Domains and “words” in proteins of different age

José Luis Villanueva-Cañas, M.Mar Albà

Oct 2022

**Data**

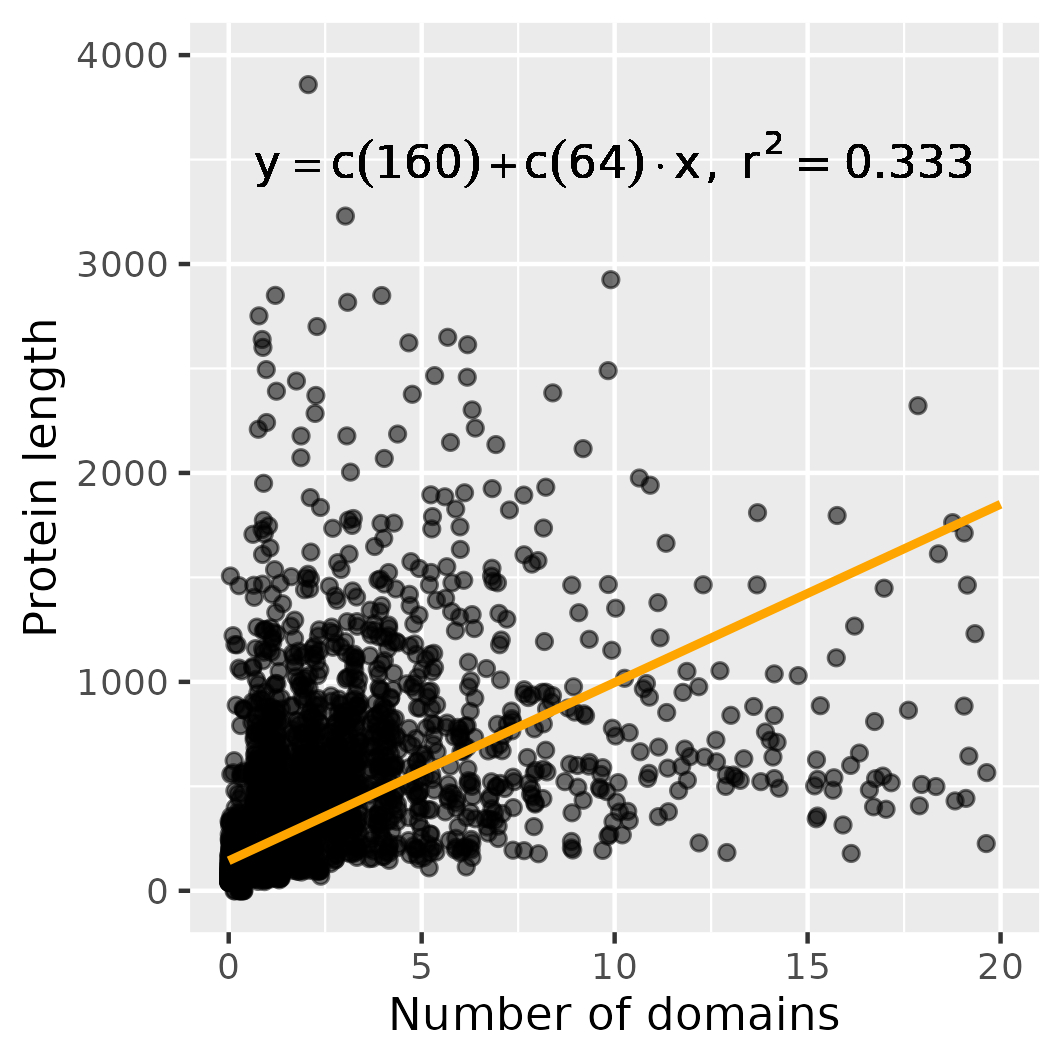
Our initial data were XML files generated by protein predictor in which each amino acid was associated with a secondary structure (H: alpha-helix; E: beta-sheet; L: loop). The analysis was done for groups of proteins of different age, as well as control sets containing intergenic and random proteins (Table 1). We added information from Pfam protein domains using hmmsearch.

