|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Length** | **Regular DFT**  | **Restrictive DFT (Uniform, Unigram)** | **+ FT selection****by intervals**(Threshold set by Unigram) |
| TITIN\_HUMAN | 34350 | 2375  | 413 (2, 59) | 1642  |
| MUC16\_HUMAN | 22152 | 1032  | 530 (2, 38) | 927  |
| SYNE1\_HUMAN | 8797 | 389  | 98 (1, 15) | 117  |
| OBSCN\_HUMAN | 7968 | 397  | 171 (1, 13) | 265  |
| DYST\_HUMAN | 7570 | 292  | 69 (1, 11) | 159  |
| MACF1\_HUMAN | 7388 | 288  | 70 (1, 12) | 166  |
| FSIP2\_HUMAN | 6907 | 238  | 59 (0, 11) | 106  |
| SYNE2\_HUMAN | 6885 | 262  | 75 (0, 10) | 133  |
| NEBU\_HUMAN | 6669 | 351  | 238 (0, 11) | 342  |
| GPR98\_HUMAN | 6306 | 207  | 46 (0, 10) | 84  |

**Table S2: Characteristics of the 10 longest human proteins**

Characteristics of the 10 longest human proteins, displaying their length, assigned DFT counts according to the ***regular*** and ***restrictive*** definitions, and error estimates on the latter, given in ‘()’, for the uniform and unigram random models respectively. In the unigram model the inhomogeneous drop with decreasing length is due to statistical fluctuations. Last column shows that amendment due to periodicity considerations (see figure S5) increases considerably the number of validated FTs.