**ICK-Data**

// add score with explanation to: Extra information and most common pos-s (VL IM)

ICK-Data is an exel-file (i.e. a grid) that contains info about:

* each annotation (Kabat, Chothia, IMGT) **numbering** for each chain (VH, V(k) and V(lambda))
* each annotation (Kabat, Chothia, IMGT) **CDRs boundaries** for each chain (VH, V(k) and V(lambda)) (the numbers of the positions located in the CDRs are surrounded by a thick black frame)
* **possible amino acids at each position** (for most of the positions) (each annotation specified)**\***

In 2 variants (based on 2 different articles, their concepts): *“conserved core concept”* and *“all-species accurate statistics for each CDR neighborhood”*

*“conserved core concept”* (ref. to Chothia et al. 1998 = Extra R) – each (almost every) antibody sequence has a conserved core (invariant, closely related and snb (with a conserved hydrophobic group) sites); it’s common for almost all the sequences and species

*“all-species accurate statistics for CDR’s neighborhood”* (ref. to North et al. 2011 = Extra RR) – the most recent and probably the most accurate all-species statistics for each CDR neighborhood, i.e. statistics = positions and their possible amino acids obtained as a result of studying a big amount of data

**\***These 2 concepts works with all the annotation schemes, but in ICK-Data this info specified according to the annotation numbering.

Also this info is supplemented by each annotation *original rules* and each annotation *Extra* (if exists).

* **score markers** that show the score (the weight) of each position and its possible amino acids according to the developed *system of heuristic weights*

I.e. shows the accuracy level of each position info – how much does it have to mean comparing to the other positions?

Also it’s necessary for the *Rule-annotator scoring system.*

All this info (each annotation numbering, CDRs boundaries, possible amino acids at each position, score markers) is linked together. So that because ICK-Data is super useful, accurate and comfortable to use. It’s the best linguistic way to annotate sequences.

Comparing the amino acids at the positions of the considered sequence with the possible amino acids of the corresponding positions (i.e. with ICK-Data) provides a super accurate linguistic annotation method that makes a user able to find the CDRs easily and for sure. It works both on human and animals and can be easily used by both human and computer.

**Important!** ***Possible amino acids of the position*** are **the most common** amino acids at the position, i.e. all the amino acids are theoretically possible to be at that position but this chance is **negligible** compared to the chance of the most common ones (best for germlines and worse for highly mutated sequences).

***ICK-Data, how it works:***

Each numbering column has another two columns to its right: they are called **R** and **RR**. **R** is for Extra R info and **RR** is for Extra RR info. So each position has its possible amino acids written to its right: in the first cell to its right – possible amino acids according to **R** and in the second – according to **RR**. Each set of possible amino acids is indicated by a sequence, each amino acid is indicated by a letter (according to generally accepted notation). Also each cell has additional type of notation: there can be two sets of possible amino acids (each indicated by a sequence) – the first set is the set of possible amino acids more common to that position and the second set – of less common. See an example:

|  |  |  |
| --- | --- | --- |
| 38 | RKQ EDN | RK |

It means that position 38 has 6 possible amino acids: more common Arg (R), Lys (K), Gln (Q) and less common Glu (E), Asp (D), Asn (N) according to **Extra R**; and has 2 possible amino acids: Arg (R) and Lys (K) according to **Extra RR**.

Also each cell has its **score marker** – its own color. Score marker generally indicates the weight of the position, i.e. the variability of that position in comparison with the other positions.   
There are three score marker colors:

|  |  |  |
| --- | --- | --- |
| white | green | yellow |

A white-colored position is the most ordinary position; green one means that it’s **mostly invariant**; and to be marked with yellow color a position has to be **super-invariant** and to **be described** in the main original rules of Kabat, Chothia or IMGT.

Orange color is a special color that isn’t connected with score and weights. It highlights the positions that *don’t correspond to the number of annotation* (written at their left) and so they have a special reverse numbering. A position marked with orange also has a negative number to its right that says about its index and about a start point of the reverse numbering. That position (i.e. its possible amino acids) should be placed on the number of its index positions back from the start point. See an example:

|  |  |  |  |
| --- | --- | --- | --- |
| 100k |  | FMGLY | -3 from W |
| 101 |  | DAGV |  |
| 102 | CVLIMFWPHYGAST | YV |  |
| 103 | FW | W |  |

It means that the set of amino acids “FMGLY” is possible not for the position 100k, but for the position located 3 positions back from 103 W.

Usually orange-colored positions appear near to the sites there a gap can be located (near to the insertion positions for example).

Also there are special positions that have orange-colored positive numbers to their right. It means the same as a negative number near to an orange-colored position but the special numbering isn’t reverse. For example, “(+2)RKEDQNPHYGAST” means that the set of amino acids “RKEDQNPHYGAST” corresponds to the position located 2 positions forward from the start position (respectively, the start position is located 2 positions back from the considered one).