**Number of Pfam domains per proteins**

The first question we have investigated is the number of Pfam domains per protein. As expected, long proteins tend to have more Pfam domains than shorter ones (Figure 1). Whereas nearly all proteins from the oldest group (Hs\_ps1-2) have at least one Pfam domain, the proportion is only 3.6% for the youngest proteins (Table 1). This precludes any comparison based on 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

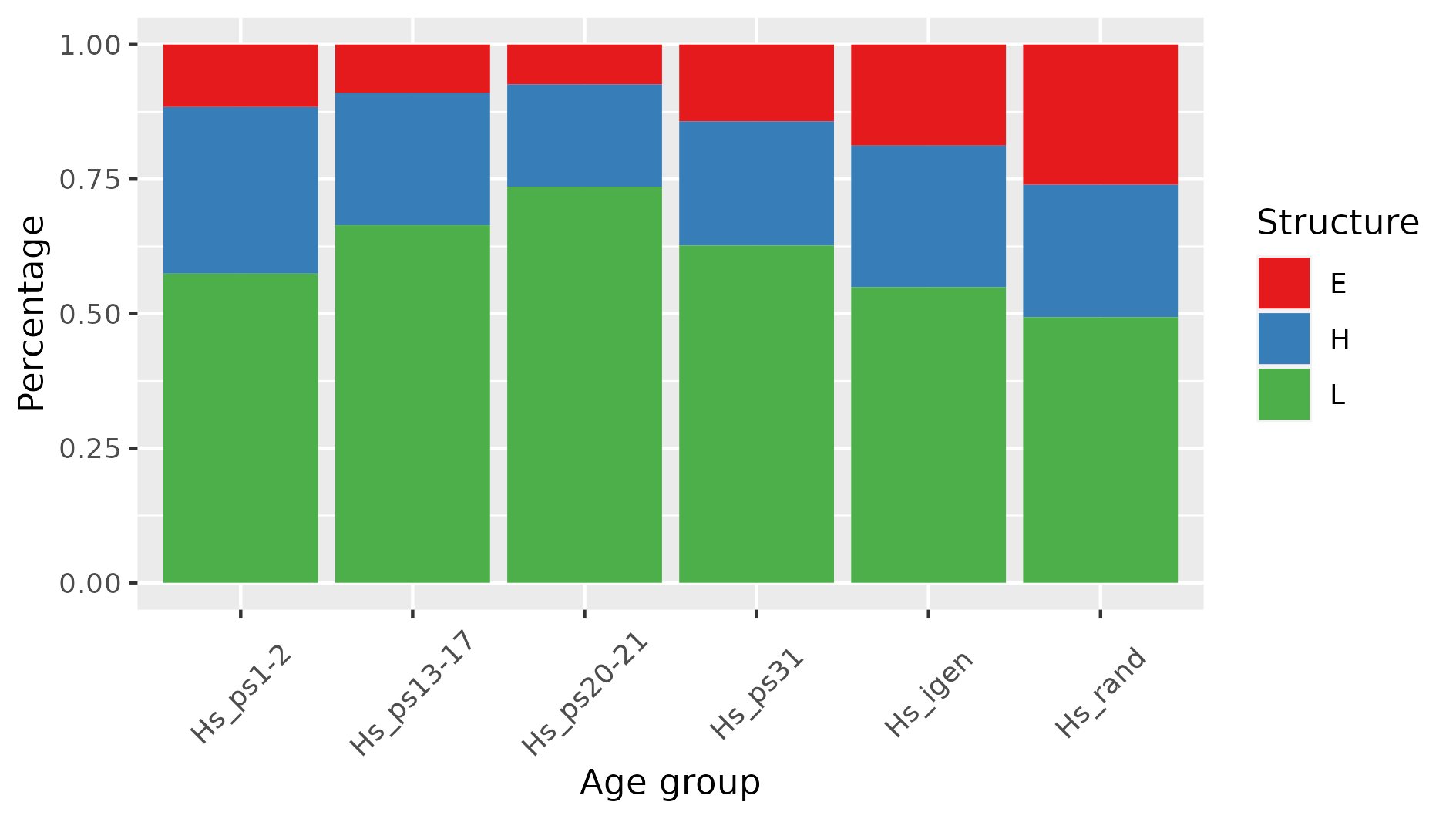
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **age** | **# proteins** | **mean #domains** | **mean protein length** | **>=1 domain** | **%** |
| Hs\_ps1-2 | 2009 | 3.62 | 609.46 | 1916 | 95.4 |
| Hs\_ps13-17 | 581 | 0.91 | 269.53 | 433 | 74.5 |
| Hs\_ps20-21 | 433 | 0.74 | 157.68 | 205 | 47.3 |
| Hs\_ps31 | 1700 | 0.06 | 73.30 | 62 | 3.6 |
| Hs\_igen | 1617 | 0.01 | 70.79 | 13 | 0.8 |
| Hs\_rand | 1658 | 0.00 | 71.24 | 0 | 0 |



