Beyond History and “on a Roll”: The List of the most Well-Studied Human Protein Structures and Overall Trends in the Protein Data Bank

Zhen-lu Li[1*] and Matthias Buck[1,2*]

[1]Department of Physiology and Biophysics, Case Western Reserve University, School of Medicine, 10900 Euclid Avenue, Cleveland, Ohio 44106, U. S. A. [2]Department of Pharmacology; Department of Neurosciences and Case Comprehensive Cancer Center, Case Western Reserve University, School of Medicine, 10900 Euclid Avenue, Cleveland, Ohio 44106, U. S. A.

- *Corresponding authors: Z.Li (zhenlu.li@case.edu) and M.Buck (matthias.buck@case.edu)*
- ORCHID IDs: Z. Li: 0000-0003-2101-8237 M. Buck: 0000-0002-2958-0403

Of the roughly 20,000 canonical human protein sequences, as of September 15, 2020, 6,937 proteins have had their full or partial, medium- to high-resolution structures determined by x-ray crystallography or other methods. Which of these proteins dominate the Protein Data Bank (the PDB) and why? In this paper, we list the 273 top human protein structures based on the number of their PDB entries. This set of proteins accounts for more than 40% of all available human PDB entries and represent past trends as well as current status for protein structural biology. We briefly discuss the relationship which some of the prominent protein structures have with protein research as a whole and mention their relevance to human diseases. The top-10 soluble and membrane proteins are all well-known (most of their first structures being deposited more than 30 years ago). Overall, there is no dramatic change in recent trends in the PDB. Remarkably, the number of structure depositions has grown nearly exponentially over the last 10 or more years (with a doubling time of 7 yrs for proteins from all organisms). Growth in human protein structures is slightly faster (at 5.9 yrs, while E.Coli and Mouse+Rat protein structures accumulate more slowly, Zebrafish protein structures are growing most, at a doubling every 3.7 years, albeit starting from only approx. 100 structure entries in 2010). The information may be informative to senior scientists but also inspire researchers who are new to protein science, providing the year 2020 snap-shot for the state of protein structural biology.

Keywords: structural biology; human disease; cancer; protein kinase; human membrane proteins, protein structures of model organisms

At currently 28%, human proteins comprise a significant fraction of all entries in the Protein Data Bank (the PDB) and a small number of proteins stand out among the human proteins: The three- dimensional structure/conformation of the polypeptide chain determines the dynamics, and then the function of an individual protein. Proteins -excluding many intrinsically disordered proteins (IDPs)- typically have at least one natively folded conformation. Structural biology techniques, principally X-ray crystallography, NMR spectroscopy and recently Cryogenic Electron Microscopy (cryo-EM) have allowed us to obtain protein structures with increased resolution and efficiency over the years.[1,2] Such structures are deposited in the Protein Data Bank, the PDB, and in 2019 alone, 10,585 protein structures were released.[3] As of September 15, 2020, there were 168,599 structures in the PDB (counting all structures from the wide range of source organisms). There are 46,795 entries for human proteins in the PDB, accounting for 28% of total entries. This number of proteins is larger than the number of human canonical proteins, as many PDB entries are the same protein with point mutation(s) and/or bound to different ligands, ranging from small molecule inhibitors to protein or other macromolecular binding partners.

The number of human protein-coding genes is estimated to be around 20,000, which would result in the same number of full length, non-modified proteins.[4] However, the real number of proteins in the human proteome increases dramatically as a consequence of alternative splicing, single amino acid polymorphisms between chromosomes and especially due to posttranslational modifications.[5] On September 15, 2020, there were 20,375 reviewed entries of full length human proteins in the Uniprot database, which is easily accessed via a web server (http://uniprot.org).[6] The analysis, based on the 20,375 proteins, indicates that the median sequence length of human proteins is 325 amino acids (Fig. 1a).

Since the launch of UniProt in 2003, the database has gathered protein sequence, structure and function information for individual protein species.[6] With respect to structure, UniProt also gives information on available PDB entries for each protein. We downloaded the database from the UniProt webserver and extracted entries for human proteins by searching with the keyword, “homo sapiens” (see a Github link at Code Availability section below). Only reviewed human protein entries (the set of 20,375 canonical proteins) were selected. From the database of these reviewed proteins, we extracted the PDB entries of each protein and sorted them by the total number of available PDB entries. The top 200 human proteins, counting the number of the PDB entries, are given in Table 1.

Separately, we also list the 100 membrane proteins with the most entries in Table 2. This list was compiled by parsing the UniProt database as above, but searching for the keyword "membrane protein". It does include many proteins which cross the membrane only once, e.g. EGFR with a single transmembrane helix, and also includes very few membrane peripheral proteins, such as the Estrogen Receptor alpha, isoform 3. Since 27 membrane proteins already appear in Table 1, in total only 273 unique proteins are listed in Table 1 and Table 2.

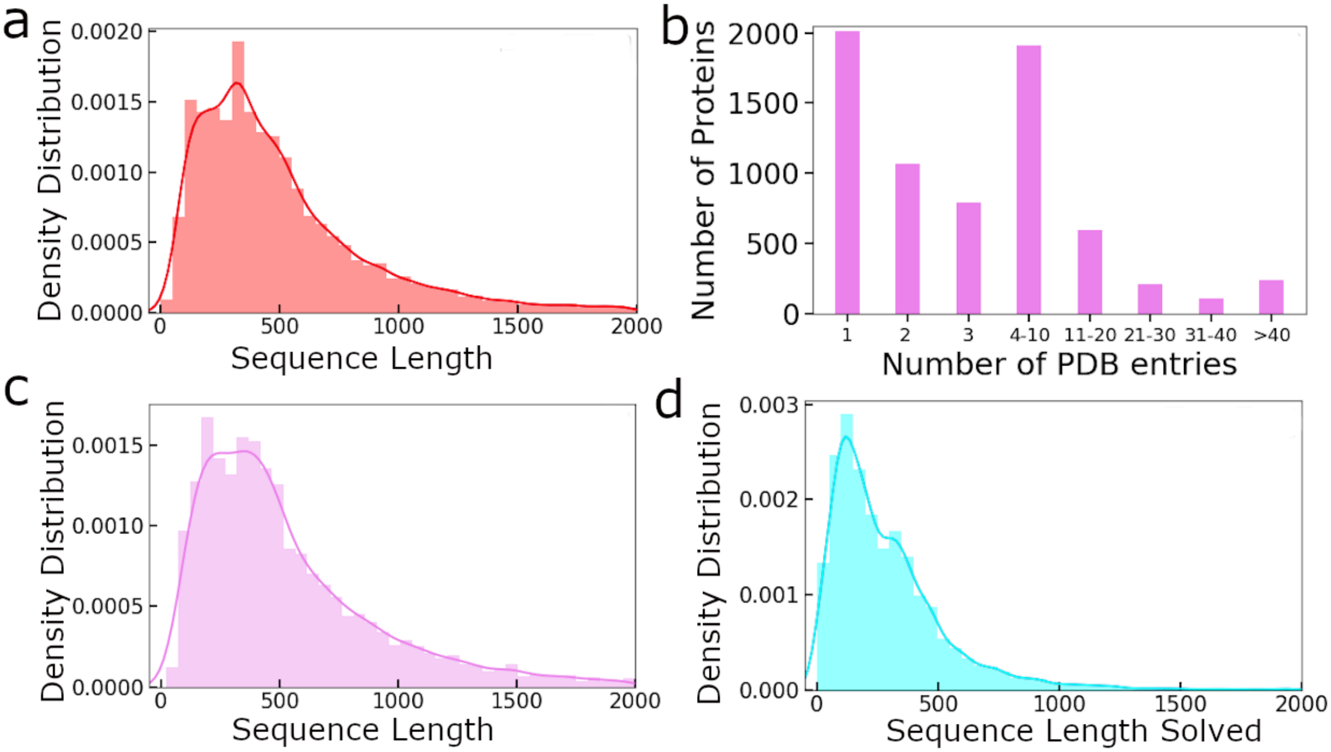

Figure 1: Protein sequence and structure statistics. (a) Distribution of sequence lengths for 20,375 human proteins. (b) PDB statistics as of September 15, 2020. The number of human proteins with 1, 2, 3 and more PDB entries. Distributions of (c) sequence lengths for 6,937 human proteins with at least one PDB entry and (d) of the actual length of solved for 6,937 human proteins. In (a), and (c)/(d) the x-axis is binned in increments of 50 and the y-axis are % of proteins in each bin, versus the total count of entries of 20375 and 6937, respectively. A line is drawn through the histogram y-values by interpolation.

