1 Supplemental Text. Derivation of an analytical solution for the theoretical number of motifs within one CDR/FR region

1.1 Objectives and motif definition

Our aim is to determine the number of possible structural interaction motifs for any motif length. A given sequence motif is defined as follows:

- An amino acid is encoded as **X**.
- A gap is encoded as integer n where n quantifies the length of the gap.
- Each motif starts and ends with an amino acid **X**.
- There can be >1 amino acids in sequential positions but not >1 gaps.

Let us give two different definitions of motif length. By simply “motif length” we mean the number of **X**s in it plus the number of gaps, we note this lengths L. By “amino acid length”, we mean the number of amino acids included in the sequence, i.e. the number of **X**s plus the sum of all gap lengths. Please, refer to the section 1.2 for a few examples.

As the interaction sequence cannot exceed the size of the CDR/FR it is located in, we need to add one more constraint:

- The amino acid length of the motif is not bigger then a predefined number.

Let us denote the number of unique motifs of lengths L and amino acid length A as $\mathrm{N}_{L,A}$ and the number of unique motifs of length L with amino acid length *not exceeding* A as $\tilde{\mathrm{N}}_{L,A} = {}^{A}_{A_1=L}\mathrm{N}_{L,A_1}$

1.2 Examples

To derive a formula for $\mathrm{N}_{L,A}$, we inspect a few examples first for intuition purposes.

$$\begin{array}{lll}
L=1, A=1 & \rightarrow \mathbf{X} & \rightarrow \mathrm{N}_{1,1}=1 \\
L=2, A=2 & \rightarrow \mathbf{XX} & \rightarrow \mathrm{N}_{2,2}=1 \\
L=3, A=3 & \rightarrow \mathbf{XXX}, \mathbf{X1X} & \rightarrow \mathrm{N}_{3,3}=2 \\
L=3, A=n>3 & \rightarrow \mathbf{X}k\mathbf{X},\ (k=n-2) & \rightarrow \mathrm{N}_{3,n}=1 \\
L=4, A=4 & \rightarrow \mathbf{XXXX}, \mathbf{X1XX}, \mathbf{XX1X} & \rightarrow \mathrm{N}_{4,4}=3 \\
L=4, A=n>4 & \rightarrow \mathbf{X}k\mathbf{XX}, \mathbf{XX}k\mathbf{X}\ (k=n-3) & \rightarrow \mathrm{N}_{4,n}=2 \\
L=5, A=5 & \rightarrow \mathbf{XXXXX}, \mathbf{X1XXX}, \mathbf{XX1XX}, \mathbf{XXX1X}, \mathbf{X1X1X} & \rightarrow \mathrm{N}_{5,5}=5 \\
L=5, A=n>5 & \rightarrow \mathbf{X}k\mathbf{XXX}, \mathbf{XX}k\mathbf{XX}, \mathbf{XXX}k\mathbf{X}, \mathbf{X}k_1\mathbf{X}k_2\mathbf{X} & \rightarrow \mathrm{N}_{5,n}=n-1
\end{array}$$

In the last line, $k = n-4$ and $k_1 + k_2 = n-3$. Let us clarify this last line: there are only 3 motifs with a single gap, but if there are two gaps, their lengths can vary: $k_1 = 1, k_2 = n-4$; $k_1 = 2, k_2 = n-5$, ... , so that we have $n-4$ double-gapped motifs in total.

1.3 General formula

Now we can proceed to derive a general formula for $\mathrm{N}_{L,A}$. Let us note the number of **X**s in a motif as n_x and the number of gaps as n_g. We can count the motifs for fixed n_x and n_g and then we will just have to sum the results over all $n_x + n_g = L$. Thus, we have n_x **X**s and $n_x - 1$ slots for gaps – between any two neighbouring **X**s there can be a gap. First, we have to choose n_g slots: the number of ways to do this is

$$\binom{n_x-1}{n_g}$$

Now we have n_g gaps of total amino acid length $A - n_x$, and we need to distribute the lengths between the gaps. In other words, we need to split the number $A - n_x$ into a sum of n_g nonzero terms. The number of ways to do this is the number of n_g*-compositions* of $A - n_x$, which equals

$$\binom{A-n_x-1}{n_g-1}$$

Now we can write down the formula for $\mathrm{N}_{L,A}$ as

$$\mathrm{N}_{L,A} = \underset{n_g+n_x=L,n_g\leq n_x-1}{} \binom{n_x-1}{n_g}\binom{A-n_x-1}{n_g-1}$$

