**IMGT Rules**

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 – 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 (the last one), (118-128)

**Afterword:** region – **information about (mostly) invariable key residues that will help to find this region (if they are exist),** information about the *canonical structure* (if exists) or extra information that may help (if exists)**,** length of the region (that will help to move from key residues and find the regions)**,** **(x-y)** – region’s limits corresponding to this type of numbering (can help if numbering has already magically happened or to check the already annotated sequence) **/////** addition positions in numbering, that are located in this region (can help somebody, maybe)

**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 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 // grey color means that the marked residue’s number not surely correct (because of the Kabat-to-IMGT translation**\*\*\*(\*\*\*\*))**

**\*the original residues’ numbers (in Kabat’s numbering) were translated to IMGT numbering using the structural alignment of these two numberings\*\*** <https://www.bioc.uzh.ch/plueckthun/antibody/Numbering/NumFrame.html>

**//received numbers with the corresponding information(listed above) were tested on several examples, so they are probably correct**

**\*\*IMGT numbering that was used in structural alignment (IMGT-1) is different from received in this work (IMGT-2) (only with 46A, B and 84A and the end’s length), so the translated numbers are corresponding to the first one (IMGT-1)**

**\*\*\*(\*\*\*\*) there are two reasons why residue’s number can be marked with a grey color: a) because of some problems with numberings, original numbers for the light and for heavy chains (in the Kabat numbering) don’t coincide while the alignment => so in this case, there are two numbers separated by “?”, corresponding to the original numbers for the light and heavy chains => it’s better to avoid using them b) because of differences in the end’s lengths of IMGT-1 and IMGT-2 (end with 120 and 128, respectively) and also some problems while the alignment, numbers after 95/99 are absent in the alignment scheme => so in this case, the IMGT numbers were placed relying on considerations of logic (and several examples) => they can be used, but also without 100% sure**

**\*\*\*\*actually, it’s caused by IMGT gaps (in FRs): because of IMGT gaps (10, 73, 81, 82), the alignment with the other numberings at these positions is indefinite // and also, there are another indefinite positions (34?39, 40?34) that appeared only because of being located in the CDR (with a lot of gaps!)**

**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 (50), KRQT, LRGTV, **L**WV, **I**VM (54), **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 (50), **E**K (51), **W**YG, IVML, GA, YW = **I**V (56), 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 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*

**\*the original residues’ numbers (in the Kabat numbering) were translated to IMGT numbering using the structural alignment of the Kabat and IMGT numberings\*\*** <https://www.bioc.uzh.ch/plueckthun/antibody/Numbering/NumFrame.html>

**//received numbers with the corresponding information (listed above) were tested on several examples, so they are probably correct**

**\*\*IMGT numbering that was used in structural alignment (IMGT-1) is different from received in this work (IMGT-2) (only with 46A, B and 84A and the end’s length), so the translated numbers are corresponding to the first one (IMGT-1)**

**\*\*\*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”**

**!!!some almost common residues for the positions could be lost, so the presence of the other residue on the considered position is saying nothing (so this information can be only an addition to the main rules)**

**!!! all this information is just about the truth, so use it carefully!!!**

**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)

**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)

**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)

**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)

**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)

**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)

**Afterword:** for each CDR you can see several amino acids sequences those 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 lengths 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**

**\*the numbers of residues corresponding to the IMGT numbering were received by testing Extra RRR on the annotated sequences (so they are probably true)**

**\*\*it’s important that the canonical structures and clusters for CDRs are not present in the reality – so they can be used only to find the hypothetical CDR or to prove the already found one**

**!!! the Extra RRR is the most contentious extra information, so use it super carefully !!!**