We found 6,937 out of the 20,375 distinct human proteins have at least one PDB entry. This number includes structures of fragments or domains, as the full-length structures are not yet solved for certain types of proteins, such as the great majority of single-pass membrane crossing receptors. In the case of a human protein in a protein-protein complex, the human segment bound may be very small, e.g. a peptide and the partner protein may not be human (e.g. in case of interactions with microorganisms). The distribution of available PDB entries per non-redundant protein is plotted in Figure 1b. Amongst these human proteins, 2,014, 1,066 and 788 proteins, have only one, two or three PDB entries respectively (together, then 56% of the 6,937 proteins have only 1-3 entries). However, at the other extreme, the 200 human proteins with the most entries (3% of 6,937) have 19,998 cumulative PDB entries, remarkably counting for around 40% of total human PDB entries. Thus, the top-200 human proteins have gathered an unusually high proportion of attention compared to the rest. In the meantime, at least 2/3rds of the structures of the human proteome remain to be determined. It is interesting to note that the distribution of lengths of proteins which have been solved (Fig. 1c) is similar in profile to the length of all human proteins. Therefore, there seems to be no preference, as far as the length of proteins is concerned, whether their structures can be determined or not. Figure 1d shows the residue length of structures actually determined, yielding a similar profile to Fig. 1c. This suggests that the shorter fragments/domains as mentioned above are not so numerous and do not significantly skew the distribution. Overall, the high frequency appearance of proteins in the PDB arises from the biological importance that they have in cellular processes, in human diseases but some also - but to an increasingly lesser extent- from their use as model systems for our understanding of protein structure and function. Below we comment on some of the most highly represented structures which have emerged, also making a note of the early history of protein structural biology.

Table 1: Top 200 proteins with the most entries in the PDB as of Sept. 15, 2020, listing common protein name, rank [1-200] (R), and number of PDB entries (N). All the PDB IDs are given in the supplement on Github in their order of listing in UniProt. References to the original papers describing these structures are given in the PDB entries.

Protein Name	N	R	Protein Name	N	R	Protein Name	N	R
Acetylcholinesterase	45	195	Cellular tumor antigen p53	185	31	Farnesyl pyrophosphate synthase	88	77
Adenosine receptor A2a	51	161	Cholinesterase	68	111	Ferritin heavy chain	58	140
ADP-sugar pyrophosphatase	62	127	Coagulation factor VII	108	56	Fibrinogen gamma chain	45	194
Aldo-keto reductase family 1 member B1	145	44	Coagulation factor X	146	42	Fibroblast growth factor 1	97	66
Aldo-keto reductase family 1 member C3	47	180	Coagulation factor XI	88	78	Fibroblast growth factor receptor 1	66	116
ALK tyrosine kinase receptor	61	131	Collagenase 3	48	174	Fibronectin	60	136
Amine oxidase [flavin-containing] B	47	181	Complement C3	47	178	Galectin-3	78	96
Amyloid-beta precursor protein	145	43	Complement factor H	46	188	Gelsolin	58	139
Androgen receptor	82	89	CREB-binding protein	96	68	Glutamate carboxypeptidase 2	77	97
Angiogenin	46	189	Cyclin-A2	94	71	Glutathione S-transferase P	65	118
Angiotensin-converting enzyme	47	179	Cyclin-dependent kinase 2	412	3	Glycogen synthase kinase-3 beta	87	80
ATPase family AAA domain-containing protein 2	85	82	Death-associated protein kinase 1	58	141	Growth factor receptor-bound protein 2	50	169
Aurora kinase A	155	37	Deoxycytidine kinase	47	176	GTP-binding nuclear protein Ran	79	93
B-cell lymphoma 6 protein	44	196	Deoxynucleoside triphosphate triphosphohydrolase SAMHD1	47	177	GTPase HRas	178	33
Bcl-2-like protein 1	80	92	Dihydrofolate reductase	79	94	GTPase KRas	165	35
Beta-2-microglobulin	828	2	Dihydroorotate dehydrogenase (quinone), mitochondrial	73	102	Guanine nucleotide-binding protein G(I)/G(S)/G(O) subunit gamma-2	46	185
Beta-secretase 1	391	5	Dipeptidyl peptidase 4	104	60	Guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-1	45	193
Bifunctional epoxide hydrolase 2	102	62	DNA cross-link repair 1A protein	312	10	Heat shock protein HSP 90-alpha	300	12
Bile acid receptor	83	85	DNA damage-binding protein 1	57	144	Hemoglobin subunit alpha	284	15
Bromodomain adjacent to zinc finger domain protein 2B	262	20	DNA polymerase beta	370	6	Hemoglobin subunit beta	278	18
Bromodomain-containing protein 1	311	11	DNA polymerase eta	129	48	Hepatocyte growth factor receptor	85	81
Bromodomain-containing protein 2	83	84	DNA polymerase iota	46	187	High affinity nerve growth factor receptor	51	160
Bromodomain-containing protein 4	357	7	DNA polymerase lambda	58	142	Histidine triad nucleotide-binding protein 1	49	173
Calmodulin-1	173	34	DNA-directed DNA/RNA polymerase mu	70	108	Histo-blood group ABO system transferase	151	38
cAMP and cAMP-inhibited cGMP 3',5'-cyclic phosphodiesterase 10A	96	67	Dual specificity mitogen-activated protein kinase 1	46	186	Histone deacetylase 8	50	168
cAMP-dependent protein kinase inhibitor alpha	103	61	Dual specificity protein kinase TTK	71	106	Histone H2A type 1-B/E	107	57
cAMP-specific 3',5'-cyclic phosphodiesterase 4D	82	86	E3 ubiquitin-protein ligase Mdm2	110	55	Histone H2B type 1-J	104	59
Carbonic anhydrase 2	837	1	E3 ubiquitin-protein ligase XIAP	66	117	Histone H3.1	258	21
Casein kinase II subunit alpha	181	32	Elongin-B	73	101	Histone H3.3	62	126
Caspase-3	100	65	Elongin-C	69	109	Histone H4	200	30
Cathepsin K	61	130	Ephrin type-A receptor 2	77	98	HLA class I histocompatibility antigen, A alpha chain	352	8
Cathepsin S	55	151	Epidermal growth factor receptor	213	26	HLA class I histocompatibility antigen, B alpha chain	212	27
Cellular retinoic acid-binding protein 2	72	103	Estrogen receptor	294	14	HLA class II histocompatibility antigen, DR alpha chain	101	63

