Domains and “words” in proteins of different age

José Luis Villanueva-Canas, M.Mar Albà

Oct 2022

**Parsing the data**

We have extracted the relevant information from the XML files generated by protein predictor and the information about pfam protein domains calculated using *hmmsearch* using in house python scripts. After crossing this information we now have the corresponding secondary structure to every protein domain identified [pfam\_structure\_dataset.txt]. We also keep the secondary structure of whole proteins.

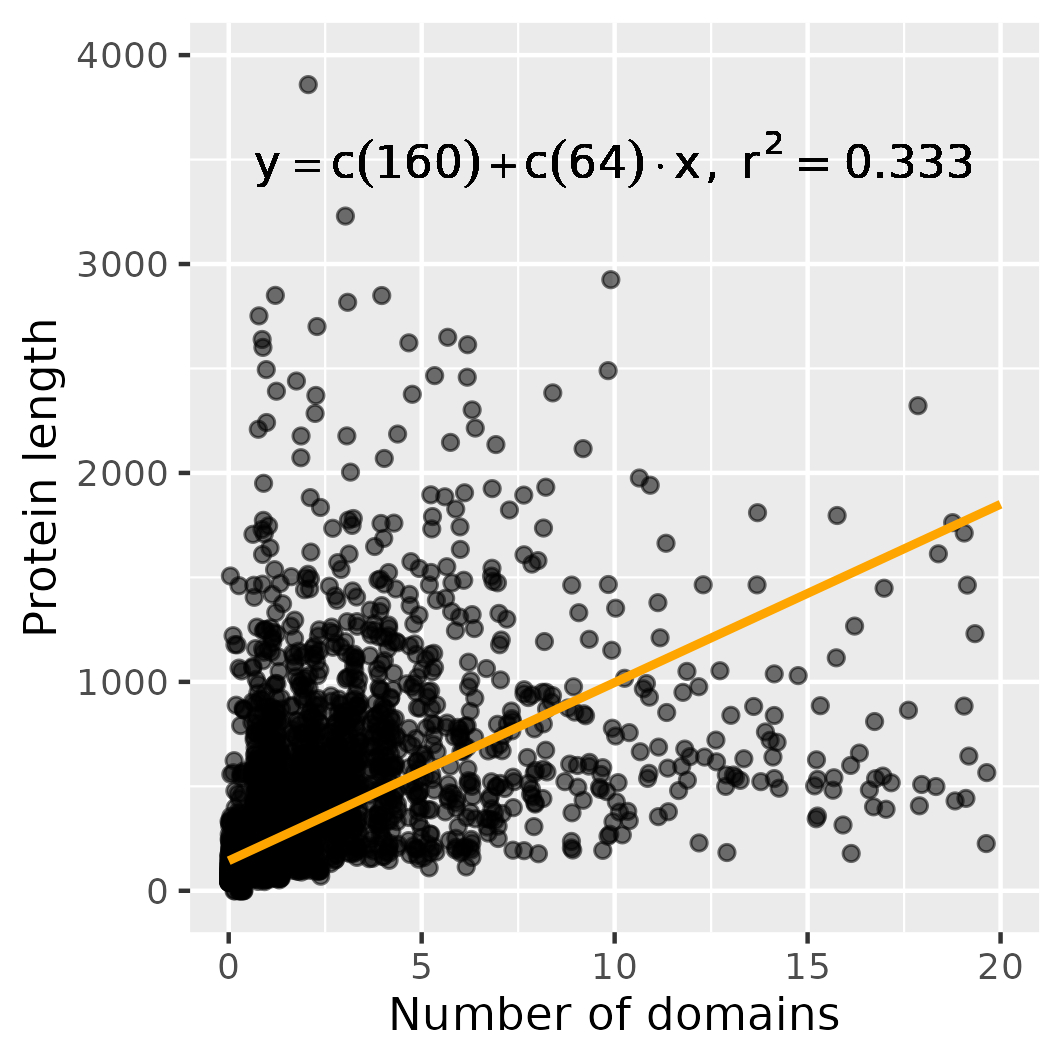
**Number of Pfam domains per proteins**

The first question we have investigated is the number of Pfam domains per protein. As can be observed in table 1, only a very small fraction of the youngest proteins has Pfam domains.