**Figure 1.** Correlation between number of domains and protein length.

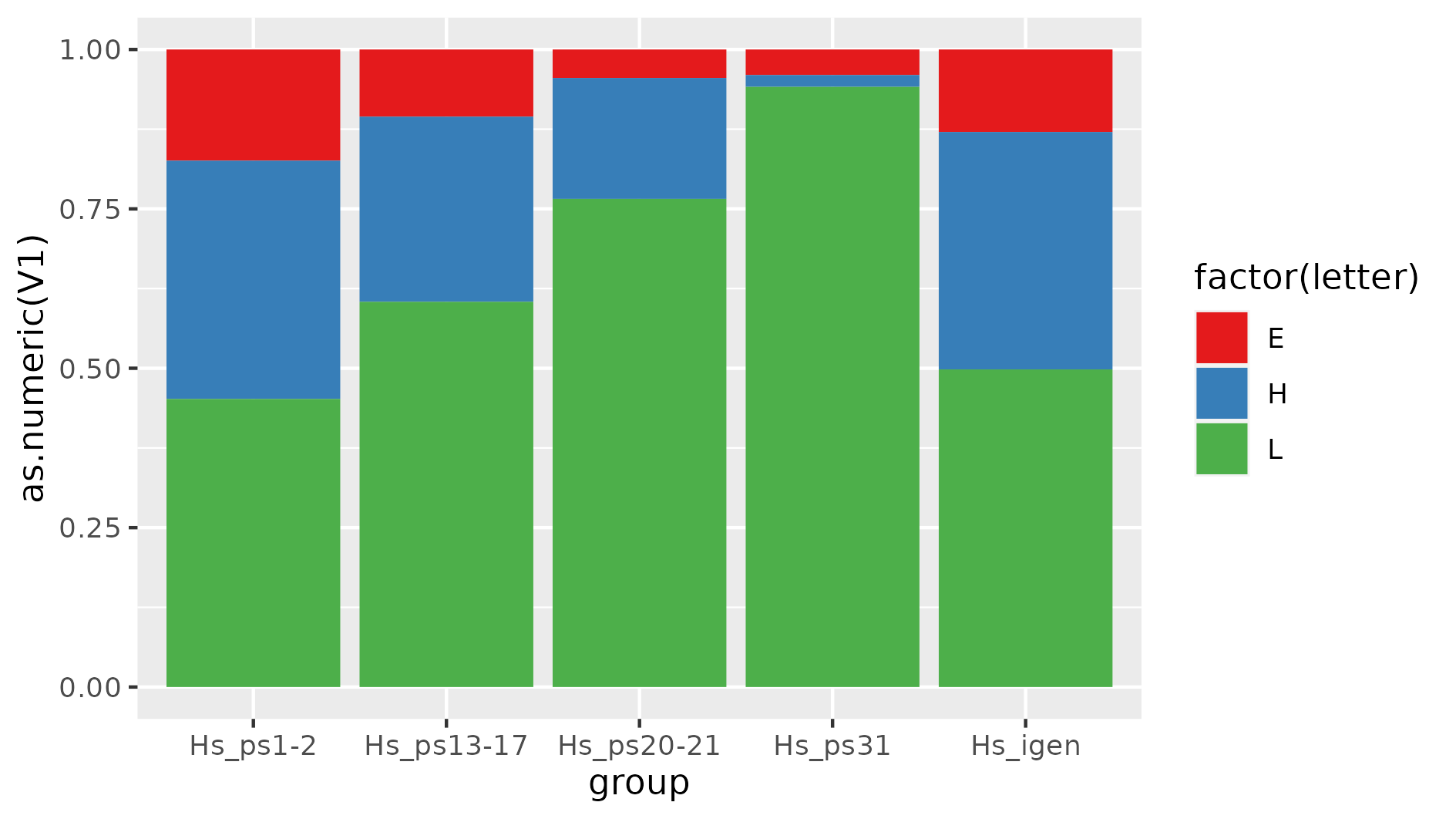
**Secondary structure composition by gene age**

Next, we counted the number of amino acids that corresponded to alpha-helix(H), beta-sheet (E) and loop (L) in the different sequence sets (Figure 2A). The youngest proteins (Hs\_ps31) have a larger proportion of L, and a lower proportion of E, than proteins in the group intergenic or random. The number of helices is larger in the oldest group.