HLA class II histocompatibility antigen, DRB1 beta chain	88	76		Nuclear receptor coactivator 2	297	13		Serotransferrin	47	175
Hypoxia-inducible factor 1-alpha inhibitor	44	200		Nuclear receptor ROR-gamma	110	54		Serum albumin	114	51
Immunoglobulin gamma-1 heavy chain	46	184		Pancreatic alpha-amylase	50	166		Small ubiquitin-related modifier 1	50	162
Immunoglobulin heavy constant gamma 1	201	29		Peptidyl-prolyl cis-trans isomerase A	135	45		Son of sevenless homolog 1	60	134
Immunoglobulin kappa constant	90	73		Peptidyl-prolyl cis-trans isomerase F, mitochondrial	49	172		Superoxide dismutase [Cu-Zn]	117	50
Induced myeloid leukemia cell differentiation protein Mcl-1	100	64		Peptidyl-prolyl cis-trans isomerase FKBP1A	51	157		T cell receptor alpha constant	133	46
Insulin	278	17		Peptidyl-prolyl cis-trans isomerase NIMA-interacting 1	82	87		T cell receptor beta constant 1	87	79
Insulin-degrading enzyme	53	153		Peregrin	62	123		T cell receptor beta constant 2	61	128
Integrin beta-3	73	100		Peroxisome proliferator-activated receptor delta	44	198		T-box transcription factor T	46	182
Interleukin-1 beta	56	148		Peroxisome proliferator-activated receptor gamma	224	24		T-cell surface glycoprotein CD4	66	115
Interleukin-1 receptor-associated kinase 4	51	159		Phosphatidylinositol 3-kinase regulatory subunit alpha	60	135		Thymidylate synthase	60	133
Kinesin-like protein KIF11	58	138		Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform	51	156		Tissue factor	49	171
Leukotriene A-4 hydrolase	62	125		Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit gamma isoform	95	69		Titin	45	190
Lysine-specific demethylase 4A	82	88		Poly [ADP-ribose] polymerase 1	63	120		Transthyretin	327	9
Lysine-specific demethylase 4D	280	16		Poly [ADP-ribose] polymerase tankyrase-2	146	41		Tyrosine-protein kinase ABL1	66	114
Lysine-specific demethylase 5A	44	199		Polyubiquitin-B	146	40		Tyrosine-protein kinase BTK	81	90
Lysine-specific demethylase 5B	56	147		Polyubiquitin-C	226	23		Tyrosine-protein kinase JAK2	93	72
Lysine-specific histone demethylase 1A	71	105		Proteasome subunit alpha type-3	44	195		Tyrosine-protein kinase Lck	56	146
Lysozyme C	208	28		Proteasome subunit beta type-1	68	110		Tyrosine-protein kinase SYK	71	104
Macrophage metalloelastase	83	83		Proteasome subunit beta type-5	50	165		Tyrosine-protein phosphatase non-receptor type 1	275	19
Macrophage migration inhibitory factor	94	70		Proteasome subunit beta type-7	51	155		Tyrosine-protein phosphatase non-receptor type 11	60	132
Major histocompatibility complex class I-related gene protein	45	192		Protein/nucleic acid deglycase DJ-1	61	129		U1 small nuclear ribonucleoprotein A	78	95
Major prion protein	57	143		Prothrombin	392	4		Ubiquitin carboxyl-terminal hydrolase 7	56	145
Mediator of RNA polymerase II transcription subunit 1	50	167		Proto-oncogene tyrosine-protein kinase Src	64	119		Ubiquitin-40S ribosomal protein S27a	55	150
Microtubule-associated protein tau	62	124		Ras-related C3 botulinum toxin substrate 1	50	163		Ubiquitin-60S ribosomal protein L40	58	137
Mitogen-activated protein kinase 1	112	52		Renin	88	75		Urokinase-type plasminogen activator	146	39
Mitogen-activated protein kinase 10	51	158		REST corepressor 1	50	164		Vascular endothelial growth factor receptor 2	52	154
Mitogen-activated protein kinase 14	240	22		Ribosyldihydronicotinamide dehydrogenase [quinone]	67	113		Vitamin D3 receptor	49	170
Neutrophil gelatinase-associated lipocalin	53	152		Retinol-binding protein 2	45	191		von Hippel-Lindau disease tumor suppressor	55	149
Nicotinamide phosphoribosyltransferase	63	121		Retinoic acid receptor RXR-alpha	89	74		WD repeat-containing protein 5	111	53
Nitric oxide synthase, brain	73	99		Serine/threonine-protein kinase B-raf	80	91		14-3-3 protein sigma	106	58
Nitric oxide synthase, endothelial	46	183		Serine/threonine-protein kinase Chk1	133	47		3-phosphoinositide-dependent protein kinase 1	68	112
Nuclear autoantigen Sp-100	120	49		Serine/threonine-protein kinase pim-1	158	36		7,8-dihydro-8-oxoguanine triphosphatase	71	107
Nuclear receptor coactivator 1	224	25		Serine/threonine-protein kinase PLK1	62	122				

Table 2: Same as Table 1, but for the Top 100 membrane proteins with the most entries in the PDB as of Sept. 15, 2020.

Protein Name	N	R	Protein Name	N	R	Protein Name	N	R
Activin receptor type-1	21	68	Ephrin type-A receptor 3	28	47	Low affinity immunoglobulin epsilon Fc receptor	23	58
Adenosine receptor A2a	51	23	Ephrin type-B receptor 4	23	59	Low-density lipoprotein receptor	33	40
Advanced glycosylation end product-specific receptor	22	63	Epidermal growth factor receptor	213	4	Low-density lipoprotein receptor-related protein 6	16	94
ALK tyrosine kinase receptor	61	20	Epithelial discoidin domain-containing receptor 1	20	71	Macrophage colony-stimulating factor 1 receptor	20	70
Amine oxidase [flavin-containing] B	47	26	Erythropoietin receptor	18	81	Major histocompatibility complex class I-related gene protein	45	27
Amyloid-beta precursor protein	145	7	Estrogen receptor	294	3	Mast/stem cell growth factor receptor Kit	26	50
ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase 1	42	31	Fibroblast growth factor receptor 1	66	19	Melanoma antigen recognized by T-cells 1	22	61
Angiotensin-converting enzyme	47	25	Fibroblast growth factor receptor 2	43	29	Neurogenic locus notch homolog protein 1	24	53
Angiotensin-converting enzyme 2	22	62	Fibroblast growth factor receptor 4	29	45	Neprilysin	18	79
Apoptosis regulator BAX	24	56	Furin	22	66	Neuropilin-1	17	85
Apoptosis regulator Bcl-2	27	49	Glucagon-like peptide 1 receptor	16	97	Platelet glycoprotein Ib alpha chain	19	75
Bcl-2 homologous antagonist/killer	25	51	Glutamate receptor ionotropic, NMDA 1	17	90	Potassium channel subfamily K member 9	16	93
Bcl-2-like protein 1	80	13	Glutamate carboxypeptidase 2	77	14	Programmed cell death 1 ligand 1	32	41
Beta-1, 4-galactosyltransferase 1	19	78	Hepatocyte growth factor receptor	85	12	Programmed cell death protein 1	18	78
Beta-2 adrenergic receptor	35	37	High affinity nerve growth factor receptor	51	22	Prostaglandin E synthase	16	92
Beta-secretase 1	391	1	Histo-blood group ABO system transferase	151	6	Proto-oncogene tyrosine-protein kinase receptor Ret	28	46
C-C chemokine receptor type 5	18	83	HLA class I histocompatibility antigen, A alpha chain	352	2	Receptor tyrosine-protein kinase erbB-2	34	39
C-type lectin domain family 4 member K	20	72	HLA class I histocompatibility antigen, B alpha chain	212	5	Receptor-type tyrosine-protein phosphatase gamma	19	74
Cadherin-1	19	77	HLA class II histocompatibility antigen, DR alpha chain	101	9	Sodium-dependent serotonin transporter	19	73
Carbonic anhydrase 9	19	76	HLA class I histocompatibility antigen, alpha chain E	16	96	Squalene synthase	29	44
Carbonic anhydrase 12	24	55	HLA class II histocompatibility antigen, DQ alpha 1 chain	17	89	Stimulator of interferon genes protein	34	38
Cation-independent mannose-6-phosphate receptor	18	82	HLA class II histocompatibility antigen, DRB1 beta chain	88	11	Suppressor of tumorigenicity 14 protein	23	57
CD81 antigen	16	99	IgG receptor FcRn large subunit p51	17	88	Synaptotagmin-1	20	69
Chloride intracellular channel protein 1	15	100	Induced myeloid leukemia cell differentiation protein Mcl-1	100	10	T-cell surface glycoprotein CD1b	16	91
Complement decay-accelerating factor	24	54	Insulin receptor	36	36	T-cell surface glycoprotein CD4	66	18
Copper-transporting ATPase 1	21	67	Insulin-like growth factor 1 receptor	30	43	Tissue factor	49	24
Corticosteroid 11-beta-dehydrogenase isozyme 1	40	32	Integrin alpha-IIb	43	28	TGF-beta receptor type-1	39	33
Coxsackievirus and adenovirus receptor	16	98	Integrin alpha-2	17	86	Toll-like receptor 8	24	52
Cystic fibrosis transmembrane conductance regulator	37	35	Integrin alpha-L	39	34	Tumor necrosis factor	27	48
Cytochrome P450 3A4	43	30	Integrin alpha-M	17	87	Vascular endothelial growth factor receptor 2	52	21
Dihydroorotate dehydrogenase (quinone), mitochondrial	73	17	Integrin alpha-V	32	42	voltage-dependent L-type calcium channel subunit alpha-1C	17	84
Dipeptidyl peptidase 4	104	8	Integrin beta-2	21	66	3-hydroxy-3-methylglutaryl-coenzyme A reductase	22	64
Disintegrin and metalloproteinase domain-containing protein 17	23	60	Integrin beta-3	73	16			
Ephrin type-A receptor 2	77	15	Leukotriene C4 synthase	16	95			

