Beyond History: The List of The Most Well Studied Human Protein Structures

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 20,000 or so canonical human protein sequences, as of July 2020, 6,747 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 database (the PDB) and why? In this paper, we list the 272 top protein structures based on the number of their PDB depositions. This set of proteins accounts for more than 40% of all available human PDB entries and represent past trend and current status for protein science. We briefly discuss the relationship which some of the prominent protein structures have with protein biophysics research and mention their relevance to human diseases. The information may inspire researchers who are new to protein science, but it also provides a year 2020 snap-shot for the state of protein science.

Main Text

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.[1] 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.[2] By July 2020, there are 20,367 reviewed entries of full length human proteins in the Uniprot database, which is easily accessed via a webserver (http://uniprot.org/). The analysis, based on the 20,367 proteins, indicates that the median sequence length of human proteins is 325 amino acids (Fig. 1a). The largest human protein is Titin with 34,350 amino acids.

Structure forms the basis of protein function: The three- dimensional structure/conformation of the polypeptide chain determines the dynamics and, then more or less directly the function of an individual protein. Proteins excluding intrinsic 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 medium to high resolution protein structures with increased accuracy and efficiency over the years. Such structures are deposited in the Protein Databank, the PDB and in 2019 alone, 10,585 protein structures were released. As of July 2020, there are 167,780 structures in the PDB. Despite this large number of PDB items, the real number of canonical/non-redundant human proteins with known structures is unclear, as many PDB entries correspond to the same protein solved by different groups, single point mutants or the same protein bound to different ligands, ranging from small molecule inhibitors to protein or other macromolecular binding partners. As mentioned above, there are 20,367 reviewed entries of full length human proteins in the Uniprot database. By reading these proteins one by one into a python script, we found 6,747 out of the 20,367 distinct human proteins have at least one PDB item for at least a segment/domain of the entire protein. Thus, at least 2/3rds of the structures of the human proteome remain to be determined. The distribution of available PDB items amongst these protein species are plotted in Fig. 1b. Amongst these proteins, 1,978, 1,051 and 775 proteins, have only one, two or three pdb entries respectively (together, then 56% of the 6,747 proteins have only 1-3 entries). However, at the other extreme, the 200 proteins with the most entries (3% of 6,747) have 21,645 cumulative PDB entries, remarkably counting at least 40% of total human PDB items (totally 46, 581) in the Uniprot database. Thus, the top human 200 proteins have gathered an unusually high proportion of attention compared to the rest.

The human proteins with the highest number of pdb entries were identified as follows: We downloaded the data of individual human protein species from the Uniprot web server. Only reviewed items (the set of 20,367 canonical proteins) were downloaded. From the downloaded file, we extracted the PDB items of each protein species and sorted them by adding the total number of available PDB items for each protein species. It should be noted that this number includes structures of fragments or domains because the determination of full length structures is still rare or not yet possible for certain types of proteins, such as the great majority of single membrane crossing receptors.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0655.v1

In 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). We also considered isoforms in the analysis, but assumed that the latin name for human "homo sapiens" was mentioned in the Uniprot entries for all human proteins. The top 200 human proteins counting the number of the PDB items is given in Table 1. Separately, we also list the 100 transmembrane proteins with the most entries (Table 2). Since 28 membrane proteins already appear in the first part of Table 1, so in total only 272 unique proteins are listed. Overall, the high frequency appearance of the proteins in the pdb arises from the biological importance that they have in cellular processes, in human diseases but some also 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 representative structures/families which have emerged, also making a note of the early history of protein structural biology.

It should be noted that the method we used to rank the top-hit proteins is completely different from the approach used in the report by Dolgin in 2017, where the top 10 genes in the human genome are identified by counting the frequency of appearance of a gene in the PubMed database.[3] By contrast, we count the absolute number of available PDB entries for each protein in the Uniprot websever. However, the two methods corroborate each other in part---some of top genes identified in the study of Dolgin, such as Cellular tumor antigen p53, Tumor necrosis factor, Epidermal growth receptor and Estrogen Receptor also appear in our lists.