**Figure 2A** Proportion of secondary structure predictions (per amino acid) in the different age

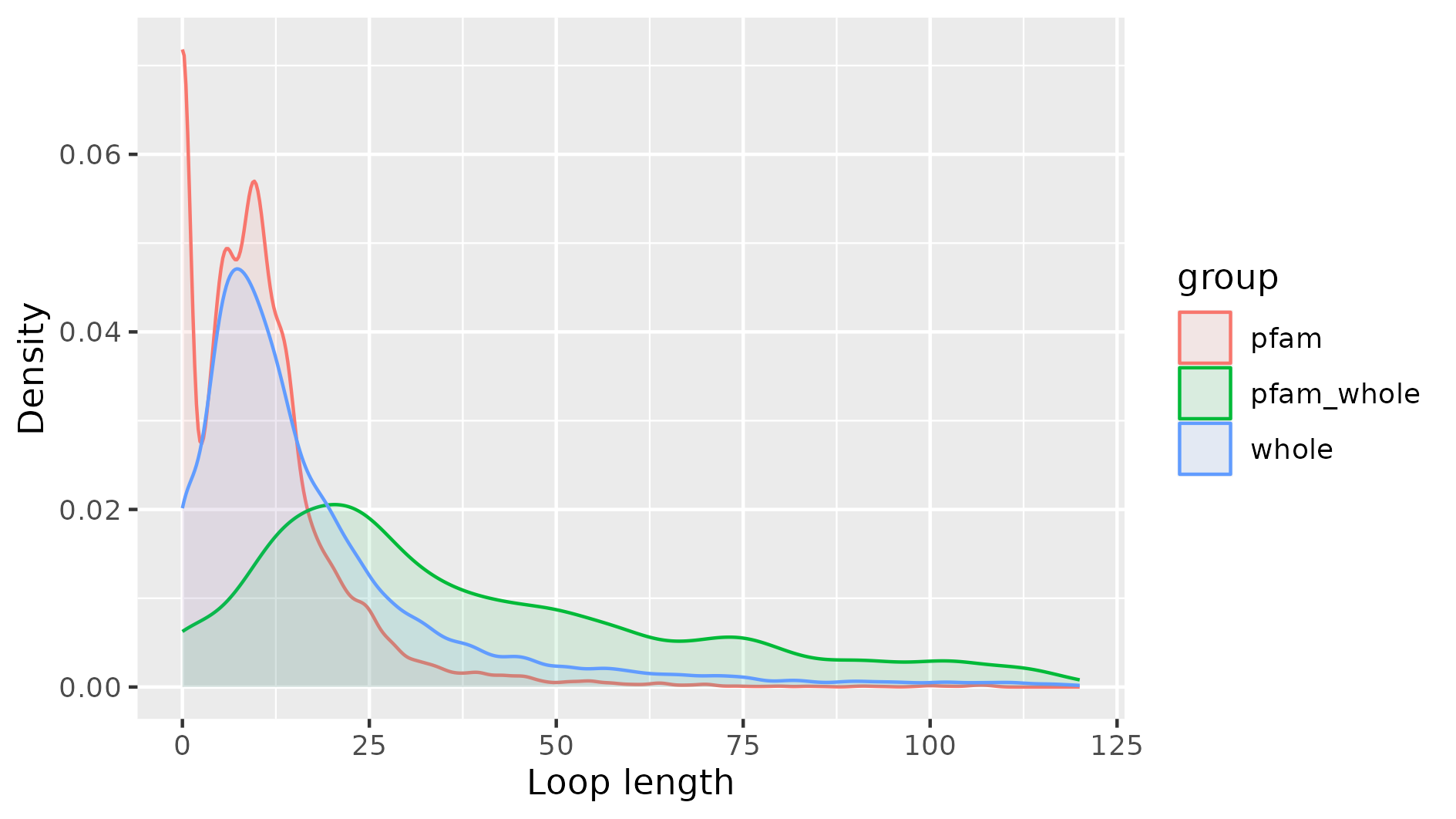
groups, simplified as letters (E=beta-sheet, H=alpha-helix, L=loop). Number of proteins in each category can be found in Table 1.