Table 3: Top 10 Human Protein Structures in the PDB by the number of entries as of Sept. 15, 2020, also giving year in which the first structure was published. And top 10 Genes, adapted from Dolgin, ref. 35, again with the approximate year the gene was discovered (taken from OMIM Data Base). Note that some of the early structures were not deposited and we are referencing the first deposited and research paper reported (as opposed to the first paper reported structure) here.

Top-10 Proteins	**Gene Name**	**PDB ID and Reference**	**Year**		**Top-10 Genes (as of 2017)**	**Year**
Carbonic anhydrase 2	CA2	4CAC;5CAC[18]	1988		TP53	1979
beta2-microglobulin	B2M	1HLA[19]	1987		TNF	1984
cyclic dependent kinase 2	CDK2	1FIN[20]	1995		EGFR	1976
Prothrombin	F2	1PPB[21]	1989		VEGFA	1989
Beta-secretase 1	BACE1	1FKN[22]	2000		APOE	1982
DNA polymerase beta	POLL	1ZQE etc[23]	1996		IL6	1986
Bromodomain containing protein 4	BRD4	2NNU[24]	2006		TGFB1	1985
HLA Class 1 histocompatibility antigen	HLA-A	1HLA[19]	1987		MTHFR	1998
Transthyretin	TTR	2PAB[25]	1978		ESR1	1986
DNA cross-link repair protein 1	DCLRE1A	4B87[26]	2012		AKT1	1987

The PDB and Progress in Structure Determination: The Protein Data Bank has seen over the last 20 years a huge increase in the number of deposited structures, in 2000 the number of total PDB entries was around 13,500 (from all source organisms). On Sep.15th, 2020, this number was 168,599. Before 1990, prior to the advent of efficient recombinant DNA/protein expression technology, most proteins were purified from natural sources. While there was a preference for working with human proteins, PDB entries were outnumbered by proteins which could be obtained and then crystallized from non-human sources, indeed, from a wide variety of organisms. For example, the earliest structure determination of hemoglobin and myoglobin were sourced from horse and whale respectively.[7,8] Over the last several decades three structural biology techniques have contributed to the vast majority of structures in Protein Data Bank. On Sept.15th, there were 149,586 and 13,120 protein structures, respectively, determined by x-ray crystallography, by solution and recently solid-state NMR, and 5903 by cryo-EM (a rapidly increasing number in the last 5 years).[1] Each technique has its well-known limitations, some of which are overcome by continued technical developments,[9] the spectacular advances in cryo-EM,[10,11] with microcrystal-electron diffraction (MicroED) and serial femtosecond crystallography (SFX) emerging for a time/ensemble resolution of structures.[12] Nevertheless, even with the advent of room temperature measurements, these techniques will likely remain most suitable for protein domains/whole length proteins whose structures are amenable to "freezing" in a few conformations/configurations, with electron densities becoming "blurry/invisible" due to higher levels of disorder/dynamics. However, in such instances, especially in cases of intrinsically disordered proteins or for the detection of weak interactions, NMR spectroscopy is the technique of choice.[e.g. 13,14] Recently, integrative computational modeling methods have emerged [15,16] which use a variety of sparse, but in some instances atom or residue specific location restraints – for example from NMR/EPR as well as from various Mass spectrometry techniques[17] in order to derive the structure of protein ensembles and complexes. Thus, for the foreseeable future several experimental methods are needed to give complementary information to the three main structural biology methods, especially when their application is limited, such as in the case of intrinsically disordered proteins/protein regions or protein aggregates/condensates.

The Top 10 List of Aqueous and Membrane Proteins: By far the most structures solved are for Carbonic anhydrase 2,[18] Beta-2-microglobulin,[19] and Cyclin-dependent kinase 2,[20] as the first, second, third place of the most deposited structures in the PDB, with 837, 828 and 412 entries respectively. Prothrombin;[21] Beta-secretase 1,[22] DNA polymerase beta;[23] Bromodomain-containing protein 4;[24] HLA class I histocompatibility antigen, A alpha Chain;[19] Transthyretin;[25] DNA cross-link repair 1A protein,[26] rank 4-10.

The top 10 for membrane proteins are Beta-secretase 1;[22] HLA class I histocompatibility antigen, A alpha chain;[19] Estrogen receptor;[27] Epidermal growth factor receptor;[28] HLA class I histocompatibility antigen, B alpha chain;[29] Histo-blood group ABO system transferase;[30] Amyloid-beta precursor protein;[31] Dipeptidyl peptidase 4;[32] HLA class II histocompatibility antigen, DR alpha chain;[33] Induced myeloid leukemia cell differentiation protein Mcl-1.[34]

Comparison of the most Prominent Human Protein Structures with the most Highly Studied Genes: It should be noted that the method we used to rank the most prominent protein structures is completely different from the approach used in the report by Dolgin in 2017,[35] where the top 10 human genes were identified by counting the frequency of mentions of their gene name in the PubMed database. By contrast, we count the absolute number of available PDB entries for each protein in the UniProt web sever. Using Pubmed entries, especially their MeSH portion for counting of gene names can in some instances be complicated by use of alternative names and by references made to homologues etc. Our method is likely more straightforward for quantification as the UniProt database has well organized information for each protein. Remarkably the two top-10 lists are completely different with not a single protein identical in both lists.

As shown in Table 3, the top-10 genes according to Dolgin, 2017 [35] are (with disease interest given in brackets): TP53 (Cancer), TNF (Cancer), EGFR (Cancer), VEGFA (pathological Angiogenesis), APOE (Alzheimer's), IL6 (Immune), TGFB1 (Cancer), MTHFR (Cancer), ESR1 Estrogen Receptor alpha (Cancer) and AKT1 (Cancer). For comparison, the top-10 list of human protein structures is also listed in Table 3.

