Current Research in Structural Biology 3 (2021) 146–152

Contents lists available at ScienceDirect

Current Research in Structural Biology

journal homepage: www.journals.elsevier.com/current-research-in-structural-biology

"Protein" no longer means what it used to

Gustavo Parisi [a,*], Nicolas Palopoli [a], Silvio C.E. Tosatto [b], María Silvina Fornasari [a], Peter Tompa [c,d,e,**]

[a] Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, CONICET, Bernal, Buenos Aires, Argentina
[b] Department of Biomedical Sciences, University of Padua, Padua, Italy
[c] VIB-VUB Center for Structural Biology (CSB), Brussels, Belgium
[d] Structural Biology Brussels (SBB), Vrije Universiteit Brussel (VUB), Brussels, Belgium
[e] Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary

ARTICLE INFO

Handling editorGlaucius Oliva

Keywords:
Native state
Protein types
Heterogeneity

ABSTRACT

Every biologist knows that the word *protein* describes a group of macromolecules essential to sustain life on Earth. As biologists, we are invariably trained under a protein paradigm established since the early twentieth century. However, in recent years, the term *protein* unveiled itself as an euphemism to describe the overwhelming heterogeneity of these compounds. Most of our current studies are targeted on carefully selected subsets of proteins, but we tend to think and write about these as representative of the whole population. Here we discuss how seeking for universal definitions and general rules in any arbitrarily segmented study would be misleading about the conclusions. Of course, it is not our purpose to discourage the use of the word *protein*. Instead, we suggest to embrace the extended universe of proteins to reach a deeper understanding of their full potential, realizing that the term encompasses a group of molecules very heterogeneous in terms of size, shape, chemistry and functions, i.e. the term *protein* no longer means what it used to.

The following passage quotes a 'certain Chinese encyclopaedia'[1] in which it is written that animals are divided into:

(a) belonging to the Emperor
(b) embalmed
(c) tame
(d) suckling pigs
(e) sirens
(f) fabulous
(g) stray dogs
(h) included in the present classification
(i) frenzied
(j) innumerable
(k) drawn with a very fine camel hair brush
(l) etcetera
(m) having just broken the water pitcher
(n) that from a long way off look like flies

1. Introduction

In his book "Other inquisitions", Jorge Luis Borges, the brilliant Argentinian writer, used an awkward classification of animals (cited above) to emphasise his idea that "obviously there is no classification of the universe that is not arbitrary and conjectural". For Borges, every classification schema will have intrinsic errors due to our ignorance of potential new categories or the arbitrariness to distinguish entities as belonging to different categories.

In the last years, it became unavoidable to realize that the word *protein* resembles Borges's famous classification of animals:

(a) Myoglobin-like
(b) Without form
(c) Small
(d) With knots
(e) Repetitives
(f) Coming from meteorites
(g) Walking proteins

* Corresponding author.
** Corresponding author. VIB-VUB Center for Structural Biology (CSB), Brussels, Belgium.
E-mail addresses: gusparisi@gmail.com (G. Parisi), peter.tompa@vub.be (P. Tompa).

[1] Jorge Luis Borges, *Otras Inquisiciones* (*Other Inquisitions*), "The Analytical Language of John Wilkins." 1952.

https://doi.org/10.1016/j.crstbi.2021.06.002
Received 12 April 2021; Received in revised form 18 June 2021; Accepted 22 June 2021

(h) Proteins that never existed on Earth
(i) Resurrected
(j) Artificially designed
(k) Circular
(l) With holes as a Gruyere cheese
(m) Containing the sequence "IADAPTEDASDIDMYCHANCES"
(n) Not even discovered

Borges would say that this classification of proteins would be no better or worse than tidier ones commonly used by scientists (for example, structural (CATH (Orengo et al., 1999)) or evolutionary (INTERPRO (Mitchell et al., 2019)) classifications). Completely disregarding the philosophical issues with classifications, during the last decades, hundreds of manuscripts purported that "Proteins evolve under …", "Protein fold stability is crucial for …", "The universe of protein structures …", etc., claiming to derive general knowledge about protein nature. The caveat with all such generalizations is that it is not (cannot be) defined which sort of proteins they are referring to. Using the word as is, implies that we are at full confidence of its meaning and coverage, i.e. that (i) we have discovered all (sort of) proteins, and (ii) we can unequivocally decide if a newly described molecule belongs into this category.

Our traditional definition of proteins grew out from early concepts formulated by Fischer, (1894) and by Pauling and Mirsky (Mirsky and Pauling, 1936). In accord, proteins are currently defined as unbranched, linear polypeptide chains of >100 amino acids in length, transcribed and translated from genomic open reading frames (ORFs). They are built by twenty genetically determined L-amino acids (except for Glicine that is neither L- nor D- and also some organisms use the additional amino acids Selenocysteine and Pyrrolysine (Zhang et al., 2005)) and fold by hydrophobic collapse into a unique native structure that carries biological activity. This native state could be built by a single polypeptide chain or by several, either identical or different ones. The combination of several chains can be necessary to express a function (the functional form is the combined one and individual chains have no biological function) or transient, but always meets well-defined stoichiometries of the constituent chains. Structural and functional information is encoded in the primary structure of coding sequences and is shaped by evolutionary processes (Kessel et al., 2010; Mathews et al., 2012). A protein thus defined is a structural, evolutionary and functional unit of the cell.

