Say it with proteins: an alphabet of crystal structures

From manual searching of the Protein Data Bank (PDB), I have curated a set of protein crystal structures corresponding to the capital letters of the Roman alphabet (**Fig. 1**). In choosing structures, I aimed to include a range of different structural motifs and to exclude nucleic acids or proteins solved while bound to nucleic acids. Sometimes these letter shapes seem to be incidental, and sometimes the shape is key to the protein's biological function. For example, the specific shape is likely to be important for *L* (from elongation factor P), which mimics the shape of tRNA; for the sinuous *W* (from DNA-binding domain from BurrH), which tracks DNA's major groove for modular sequence recognition; and for proteins with holes that enclose DNA (*A*, from DNA gyrase) or puncture the membrane (*O*, from the toxin cytolysin A). PDB accession codes and descriptions of function for all proteins are provided in **Supplementary Table 1**. This set may be useful for outreach and teaching, by drawing attention to the diversity of protein structures attained by natural selection. It is conceivable that the set may also have value in bionanotechnology and synthetic biology, in which at times molecular assembly needs a specific shape more than a specific function.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nsmb.3011).

ACKNOWLEDGMENTS
Funding was provided by Worcester College Oxford. I thank E. Lowe (University of Oxford) for the diffraction image.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

Mark Howarth
Department of Biochemistry, Oxford University, Oxford, UK.

e-mail: mark.howarth@bioch.ox.ac.uk

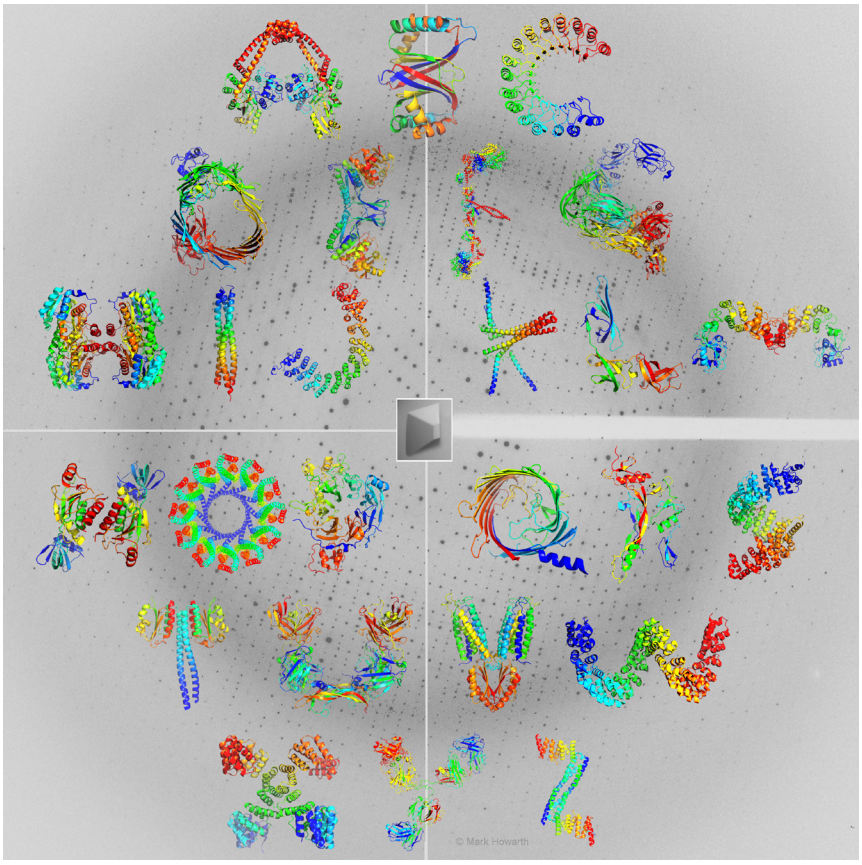


Figure 1 A protein alphabet. Selected protein crystal structures from the PDB in cartoon format and alphabetical order, overlaid on a diffraction image (provided by E. Lowe), with a central bright-field image of a protein crystal. Proteins are colored with the chainbows format, with the N terminus of each chain in blue through to the C terminus in red. Proteins are shown in monomeric (*C*, *D*, *G*, *J*, *L*, *P*, *Q*, *W*), homodimeric (*A*, *B*, *E*, *M*, *N*, *T*, *V*, *X*), heterodimeric (*R*, *S*), homotrimeric (*I*), heterotrimeric (*Z*), homotetrameric (*H*, *K*), heterotetrameric (*Y*), heterohexameric (*F*, *U*) or homododecameric (*O*) forms.