Several of the 10 top genes identified in the study of Dolgin, -that is the Cellular tumor antigen p53/TP53, Epidermal growth factor receptor/EGFR, Estrogen receptor/ESR1 and Tumor necrosis factor/TNF- also appear in our lists, but further down at positions 31, 26, and 14 respectively in the top-200 list (and with TNF at position 48 in the list of top-100 membrane proteins). It is noticeable that the majority of proteins whose structures have been determined multiple times are related to cancer (mostly bound to different protein binding partners or small molecule inhibitors in drug screening/design projects) whereas the list of popular genes is more diverse and includes proteins such as oncogenic mutants of TP53 (tumor antigen p53)[36] and regions of the Estrogen Receptor alpha,[37] which have considerable internal dynamics /aggregate or are partially unstructured and have been hard to crystallize. Another factor is that many of the reports on genes are from research on the genetics/genetic linkages, mutations (which are silent at the protein level or occur in introns) and other aspects, driven more by the interest in the genes rather than their protein products.

Protein Classification of the most popular Structures: Assembly of the human proteome in its current form has indicated that there are around 20,000 human genes and correspondingly, at least 20,000 non-modified (canonical) human proteins.[4,5,38,39] Of these non-modified human proteins, noticeable protein groupings include 1,653 metabolic enzymes; 1,089 non-metabolic enzymes such as kinases and GTPases; 1,600 transcription factors; at least 1,555 transporters and channels; and 831 G-Protein Coupled Receptors (GPCRs). By contrast, among the 273 top- protein structures listed in the tables, there are 72 metabolic enzymes, 7 GTP- /ATPases/G proteins, 42 kinases, 16 transcription factors, 6 ion channels and 4 G-protein-coupled receptors (GPCRs). Moreover, there are 10 human leukocyte antigens, 5 histone proteins, 5 bromodomain (BRD) containing proteins, 4 Hormone and Growth factors, 10 cell adhesion molecules, and 6 cystic fibrosis family proteins. The other 86 of 273 proteins are not classified into major protein families, but all of them have important biological functions. Here we comment on several of the families.

The Relevance of Structures to Human Disease: The great majority of the 273 proteins are important for their involvement in human diseases and remain a focus of research, likely for some time to come. Below we briefly list several of the proteins grouped by the relevance to major diseases:

Cancers: Proteins such as p53, Ras GPTases, EGFR, Estrogen receptor, and tumor necrosis factor are crucial proteins either in cancer development and/or metastasis.[36,40-43]
Metabolic disorders: Low-density lipoprotein/and its receptor, Insulin/Insulin receptor regulate the metabolism of carbohydrates or fats, and are major biomarkers for human health.[44,45]

Cardiovascular diseases: Angiotensin-converting enzyme (ACE) controls blood pressure by altering the blood vessels and volume of fluids.[46] Vascular endothelial growth factor receptor as well as galectin-3 are associated with regulation of angiogenesis, vascular development, and heart failure.[47,48]
Neurological disorders: Fibroblast growth factors and 14-3-3-protein are vital factors for neuronal development.[49,50] Amyloid-beta precursor proteins, Microtubule-associated protein tau, the prion protein and transthyretin are associated with the formation of amyloid fibrils.[51-54]
Immunology and infectious diseases: Human leukocyte antigen, T-cell surface glycoprotein CD4 as well as adenosine A2A receptor plays a regulatory role in adaptive immunity.[55-57] Many receptor proteins in the list are host factors for different viruses. Noticeably, Angiotensin-converting enzyme 2 (ACE2) and recently Neuropilin-1 were identified as entry receptors for coronavirus SARS-COV-2.[58,59]

Kinases are among the Best Studied Proteins: The high frequency of appearance of kinases (44 of 273) is remarkable in contrast to its low fraction amongst the 20,000 canonical human proteins (518 of 20,000). Protein kinases are key proteins in cell signaling and are thought to modify up to 30% of all human proteins by tyrosine or serine/threonine phosphorylation.[60,61] Many of them such as Raf kinase, Aurora kinase A, Ephrin type-A receptor 2 and Epidermal growth factor receptor (EGFR) can become easily dysregulated and have a crucial role in diseases, especially in cancer.[62] Clinically, more than 250 kinase inhibitors are undergoing clinical trials and 37 are already approved as therapeutics.[63] Due to this biomedical significance, kinases are one of the most well studied families of human proteins.

Membrane Protein Structures: Membrane proteins represent 20-30% of human proteins. In many earlier reports, it was noted that the membrane proteins are largely underrepresented (only ~2% of all PDB entries) in structure determination by comparison to their number in genomes. This number is inaccurate today, however, especially for human proteins. If we count all peripheral-, transmembrane and integral membrane proteins, 2,340 distinct membrane proteins have at least one structure, corresponding to 34% of all available human protein structures. By counting single-pass and multi-pass transmembrane proteins only, 1,207 of 6,937 (17%) proteins with available structures are membrane proteins. In both cases, this is close to the proportional number of membrane proteins in the human genome. However, it is true that integral membrane proteins such as transporters, ion channels and GPCRs, are still not presented well in the top 200 proteins with the most PDB entries. This is despite the fact that GPCRs for example account for approx. 25-30% of all drug targets.[64] Several proteins of intense research interest are in the top-100 table for membrane proteins and others are catching up. Until recently integral membrane proteins, proteins where most of the polypeptide chain is inside the lipid bilayer, typically had much fewer PDB entries than the average soluble protein. This is at least partially due to difficulties in protein expression and purification. Single-pass transmembrane proteins such as Receptor Tyrosine Kinases (such as EGF receptors and Eph receptors) and Cell Adhesion proteins (such as Integrin) are prominently represented in the lists. However, these proteins have exceptionally high relevance to cancer cell signaling and have the majority of their domains exposed in solvent. It is these domains whose structure has been determined, excluding the single membrane crossing segment; there are only a few structures available for the membrane crossing regions typically from NMR (about 27 as of 2017).[65] Due to the technical challenges with sample preparation and the likely dynamic nature of the structures, the determination of full length transmembrane protein structures remains a frontier of structural biology for proteins. However, increasingly structures are reported for the transmembrane segment of such proteins, by the use of NMR and also recently by crystallography.[66] While methods for the determination of the structure of membrane integral protein are becoming well established by cryo-EM incl. cryo-electron tomography (cryo-ET),[67] again data from several complementary methods will likely need to be combined in an integrated computational modeling approach in order to solve the structure of full length transmembrane proteins and their complexes.

Historical Implication and Model Proteins for Protein Science: Several of the proteins listed in the tables have historical contexts and/or have become model proteins for structural biology and protein biophysics research.9 However, it should be noted that some well-known proteins (from other organisms) in protein history do not appear in the tables, as here we have ranked only human proteins. Due to the challenge of crystallization especially of eukaryotic proteins, traditionally crystallographers pursed "a wide range species approach", especially in the days when proteins had to be purified from the organism itself. With the advent of recombinant protein expression, the focus shifted to prokaryotic homologues of human proteins and then with the mandate of several structural genomics efforts work on human proteins, exclusively to human proteins.eg.[68] (Readers may refer to the Structural Genomics Consortium (SGC) website (http://www.thesgc.org/structures) to find detailed functional and disease relevance of the human proteins which have been crystalized[69]). In part through such consortia, the coverage of human proteins in the PDB received a significant boost, but the focus on particular individual proteins, which increased the count of their PDB entries has been due to their central role in diseases and the research community at large. As a reference, if counting all the species, the number of PDB entries for proteins with the largest representation are the following (12 are listed): (Lysozyme C – chicken, 867 entries); (Carbonic anhydrase 2 - human, 837); (Beta-2-microglobulin – human, 828); (Endolysin/Lysozyme - bacteriophage, 707); (Endothiapepsin - endothia parasitica, 557); (Cationic trypsin - bovine, 512); (Cyclin-dependent kinase 2 –human, 412); (Prothrombin - human, 392); (Beta-secretase 1 - human, 391); (DNA polymerase beta –human, 370); (Bromodomain-containing protein 4 –human, 357); (Green fluorescent protein - jellyfish, 354). Only five of these twelve proteins are from non-human organisms. It is unlikely that non-human proteins will come up with a large number of PDB entries to compete with this list soon.

