**Kabat 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”***.

**Light chain**

FR1 – **ends with Cys23**, length = 23-24 for V(k) and 22-23 for V(lambda), (1-23) ///// 0, del 10 in V(lambda)

CDR1 – **begins after Cys23 and is followed by Trp35**, length <= 17, (24-34) ///// 27A, B, C, D, E, F

FR2 – **begins with Trp35** (and ends with Tyr49), length = 15, (35-49) // length = 19 for some lambda rabbits (4 insertions 48A, B, C, D)

CDR2 – (begins after Tyr49) and **is followed by Gly 57**, length <= 7, (50-56) // 48-56 for V(lambda) mice

FR3 – **begins with Gly 57 and** **ends with Cys88**, length = 32, (57-88)

CDR3 – **begins after Cys88 and is followed by Phe98**, length <= 15, (89-97) ///// 95A, B, C, D, E, F

FR4 – **begins with Phe98**, (nearly invariant Gly99 and Gly101), length = 10-11, (98-107) ///// 106A

**Heavy chain**

FR1 – **ends *nearly* after Cys22**, (Ser/Thr 25, Gly26), length = 30 (31 when 0 is present), (1-30) ///// 0 // Ser25 mostly (1384/1521) + Thr25 (107/1521); Gly26 (1445/1474)

CDR1 – **begins *nearly* after Cys22 and is followed by Trp36**, length <= 7, (31-35B) ///// 35A, B

FR2 – **begins with Trp36**, length = 14, (36-49)

CDR2 – length <= 19, (50-65) (65 in FR3 for rabbits) ///// 52A, B, C

FR3 – **ends *nearly* after Cys92**, length = 29-32, (66-94) ///// 82A, B, C

CDR3 – **begins *nearly* after Cys92 and is followed by Trp103**, length <= 19, (95-102) ///// 100A, B, C, D, E, F, G, H, I, J, K

FR4 – **begins with Trp103**, (Gly104, Gly106), length = 11, (103-113)

Numbering:

**Light-chain:** (0) 1 … 26 27 27A 27B 27C 27D 27E 27F 28 … 94 95 95A 95B 95C 95D 95E 95F 96 … 105 106 106A 107…

**Heavy-chain:** (0) 1 … 34 35 35A 35B 36 … 51 52 52A 52B 52C 53 … 81 82 82A 82B 82C 83 … 99 100 100A 100B 100C 100D 100E 100F 100G 100H 100I 100J 100K 101 … 113…

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

**Afterword:** CDR / FR – **information about (mostly) invariable key residues that helps to find this region (if they exist),** **(**information about *less* invariable residues but also possible for use (if they exist)**),** length of the region (that help to move from the key residues and find the regions)**,** **(x-y)** – region’s boundaries corresponding to the Kabat 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) **//** some commentaries

*<< xvi, Sequences of Proteins of Immunological Interest, 1983 >> regions’ boundaries, lengths and additional positions; commentaries about insertions between positions 48-49 in some rabbits V(lambda); commentaries about position 65 in rabbits’ VH*

*<< xvii, Sequences of Proteins of Immunological Interest, 1983 >> insert positions of CDR3*

*<< xlvi, Sequences of Proteins of Immunological Interest, 1983 >> Trp36 in VH and Trp35 in VL*

*<< lxx-lxxi, Sequences of Proteins of Immunological Interest, 1983 >> key residues in VL and VH; mostly invariable residues; commentaries about mice V(lambda) CDR2*

*<< Sequences of Proteins of Immunological Interest, 1983 >> numbering*

*<< 4, 1970 >> boundaries and key residues in L1, L3*

*<< 28, 1970 >> boundaries and key residues of L1, L2, L3*

*<< 4, 1971 >> boundaries of H1, H2, H3*

**Extra (invariant residues and Gly)**

VL – invariant residues: Gln6, Cys23, Trp35, Pro59, Arg61, Asp/Asx82, Tyr86, Cys88, Phe98, Gly99, Gly101 = 6 Q, 23 C, 35 W, 59 P, 61 R, 82 DN, 86 Y, 88 C, 98 F, 99 G, 101 G

