k, HM)

L2-12-2 – ELKKDGSHSTGD (2, lambda, M)

*<< 8, North >> L2’s clusters*

*<< 13, North >>* ***//add more detailed commentaries later***

**L3** – length 7-13 (mostly 9)

**///// 105-117 (CDR3)**

L3-9-cis7-1 (mostly this) – Q**Q**GSS-**P**L**T** (219, k, HM)

L3-9-1 – ALW-SNHWV (22, k/lambda, HM)

L3-8-1 – L**Q**YYNLR**T** (15, k, HM)

L3-9-2 – Q**Q**STH-PP**T** (12, k, HM)

L3-11-1 – AAWDSSLDAVV (9, lambda, H)

L3-9-cis7-2 – **QH**FWS**TP**R**T** (8, k, HM)

L3-10-1 – QSYDSS-SVV (6, lambda, H)

L3-8-2 – QQFWRTP**T** (4, k, M)

L3-8-cis6-1 – **Q**QWNY**P**F**T** (3, k, M)

L3-13-1 – AAW**D**DSRGGPDW**V** (3, lambda, HM)

L3-7-1 – Q**Q**YN**SY**S (2, k, HM)

L3-9-cis7-3 – Q**Q**YYIY**P**Y**T** (2, k, HM)

L3-10-cis8-1 – LYSREF**PP**W**T** (2, k, M)

L3-9-cis6-1 – **QQWTYPLIT** (1, k, M)

L3-10-cis7,8-1 – **SQSTHVPPLT** (1, k, M)

L3-11-cis7-1 – **QQYNNWPPRYT** (1, k, H)

L3-12-1 – **ATWDSGLSADWV** (1, lambda, H)

*<< 9, North >> L3’s clusters*

*<< 13-14, North >>* ***//add more detailed commentaries later***

**H1** – length 7-10 (mostly 8)

**///// 24-40 ([24-26] FR1 + CDR1 + [39-40] FR2)**

H1-13-1 (mostly this) – KA**SG**FTFTDYYMH (267, HM)

H1-14-1 – **TVTGYSIT**SG**Y**A**W**N (11, M)

H1-15-1 – SF**SGFS**LSTSGMG**V**G (9, HM)

H1-13-2 – KA**S**GFNITDYYIS (7, HM)

H1-13-3 – KA**SG**YT**F**TTYAMN (5, HM)

H1-13-4 – AVS**G**FSFSGYYWS (4, HM)

H1-13-5 – A**ASG**FTYSINYMG (4, HM)

H1-13-6 – **A**A**SG**YKYTNYCM**G** (4, C)

H1-13-7 – SVT**G**DSI**TS**GYWN (3, M)

H1-13-8 – KA**SG**YTFTTYDMG (3, M)

H1-13-9 – **A**A**SG**N**T**LSTYDMG (3, CL)

H1-13-10 – **KASGGTFS**M**Y**GFN (2, H)

H1-13-11 – K**AS**EY**T**LTSYLFQ (2, M)

H1-13-cis9-1 – **AASGYTIGPYCMG** (2, C)

H1-10-1 – A**AS**T**YT**DTV**G** (2, C)

H1-12-1 – **KLWYTFTDYGMN** (1, M)

H1-16-1 – **AASGRAASGHGHYGMG** (1, L)

*<< 10, North >> H1’s clusters*

*<< 14, North >>* ***//add more detailed commentaries later***

**H2** – length mostly 8 or 9 (but 7-14 are present)

**///// 56-66 (CDR2 + [66] FR3)**

H2-10-1 – -**I**YPGNG-T- (155, HM) ///// AVLISTQ 71 (mostly)

H2-9-1 – Y**I**WYS**G**STY (77, HM)

H2-10-2 – -**I**SSGGGNTY (42, HM) ///// **R**D 71 (mostly)

H2-12-1 – E**IR**N**K**ANNYT**T**E (26, M)

H2-10-3 – E**I**L**PG**SGSTN (11, HM)

H2-10-4 – T**I**SSG**G**GYTN (7, M)

H2-10-5 – **A**ISG**GG**TYIH (3, MC)

H2-10-6 – RIDPN**G**GG**TK** (3, HM)

H2-10-7 – **T**TLS**G**GGF**T**F (2, HM)

H2-10-8 – G**I**D**P**HN**GG**GA (2, HM)

H2-10-9 – G**I**DPHNGGPV (2, HM)

H2-8-1 – TILG**GS**TY (2, H)

H2-9-2 – S**I**YNGFRIH (2, M)

H2-9-3 – Y**I**RYG**G**GT**Y** (2, MC)

H2-15-1 – **TIGRNLVGPSDFYTR** (1, L)

*<< 11, North >> H2’s clusters*

*<< 14-16, North >> commentaries for some H2 clusters*

***//add more detailed commentaries later***

**H3** – length 5-26 (mostly 7-16)

**///// 105-117 (CDR3)**

H3-anchor-1 (mostly this) – A**R**- … YFDY (204) ///// bulged (e.g. mostly when K/R 94 and D101?)

H3-anchor-2 – ARY … DFD**Y** (35) /////(non-bulged)

H3-anchor-3 – ARG … YFDY (25)

H3-anchor-4 – ANW … DG**D**Y (24)

H3-anchor-5 – VR- … -RDY (12)

H3-anchor-6 – AS- … SFAY (6)

H3-anchor-7 – **AR**R … GFDY (4)

H3-anchor-cis4-1 – **AR**E … **P**F**D**Y (2)

(Anchor = first 3 and last 4 residues of H3)

*<< 12, North >> H3’s anchors*

*<< 16-17, North >> H3’s length and commentaries for some H3 anchors*

***//add more detailed commentaries later***

**Afterword:** for each CDR you can see several amino acids sequences that usually appear there (in the CDR), e.g. the CDR’s clusters

///// cluster’s name – XYZ…RG (amino acids sequence, CDRs are marked with a red color**\***; the residue written **in bold** means that within the framework of the cluster, it’s usually invariant(>90%) // then (x, k/lambda, HMCL): x – the number of proteins used in the work**\*** and having CDRs similar to the considered cluster (it’s useful to compare the occurrences of the clusters (of the considered CDR)) // k/lambda – type of the light chain (if it is) //the organisms-owners of the proteins used in the work(**\***) – H = Human, M = Mouse, C = Camel, L = Llama // + (///// some commentaries)

///// lengths that are written nearly the CDR’s name are just above the truth; their real sizes corresponding to the IMGT numbering can be found above, in the main rules

///// don’t forget that the CDR of the real sequence can differ from its expected cluster a lot // also it’s possible (and happens frequently) that none of these clusters suits the sequence’s CDR

Therefore, use these clusters **only** to find the hypothetical CDR or to prove the already found one (and in no case for a refutation)!

**\*North, B. et al. (2011). A New Clustering of Antibody CDR Loop Conformations.**

**\*CDRs are marked according to the IMGT numbering**

**\*\*it’s important that the canonical structures and clusters for CDRs are just an approximation – so they can be used only to find the hypothetical CDR or to prove the already found one**