In terms of protein model systems – a relatively subjective label for proteins with key biomedical importance-, the studies of Hemoglobin, Insulin, G-proteins, Na-K-ATPase, Prion, Cyclin dependent kinase, Ion channels, Ubiquitin, GPCRs and PD-L1 have been recognized with the Nobel prize. For example, whale Myoglobin and horse Hemoglobin were the earliest proteins to have their 3D structure revealed by x-Ray crystallography.[7,8] Hemoglobin was also the first well known allosteric protein complex identified in the 1960s and a key advance in our understanding of cooperativity.[70] Myoglobin and Cytochrome are early known examples of structure-based allostery for an individual protein.[71,72] In biophysical research, Ubiquitin, individually or as a multi-protein chain, is a model protein for studying protein conformational as well as configurational ensembles, protein dynamics and protein association/recognition.[e.g.73] Calmodulin and Lysosome were widely used in the earlier NMR characterization of protein dynamics and conformational entropy.[74] H- and KRas are recently used as model proteins for investigating the multi-orientational nature of protein configurations at the cell's plasma membrane.[75,76] Recently, p53 and Estrogen receptor have also been studied concerning their likely changes over the course of evolution.[77,78]

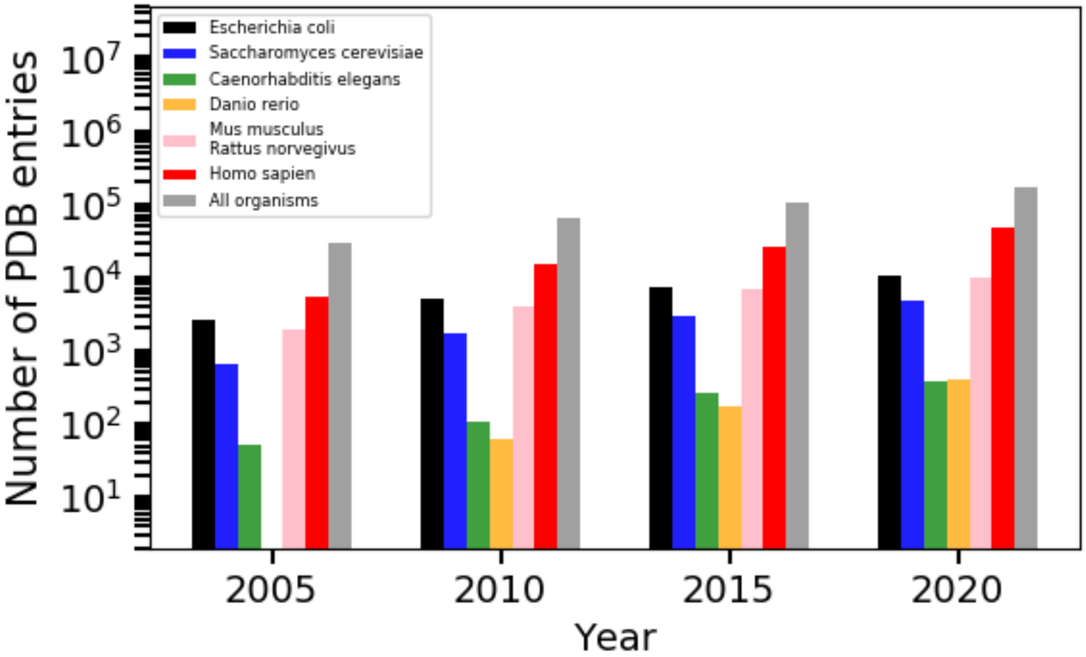

Figure 2: The number of PDB entries for the Data Bank as a whole (grey) and for specific organisms plotted in 5 year increments. Please note the log10-scale as the y-axis. The data points (2010-2020) were linear-fit with the following values as gradients (i.e. fractional change/year) as 0.031 ± 0.001 (black, E. Coli), 0.052 ± 0.012 (green C. Elegans), 0.081 ± 0.005 (orange, D. Rerio), 0.039 ± 0.002 (pink, Mouse+Rat), 0.031 ± 0.001 (red, H. Sapiens), 0.043 ± 0.002 for PDB as a whole (grey, all organisms)

The PDB "on a roll" - Structural Biology and Model Organisms...paving the way towards all cell studies: As noted above, the PDB has increased tremendously in size over several decades and human protein structures comprise nearly 1/3rd of its entries. However, the biomedical but also recently strengthening gene/protein evolution community has an interest in the use of model systems for a wide range of reasons. Organisms such as E. Coli (bacteria), S. Cerevisiae (a species of yeast), C. Elegans (a nematode/worm), D. Rerio (a zebrafish) as well as mice and rats are popular in many types of studies. While not as prominent as human proteins in the PDB, the number of entries listing these organisms in the PDB headers is increasing as well. In order to quantitate the number of entries as a function of time, we examined the source index file at ftp://snapshots.rcsb.org/ which saves information on the current state of the PDB, including all its entries periodically since 2005. Counting up the number of times the above organisms are listed in this file, for the year 2005, 2010, 2015 and 2020, we plotted Fig. 2. On a log10 scale, 2005 appears to represent a different era of the PDB with most numbers below those seen for the period 2010-2020, which show a closely exponential growth in the number of entries. In fact, linear fits can be used to estimate the time that was needed to double the number of entries over this time period: This gives approximately 7 yrs for a doubling of the number of PDB entries of proteins from all organisms. The Growth in human protein structures is slightly faster (at 5.9 yrs, while E.Coli (E. Coli + K-12 E.Coli) and Mouse+Rat protein structures accumulate more slowly, Zebrafish protein structures are growing most, at a doubling every 3.7 years, albeit starting from only approx. 100 structure entries in 2010. These data show that despite the focus on human proteins, the scientific community also has a strong interest in determining protein structures of model organisms. With the possible exception of C. Elegans proteins, the fits have small uncertainties, indicating a near exponential growth. If any trend could be indicated, it is one of the consistent growth and continuity. Similar to Moore's Law for computer speed, one may envisage that the number of protein structure determination for some organisms may slow down. By analogy, such a prediction will likely be erroneous, as technical developments keep on coming which will lead to a faster and more efficient determination of protein structures. For the PDB as a whole it could be envisaged that once the coverage of human proteins is deemed relatively complete, attention may shift focus on the structure determination of non-human proteins, especially if very high throughput structure determination efforts become available. Added to this is a desire to have an as complete set of protein structures for an organism, if not cell/tissue-type, allowing full-cell scale cryo-Electron Tomography (cryo-ET) in the future which will likely be coupled with extensive molecular dynamics simulations on the cell-scale.