****

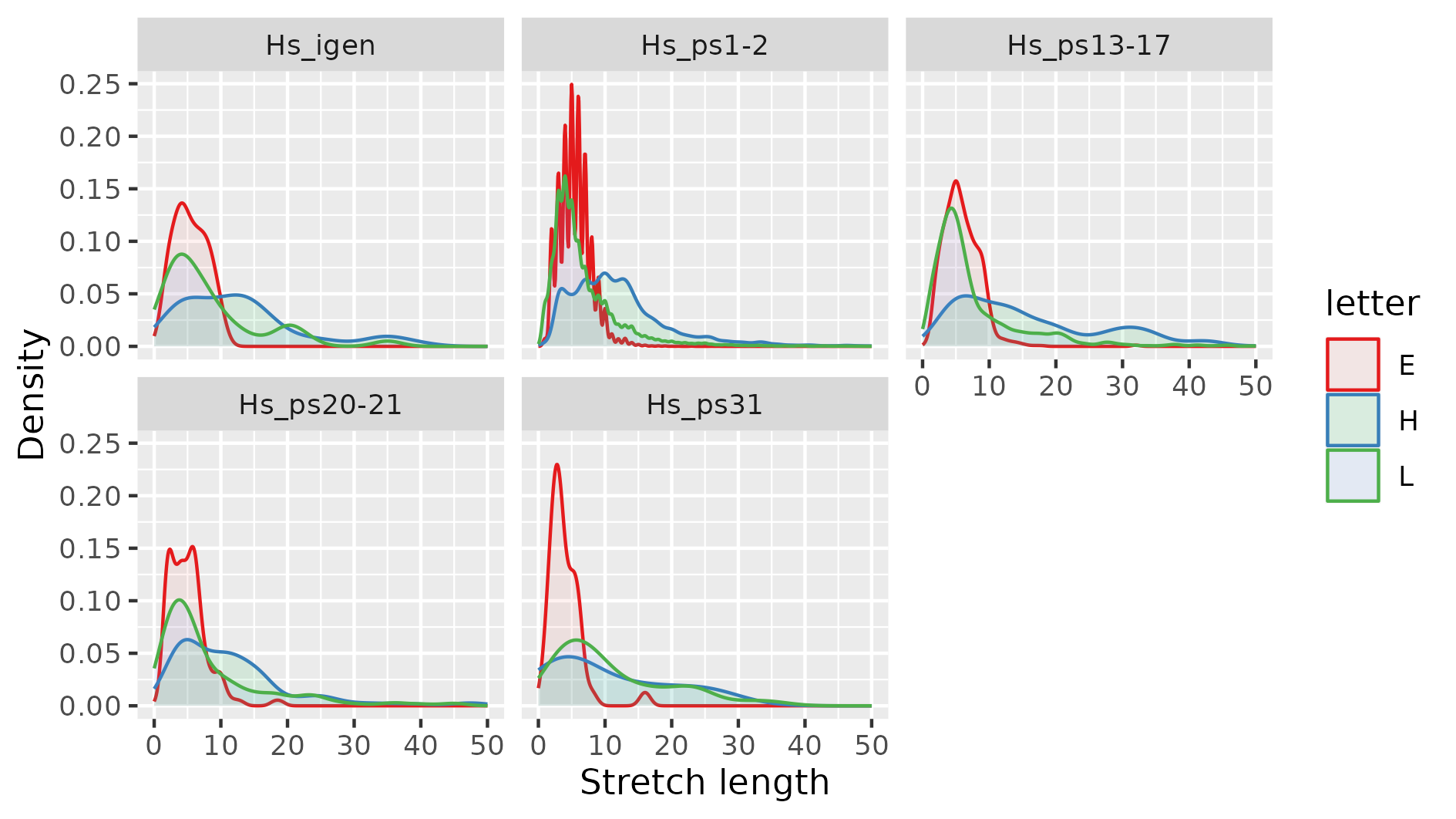
**Figure 2B** Proportion of secondary structure predictions (per amino acid) **within Pfam domains** for the different age groups, simplified as letters (E=beta-sheet, H=alpha-helix, L=loop).