Does this definition cover all macromolecules considered "proteins"? Do the rules and knowledge derived in the aforementioned studies apply to all known cases? Have we already discovered all types of proteins? Soberingly, the answer to these questions is almost always "no". As outlined in this opinion article, proteins are widely heterogeneous in terms of their length, form, structure, chemical composition, (evolutionary) origin and structure-function relationship, and relevant discoveries or technological innovations continuously expand their universe.

2. The origin of the paradigm

Most of us take it for granted that "proteins" are globular (Fig. 1A), more or less compact spherical particles with most of their charged residues on their surface and a core mostly composed of hydrophobic amino acids. This paradigm originated in the early twentieth century, with studies on serum proteins and other animal and plant proteins (Block, 1935). Between 1920 and 30, a series of diffusion, sedimentation and viscosity experiments starting with the Linderstrøm-Lang adaptation of the Debye-Huckel model to multi-charged particles, such as proteins, (Comptes rendus des traveaux du Laboratoire Carlsberg, 1924), showed that most proteins being studied at that time (hemoglobin, egg albumin, and serum albumin) were "globular" (for an excellent review on the history of proteins, see (Tanford and Reynolds, 2001)). During the same period, increasing evidence on the existence of fibrous proteins (Fig. 1B) accumulated, both as deviations from globularity when studied in solution (casein and gelatin) or as direct observation of fibres under the microscope (keratin, silk fibroin, and collagen) or through x-ray crystallography (Astbury and Woods, 1930). By the mid 1930's, after crystallizing pepsin and confirming its globularity, the dichotomy of the proteins "bestiary" was established (Philpot and Eriksson-Quensel, 1933; Bernal and Crowfoot, 1934; Astbury, 1937).

In spite of this dichotomy, we still largely equate "proteins" with "globularity", due to their predominance in proteins studied and deposited in structure-based databases. For example, using the structural classification of proteins in SCOP (Andreeva et al., 2020), ~95% of the known structural space of proteins belongs to the "globular" category. Accumulation of protein structures also allowed their comparison in search for similarities. Recurrent structural patterns (folds) received names such as Rossman fold or TIM barrel, defined by their topologies (Rao and Rossmann, 1973; Banner et al., 1975; Lasters et al., 1988). Fold types were assigned by Levitt and Chothia based on the content and organization of their secondary structures, defining the now generally accepted α, α+ß, α/ß and ß classes (Levitt and Chothia, 1976). The term "domain" also arose, allowing to segment, visualise and further classify protein structures (Fig. 1C). By that time, repetitive proteins (Fig. 1D) started to be studied (Ycas, 1976) and accumulated in the "Atlas of protein sequences and structure" by Margaret Dayhoff (Dayhoff, 1965), which inspired the development of tools to study them (Barker et al., 1978). It became clear that different globular arrangements can emerge from repeating a given unit allowing a further classification of repetitive globular proteins (Kajava, 2001) (Fig. 1E). Interestingly, proteins with a given repeated unit could embrace both canonical structural types (globular and fibrous), highlighting that classification schemes are often arbitrary and non-orthogonal (Di Domenico et al., 2014).

3. Strange things

As the number of proteins with known structures steadily rose, odd folds and arrangements started to appear, contributing to further diversification of the protein repertoire. The left-handed twist of connections between parallel β-strands (Fig. 1F) or loops crossing multiple (more than four) or mixed layers, or connecting antiparallel adjacent regions, are very rarely observed in proteins (Finkelstein and Ptitsyn, 2016). Intertwined oligomers with "swapped domains" have been also described (Fig. 1G). 3D domain swapping in oligomers occurs by adding individual monomers that exchange from secondary elements up to whole domains (Bennett et al. 1995, 2006). Naturally occurring circular proteins (Fig. 1H) came on the scene in the 1990s (for a review see (Trabi and Craik, 2002)) and have, by now, been observed in plants, fungi, bacteria and also in mammals (Craik et al., 2003; Göransson et al., 2012). They can be as long as 78 residues (Conlan et al., 2010) and tend to be very stable, since their circularity could reduce their sensitivity to proteolytic degradation. Mechanism of cyclization of well characterized systems involves a specific asparaginyl endopeptidase both for the processing of the polypeptide chain and its cyclization by transpeptidation (Conlan et al., 2010; Du et al., 2020). Another class of strange folds are those present in proteins with knots. Although they are not true mathematical knots (which require presence of no ends as in proteins), these proteins show entangled arrangements generally described as knotoids to adapt the name to polymers with free ends (Dabrowski-Tumanski et al., 2019). The first characterized knotted protein (Fig. 1I) was carbonic anhydrase B (Mansfield, 1994). Knots in proteins are infrequent (there are currently about 1800 entries in the KnotProt 2.0 database (Dabrowski-Tumanski et al., 2019)) but, interestingly, inspection of evolutionary relationships of knotted proteins indicates that knots are conserved during evolution (Sułkowska et al., 2012). It has been suggested that knots could prevent degradation, increase thermal and mechanical stability, help in building binding sites and also promote the formation of polymeric filaments (for a review see (Faísca, 2015)).