**Table 1.** Summary table for the number of domains in each protein. *Mean* *#domains:* Average number of domains per protein. *>=1 domain*: Number of proteins with at least one domain

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **age** | **# proteins** | **mean #domains** | **mean protein length** | **>=1 domain** |
| Hs\_ps1-2 | 2009 | 3.62 | 609.46 | 1916 |
| Hs\_ps13-17 | 581 | 0.91 | 269.53 | 433 |
| Hs\_ps20-21 | 433 | 0.74 | 157.68 | 205 |
| Hs\_ps31 | 1700 | 0.06 | 73.30 | 62 |
| Hs\_igen | 1617 | 0.01 | 70.79 | 13 |
| Hs\_rand | 1658 | 0.00 | 71.24 | 0 |

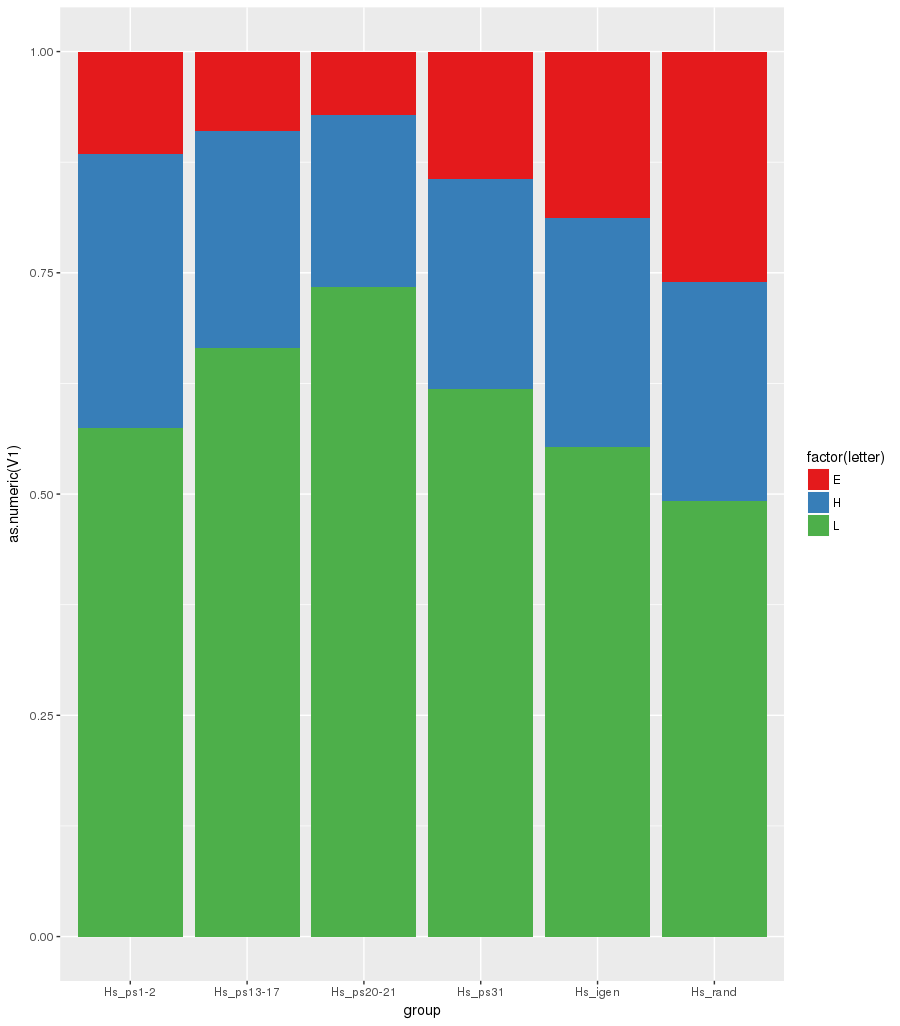
The number of domains increases is higher in older proteins, possibly because they are longer. There is certain correlation between protein length and number of domains (Figure 1).



**Figure 1. Correlation between number of domains and protein\_length.**

**Secondary structure composition by gene age**

We counted the number of amino acids that corresponded to alpha-helix(H), beta-sheet (E) and loop (L) in the different sequence sets (Table S1, Figure 2). There is a clear inverse correlation between age and the fraction of Hs.

