Summary of Human Protein Data (2020-10-13)

Data tabulated (Oct 14 PDB Release) for human (taxId=9606) or Humanized antibody proteins are annotated as human taxonomy (taxId=9606).

```
Total protein entity count: 70856  (human taxonomy taxId=9606)
Unique reference sequence (UniProt) assignments: 7105
Protein entities without reference sequence assignments: 4876
```

Protein entities with multiple distinct taxonomies have been excluded. Proteins with multiple human taxonomy reference sequence assignments (e.g., 2 UniProt ID within human taxonomy scope) are NOT excluded. Some artifacts in which PDB taxonomy assignment differs from the SIFTS reassignment have been excluded.

```
Multi-taxonomy entities: 1106 (excluded)
Bad SIFTS assignments (switched taxonomies): 112 (excluded)
```

Clustering performed on all protein polymer entities. The following table gives the essential statistics for the clustering solution.

% Sequence Identity	Distinct Groups in the Cluster
30	36477
40	41428
50	45863
70	53376
90	61734
95	65719
100	94393

For the current human protein entity cohort, the populated cluster statistics are summarized below.

% Sequence Identity	Cluster Groups Containing a Human Protein
100	21814
95	14392
90	13091

Protein Entity-level Data Sets

Data files containing the leading human protein representative of each cluster group are attached in `100-first-human-entity-abbrev.csv`, `95-first-human-entity-abbrev.csv`, and `90-first-`

`human-entity-abbrev.csv` for sequence identities 100, 95, and 90, respectively. Some cluster groups have been extensively annotated by UniProt and these groups may contain multiple reference sequence assignments. A separate collection of file is provided which contains the leading representative of each cluster as well as the leading representative for each reference sequence assignment. These file are attached for each level of sequence identity in `100-first-human-entity-full.csv`, `95-first-human-entity-full.csv`, and `90-first-human-entity-full.csv`. These files are ordered by the size of the cluster group largest to smallest. To look at the leading examples chronologically, sort the data set by PDB release year column.

Column Name	Description
Cluster_ID	identifer for the cluster group
Cluster_Members_Total	total number of polymer entities in the cluster group
Cluster_Members_Human	number of human protein members in the cluster group
Cluster_UniProt_Count	numer of distinct reference sequences assigned in the cluster group
PDB_Entity_ID	PDB Entity ID for the leading human protein in the cluster group. If multiple reference sequences are assigned to the cluster group, the leading Entity ID is for each reference sequence is given in files labeled `full`. Files labeled `abbrev` contain only the leading entity in each cluster group.
PDB_Release_Year	release year for the entry containing the protein entity
UniProt_IDs	reference sequence assignments in the cluster group
Assign_Count	count of reference sequences assigned in the cluster group
PDB_Struct_title	PDB entry structure title
PDB_Struct_Descr	PDB entry structure description
PDB_Entity_Descr	PDB entity description
UniProt_Name	UniProt recommended protein name
Uniprot_Gene	UniProt primary gene name

Release statistics for Human Proteins

Counts of entries containing human taxonomy proteins released by year are summarized in the following table. These data are include in attached file `human-containing-entries-by-year.csv`.

Year	Entries_Containing_Human_Proteins
1976	2
1977	1
1979	2
1981	2
1982	1

Year	Entries_Containing_Human_Proteins
1983	3
1984	4
1985	1
1987	1
1988	2
1989	3
1990	17
1991	10
1992	46
1993	110
1994	235
1995	178
1996	273
1997	273
1998	426
1999	505
2000	538
2001	554
2002	614
2003	832
2004	1095
2005	1423
2006	1814
2007	1974
2008	1674
2009	1857
2010	2006
2011	2132
2012	2470
2013	2655

Year	Entries_Containing_Human_Proteins
2014	2722
2015	2842
2016	3421
2017	4268
2018	4017
2019	3832
2020	3588

Release Statistics for Leading Examples of Human Proteins

Counts of entries containing the leading example of a human protein entity by year are summarized in the following tables cluster sequence identity 100, 95 and 90 percent. These data are included in attached files `100-leading-human-containing-entities-by-year.csv`, `95-leading-human-containing-entities-by-year.csv`, and `90-leading-human-containing-entities-by-year.csv`.

Sequence Identity 100%

Year	Entries_With_Leading_Human_Protein
1976	2
1977	1
1979	1
1981	1
1982	1
1983	2
1984	1
1985	1
1987	1
1988	2
1989	3
1990	9
1991	7
1992	31
1993	66
1994	98

Year	Entries_With_Leading_Human_Protein
1995	90
1996	88
1997	160
1998	236
1999	276
2000	314
2001	318
2002	291
2003	383
2004	529
2005	821
2006	979
2007	1120
2008	800
2009	784
2010	798
2011	816
2012	870
2013	939
2014	1012
2015	953
2016	1157
2017	1127
2018	1162
2019	1259
2020	1120

Sequence Identity 95%

Year	Entries_With_Leading_Human_Protein
1976	2

Year	Entries_With_Leading_Human_Protein
1977	1
1979	1
1981	1
1983	2
1984	1
1985	1
1987	1
1988	2
1989	3
1990	8
1991	7
1992	14
1993	37
1994	65
1995	65
1996	61
1997	107
1998	148
1999	192
2000	202
2001	205
2002	201
2003	258
2004	380
2005	689
2006	750
2007	884
2008	568
2009	528
2010	547

Year	Entries_With_Leading_Human_Protein
2011	534
2012	546
2013	562
2014	593
2015	579
2016	671
2017	712
2018	699
2019	788
2020	738

Sequence Identity 90%

Year	Entries_With_Leading_Human_Protein
1976	2
1977	1
1979	1
1981	1
1983	2
1984	1
1985	1
1987	1
1988	2
1989	3
1990	8
1991	7
1992	13
1993	36
1994	64
1995	62
1996	60

Year	Entries_With_Leading_Human_Protein
1997	105
1998	142
1999	187
2000	199
2001	201
2002	200
2003	253
2004	365
2005	678
2006	725
2007	855
2008	540
2009	507
2010	516
2011	497
2012	512
2013	519
2014	525
2015	520
2016	600
2017	625
2018	601
2019	712
2020	630