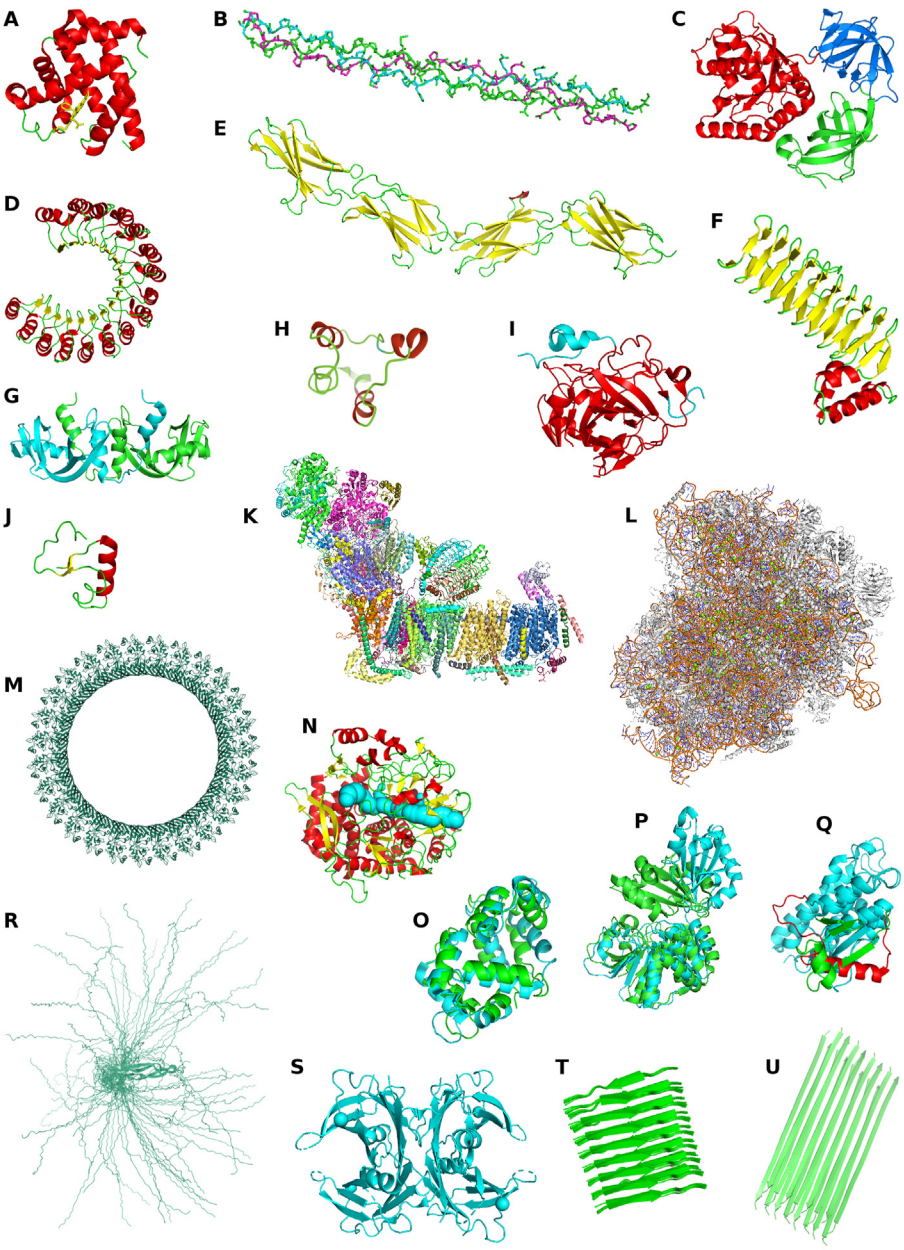

Fig. 1. Major protein types described in the text. A. Wild type sperm whale myoglobin as a typical example of globular proteins, well populated in secondary structure as well as in a hydrophobic core (PDB ID 5iks). B. Type III collagen containing three interwounded helix, as an example of fibrillar protein (3dmw). C. Elongation factor Tu showing three different domains. Each of these domains are segments that can adopt its fold independently of the rest of the protein (2c78). D. The mouse ribonuclease inhibitor is an example of a repetitive protein of the class α/ß solenoid (3tsr). E. Human fibronectin as an example of a globular protein with domain repeats (3t1w). F. Structure of the UDP-N-acetylglucosamine acyltransferase as an example of a protein containing a left-handed β-helix with unusual left-handed connections (1lxa). G. Crystal structure of bovine pancreatic ribonuclease A as an example of 3D domain-swapp. The N-terminal helix of each subunit (green and cyan) is swapped into the major domain of the other subunit. H. Structure of Acidocin B, a circular bacteriocin, an antimicrobial ribosomally synthesized peptide, from *Lactobacillus acidophilus* M46 (2mwr). I. Human carbonic anhydrase IX catalytic domain as an example of a knotted protein displaying a trefoil knot (6y74). Knot regions (cyan) and knot range (red) are displayed following its annotation in KnotProt 2.0 database. J. Crystal structure of crambin, a small seed storage protein (just 46 amino acid long) (3nir). K. Cryo-electron microscopy structure of plant mitochondrial respiratory complex I from *Brassica oleracea* as an example of large multi-chain assembly containing 44 unique proteins (7a23). L An example of a supramolecular structure, the ribosome 80 S subunit from *Homo sapiens* with 76 different protein chains and 5 RNAs (6ek0). M. Highly symmetric protein with a 27-fold symmetric pore known as Gasdermin A3 (6cb8). N. Cellulose cel48 F from *Clostridium cellulolyticum* is an example of a rigid protein, showing conformational diversity only at the residue level that allows open and close of tunnels (in cyan) for the transit of the substrate (1f9d). O. Calmodulin, a Ca2++ sensor protein, is a hub protein that can interact with more than 350 partners and display large conformational diversity, although it is commonly considered an ordered protein (two different conformers, 1niw in green, 1lin in cyan). P. Higher conformational changes can be obtained by hinge motions as between the open and closed structures of the type-C inorganic pyrophosphatases from *Streptococcus gordonii* (1k20 in green or closed conformer, 1k23 in cyan or open conformer). Q. Alternative extreme conformational changes involving secondary structural elements in CLIC1 protein from *Homo sapiens* as an example of fold-switching proteins (1rk4, 1k0n). Both structures are represented in cyan while their structural differences are colored in red and green. R. The NMR derived conformational ensemble of sclerostin, a secreted glycoprotein