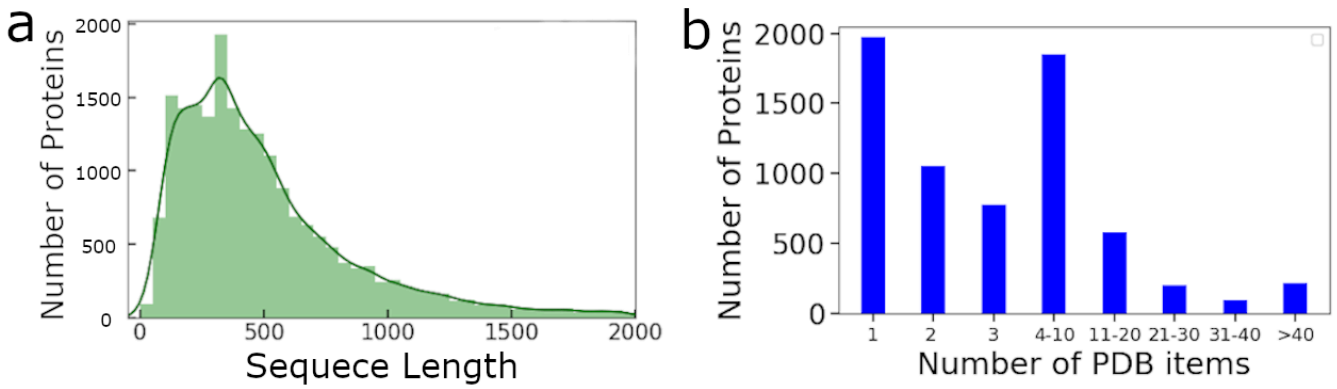

Figure 1: Protein sequence and structure statistics. (a) Distribution of sequence lengths for 20,367 human proteins. (b) PDB statistics as of July 2020. Number of proteins with 1, 2, 3 and more PDB items.

Top 10 Aqueous And Membrane Proteins Identified: By far the most structures solved are for Beta-2-microglobulin, Carbonic anhydrase 2 and Cyclin-dependent kinase 2 as the first, second, third place of the most deposited structures in the PDB, with 770, 766 and 410 entries respectively. Prothrombin; Beta-secretase 1; DNA polymerase beta; HLA class I histocompatibility antigen, A alpha Chain; Transthyretin; Bromodomain-containing protein 4; DNA cross-link repair 1A protein ranks 4-10.

The top 10 for membrane proteins (with transmembrane regions) are listed below: Beta-secretase 1; HLA class I histocompatibility antigen, A alpha chain; Estrogen receptor; HLA class I histocompatibility antigen, B alpha chain; Epidermal growth factor receptor; Histo-blood group ABO system transferase; Amyloid-beta precursor protein; Dipeptidyl peptidase 4; HLA class II histocompatibility antigen, DR alpha chain; Hepatocyte growth factor receptor.

Protein Classification: Of all the 20,000 canonical human proteins, noticeable protein groupings include 1,653 metabolic enzymes; 1,089 non-metabolic enzyme such as kinase and GTPases; 1,600 transcription factors; at least 1,555 transporters and channels; and 831 GPCRs.[4-6] By contrast, among the 272 top-hit proteins listed in the tables, there are 73 metabolic enzymes, 5 GTP-/ATPases, 44 kinases, 16 transcription factors, 5 ion channels and 4 G-protein-coupled receptors (GPCRs). Moreover, there are 11 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 84 of 272 proteins are not classified into major protein families, but all of them have important biological functions. Here we comment on several of the families.

Kinase as One of The Best Studied Protein Family: The high frequency of appearance of kinases (44 of 272) is remarkable in contrast to its low fraction amongst the 20,000 canonical human proteins (518 of 20,000). Protein kinases are thought to modify up to 30% of all human proteins and many of them such as Raf kinase, Akt kinase, Ephrin type-A receptor 2 and Epidermal growth factor receptor (EGFR) have a crucial role in disease development,