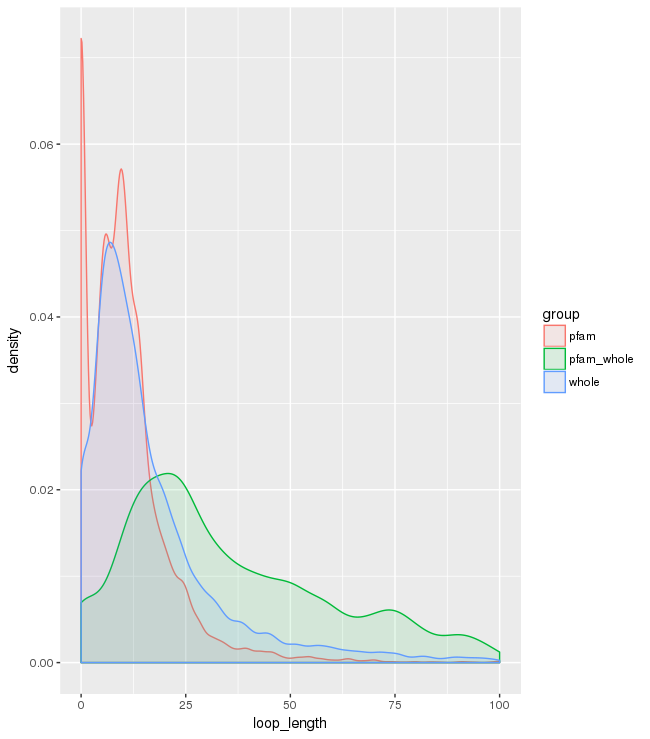


**Figure 2.** Proportion of secondary structure predictions (per amino acid) in the different age groups, simplified as letters (E=beta-sheet, H=alpha-helix, L=loop). Number of proteins in each category can be found in Table 1.

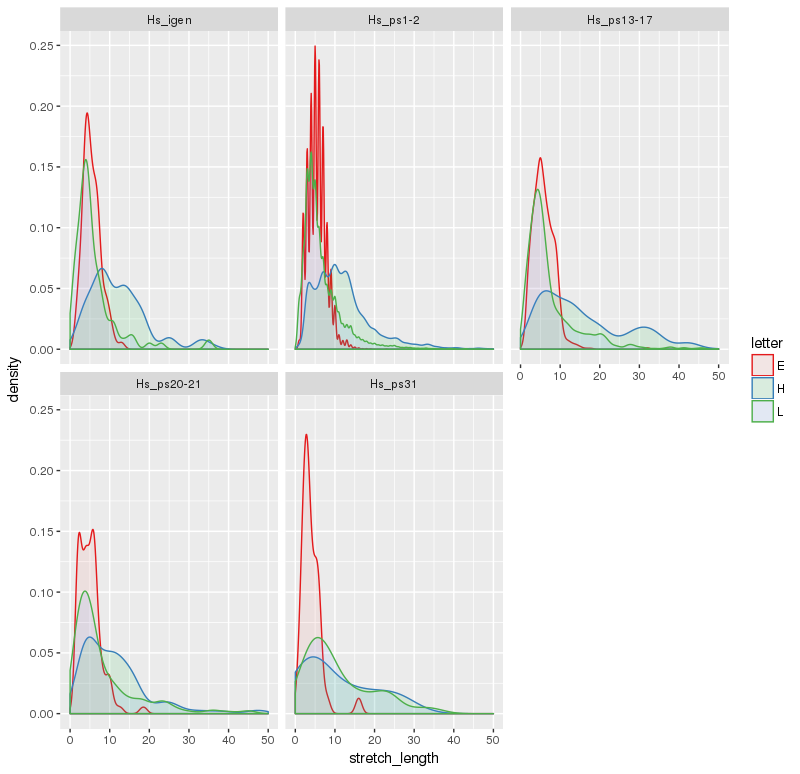
**Size of the different secondary structures**

We calculated that the median secondary structure length in the set of proteins analysed is 11 for H and 5 for either E or L. Loops up to 5 amino acid in length correspond to turns between other secondary structures. However, longer loops may take functions of their own [Papaleo et al 2016].

We investigated the size distribution of the loops in several sequence sets (see below), selecting the longest loop in each protein/domain, and excluding loops in the N- and C-termini of proteins. We established that loops shorter than 30 amino acids are unlikely to be found within a Pfam domain (p-value < 0.05). Therefore, loops of this size may be used to partition the proteins into different stretches when looking for “words”.



**Figure 3. Length of secondary structures. Red:** Longest loop inside a pfam domain, excluding terminal loops. **Green:** Longest loop in a protein that contains at least one pfam domain, again excluding -terminal loops. **Blue:** Longest loop inside a protein, independently of the presence of pfam domains in the protein.