with a key negative regulatory role in Wnt signaling in bone. Sclerostin has two highly flexible N- and C-terminal regions with more than 50% of the protein being disordered (2k8p). S and T. Transthyretin, a thyroid hormone-binding protein that can adopt two very different conformations, a wild-type tetrameric form (4mrb in cyan) and one found in human diseases adopting an amyloid fibril (2m5n in green). U. Several proteins can adopt the same amyloid fibrils but as their main functional state, such as the human peptide hormone glucagon (6nzn).

4. A matter of sizes

The average length of proteins in UniProt (Wu et al., 2006) is ~337 residues, with the great majority of them falling between ~100 and 500, resulting in part from genome annotations using an arbitrary cutoff of a minimum of 100 codons for ORFs. Whereas this cutoff may reduce the level of spurious annotations (Orr et al., 2020), it makes small ORFs largely overlooked and biases the length distribution of known proteins. Actually, in the last few years, there is increasing support for the existence of small proteins, with less than 100 - or even 50 - amino acids, depending on their definition (Su et al., 2013; Storz et al., 2014). The ambiguity in this matter is also stressed by often calling such short proteins "peptides", also referring to the fact that sometimes they might not even be ribosomally synthesized (Finking and Marahiel, 2004). Recently, several publications have suggested that small proteins (Fig. 1J) can be found in all classes of organisms and are involved in various important functional roles such as information storage and processing, cellular signaling and metabolism (Su et al., 2013).

At the other extreme of the size spectrum, large supramolecular protein assemblies (Fig. 1K and L) are also increasingly being characterized (Pieters et al., 2016; Steven et al., 2016). Such structural associations are part of huge macromolecular complexes, like chromosomes, ribosomes, spliceosomes or DNA replication complexes; large protein assemblies such as complex I in membrane mitochondria, chaperones, chaperonins and the proteasome; other assemblies including fibers, tubes, catenanes, and cages; or molecular motors like myosin- and kinesin-associated motors or proton-driven motors. Assemblies can even lack strict stoichiometries and symmetry (important exceptions to this rule exist, as illustrated in Figure M). Recently, it has been recognized that many proteins, often along with RNA, form highly dynamic, non-stoichiometric liquid-like assemblies (droplets) *in vitro* and in cells, by a process of liquid-liquid phase separation (LLPS). The resulting bodies represent a newly recognized cellular organizational principle by membraneless organelles, such as stress granules and nucleoli (Shin and Brangwynne, 2017). These, and the other noted higher-order assemblies showcase the collective (emergent) functioning of proteins (Pancsa et al., 2019), highlighting that many proteins may not behave as evolutionary and/or functional units of the cell.