**Size of the different secondary structures**

Next we investigated the length of stretches composed of consecutive amino acids from the same secondary structure class (H,E or L) in the different sequence sets (Figure 4). We did not consider Ls at the N- or C-termini of the protein. In general, beta-sheets tend to be shorter than the other secondary structures (median size 5 amino acids), alpha-helices are the longest (median size 11 amino acids). The loops were longer in the youngest proteins (Hs\_ps31) than in intergenic proteins.



**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.**  Each secondary structure is made of consecutive amino acids of the same type.

**Defining “words” in proteins**

The next aim was to define “words” made up of one or more secondary structures. For example an alpha-helix (one or more consecutive Hs) followed by a beta-sheet (one or more consecutive Es) would be the word HE. We did not consider Ls at the N- or C-termini of the protein.

Internal loops were only considered independent structures if they were longer than 5 amino acids. This is because loops up to 5 amino acid in length correspond to turns between other secondary structures, whereas longer loops may take functions of their own [Papaleo et al 2016].

For example:

sequence: **LLLLLL**HHHHHH**LL**HHHHH**LLL**EEE**LLLLLLLL**HHHHH**LLLLLLLLLLLLLL**

word: HHELH

We also decided that loops longer than 30 amino acids would separate different “words”. This was based on the observation that the probability to find a loop longer than 30 amino acid within a Pfam domain was less than 5%.

For example:

sequence: **L**HHHHHH**LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL**HHHH**LL**EEEE**LL**

words: H; HE

(the longest loop is 36 amino acids long)

Size of the different secondary structures

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].

**Occurrence of “words”**

In the first instance, we identified the 20 most frequently occurring words in the different groups (Figure 4). Some trends can be identified, for example words with H become more frequent in ps31 when compared to intergenic/random. However, in general, the comparison between the sets is not straightforward due to differences in dataset size and protein length.

Proteins of different age show important difference in size (Table 1). Everything else being equal, this will result in longer words being more frequent in the classes with longer proteins (oldest). In order to avoid this bias we decided to focus on the same set of words in all classes. We analyzed the frequency of all possible words of size 1 to 3 (Figures 5 and 6). We called these words “simple words” (Table 2). For simplicity, and to increase word counts, we considered that HE would be the same as EH, or HHL the same as LHH, etc.