 doi:10.20944/preprints202008.0655.v1

especially in cancer.[7] Clinically, more than 250 kinase inhibitors are undergoing clinical trials and 37 are already approved as therapeutics.[8] 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 items) 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, 2,237 distinct membrane proteins have at least one structure, corresponding to 33.2% of all available human protein structures. By counting single-pass and multi-pass transmembrane proteins only, 1,132 of 6,747 (16.8%) 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 272 proteins with most of pdb items. This is despite the fact that GPCRs for example account for approx. 30% of all drug targets. 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 typically had much fewer pdb items than soluble proteins. This is at least partially due to difficulties in protein expression and purification. 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 the majority of domains exposed in solvent, and it is these domains whose structure has been mostly 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).[9] Due to the technical challenges with sample preparation and likely the dynamic nature of the structures, the determination of full length TM protein structures remains a frontier of structural biology, with increasing success reported by use of NMR, molecular modeling and cryo-EM incl. cryo-electron tomography (cryo-ET).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0655.v1

Table 1: Human proteins with the most PDB entries. The top 200 ranked proteins and are listed alphabetically (the top 10 of all proteins in blue).

Protein Name; Number of PDB items; Rank									
Acetylcholinesterase	42	200	Cellular retinoic acid-binding protein 2	71	93	Ephrin type-A receptor 2	70	96
Adenosine receptor A2a	49	160	Cellular tumor antigen p53	177	30	Epidermal growth factor receptor	189	29
ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase 1	42	199	Cholinesterase	64	114	Estrogen receptor	278	14
Aldo-keto reductase family 1 member	145	39	Coagulation factor VII	108	52	Farnesyl pyrophosphate synthase	82	79
ADP-sugar pyrophosphatase	53	146	Coagulation factor X	144	40	Ferritin heavy chain	51	152
Aldo-keto reductase family 1 member C3	47	171	Coagulation factor XI	81	83	Fibrinogen gamma chain	45	179
ALK tyrosine kinase receptor	61	125	Collagenase 3	48	165	Fibroblast growth factor 1	97	61
Amine oxidase [flavin-containing] B	44	184	Complement C3	47	170	Fibroblast growth factor receptor 1	65	110
Amyloid-beta precursor protein	140	44	Complement factor H	46	174	Fibronectin	58	129
Androgen receptor	82	81	CREB-binding protein	91	66	Galectin-3	76	89
Angiogenin	46	175	Cyclin-A2	94	63	Gelsolin	53	145
Angiotensin-converting enzyme	44	183	Cyclin-dependent kinase 2	410	3	Glucocorticoid receptor	43	190
ATPase family AAA domain-containing protein 2	57	134	Cytochrome P450 3A4	43	192	Glutamate carboxypeptidase 2	75	90
Aurora kinase A	155	36	Cytosolic purine 5'-nucleotidase	43	191	Glutathione S-transferase P	64	113
Bcl-2-like protein 1	79	87	Death-associated protein kinase 1	57	133	Glycogen synthase kinase-3 beta	84	76
Beta-2-microglobulin	770	1	Deoxycytidine kinase	47	169	Growth factor receptor-bound protein 2	49	159
Beta-secretase 1	380	5	Deoxynucleoside triphosphate triphosphohydrolase SAMHD1	47	168	GTP-binding nuclear protein Ran	79	85
Bifunctional epoxide hydrolase 2	102	57	Dihydrofolate reductase	79	86	GTPase HRas	175	32
Bile acid receptor	82	80	Dihydroorotate dehydrogenase (quinone), mitochondrial	66	108	GTPase KRas	142	42
Bromodomain adjacent to zinc finger domain protein 2B	262	19	Dipeptidyl peptidase 4	104	54	Heat shock protein HSP 90-alpha	295	12
Bromodomain-containing protein 1	308	11	DNA cross-link repair 1A protein	312	10	Hemoglobin subunit alpha	263	18
Bromodomain-containing protein 2	71	94	DNA damage-binding protein 1	50	156	Hemoglobin subunit beta	257	20
Bromodomain-containing protein 4	313	9	DNA polymerase beta	355	6	Hepatocyte growth factor receptor	85	75
Calmodulin-1	160	34	DNA polymerase eta	123	48	High affinity nerve growth factor receptor	45	178
cAMP and cAMP-inhibited cGMP 3',5'-cyclic phosphodiesterase 10A	94	62	DNA polymerase iota	46	173	Histidine triad nucleotide-binding protein 1	48	164
cAMP-dependent protein kinase inhibitor alpha	102	56	DNA polymerase lambda	58	130	Histo-blood group ABO system transferase	146	38
cAMP-specific 3',5'-cyclic phosphodiesterase 4B	43	185	DNA-directed DNA/RNA polymerase mu	70	97	Histone deacetylase 8	47	167
cAMP-specific 3',5'-cyclic phosphodiesterase 4D	67	102	Dual specificity mitogen-activated protein kinase kinase 1	42	194	Histone H2A type 1-B/E	88	72
Carbonic anhydrase 2	766	2	Dual specificity protein kinase TTK	67	105	Histone H2B type 1-J	91	65
Casein kinase II subunit alpha	168	33	E3 ubiquitin-protein ligase Mdm2	103	55	Histone H3.1	235	22
Caspase-3	100	59	E3 ubiquitin-protein ligase XIAP	66	107	Histone H3.3	61	124
Cathepsin K	56	135	Elongin-B	68	100	Histone H4	176	31
Cathepsin S	55	140	Elongin-C	65	111	HLA class I histocompatibility antigen, A alpha chain	337	7