Conclusion

In summary, we considered the number of human proteins with fully or partially available medium to high resolution structures among the 20,000 or so canonical human proteins. From this set, we listed the 273 proteins with the most number of structure entries in the PDB. Unsurprisingly, these proteins are also the ones which have been a focus of intense studies, either because of their history as model systems, and more recently as proteins with high biomedical importance. Many of these proteins have influenced our understanding of protein structural and functional biology as well as biophysics. Remarkably, the increase in PDB entries has been growing exponentially over the last 10 years at least, not just for the PDB as a whole, but also for human proteins as well as for those of other organisms examined. Given the numbers involved and the long-term cumulative nature of the PDB, it is unlikely to be influenced by short term trends, if any. The information we provided here should be particularly helpful to researchers who are new to protein science, as in a sea of proteins, the top-studied proteins may serve as "Lighthouses" for future investigations. However, our analysis may also interest seasoned structural biologists, as a "Stamp in Time", showing how far Protein Science has moved and "the Waters which may lie ahead".

Code availability

The raw data, results and codes can be found at https://github.com/sdlzlcase2015/buck_lab_protein_ranking.

Acknowledgements

We thank Andrew Donshik for help with the analysis scripts and Drs. Nami Tajima and Xinghong Dai (both at CWRU) for stimulating discussion. This work is supported by a NIH R01 grant from the National Eye Institute R01EY029169 and previous grants from NIGMS (R01GM073071 and R01GM092851) to the Buck lab.

Competing interests

There is no conflict of interest declared.

Author contribution

Z.L. analyzed the ranking of protein based on available PDB entries. Z.L. and M.B. wrote the manuscript.

References

1. Berman HM, Vallatc B, Lawson CL. The data universe of structural biology. *IUCrJ* 2020;7:630-8.
2. Xu B, Liu L. Developments, applications, and prospects of cryo-electron microscopy. *Protein Sci*. 2020;29(4): 872–82.
3. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* 2019;47(D1):D520-28.
4. Kim M-S, Pinto SM, Getnet D, et al. A draft map of the human proteome. *Nature* 2014;509(7502):575–81.
5. Ponomarenko EA, Poverennaya EV, Ilgisonis EV, et al. The Size of the Human Proteome: The Width and Depth. *Int J Anal Chem* 2016;2016:7436849.
6. The UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res* 2018;47(D1):D506-15.
7. Perutz MF, Rossmann MG, Cullis AF, et al. Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. *Nature* 1960;185:416–22.
8. Kendrew JC, Bodo G, Dintzis HM, et al. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 1958;181(4610):662–6.
9. Shi Y. A glimpse of structural biology through X-ray crystallography. *Cell* 2014;59(5):995-1014.
10. Kuhlbrandt W. The resolution revolution. *Science* 2014;343(6178):1443-4.
11. Bai X-C, McMullan G, Scheres SHW. How cryo-EM is revolutionizing structural biology. *Trends Biochem Sci* 2015;40(1):49-57.
12. Zatsepin NA, Li C, Colasurd P, et al. The complementarity of serial femtosecond crystallography and MicroED for structure determination from microcrystals. *Curr Opin Struct Biol* 2019;58:286-93.
13. Vinogradova O, Qin J. NMR as a Unique Tool in Assessment and Complex Determination of Weak Protein–Protein Interactions. *Topics Curr Chem* 2012;326:35-45.
14. Dyson HJ, Wright PE. Perspective: the essential role of NMR in the discovery and characterization of intrinsically disordered proteins. *J Biomol NMR* 2019;73(12):651–9.
15. Rodrigues JPGLM, Bonvin AMJJ. Integrative computational modeling of protein interactions. *FEBS J* 2014;281(8):1988-2003.
16. Alber F, Forster F, Korkin D, et al. Integrating diverse data for structure determination of macromolecular assemblies. *Annu Rev Biochem* 2008;77:443-77.
17. Allison TM. Structural mass spectrometry comes of age: new insight into protein structure, function and interactions. *Biochem Soc Trans* 2019;47(1):317–27.
18. Eriksson AE, Jones TA, Liljas A. Refined structure of human carbonic anhydrase II at 2.0 A resolution. *Proteins* 1988;4:274-82.
19. Bjorkman PJ, Saper MA, Samraoui B, et al. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* 1987;329:506-12.
20. Jeffrey PD, Russo AA, Polyak K, et al. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature* 1995;376:313-20.
21. Bode W, Mayr I, Baumann U, et al. The refined 1.9 A crystal structure of human alpha-thrombin: interaction with D-Phe-Pro-Arg chloromethylketone and significance of the Tyr-Pro-Pro-Trp insertion segment. *EMBO J* 1989;8:3467-75.
22. Hong L, Koelsch G, Lin X, et al. Structure of the protease domain of memapsin 2 (beta-secretase) complexed with inhibitor. *Science* 2000;290:150-3.
23. Pelletier H, Sawaya MR, Wolfle W, et al. A structural basis for metal ion mutagenicity and nucleotide selectivity in human DNA polymerase beta. *Biochemistry* 1996;35:12762-77.
24. Abbate EA, Voitenleitner C, Botchan MR. Structure of the Papillomavirus DNA-Tethering Complex E2:Brd4 and a Peptide that Ablates HPV Chromosomal Association. *Mol Cell* 2006;24:877-89.
25. Blake CC, Geisow MJ, Oatley SJ, et al.Structure of prealbumin: secondary, tertiary and quaternary interactions determined by Fourier refinement at 1.8 A. *J Mol Biol* 1978;121:339-56.
26. Allerston CK, Lee SY, Newman, JA, et al. The Structures of the Snm1A and Snm1B/Apollo Nuclease Domains Reveal a Potential Basis for Their Distinct DNA Processing Activities. *Nucleic Acids Res* 2015;43:11047-60.
27.Schwabe JW, Chapman L, Finch JT, et al. The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: how receptors discriminate between their response elements. *Cell* 1993;75:567-78.
28. Ogiso H, Ishitani R, Nureki O, et al. Crystal Structure of the Complex of Human Epidermal Growth Factor and Receptor Extracellular Domains. *Cell* 2002;110:775-87.
29.Madden DR, Gorga JC, Strominger JL, et al. The three-dimensional structure of HLA-B27 at 2.1 A resolution suggests a general mechanism for tight peptide binding to MHC. *Cell* 1992;70:1035-48.
30.Patenaude SI, Seto NO, Borisova SN, et al. The structural basis for specificity in human ABO(H) blood group biosynthesis. *Nat Struct Biol* 2002;9:685-90.
31.Hynes TR, Randal M, Kennedy LA, et al. X-ray crystal structure of the protease inhibitor domain of Alzheimer's amyloid beta-protein precursor. *Biochemistry* 1990;29:10018-22.
32.Rasmussen HB, Branner S, Wiberg FC, et al. Crystal structure of human dipeptidyl peptidase IV/CD26 in complex with a substrate analogue. *Nat Struct Biol* 2003;10:19-25.
33. Stern LJ, Brown JH, Jardetzky TS, et al. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* 1994;368:215-21.
34.Czabotar PE, Lee EF, van Delft MF, et al. Structural insights into the degradation of Mcl-1 induced by BH3 domains. *Proc Natl Acad Sci U S A* 2007;104:6217-6222.