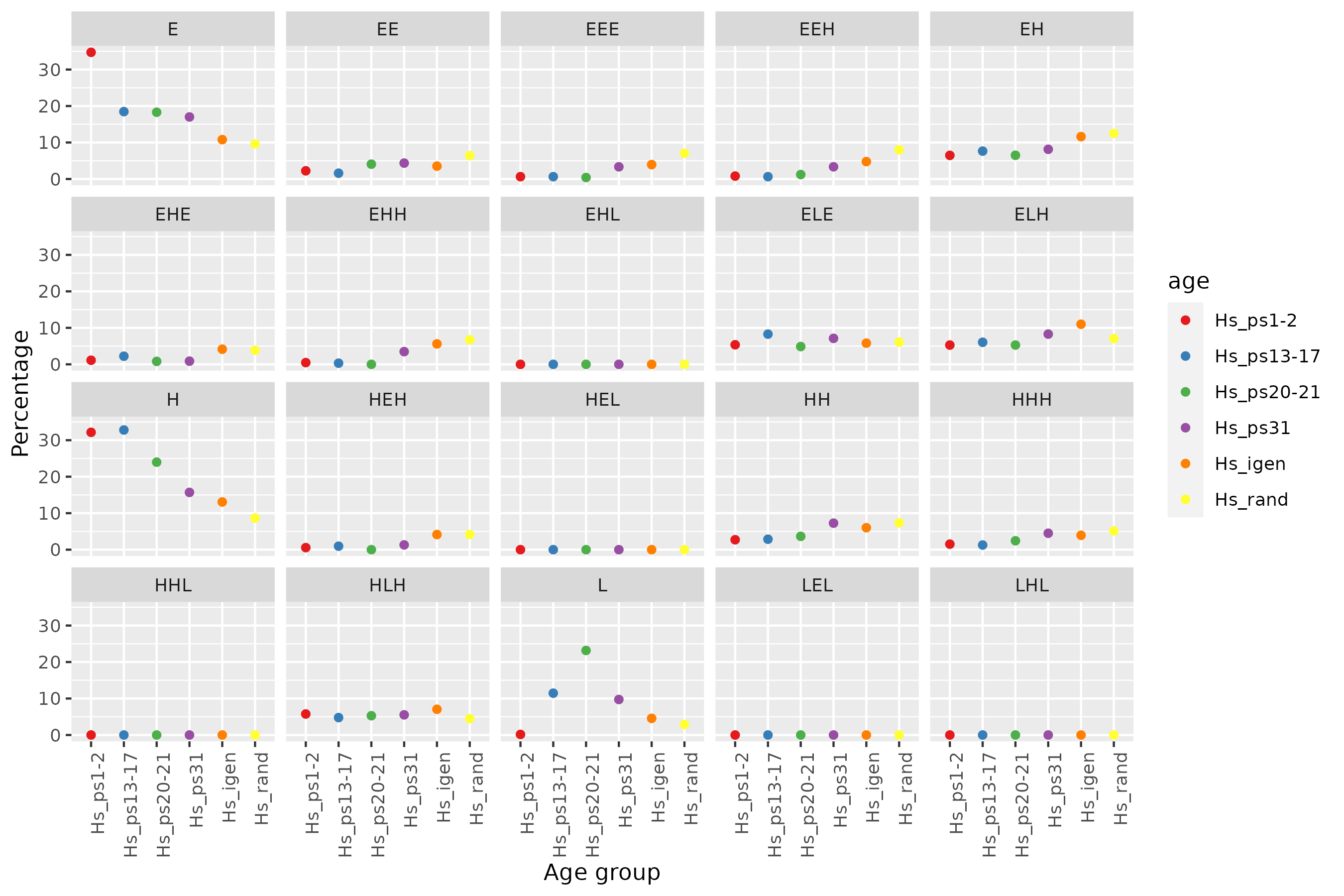
**Table 2.** Number and percentage of proteins in the different age groups with simple words.

|  |  |  |  |
| --- | --- | --- | --- |
| **age** | **# proteins** | **# proteins with simple words** | **%** |
| Hs\_ps1-2 | 2009 | 1250 | 62.22 |
| Hs\_ps13-17 | 581 | 314 | 54.04 |
| Hs\_ps20-21 | 433 | 246 | 56.81 |
| Hs\_ps31 | 1700 | 688 | 40.47 |
| Hs\_igen | 1617 | 482 | 29.81 |
| Hs\_rand | 1658 | 312 | 18.82 |

We have found that none of the “simple words” appears to be banned from real or fake proteins. The yougest group (ps31) is relatively similar to intergenic. The differences go in the direction observed for older proteins (for example excess of H).



**Figure 5. Short word occurrences in different sequence sets.** The size of the words is relative to their frequency.



**Figure 6.** Short word combination frequency (percentage) by age group.

**Over-represented words**

Next we decide to focus on words of sizes 3 to 5 with at least 10 or more occurrences. The list of words, and frequencies in the different protein groups, is in the file “word\_occurrences.xlsx”. We cannot perform statistical tests on all the words because this would involve massive multiple test correction and we would not be able to detect any trends. This happens because the protein groups are relatively small and the number of words high, so many cells have low counts and our statistical power is quite limited.

However, focusing on the words that are more frequent, repeated patterns, and cases with p-value < 0.01, we can conclude the following:

1. There is an under-representation of EEE in old proteins versus young proteins

The following words are under-represented in ps1-2 with respect to ps31:

EEE

ELEEE

2. Under-representation of EEEE in young proteins versus random proteins

The frequency of EEEE is very low in ps1-2, ps31, but increases in intergenic and random. The difference between ps31 and random is highly significant (p-value = 0.0001615). The latter effect can also be seen, to some extent, for EEE.

3. Several words that contain HHLH/HLHH are over-represented in old proteins versus young proteins

HHLH

HLHH

EHHLH

4. Other cases

Of the cases tested, only two other words show significant differences:

HLH is significantly over-represented in ps1-2 with respect to ps31

HHE is significantly over-represented in ps31 with respect to ps1-2

These cases may represent false positives.

**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