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0655.v1

HLA class I histocompatibility antigen, B alpha chain	201	28	Pancreatic alpha-amylase	48	162	Son of sevenless homolog 1	60	126
HLA class II histocompatibility antigen, DR alpha chain	100	58	Peptidyl-prolyl cis-trans isomerase A	130	47	Stromelysin-1	43	187
Hypoxia-inducible factor 1-alpha inhibitor	42	198	Peptidyl-prolyl cis-trans isomerase F, mitochondrial	42	196	Superoxide dismutase [Cu-Zn]	112	50
Immunoglobulin heavy constant gamma 1	201	27	Peptidyl-prolyl cis-trans isomerase FKBP1A	50	155	T cell receptor alpha constant	133	45
Immunoglobulin kappa constant	90	67	Peptidyl-prolyl cis-trans isomerase NIMA-interacting 1	80	84	T cell receptor beta constant 1	87	73
Induced myeloid leukemia cell differentiation protein Mcl-1	83	77	Peregrin	61	122	T cell receptor beta constant 2	61	120
Insulin	272	17	Peroxisome proliferator-activated receptor delta	44	181	T-box transcription factor T	46	172
Insulin-degrading enzyme	53	144	Peroxisome proliferator-activated receptor gamma	206	26	T-cell surface glycoprotein CD4	62	115
Integrin alpha-IIb	42	197	Phosphatidylinositol 3-kinase regulatory subunit alpha	60	127	Thymidylate synthase	59	128
Integrin beta-3	69	99	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform	51	150	Tissue factor	48	161
Interleukin-1 receptor-associated kinase 4	47	166	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit gamma isoform	93	64	Titin	45	176
Kinesin-like protein KIF11	55	139	Poly [ADP-ribose] polymerase 1	62	117	Transforming protein RhoA	43	186
Leukotriene A-4 hydrolase	61	123	Poly [ADP-ribose] polymerase tankyrase-2	146	37	Transitional endoplasmic reticulum ATPase	42	193
Lysine-specific demethylase 4A	82	78	Polycomb protein EED	43	189	Transthyretin	323	8
Lysine-specific demethylase 4D	274	15	Polyubiquitin-B	141	43	Tyrosine-protein kinase ABL1	66	106
Lysine-specific demethylase 5A	44	182	Polyubiquitin-C	211	23	Tyrosine-protein kinase BTK	78	88
Lysine-specific demethylase 5B	55	138	Proteasome subunit alpha type-1	42	195	Tyrosine-protein kinase JAK2	88	70
Lysine-specific histone demethylase 1A	62	119	Proteasome subunit alpha type-3	43	188	Tyrosine-protein kinase Lck	55	137
Lysozyme C	208	25	Proteasome subunit beta type-1	67	104	Tyrosine-protein kinase SYK	70	95
Macrophage metalloelastase	81	82	Proteasome subunit beta type-5	49	158	Tyrosine-protein phosphatase non-receptor type 1	273	16
Macrophage migration inhibitory factor	89	69	Proteasome subunit beta type-7	50	154	Tyrosine-protein phosphatase non-receptor type 11	57	131
Major prion protein	54	142	Protein/nucleic acid deglycase DJ-1	61	121	U1 small nuclear ribonucleoprotein A	69	98
Mediator of RNA polymerase II transcription subunit 1	48	163	Prothrombin	391	4	Ubiquitin carboxyl-terminal hydrolase 7	50	153
Microtubule-associated protein tau	54	141	Proto-oncogene tyrosine-protein kinase Src	64	112	Ubiquitin-40S ribosomal protein S27a	53	143
Mitogen-activated protein kinase 1	110	51	Renin	88	71	Ubiquitin-60S ribosomal protein L40	55	136
Mitogen-activated protein kinase 10	51	151	REST corepressor 1	45	177	Urokinase-type plasminogen activator	143	41
Mitogen-activated protein kinase 14	236	21	Ribosyldihydronicotinamide dehydrogenase [quinone]	67	103	Vascular endothelial growth factor receptor 2	52	148
Neutrophil gelatinase-associated lipocalin	52	149	Retinoic acid receptor RXR-alpha	87	74	Vitamin D3 receptor	49	157
Nicotinamide phosphoribosyltransferase	62	118	Serine/threonine-protein kinase B-raf	74	91	von Hippel-Lindau disease tumor suppressor	52	147
Nitric oxide synthase, brain	57	132	Serine/threonine-protein kinase Chk1	133	46	WD repeat-containing protein 5	73	92
Nuclear autoantigen Sp-100	120	49	Serine/threonine-protein kinase pim-1	156	35	14-3-3 protein sigma	90	68
Nuclear receptor coactivator 1	210	24	Serine/threonine-protein kinase PLK1	62	116	3-phosphoinositide-dependent protein kinase 1	68	101
Nuclear receptor coactivator 2	288	13	Serotransferrin	44	180	7,8-dihydro-8-oxoguanine triphosphatase	66	109
Nuclear receptor ROR-gamma	97	60	Serum albumin	107	53			