***(Mostly) invariant Gly:***

V(k) - G: 16, 41(39), 57, 64, 66, 68, 99, 101

V(k) mouse – G: 16, 41, 57, 64, 68, 99, 100, 101

V(lambda) – G: 16, 25, 41, 57, 64, 68, 77, 99, 100, 101

VH – G: 8, 26, 42, 104, 106

*<< 22-23, 1970 >> invariant (90%) Gly in VL*

*<< 23, 1970 >> invariant residues in VL*

*<< 8, 1971 >> invariant Gly in VH*

**Extra R (Kabat specified)\***

**invariant residues: 23/22 C, 35/36 W, 38/39 Q, 82/86 D, 86/90 Y, 88/92 C, 99/104 G, 101/106 G, 102/107 T**

**closely related residues: 4/4 LM, 6/6 QE, 16/15 GS, 21/20 VLIM, 22/21 ST, 47/48 VLIM, 61/66 RK, 63/68 ST, 65/70 ST, 73/80 LIMF, 75/82 LIM, 78/82c VLM, 84/88 GA, 87/91 YF, 98/103 FW, 104/109 VL, 106/111 VLI**

***bn* sites (C, V ,L, I, M, F, W + P, H ,Y ,G, A, S, T) = 11/10 (+RKE), 12/11, 13/12 (+RKE), 15/14 (P), 19/18 (VLIM), 25/24, 33/(36-2), 36/37, 44/45, 46/47, 48(I)/49(GAS), 62/67 (*b*), 64(G)/69(MI), 71(A if (lambda), FY if (k))/78, 83/87, 85/89, 97/102**

***n* sites (P, H, Y ,G, A, S, T) = 40/41 (PS), 72(STA)/79(YF)**

***sn* sites (R, K, E, D, Q, N + P, H ,Y ,G, A, S, T) = 5/5, 14/13, 17/16, 18/17, 20/19, 24/23, 34/(36-1), 39/40, 41/42 (G), 42/43, 43/44 (*n*), 60/65, 66/71, 67(S)/72(D), 68/73, 69/74, 70/77, 74/81, 76/82a, 77/82b, 79/83, 80/84, 81/85, 89/93, 90/94, 100/105, 103/108, 105/110, 107/112**

***s* sites (R, K, E, D, Q, N) = 37/38 (RKQ), 45/46**

**(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/n – light/heavy** chain in the Kabat numbering ///// y(X) means that X (or X-group) is present at y residue rather than other possible ((if there a space before (X), it means that it concerns to both light and heavy chains))

**\*numbering considered in this article is different the from the Kabat one, so the obtained info was accurately translated into the Kabat annotation**

*<< 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 (Kabat specified)\***

**L1** – **G**, DEQ, RKQTS, VA, TS, ILM, STN, **C (23)** = RKST, ASG, **S**T … YNFAW, LVMIA, ANH = **W (35)**, **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 (49) = YKW, ATV … LRS, AF, STDP = **G (57)**, **V**I, **P**, DSAV, **R**, **F**, **S**T, **G**

**L3** – **E**D, **D**, LFEIAV, AG, TVDI, **Y**, YF, **C (88)** = QFLAS, QHL, YGSWH … **P**LH, YLPRWF, **T**V = **F (98)**, **G**, GQAST, **G**, **T**, **K**R, **L**V, **E**TD

**H1** – **G**S, GAQ, **S**T, LVM, KRS, LIMV, **S**T, **C (22)**, KATS, AVT, **S**T (25) … YWGA, MIWV, HNSG = **W (36)**, **V**IF, RK, **Q**K, ARPFST, **P**H, **G**E, KQN