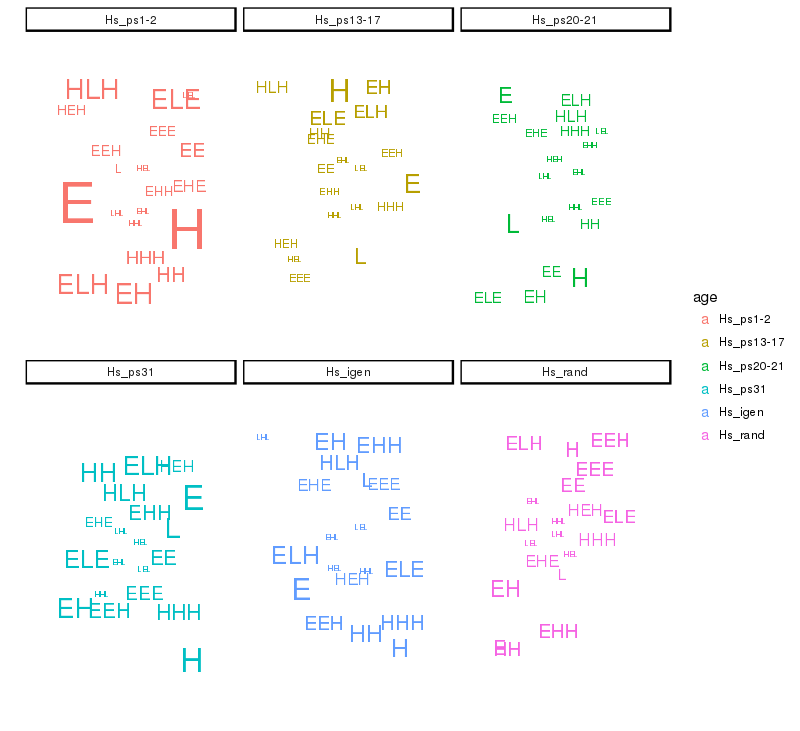


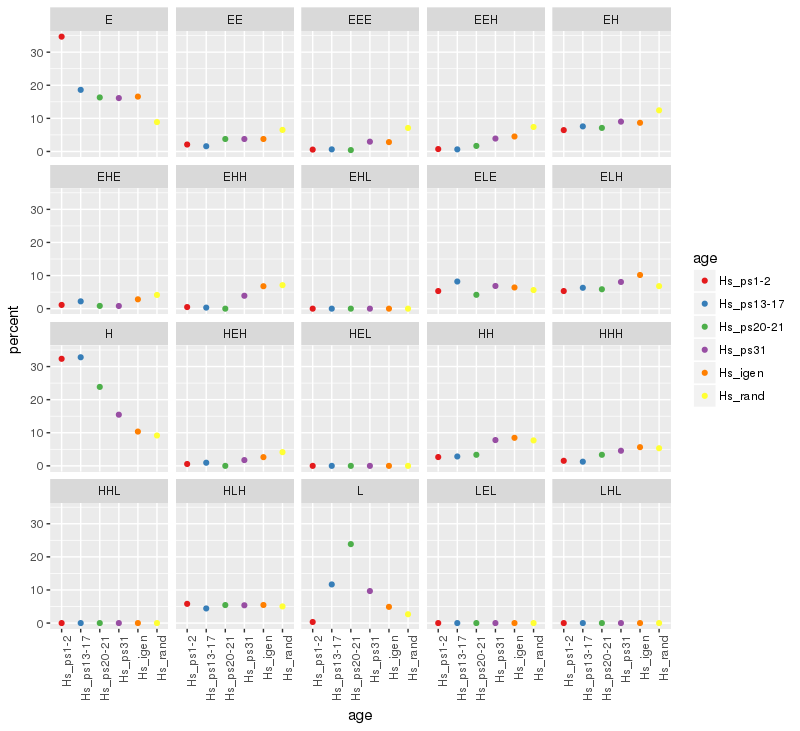
**Figure 4. Length of secondary structures in different sequence sets.**

**Occurrence of “words”**

Words in proteins will be defined as sequences of secondary elements. We take all the protein sequences and collapse identical contiguous letters (for example EEEEEE in the original file becomes E). We remove the terminal loops but consider loops longer than 5 amino acids as a loop element (L). Any loops of size 30 amino acids or longer are used to divide the sequences into different putative folds, as loops of this size are very rarely found within Pfam domains. Notice that loops shorter than 5 are turns and can interrupt a stretch of H or E. That's how we can obtain HHLEELH words. We also generate another dataset where we construct the words only in the sequences that overlap with a pfam domain prediction. The analysis results in lists of words with different frequency in each gene age class.

First we analyse all short words (3 letters or less) to make them comparable across gene age groups (Table S2). Figure 5 shows the secondary structure composition by gene age, which confirms that young proteins have more alpha-helices than the oldest set.

 **Figure 5. Short word occurences (1 to 3 stretches).** The size of the words is relative to their frequency (number of <=3 stretches words).



**References**

#### The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery.Elena Papaleo, Giorgio Saladino, Matteo Lambrughi, Kresten Lindorff-Larsen, Francesco Luigi Gervasio, and Ruth Nussinov

*Chemical Reviews* 2016 *116* (11), 6391-6423

DOI: 10.1021/acs.chemrev.5b00623