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0655.v1

Table 2: Top 100 membrane proteins with most of the pdb items. The top 28 proteins also appear in Table 1.

Activin receptor type-1	16	90		Disintegrin and metalloproteinase domain-containing protein 17	23	58		Integrin alpha-V	21	78
Adenosine receptor A2a	49	21		Ephrin type-A receptor 2	70	14		Integrin alpha-2	17	67
Advanced glycosylation end product-specific receptor	22	60		Ephrin type-A receptor 3	27	54		Integrin beta-3	69	15
ALK tyrosine kinase receptor	61	19		Ephrin type-A receptor 4	15	55		Leukotriene C4 synthase	16	84
Amine oxidase [flavin-containing] B	44	25		Ephrin type-B receptor 4	23	57		Low affinity immunoglobulin epsilon Fc receptor	23	56
Amyloid-beta precursor protein	140	7		Epidermal growth factor receptor	189	5		Low-density lipoprotein receptor	33	35
ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase 1	42	28		Epithelial discoidin domain-containing receptor 1	20	65		Low-density lipoprotein receptor-related protein 6	16	83
Angiopoietin-1 receptor	15	91		Erythropoietin receptor	18	72		Macrophage colony-stimulating factor 1 receptor	19	69
Angiotensin-converting enzyme	44	24		Estrogen receptor	278	3		Major histocompatibility complex class I-related gene protein	31	38
Angiotensin-converting enzyme 2	16	89		Fibroblast growth factor receptor 1	65	17		Mast/stem cell growth factor receptor Kit	23	55
Antigen-presenting glycoprotein CD1d	15	92		Fibroblast growth factor receptor 2	41	29		Melanoma antigen recognized by T-cells 1	22	59
Apoptosis regulator BAX	24	52		Fibroblast growth factor receptor 4	25	45		Neurogenic locus notch homolog protein 1	24	49
Apoptosis regulator Bcl-2	26	42		Furin	21	63		Neuropilin-1	17	77
Bcl-2 homologous antagonist/killer	25	47		Glucagon-like peptide 1 receptor	15	96		Platelet glycoprotein Ib alpha chain	19	68
Bcl-2-like protein 1	79	12		Glutamate receptor ionotropic, NMDA 1	17	79		Potassium channel subfamily K member 9	16	82
Beta-1, 4-galactosyltransferase 1	19	71		Glutamate receptor ionotropic, NMDA 2A	15	99		Programmed cell death 1 ligand 1	31	37
Beta-2 adrenergic receptor	25	46		Glutamate carboxypeptidase 2	75	13		Prostaglandin E synthase	16	81
Beta-secretase 1	380	1		Hepatocyte growth factor receptor	85	10		Proto-oncogene tyrosine-protein kinase receptor Ret	25	43
Butyrophilin subfamily 3 member A1	15	93		High affinity nerve growth factor receptor	45	23		Receptor tyrosine-protein kinase erbB-2	34	34
C-C chemokine receptor type 5	18	76		Histo-blood group ABO system transferase	146	6		Receptor-type tyrosine-protein phosphatase gamma	19	67
C-type lectin domain family 4 member K	20	66		HLA class I histocompatibility antigen, A alpha chain	337	2		Squalene synthase	26	41