5. And yet, it moves (*E pur si muove*)

Changes in oxy- and deoxy-hemoglobin crystals were the first evidence that protein structure is not rigid (Haurowitz, 1938). In 1950, Fred Karush proposed that the native state of albumin encompasses different conformers in equilibrium, in order to explain its unusual binding isotherm (Karush, 1950). This early idea was further developed by the induced-fit binding model of Daniel Koshland (Koshland, 1958) followed by the model of pre-equilibrium by Jacques Monod (Monod et al., 1965). By the end of the century, a spectacular increase in the number of crystallographic structures and NMR models allowed the comparison of different conformers which help understanding protein function (James and Tawfik, 2003; Wei et al., 2016) and also improved the classification of conformational arrangements (Gerstein and Krebs, 1998; Echols et al., 2003; Flores et al., 2007). In order to study the dynamic behaviour of proteins, large datasets of proteins were analyzed comprehensively by comparing conformers of the same protein (holo and apo forms, for example) (Burra et al., 2009; Monzon et al., 2017). It was found that the majority of protein crystal structures show no evidence of significant atomic movements (~60% of the proteins show a conformational diversity ~0.8 Å measured as their maximum RMSD between any available conformers (Monzon et al., 2017)), an observation in agreement with previous results (Gutteridge and Thornton, 2005). Although the peak of the RMSD distribution contains these "rigid" proteins (Figure 1N), the distribution shows a large skew towards proteins comprising higher RMSD values, indicative of different sorts of movements. Although these movements can be tiny and local, such as the rotation of individual residues in opening tunnels or enlarging cavities (Gora et al., 2013; Kingsley and Lill, 2015; Pravda et al., 2018), they can be much larger, ranging from the movement of loops (Figure 1O) (Gu et al., 2015) or rearrangement of secondary or tertiary structure elements, to the displacement of entire domains (Figure 1P) (Gerstein et al., 1994). More recently, conformers that can change their secondary structure and then turn to different tertiary structures (fold-switching proteins) have been described to be more common than previously thought (Figure 1Q) (Porter and Looger, 2018).

Protein movements were classified using different criteria (for example (Qi et al., 2005; Amemiya et al., 2012) and following those criteria it was found that the dynamical behaviour is mostly not conserved during evolution (Marino-Buslje et al., 2019).

The apotheosis of flexibility is reached in intrinsically disordered proteins (IDPs) (Figure 1R). IDPs were discovered around the turn of the century (Wright and Dyson, 1999) as proteins characterized by a lack of a stable three dimensional structure under native conditions (Tompa, 2002). For that reason, IDPs challenge the well-established foundational idea of structure-function relationship in molecular biology (Chouard, 2011). Their native states are represented by a large collection of conformers describing a complex ensemble. A reference collection of them is included in the PED database (Varadi et al., 2014). The behaviour of these ensembles is far from random and is very sensitive to environmental conditions, as well as post-translational modifications (Shimojo et al., 2016; Davey, 2019). Rather than an oddity, disordered proteins have a wide phylogenetic distribution and are particularly abundant in eukaryotes, with high variation among their different taxa (Pancsa and Tompa, 2012). They have a broad range of biological activities by regulating cell division, signal transduction and transcription, serving as targets of post-translational modifications, assisting protein folding as chaperones, or even enabling the self-assembly of large multiprotein complexes such as the ribosome (Jakob et al., 2014). Due to their structural flexibility and adaptability, they can engage with distinct partners and fulfil several distinct functions (i.e., moonlighting) (Tompa et al., 2005), representing a prime example of proteins defying the one protein-one function paradigm.

6. Ever darker shades of darkness

In 2002, Donald Rumsfeld, then the U.S. Secretary of Defense, stated that: "There are known knowns: there are things we know that we know. There are known unknowns. That is to say, there are things that we now know we don't know. But there are also unknown unknowns: there are things we do not know we don't know" (https://en.wikipedia.org/wiki/There_are_known_knowns). This overarching concept has been adapted to several and different areas such as risk evaluation, project management and of course, Biology. In particular, and related to proteins, the concept impinges on the idea that some structure-function relationships cannot be predicted from existing data. These proteins belong to the so-called dark proteome, because they cannot be modeled on existing PDB structures and might not fit in the classical structure-function relationship (Perdigão et al., 2015). The reasons for "darkness" may be inferred from the foregoing sections but are largely still not fully understood. We can easily predict what dark proteins are not (briefly, they are not mainly disordered (Perdigão et al., 2020) nor transmembrane, whereas they have slight compositional biases (Perdigão et al., 2015)). Although dark proteins fit the general requirements of proteins in the proteome, our computational tools (trained with "classical" proteins) overwhelmingly fail to characterize what they actually are. They could certainly be represented by several of the atypical arrangements we described before (scarcely included in protein *Zeitgeist*), but also by folds we have not seen yet, or simply, by the lack of a structure or a given structure-function relationship. As their frequency in eukaryotic and viral proteomes is estimated to be between 26 and 55%, many of the proteins in any proteome-level study are likely to be unknown unknowns (Perdigão et al., 2020).