35. Dolgin E. The Most Popular Genes in the Human Genome. *Nature* 2017; 551(7681):427-31.
36. Joerger AC, Fersht AR. The p53 Pathway: Origins, Inactivation in Cancer, and Emerging Therapeutic Approaches. *Annu Rev Biochem* 2016;85:375-404.
37. Peng Y, Cao S, Kiselar J, et al. A Metastable Contact and Structural Disorder in the Estrogen Receptor Transactivation Domain. *Structure* 2019;27(2):229-40.
38. Wilhelm M, Schlegh J, Hahne H, et al. Mass-spectrometry-based draft of the human proteome. Nature, 2014;509:582-7.
39. Rubin GM. Comparative Genomics of the Eukaryotes. Science 2000; 287(5461): 2204-15.
40. Levine AJ, Lane DP, eds. The p53 family. Cold Spring Harbor Perspectives in Biology. Cold Spring Harbor, N.Y.: *Cold Spring Harbor Laboratory Press*. ISBN 978-0-87969-830-0.
41. Prior IA, Lewis P.D, Mattos C. A Comprehensive Survey of Ras Mutations in Cancer. *Cancer Res* 2012;72(10):2457-67.
42. Nilsson S, Mäkelä S, Treuter E, et al. Mechanisms of estrogen action. *Physiolog Rev* 2001; 81(4):1535-65.
43. Locksley RM, Killeen N, Lenardo MJ. The TNF and TNF receptor superfamilies: integrating mammalian biology. *Cell* 2001;104(4):487–501.
44. Goa G-W, Mania A. Low-Density Lipoprotein Receptor (LDLR) Family Orchestrates Cholesterol Homeostasis. *Yale J Biol Med* 2012;85(1):19–28.
45. Boucher J, Kleinridders A, Kahn R. Insulin Receptor Signaling in Normal and Insulin-Resistant States. *Cold Spring Harb Perspect Biol* 2014;6(1):a009191.
46. Natesh R, Schwager SL, Sturrock ED, et al. Crystal structure of human angiotensin-converting enzyme-lisinopril complex. *Nature* 2003;421(6922):551–4.
47. Shibuya M. Vascular Endothelial Growth Factor (VEGF) and Its Receptor (VEGFR) Signaling in Angiogenesis. *Genes Cancer* 2011;2(12):1097–105.
48. Lok DJ, Van Der Meer P, de la Porte PW, et al. Prognostic value of galectin-3, a novel marker of fibrosis, in patients with chronic heart failure: data from the DEAL-HF study. *Clin Res Cardiol* 2010;99(5):323–8.
49. Yun YR, Won JE, Jeon E, et al. Fibroblast Growth Factors: Biology, Function, and Application for Tissue Regeneration. *J Tissue Eng* 2010;2010:218142.
50. Foote M, Zhou Y. 14-3-3 proteins in neurological disorders. *Int J Biochem Mol Biol* 2012;3(2):152–164.
51. O'brien RJ, Wong PC. Amyloid Precursor Protein Processing and Alzheimer's Disease. *Annu Rev Neurosci* 2011;34:185-204.
52. Lei P, Ayton S, Finkelstein DI, et al. Tau protein: relevance to Parkinson's disease. *Int J Biochem Cell Biol* 2010;42 (11):1775–8.
53. Laurén J. Cellular prion protein as a therapeutic target in Alzheimer's disease. *J Alzheimers Dis* 2014;38(2):227–244.
54. Zeldenrust SR, Benson MD. Familial and senile amyloidosis caused by transthyretin. In Ramirez-Alvarado M, Kelly JW, Dobson C (eds.). Protein misfolding diseases: current and emerging principles and therapies. New York: Wiley. 2010; pp.795–815.
55. Choo SY. The HLA System: Genetics, Immunology, Clinical Testing, and Clinical Implications. *Yonsei Med J* 2007; 48(1): 11–23.
56. Brady RL, Dodson EJ, Dodson GG, et al. Crystal structure of domains 3 and 4 of rat CD4: relation to the NH2-terminal domains. *Science* 1993;260(5110):979–83.
57. Hasko G, Pacher P. A2A receptors in inflammation and injury: lessons learned from transgenic animals. *J Leukoc Biol* 2008;83(3):447–55.
58. Shang J, Ye G, Shi K, et al. Structural Basis of Receptor Recognition by SARS-CoV-2. *Nature* 2020;581(7807): 221-24.
59. Daly JL, Simonetti B, Antón-Plágaro C, et al. Neuropilin-1 Is a Host Factor for SARS-CoV-2 Infection. *BioRxiv* 2020; doi:10.1101/2020.06.05.134114.
60. Manning G, Whyte DB, Martinez R, et al. The Protein Kinase Complement of the Human Genome *Science* 2002:298(5600):1912-34.
61. Scheeff ED, Bourne PE. Structural Evolution of the Protein Kinase–Like Superfamily. *Plos Comput Biol* 2005;1(5):e49.
62. Lahiry P, Torkamani A, Schork N, et al. Kinase mutations in human disease: interpreting genotype–phenotype relationships. *Nat Rev Genet* 2010;11:60–74.
63. Wu P, Nielsen TE, Clausen MH. FDA-approved Small-molecule Kinase Inhibitors. *Trends Pharmacol Sci* 2015;36(7):422-39.
64. Hauser AS, Chavali S, Masudo I. Pharmacogenomics of GPCR Drug Targets. *Cell* 2018;172(1-2):41–54.e19.
65. Bocharov EV, Mineev KS, Pavlov KV, et al. Helix-helix interactions in membrane domains of bitopic proteins: Specificity and role of lipid environment. *Biochim Biophys Acta Biomembr* 2017;1859(4):561-76.
66. Trenker R, Call MJ, Call ME. Progress and prospects for structural studies of transmembrane interactions in single-spanning receptors. *Curr Opin Struct Biol* 2016;39:115-123.
67. Autzen HE, Julius D, Cheng Y. Membrane mimetic systems in CryoEM: keeping membrane proteins in their native environment. *Curr Opin Struct Biol* 2019;58:259-68.
68. Jones, MM, Castle-Clarke S, Jones, Brooker D, et al. The Structural Genomics Consortium. A Knowledge Platform for Drug Discovery: A Summary. *Rand Health Q*. 2014;4(3):19.
69. Gileadi O, Knapp S, Lee WH, et al. The scientific impact of the Structural Genomics Consortium: a protein family and ligand-centered approach to medically-relevant human proteins. *J Struct Funct Genomics* 2007; 8(2-3): 107–19.
70. Monod J, Changeux JP, Jacob F. Allosteric Proteins and Cellular Control Systems. *J Mol Biol* 1963;6(4): 306-29.

71. Frauenfelder H, McMahon BH, Austin RH, et al. The role of structure energy landscape dynamics and allostery in the enzymatic function of myoglobin. *Proc Natl Acad Sci USA* 2001;98:2370–4.
72. Johnson EF, Schwab GE, Dieter HH. Allosteric regulation of the 16α-hydroxylation of progesterone as catalyzed by rabbit microsomal cytochrome P-450 3b. *J Biol Chem* 1983;258:2785–8.
73. Lange OF, Lakomek NA, Fares C, et al. Recognition Dynamics Up to Microseconds Revealed from an RDC-Derived Ubiquitin Ensemble in Solution. *Science* 2008;320(5882):1471-5.
74. Frederick KK, Marlow M.S, Valentine KG, et al. Conformational Entropy in Molecular Recognition by Proteins. *Nature* 2007;448(7151):325-9.
75. Prakash P, Zhou Y, Liang H, et al. Oncogenic K-Ras Binds to an Anionic Membrane in Two Distinct Orientations: A Molecular Dynamics Analysis. *Biophys J* 2016;110(5):1125-38.
76. Li Z-L, Buck M. Computational Modeling Reveals that Signaling Lipids Modulate the Orientation of K-Ras4A at the Membrane Reflecting Protein Topology. *Structure* 2017;25(4):679-89.
77. Zhao Y, Ren J-L, Wang M-Y, et al. Codon 104 Variation of P53 Gene Provides Adaptive Apoptotic Responses to Extreme Environments in Mammals of the Tibet Plateau. *Proc Natl Acad Sci* 2013;110(51):20639-44.
78. Harms MJ, Eick GN, Goswami D, et al. Biophysical Mechanisms for Large-effect Mutations in the Evolution of Steroid Hormone Receptors. *Proc Natl Acad Sci* 2013;110(28): 11475-80.