Cadherin-1	18	75		HLA class I histocompatibility antigen, B alpha chain	201	4		Stimulator of interferon genes protein	30	39
Carbonic anhydrase 12	18	74		HLA class II histocompatibility antigen, DR alpha chain	100	9		Suppressor of tumorigenicity 14 protein	23	54
Cation-independent mannose-6-phosphate receptor	18	73		HLA class II histocompatibility antigen gamma chain	15	97		Synaptotagmin-1	15	98
CD81 antigen	15	94		HLA class I histocompatibility antigen, alpha chain E	16	87		T-cell surface glycoprotein CD1b	16	80
Chloride intracellular channel protein 1	15	95		HLA class II histocompatibility antigen, DQ alpha 1 chain	16	86		T-cell surface glycoprotein CD4	62	18
Complement decay-accelerating factor	24	51		HLA class II histocompatibility antigen, DRB1 beta chain	19	70		Tissue factor	48	22
Copper-transporting ATPase 1	21	64		Induced myeloid leukemia cell differentiation protein Mcl-1	83	11		TGF-beta receptor type-1	37	32
Corticosteroid 11-beta-dehydrogenase isozyme 1	40	30		Insulin receptor	35	33		Toll-like receptor 8	24	48
Coxsackievirus and adenovirus receptor	16	88		Insulin-like growth factor 1 receptor	25	44		Tumor necrosis factor	23	53
Cystic fibrosis transmembrane conductance regulator	32	36		Integrin alpha-IIb	42	27		Vascular endothelial growth factor receptor 2	52	20
Cytochrome P450 3A4	43	26		Integrin alpha-2	39	31		3-hydroxy-3-methylglutaryl-coenzyme A reductase	22	61
Dihydroorotate dehydrogenase (quinone), mitochondrial	66	16		Integrin alpha-L	16	85				
Dipeptidyl peptidase 4	104	8		Integrin alpha-M	24	50				

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. However, it should be noted that some of well-known proteins (from the other organisms) in protein history do not appear in the tables, as here we have adhered to human proteins. Due to the challenge of crystallization especially of eukaryotic proteins, traditionally crystallographers have tried their luck with 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 to work on human proteins. The number of human proteins in the PDB received a significant boost. As an reference, if counting all the species, the number of pdb items for proteins with the largest representation are the following (12 are listed): (Lysozyme C – chicken, 834 items); (Beta-2-microglobulin – human, 770); (Carbonic anhydrase 2 - human, 766); (Endolysin - bacteriophage, 702); (Endothiapepsin - endothia parasitica, 532); (Cationic trypsin - bovine, 487); (Cyclin-dependent kinase 2 –human, 410); (Prothrombin - human, 391); (Beta-secretase 1 - human, 380); (DNA polymerase beta –human, 355); (HLA class I histocompatibility antigen, A alpha chain –human, 337); (Green fluorescent protein - jellyfish, 332). Seven of these twelve proteins have human source.