Unexpectedly, there are even darker shades of protein darkness, i.e. macromolecules that fall even farther from our traditional protein concept and definition. Without attempting to be fully comprehensive in coverage, we should mention synthetic/artificial proteins manufactured for biotech purposes (Langan et al., 2019) that may even incorporate amino acids beyond the canonical twenty (Elsässer et al., 2016), non-evolved fusion proteins generated in cancer cells (Hegyi et al., 2009), proteins generated by *de novo* gene birth or intron exonization (Van Oss and Carvunis, 2019) or proteins made up entirely of D-amino acids (at least, *in silico*) to develop peptides made of D-amino acids that would effectively bind natural proteins (Garton et al., 2018). Beyond these examples of a broad variety noted, there are even more curious ones, such as peptide nucleic acids (PNAs), synthetic mimics of DNA derived from (poly)peptide chemistry (Ricciardi et al., 2018). Being powerful tools in DNA manipulation and analytics, they may represent the ultimate challenge to our conceptualization of proteins.

7. Discussion

The types of proteins listed above are not intended to be exhaustive. We deliberately omitted proteins e.g. with biased composition (Pascal et al., 2005; Cascarina and Ross, 2018) or with post-translational modifications, proteins affected by particular metabolic/physiological processes (expression level (Lemos et al., 2005), essentiality (Alvarez-Ponce et al., 2016), export (Loos et al., 2019), etc), proteins evolving under given evolutionary processes (positive selection, conformational epistasis (Ortlund et al., 2007), etc), or proteins with transient and/or permanent oligomers. Our intention is to challenge the reader by offering a glimpse of their astonishing heterogeneity. As we mentioned before, hundreds of scientific works claim to describe the influence of certain processes, parameters, and factors on "protein" biology as a whole, as if the latter represented a homogeneous group of molecules. Apparently, these considerations about the apparent *homogeneity of proteins* could significantly impair and restrain our knowledge about protein biology.

Notwithstanding these reservations, paradigms in science are useful frames of thought which arrange and rationalize established knowledge on a given subject, although they inevitably delay the acceptance and development of new ideas. In protein history (Table 1), paradigm shifts turned the idea of proteins as colloids into their acceptance as true macromolecules (Tanford and Reynolds, 2001). In a similar way, the "lock and key" model to explain the structure-function relationship postulated early by Emil Fischer in 1894 (Fischer, 1894) was challenged twice, with conceptually different ideas. The first one, promoted by Daniel Koshland in the late fifties, was the consideration of protein movements to explain some anomalies in the kinetic behaviour of different enzymes. These movements were absent in the model of Fisher, who considered that native states were rigid "negatives" of their ligands with perfect complementarity molded in their binding sites. The induced-fit model and subsequent pre-equilibrium model (Monod et al., 1965) formalized the use of protein motions and gave a solid background for the development of current models of structure-function relationship (Wei et al., 2016). The second challenge to the "lock and key" model was much more recent, and then the observed "abnormalities" were the absence of a folded three-dimensional structure as a prerequisite for the protein to be functional (Romero et al., 1998; Wright and Dyson, 1999). Beside this, circular proteins and the presence of knots challenge the classic view of proteins as linear polymers, while non-stoichiometric supramolecular assemblies and their emerging functional properties challenge the established idea of a well defined functional native state (Pancsa et al., 2019). The recent characterization of functional amyloid fibrils and prions (Varadi et al., 2018), structures commonly found in 'protein misfolding' disorders (Soto et al., 2006) (Fig. 1S, T and U) challenge the central paradigm of how sequence information encodes structural and functional information. Proteins incorporating non-canonical amino acids broaden also the chemical horizon of proteins.

Table 1
Timeline of major concepts in protein structure-function relationships.

1884: Fischer's lock and key model.
1924: First insights about protein globularity.
1930: First x-ray interpretation of fibrillar nature of keratin.
1936: Native state definition by Minsky & Pauling.
1950: Karush proposed a native state with several conformers.
1958: Koshland's induced-fit model.
1960: Structure of hemoglobin
1965: Monod's pre-equilibrium model.
1976: Fold types are classified based on secondary structure content and organization.
1978: Methods to detect internal repetitions in proteins.
1987: Folding funnel hypothesis predicts that native state is its free energy minimum
1994: First characterized knotted protein.
1995: First structural evidence of circular proteins.
1997: Phase transition proteins are associated with human diseases.
1998: Conformational movements in ordered proteins start to be classified.
1999: Intrinsically disordered proteins start to shift the structure-function paradigm.
2006: First characterization of functional amyloid in mammals.
2007: Shortest protein (11 amino acids) characterized.
2015: Notion of "dark" proteins.
2018: Fold-switching proteins are proposed to be widespread

Apparently, with all these transitions based on revolutionary discoveries, the "protein paradigm" has not come to a standstill. Switching our view of proteins into a heterogeneous group of shapes, forms, motions, lengths, and compositions will allow us to uncover new properties and relationships that are yet hidden by the effect of our over-averaging view of protein entities. Why is this immense structural diversity of proteins required to sustain life? What is its origin? Which are the missing pieces that remain undiscovered in the dark proteome? Breaking dogmas is never easy and comfortable but, eventually, always has its rewards. Predictably, putting aside the concept of homogeneity will allow us to freely explore new compositions, forms, mechanisms and principles in the vast universe of proteins.