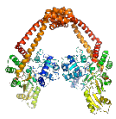
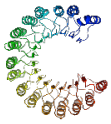
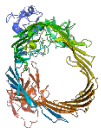
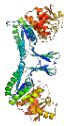
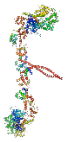
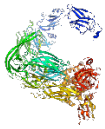
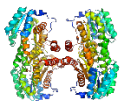
Supplementary Information

Say it with proteins:
an alphabet of crystal structures

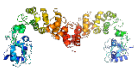
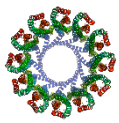
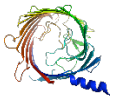
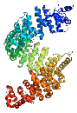
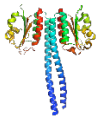
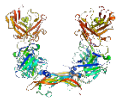
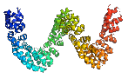
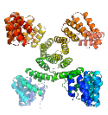
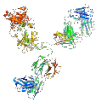
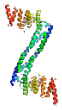
Mark Howarth,
Department of Biochemistry, University of Oxford,
South Parks Road, Oxford, OX1 3QU, UK.
mark.howarth@bioch.ox.ac.uk

Nature Structural & Molecular Biology doi:10.1038/nsmb.3011

Supplementary Table 1 PDB code, general function and distinctive features of alphabetical protein structures.

Letter	Image	PDB code	Function	Comments
A		3ifz	DNA topology	DNA gyrase reaction core from *Mycobacterium tuberculosis*; target of antibiotics.
B		2qyc	Unknown	Ferredoxin-like protein from *Bordetella bronchiseptica*.
C		2bnh	Blocking RNA degradation	Ribonuclease inhibitor from pig (*Sus scrofa*); contains leucine-rich repeats; RNase binds to the center of the protein with exceptional affinity.
D		4j3o	Pore for export of adhesion proteins	FimD from *Escherichia coli* forms the usher pore; contains a 24-stranded β-barrel; non-pore subunits were removed to generate the image.
E		2q5r	Milk sugar metabolism	Tagatose-6-phosphate kinase from *Staphylococcus aureus*; β-clasp dimer interface.
F		3j04	Smooth muscle contraction	Fragments of myosin-11, myosin regulatory light chain 2, and myosin light polypeptide 6 from chicken (*Gallus gallus*); structure from electron crystallography of 2D array.
G		4u48	Protease inhibitor	α2-macroglobulin from *Salmonella enterica*; mimics the α2-macroglobulins in eukaryotic innate immunity.
H		1xu9	Steroid metabolism	11β-hydroxysteroid dehydrogenase type I from *Homo sapiens*; catalyzes interconversion of cortisone and cortisol; 4-helix bundle mediates tetramerization.
I		3h7x	Bacterial adhesion	Stalk domain of YadA adhesin from *Yersinia enterocolitica*; trimeric coiled-coil.
J		1b3u	Intracellular signaling	PR65/A subunit of Protein Phosphatase 2A from *Homo sapiens*; contains 15 HEAT motifs.
K		4ox0	Gene regulation	Keratin-like domain of transcription factor SEPALLATA3 from *Arabidopsis thaliana*.
L		1ueb	Protein synthesis	Elongation Factor P from *Thermus thermophilus*; three β-barrels, mimicking charge and L-shape of transfer RNA.

Nature Structural & Molecular Biology doi:10.1038/nsmb.3011

M		1ou5	Protein synthesis	ATP(CTP):tRNA nucleotidyltransferase from *Homo sapiens*; adds CCA trinucleotide to 3' of transfer RNA.
N		1z85	RNA methyl-transferase (predicted)	Hypothetical protein TM1380 from *Thermotoga maritima*; β-barrel and 3-layer sandwich.
O		2wcd	Bacterial toxin	Cytolysin A from *Escherichia coli*; 12 copies of 3-helix bundle.
P		3afc	Development of nervous system and blood vessels	Semaphorin 6A extracellular domain from mouse (*Mus musculus*); contains β-propeller fold.
Q		3szv	Membrane channel	OccK3 outer membrane channel from *Pseudomonas aeruginosa*; 18-stranded β-barrel.
R		2arp	Regulation of differentiation and inflammation	Activin A from *Homo sapiens* bound to a fragment of follistatin from rat (*Rattus norvegicus*).
S		2ot8	Nuclear import	Transportin-1 from *Homo sapiens* recognizing a nuclear localization signal; contains HEAT repeats.
T		3e98	Unknown	GAF domain from *Pseudomonas aeruginosa*.
U		2vwe	Control of blood vessel formation	Vascular Endothelial Growth Factor-B from *Homo sapiens* bound to neutralizing antibody fragment.
V		3h90	Metal ion transport	YiiP zinc transporter from *Escherichia coli*.
W		4cj9	DNA-binding protein	DNA-binding domain of BurrH from *Burkholderia rhizoxinica*; helix-loop-helix repeats with modular DNA-binding specificity; potentially useful for genome editing.
X		1w3b	Protein glycosylation	Tetratricopeptide repeat domain of O-linked N-acetylglucosamine transferase from *Homo sapiens*.
Y		1igt	Immune defence	IgG2a antibody from mouse (*Mus musculus*); flexibility of arms relates to recognition of targets.
Z		4bta	Collagen stabilization	Part of collagen prolyl 4-hydroxylase from *Homo sapiens*; 4-helix bundle for dimerization; substrate peptide bound.

Nature Structural & Molecular Biology doi:10.1038/nsmb.3011