**H2** – **G**E, KQNH, GRKEA, **L**R, **E**K, **W**YG, IVML, GA (49) = YW (50), **I**V, SNYD … SNYDT (56), TIPSKA, YN, **Y**F, NASVG, DPEQA, SKADT, FVL, **K**QR, GSD (65) = RK (66)

**H3** – **E**DA, **D**, TS, **A**G, VTIML, **Y**, **Y**F, **C (92)**, ATV, RASN = GW … FMGLY, DAGV, YV = **W (103)**, **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 doesn’t completely correlate between different positions

///// 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 original rules (Kabat, Chothia and IMGT)

///// **the region marked with a red color corresponds to the CDR (=\*)**

///// some positions are numbered (in the Kabat/Chothia numberings (at these positions they are identical)) 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)**\*\***)

**\*numbering and CDRs boundaries considered in this article are different from the Kabat ones, so the obtained info was accurately translated into the Kabat annotation**

**\*\*it can be done, because regions with structural insertions (deletions) those make all the differences in the numberings, are absent here or located in the region replaced with ellipsis**

**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 (Kabat specified)\***

**//put them in the length order and add commentaries**

**Clusters for CDRs \*\***

**L1** – length 10-17 (mostly 11 or 16)

**///// 24-34 (CDR1)**

(L1-2A) L1-11-1 – R**ASQ**DISNYLA (76, k, HM) ///// F71

(L1-4) L1-16-1 – **RSS**QSLVHSN**G**N**TYL**E (68, k, HM)

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

(L1-3) L1-17-1 – **KSSQSL**LN**S**RTRK**NYLA** (21, k, HM)

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

(L1-3λA, L1-3λB) L1-14-1 – S**A**S**SSV**S**Y**MH (14, lambda, H)

(L1-5) L1-15-1 – R**AS**E**SVD**SY**G**N**S**F**M**N (11, k, HM)

(L1-1λ) 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)

(L1-4λ) L1-11-3 – SGNNLGS-SVH (5, lambda, H)

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

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

L1-10-2 – **SASSSVSY**MY (2, k, M)

(L1-6) 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 7 or 11

**///// 49-56 ([49] FR2 + CDR2)**

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

(L2-1) 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)

**///// 89-97 (CDR3)**

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

(L3-1λA, L3-1λB, L3-1λC) 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)

(L3-2λ) L3-11-1 – AAWDSSLDAVV (9, lambda, H)

(L3-1) L3-9-cis7-2 – **QH**FWS**TP**R**T** (8, k, HM)

L3-10-1 – QSYDSS-SVV (6, lambda, H)

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

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

(L3-4) L3-7-1 – Q**Q**YN**SY**S (2, k, HM)

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

(L3-5) L3-10-cis8-1 – LYSREF**PP**W**T** (2, k, M)

(L3-2) 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 2-8 (mostly 5)

**///// 23-35B ([23-30] FR1 + CDR1)**

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

(H1-2) H1-14-1 – **TVTGYSIT**SG**Y**A**W**N (11, M)

(H1-3) H1-15-1 – SF**SGFS**LSTSGMG**V**G (9, HM)

(H1-1) 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** – …

**///// 51-58 ([51-58] CDR2)**

(H2-2A) H2-10-1 – -**I**YPGNG-T- (155, HM) ///// AVLISTQ 71 (mostly)

(H2-1) H2-9-1 – Y**I**WYS**G**STY (77, HM)

(H2-3A, H2-3C) H2-10-2 – -**I**SSGGGNTY (42, HM) ///// **R**D 71 (mostly)

(H2-4) H2-12-1 – E**IR**N**K**ANNYT**T**E (26, M)

(H2-2B) 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)

(H2-3B) 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 3-24 (mostly 5-14)

**///// 93-102 ([93-94] FR3 + 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 different CDR’s clusters (idea similar to the canonical structures concept)

///// (name of Chothia canonical structure if exists) cluster’s name (according to this article) – 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 Kabat 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 Kabat 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**