CRediT authorship contribution statement

Gustavo Parisi: Conceptualization, Writing - review & editing. **Nicolas Palopoli:** Writing - review & editing. **Silvio C.E. Tosatto:** Writing - review & editing. **María Silvina Fornasari:** Writing - review & editing. **Peter Tompa:** Conceptualization, Writing - review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Gustavo Parisi reports financial support was provided by National University of Quilmes.

Acknowledgements

NP, MSF and GP are researchers from CONICET. This work was supported by Universidad Nacional de Quilmes (PUNQ 1004/11), ANPCyT (PICT-2014-3430) and the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreements No. 778247 and No. 823886). PT acknowledges grants K124670 and K131702 from the Hungarian Scientific Research Fund (OTKA) and a Spearhead grant (SRP51, 2019–24) from VUB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We would like to thank Antonio Lagares (IBBM-CONICET) for useful comments during the preparation of this manuscript.

References

Alvarez-Ponce, D., Sabater-Muñoz, B., Toft, C., Ruiz-González, M.X., Fares, M.A., 2016. Essentiality is a strong determinant of protein rates of evolution during mutation accumulation experiments in Escherichia coli. Genome Biol. Evol. 8, 2914–2927.

Amemiya, T., Koike, R., Kidera, A., Ota, M., 2012. PSCDB: a database for protein structural change upon ligand binding. Nucleic Acids Res. 40, D554–D558.
Andreeva, A., Kulesha, E., Gough, J., Murzin, A.G., 2020. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. 48, D376–D382.
Astbury, W.T., Woods, H.J., 1930. The X-ray interpretation of the structure and elastic properties of hair keratin. Nature 126, 913–914.
Astbury, W.T., 1937. Relation between "fibrous" and "globular" proteins. Nature 140, 968–969.
Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Pogson, C.I., Wilson, I.A., Corran, P.H., Furth, A.J., Milman, J.D., Offord, R.E., et al., 1975. Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 angstrom resolution using amino acid sequence data. Nature 255, 609–614.
Barker, W.C., Ketcham, L.K., Dayhoff, M.O., 1978. A comprehensive examination of protein sequences for evidence of internal gene duplication. J. Mol. Evol. 10, 265–281.
Bennett, M.J., Sawaya, M.R., Eisenberg, D., 2006. Deposition diseases and 3D domain swapping. Structure 14, 811–824.
Bennett, M.J., Schlunegger, M.P., Eisenberg, D., 1995. 3D domain swapping: a mechanism for oligomer assembly. Protein Sci. 4, 2455–2468.
Bernal, J.D., Crowfoot, D., 1934. X-ray photographs of crystalline pepsin. Nature 133, 794–795.
Block, R.J., 1935. On the nature and origin of proteins. Yale J. Biol. Med. 7, 235–252.
Burra, P.V., Zhang, Y., Godzik, A., Stec, B., 2009. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. Proc. Natl. Acad. Sci. U.S.A. 106, 10505–10510.
Cascarina, S.M., Ross, E.D., 2018. Proteome-scale relationships between local amino acid composition and protein fates and functions. PLoS Comput. Biol. 14, e1006256.
Chouard, T., 2011. Structural biology: breaking the protein rules. Nature 471, 151–153.
Comptes rendus des traveaux du Laboratoire Carlsberg L-L K (Ed.), 1924. On the Ionisation of Proteins.
Conlan, B.F., Gillon, A.D., Craik, D.J., Anderson, M.A., 2010. Circular proteins and mechanisms of cyclization. Biopolymers 94, 573–583.
Craik, D.J., Daly, N.L., Saska, I., Trabi, M., Rosengren, K.J., 2003. Structures of naturally occurring circular proteins from bacteria. J. Bacteriol. 185, 4011–4021.
Dabrowski-Tumanski, P., Rubach, P., Goundaroulis, D., Dorier, J., Sulkowski, P., Millett, K.C., Rawdon, E.J., Stasiak, A., Sulkowska, J.I., 2019. KnotProt 2.0: a database of proteins with knots and other entangled structures. Nucleic Acids Res. 47, D367–D375.
Davey, N.E., 2019. The functional importance of structure in unstructured protein regions. Curr. Opin. Struct. Biol. 56, 155–163.
Dayhoff, M.O., 1965. Atlas of Protein Sequence and Structure, vol. I. National Biomedical Research Foundation.