***System of heuristic weights, how it works:***

The main goal of the system of heuristic weights is to make possible for a machine **to score matches** of input sequence positions in the corresponding sets of possible amino acids. So this system has to be sensitive to the general variability of a position (in comparison with the other positions, i.e. invariant/mostly invariant/ordinary etc.) and to the number of its possible amino acids (imagine that there is a position that has 19 (out of 20 theoretically possible for human and animals) possible amino acids; then almost every input sequence positions will fit the 19-amino-acids-set easily despite some shifts and mutations).

So the first step was to determine the weight system sensitive to the variability of a position. Each position in ICK-Data already has its own score marker indicating its variability. So let’s give them numerical values:

|  |  |  |  |
| --- | --- | --- | --- |
| Extra R *“conserved core concept”* | ordinary = 1 | mostly invariant = 3 | Invariant = 9 |
| Extra RR *“all-species accurate statistics for CDR’s neighborhood”* | ordinary = 1 | mostly invariant = 6 | Invariant = 12 |

***Explanation:***

* **Extra RR information in most cases is more accurate than Extra R information.** Maybe it’s connected with their concepts – mathematical statistics against studying protein structures. The first was made based on the positions of a huge amount of sequences and the second was made based on a smaller amount of data and great and accurate but still thinking about protein structures. But this fact **can be proved only by a number of experiments** on real germlines.
* On average, **RR** is more accurate and respectively more important for matching than **R twice**. But considering the individual categories, it becomes clear that **the weights of ordinary positions of R and RR** actually don’t differ by half and *should be above equal*.  
  Maybe using fractional numbers is more accurate (like 1.1 and 1.5 etc.) but it becomes much less comfortable to work with them; and also in comparison with the mostly invariant positions weights these fractional additions will be negligible.   
  Also the ordinary positions are *the most simple* ones and have *the highest variability* so it’s reasonable to give them a minimal **weight = 1**.
* **R mostly invariant positions** on average are *three times more* important for matching than the ordinary ones. So their weights are **equal to 3**.   
  **RR mostly invariant positions** are even comparable with the invariant ones of R and on average are two times more important for matching than R mostly invariant. So they get **weights = 6**.
* **RR invariant positions** are the most important ones and on average are more important than RR mostly invariant ones twice. So they **weight 12 points**.   
  **R invariant positions** are comparable with the RR mostly invariant ones but on average are more important than them and three times more important than R mostly invariant positions. So each **gets 9 points**.
* **Note!** All this information is firstly described in the source articles. **Extra R** has invariant residues, closely related residues and others; **Extra RR** has positions with high possibility of the considered set to match, with super-high possibility and with super-high possibility also proved by the original main rules. That because it’s accurate and proved.   
  But the **relation between Extra R and Extra RR** information **is undefined theoretically**.  
  So all the information about this relation and also about **relations between R variability groups** was obtained in the course of a large number of experiments on real germlines and based on the experience of a person who has analyzed a lot of different annotations and articles on the relevant topic. Therefore the **system of weights** is **heuristic**.

And the second step was to determine the weight system sensitive to the number of its possible amino acids. Actually that wasn’t really hard and the solution is based only on math.  
The more possible amino acids a position has, the more likely it is to accidentally match the corresponding input sequence position. So if a position has **X possible amino acids**, then it gets **(N-X) score points**, where N is the maximum number of theoretically possible amino acids (N = 20 for human and animals). This idea is fully consistent with the probability theory.

For now each position has two score points number: one for its variability and other for the number of its possible amino acids. But to obtain the system of weights each position has to have only one score point number = its weight. So the last step is to link them together – just **multiply them**.

This solution provides a super good and accurate system of heuristic weights that was **checked with a lot of real germlines** examples – each annotation the program gave as output the smallest difference between scores of the positions; this difference was less than half only in a case of undefined CDR and **never** in other cases. Also on average it has total score **above 95%** out of the maximum.

Kabat refs:

*<< 1 // Kabat Rules, Kabat R >>* the main original Kabat rules (i.e. numbering, CDRs boundaries and some key residues)

*<< 2-3 // Extra R, Kabat R >>* Extra R = *“conserved core concept”*

*<< 3-4 // Extra RR, Kabat R >>* Extra RR = *“all-species accurate statistics for CDR’s neighborhood”*

*<< 2 // Extra, Kabat R >>* Extra = mostly and nearly invariant residues (additional to Extra R and Extra RR information)

Chothia refs:

*<< 1-2 // Chothia Rules, Chothia R >>* the main original Chothia rules (i.e. numbering, CDRs boundaries and some key residues)

*<< 12-13 // Extra R, Chothia R >>* Extra R = *“conserved core concept”*

*<< 13-14 // Extra RR, Chothia R >>* Extra RR = *“all-species accurate statistics for CDR’s neighborhood”*

*<< 2-3 // Extra, Chothia R >>* Extra = mostly and nearly invariant residues (additional to Extra R and Extra RR information)

IMGT refs:

*<< 1 // IMGT Rules, IMGT R >>* the main original IMGT rules (i.e. numbering, CDRs boundaries and some key residues)

*<< 1-2 // Extra R, IMGT R >>* Extra R = *“conserved core concept”*

*<< 2-3 // Extra RR, IMGT R >>* Extra RR = *“all-species accurate statistics for CDR’s neighborhood”*