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, Myoglobin and Hemoglobin were the earliest proteins to have their 3D structure revealed by x-Ray crystallography. Hemoglobin was also the first well known allosteric protein complex identified in the 1960s and a key advance in our understanding of cooperativity.[10] Myoglobin and Cytochrome are early known examples of structure-based allostery for an individual protein. In biophysical research, Ubiquitin, individually or as a multi-protein chain, are is a model protein for studying protein conformational as well as configurational ensembles, protein dynamics and protein association/recognition.[11] Calmodulin and Lysosome were widely used in the earlier NMR characterization of protein dynamics and conformational entropy.[12] H- and KRas are recently used as model proteins for investigating the multi-orientational nature of protein configurations at the cells plasma membrane.[13] Recently, p53 and Estrogen receptor have also been studied with respect to their likely changes over the course of evolution.[14]

Relevance to Human Disease: Many of the 272 proteins are important for their involvement in human diseases and remain a current focus of research. For example, the (Low-density lipoprotein receptor) LDL receptor is vital for the regulation of the concentration of human lipoprotein which tracks human fat content. Proteins such as p53, Ras GPTases, Estrogen receptor and 14-3-3- proteins are crucial proteins either in cancer development or cancer metastasis.[15] Fibroblast growth factors and Neuropilin-1 are vital factors for human cell development. Cytokines and Cell adhesion molecules such as human leukocyte antigen, Tumor necrosis factor and T-cell surface glycoprotein CD4, are important for immunity. Amyloid-beta precursor proteins, Microtubule-associated protein tau and the prion protein are crucial for development of neuronal diseases.[16] Angiotensin-converting enzyme 2 (ACE2) and recently Neuropilin-1 were identified as entry receptor for coronavirus SARS-COV-2.[17,18]

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 identified 272 protein structures with the most number of items in the PDB. 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 relevance. Many of these proteins have influenced our understanding of protein structural and functions biology as well as biophysics. The information we provided here should be 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 structural biologists, as a "Stamp in Time", showing how far Protein Science has moved and "the Waters which may lie ahead".

Acknowledgements

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.

Author Contribution

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

Competing Interests

The authors declare no competing interests.

