Coverage of protein domain families with structural protein-protein interactions: current progress and future trends

Alexander Goncearenco, Benjamin A. Shoemaker, Dachuan Zhang, Alexey Sarychev and Anna R. Panchenko*

Computational Biology Branch of the National Center for Biotechnology Information in Bethesda, Maryland, United States of America

* panch@ncbi.nlm.nih.gov

Supplementary Information

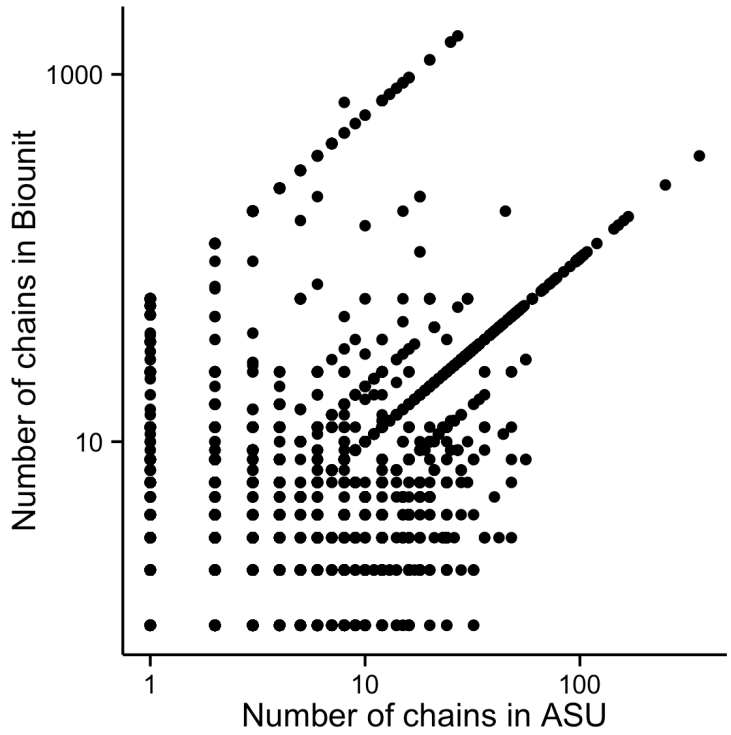

Figure S1. The correspondence between the number of chains in the asymmetric unit (ASU) and in the biological assembly (Biounit). Each point represents a structure in MMDB/PDB database.

Table S1. Estimates of structural and PPI coverage yearly growth rates and the yearly database growth rates for families and superfamilies in CDD and Pfam.

A. Structural coverage is the numbers of families having at least one representative struture deposited by a given year. PPI coverage is the is the number of families having at least one structure with at least one protein-binding site. We consider the families and superfamilies in CDD 3.11 and families in Pfam 27 and take the numbers of structural and PPI coverage for the years of 2012 and 2007. Based on the increase of coverage in this period we estimate the yearly growth rate. We assume that the coverage growth is linear and the growth rate will not change in the future. For this estimate we assume that the number of families and superfamilies is constant. We take the most recent coverage (in year 2014 for CDD and year 2013 for Pfam) and estimate the projected coverage in the year of 2012 given the current coverage growth rates.

Database		Type of coverage	2012	2007	Yearly increase in coverage	Current coverage	Projected coverage in 2020
CDD	**Superfamilies**	**Structural**	918	765	30.6	927	1110.6
CDD	**Superfamilies**	**PPI**	789	648	28.2	794	963.2
CDD	**Families**	**Structural**	4657	3568	217.8	4750	6056.8
CDD	**Families**	**PPI**	2841	2057	156.8	2912	3852.8
PFAM	**Families**	**Structural**	5749	4700	209.8	5749	7217.6
PFAM	**Families**	**PPI**	3740	2936	160.8	3740	4865.6

B. Yearly increment in the number of (super)families covered by strucrtures or PPI interactions by strucrures deposited during the year, starting with year 2000.

Year	Pfam families, structural coverage	Pfam families, PPI coverage	CDD families, structural coverage	CDD families, PPI coverage	CDD superfam. structural coverage	CDD superfam. PPI coverage
2000	322	208	253	161	63	52
2001	352	218	237	140	50	40
2002	305	182	206	104	28	29
2003	400	238	269	149	29	27
2004	466	276	344	181	24	28
2005	445	269	415	193	26	32
2006	421	301	359	198	23	35
2007	430	294	320	210	19	30
2008	297	202	203	154	15	17
2009	274	228	290	205	27	38
2010	259	197	222	159	19	18
2011	172	153	202	135	41	31
2012	47	24	172	131	29	37
2013			85	69	8	5
2014			8	2	1	

C. Increment in the number of families and superfamilies defined in different releases in CDD and Pfam. For CDD we take the average number of new (super)families (listed in section D below) between the years 2009 and 2013 as the yearly growth rate. For Pfam we take the difference in size of the database between years 2008 and 2013 and calculate the average yearly growth rate. Based on the number of families in the latest versions of the databases we estimate the projected database sizes in the year of 2020.

Database		Yearly database growth rate	Current database size	Projected database size in 2020
CDD	**Superfamilies**	75	1060	1510
CDD	**Families**	1220.4	9860	17182.4
PFAM	**Families**	898.2	14831	21118.4

D. The number of CDD families and superfamilies added every year between 2008 and 2014. The number includes singleton superfamilies and only counts the models currently marked as active and the superfamilies with at least one NCBI-curated model.

Year	New superfamilies	New families
2008	556	752
2009	40	673
2010	36	1623
2011	111	1498
2012	85	1045
2013	103	1263
2014	83	