We did not take into account the all-**X** case, so for $A = L$ we should have

$$\mathrm{N}_{L,L} = 1 + \underset{n_g+n_x=L,n_g\leq n_x-1}{} \binom{n_x-1}{n_g}\binom{A-n_x-1}{n_g-1}$$

Similarly, the formula for $\bar{\mathrm{N}}_{L,A}$ is

$$\bar{\mathrm{N}}_{L,A} = \overset{A}{\underset{A_1=L}{}} \mathrm{N}_{L,A_1} = 1 + \overset{A}{\underset{A_1=L}{}} \underset{n_g+n_x=L,n_g\leq n_x-1}{} \binom{n_x-1}{n_g}\binom{A_1-n_x-1}{n_g-1}$$

Figures 1 and 2 show the growth of $\bar{\mathrm{N}}_{L,A}$ for L in $1,...,10$. We set 10 as maximum motif length based on our observations (see Fig. 2B in the main text).

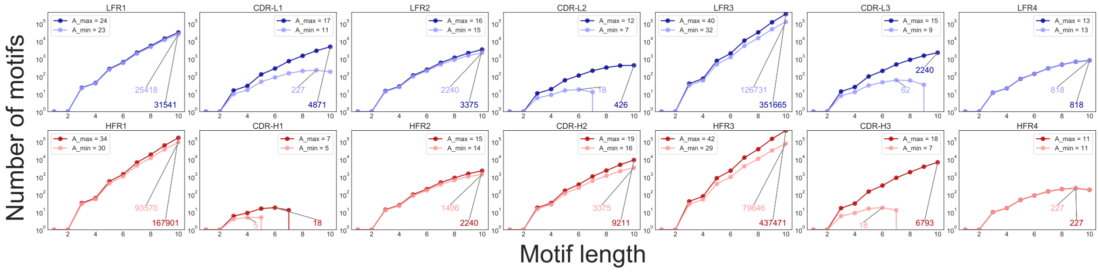

Figure 1: The number of unique motifs (Y axis) for a given motif length (X axis) that could be located in a certain FR/CDR (see possible FR/CDR lengths in Supplementary Table S1). The amino acid length of the motifs is bounded by the minimum and maximum possible region length (Supplementary Table S1).

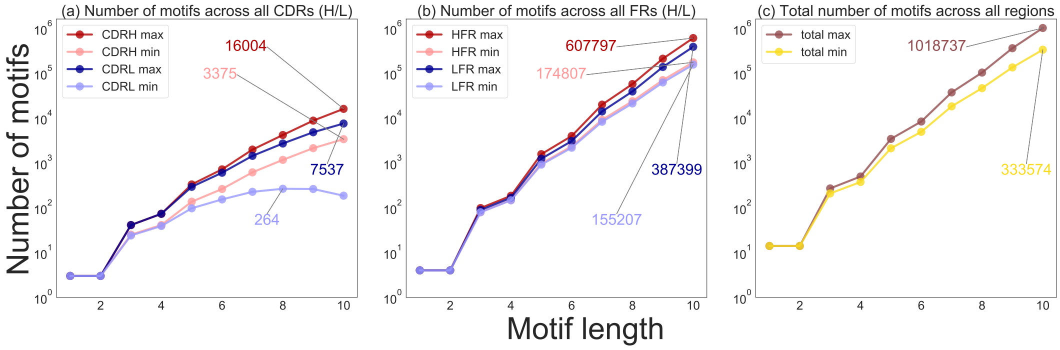

Figure 2: The total number of unique motifs (Y axis) for a given length (X axis) across all CDR-Ls and CDR-Hs **(a)**, across all LFRs and HFRs **(b)**, across all regions **(c)**.