Reference

1. Kim Min-Sik, et al "A draft map of the human proteome." *Nature*. 509, no 7502 (2014): 575–581. doi: 10.1038/nature13302.
2. Ponomarenko, Elena A., Ekaterina V. Poverennaya, Ekaterina V. Ilgisonis, Mikhail A. Pyatnitskiy, Arthur T. Kopylov, Victor G. Zgoda, Andrey V. Lisitsa, and Alexander I. Archakov. "The Size of the Human Proteome: The Width and Depth." *International Journal of Analytical Chemistry* 2016 (2016): 1-6. doi:10.1155/2016/7436849.
3. Dolgin, Elie. "The Most Popular Genes in the Human Genome." *Nature* 551, no. 7681 (2017): 427-31. doi:10.1038/d41586-017-07291-9.
4. Rubin, G. M. "Comparative Genomics of the Eukaryotes." *Science* 287, no. 5461 (2000): 2204-215. doi:10.1126/science.287.5461.2204.
5. Romero, Pedro, Jonathan Wagg, Michelle L. Green, Dale Kaiser, Markus Krummenacker, and Peter D Karp. "Computational prediction of human metabolic pathways from the complete human genome." *Genome Biology* 6, no. 1 (2005): R2. doi: 10.1186/gb-2004-6-1-r2.
6. Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. "The Human Transcription Factors." *Cell* 175, no. 2 (2018): 598-99. doi:10.1016/j.cell.2018.09.045.
7. Miao, Hui, Da-Qiang Li, Amitava Mukherjee, Hong Guo, Aaron Petty, Jennifer Cutter, James P. Basilion, John Sedor, Jiong Wu, David Danielpour, Andrew E. Sloan, Mark L. Cohen, and Bingcheng Wang. "EphA2 Mediates Ligand-Dependent Inhibition and Ligand-Independent Promotion of Cell Migration and Invasion via a Reciprocal Regulatory Loop with Akt." *Cancer Cell* 16, no. 1 (2009): 9-20. doi:10.1016/j.ccr.2009.04.009.
8. Wu, Peng, Thomas E. Nielsen, and Mads H. Clausen. "FDA-approved Small-molecule Kinase Inhibitors." *Trends in Pharmacological Sciences* 36, no. 7 (2015): 422-39. doi:10.1016/j.tips.2015.04.005.
9. Bocharov, Eduard V., Dmitry M. Lesovoy, Konstantin V. Pavlov, Yulia E. Pustovalova, Olga V. Bocharova, and Alexander S. Arseniev. "Alternative Packing of EGFR Transmembrane Domain Suggests That Protein–lipid Interactions Underlie Signal Conduction across Membrane." *Biochimica Et Biophysica Acta (BBA) - Biomembranes* 1858, no. 6 (2016): 1254-261. doi:10.1016/j.bbamem.2016.02.023.
10. Monod, Jacques, Jean-Pierre Changeux, and François Jacob. "Allosteric Proteins and Cellular Control Systems." *Journal of Molecular Biology* 6, no. 4 (1963): 306-29. doi:10.1016/s0022-2836(63)80091-1.
11. Lange, O. F., N.-A. Lakomek, C. Fares, G. F. Schroder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmuller, C. Griesinger, and B. L. De Groot. "Recognition Dynamics Up to Microseconds Revealed from an RDC-Derived Ubiquitin Ensemble in Solution." *Science* 320, no. 5882 (2008): 1471-475. doi:10.1126/science.1157092.
12. Frederick, Kendra King, Michael S. Marlow, Kathleen G. Valentine, and A. Joshua Wand. "Conformational Entropy in Molecular Recognition by Proteins." *Nature* 448, no. 7151 (2007): 325-29. doi:10.1038/nature05959.
13. Prakash, Priyanka, Yong Zhou, Hong Liang, John F. Hancock, and Alemayehu A. Gorfe. "Oncogenic K-Ras Binds to an Anionic Membrane in Two Distinct Orientations: A Molecular Dynamics Analysis." *Biophysical Journal* 110, no. 5 (2016): 1125-138. doi:10.1016/j.bpj.2016.01.019.
14. Harms, M. J., G. N. Eick, D. Goswami, J. K. Colucci, P. R. Griffin, E. A. Ortlund, and J. W. Thornton. "Biophysical Mechanisms for Large-effect Mutations in the Evolution of Steroid Hormone Receptors." *Proceedings of the National Academy of Sciences* 110, no. 28 (2013): 11475-1480. doi:10.1073/pnas.1303930110.
15. Prior, I. A., P. D. Lewis, and C. Mattos. "A Comprehensive Survey of Ras Mutations in Cancer." *Cancer Research* 72, no. 10 (2012): 2457-467. doi:10.1158/0008-5472.can-11-2612.
16. O'brien, Richard J., and Philip C. Wong. "Amyloid Precursor Protein Processing and Alzheimer's Disease." *Annual Review of Neuroscience* 34, no. 1 (2011): 185-204. doi:10.1146/annurev-neuro-061010-113613
17. Shang, Jian, Gang Ye, Ke Shi, Yushun Wan, Chuming Luo, Hideki Aihara, Qibin Geng, Ashley Auerbach, and Fang Li. "Structural Basis of Receptor Recognition by SARS-CoV-2." *Nature* 581, no. 7807 (2020): 221-24. doi:10.1038/s41586-020-2179-y.
18. Daly, James L., Boris Simonetti, Carlos Antón-Plágaro, Maia Kavanagh Williamson, Deborah K. Shoemark, Lorena Simón-Gracia, Katja Klein, Michael Bauer, Reka Hollandi, Urs F. Greber, Peter Horvath, Richard B. Sessions, Ari Helenius, Julian A. Hiscox, Tambet Teesalu, David A. Matthews, Andrew D. Davidson, Peter J. Cullen, and Yohei Yamauchi. "Neuropilin-1 Is a Host Factor for SARS-CoV-2 Infection." 2020. *BioRxiv* doi:10.1101/2020.06.05.134114.