Di Domenico, T., Potenza, E., Walsh, I., Parra, R.G., Giollo, M., Minervini, G., Piovesan, D., Ihsan, A., Ferrari, C., Kajava, A.V., et al., 2014. RepeatsDB: a database of tandem repeat protein structures. Nucleic Acids Res. 42, D352–D357.
Du, J., Yap, K., Chan, L.Y., Rehm, F.B.H., Looi, F.Y., Poth, A.G., Gilding, E.K., Kaas, Q., Durek, T., Craik, D.J., 2020. A bifunctional asparaginyl endopeptidase efficiently catalyzes both cleavage and cyclization of cyclic trypsin inhibitors. Nat. Commun. 11, 1575.
Echols, N., Milburn, D., Gerstein, M., 2003. MolMovDB: analysis and visualization of conformational change and structural flexibility. Nucleic Acids Res. 31, 478–482.
Elsässer, S.J., Ernst, R.J., Walker, O.S., Chin, J.W., 2016. Genetic code expansion in stable cell lines enables encoded chromatin modification. Nat. Methods 13, 158–164.
Faísca, P.F.N., 2015. Knotted proteins: a tangled tale of Structural Biology. Comput. Struct. Biotechnol. J. 13, 459–468.
Finkelstein, A.V., Ptitsyn, O.B. (Eds.), 2016. Protein Physics. A Course of Lectures. Elsevier.
Finking, R., Marahiel, M.A., 2004. Biosynthesis of nonribosomal peptides1. Annu. Rev. Microbiol. 58, 453–488.
Fischer, E., 1894. Einfluss der Configuration auf die Wirkung der Enzyme. Ber. Dtsch. Chem. Ges. 27, 2985–2993.
Flores, S.C., Lu, L.J., Yang, J., Carriero, N., Gerstein, M.B., 2007. Hinge Atlas: relating protein sequence to sites of structural flexibility. BMC Bioinf. 8, 167.
Garton, M., Nim, S., Stone, T.A., Wang, K.E., Deber, C.M., Kim, P.M., 2018. Method to generate highly stable D-amino acid analogs of bioactive helical peptides using a mirror image of the entire PDB. Proc. Natl. Acad. Sci. U.S.A. 115, 1505–1510.
Gerstein, M., Krebs, W., 1998. A database of macromolecular motions. Nucleic Acids Res. 26, 4280–4290.
Gerstein, M., Lesk, A.M., Chothia, C., 1994. Structural mechanisms for domain movements in proteins. Biochemistry 33, 6739–6749.
Gora, A., Brezovsky, J., Damborsky, J., 2013. Gates of enzymes. Chem. Rev. 113, 5871–5923.
Göransson, U., Burman, R., Gunasekera, S., Strömstedt, A.A., Rosengren, K.J., 2012. Circular proteins from plants and fungi. J. Biol. Chem. 287, 27001–27006.
Gu, Y., Li, D.-W., Brüschweiler, R., 2015. Decoding the mobility and time scales of protein loops. J. Chem. Theor. Comput. 11, 1308–1314.
Gutteridge, A., Thornton, J., 2005. Conformational changes observed in enzyme crystal structures upon substrate binding. J. Mol. Biol. 346, 21–28.
Haurowitz, F., 1938. Das Gleichgewicht zwischen Hämoglobin und Sauerstoff. Hoppe-Seyler's Zeitschrift für physiologische Chemie 254, 266–274.
Hegyi, H., Buday, L., Tompa, P., 2009. Intrinsic structural disorder confers cellular viability on oncogenic fusion proteins. PLoS Comput. Biol. 5, e1000552.
Jakob, U., Kriwacki, R., Uversky, V.N., 2014. Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function. Chem. Rev. 114, 6779–6805.
James, L.C., Tawfik, D.S., 2003. Conformational diversity and protein evolution–a 60-year-old hypothesis revisited. Trends Biochem. Sci. 28, 361–368.
Kajava, A.V., 2001. Review: proteins with repeated sequence–structural prediction and modeling. J. Struct. Biol. 134, 132–144.
Karush, F., 1950. Heterogeneity of the binding sites of bovine serum albumin. J. Am. Chem. Soc. 72, 2705–2713.
Kessel, Amit, Ben-tal, Nir, 2010. Introduction to Proteins: Structure, Function, and Motion (Chapman & Hall/crc Mathematical and Computational Biology), first ed. Crc Press.
Kingsley, L.J., Lill, M.A., 2015. Substrate tunnels in enzymes: structure-function relationships and computational methodology. Proteins 83, 599–611.
Koshland, D.E., 1958. Application of a theory of enzyme specificity to protein synthesis. Proc. Natl. Acad. Sci. U.S.A. 44, 98–104.
Langan, R.A., Boyken, S.E., Ng, A.H., Samson, J.A., Dods, G., Westbrook, A.M., Nguyen, T.H., Lajoie, M.J., Chen, Z., Berger, S., et al., 2019. De novo design of bioactive protein switches. Nature 572, 205–210.
Lasters, I., Wodak, S.J., Alard, P., van Cutsem, E., 1988. Structural principles of parallel beta-barrels in proteins. Proc. Natl. Acad. Sci. U.S.A. 85, 3338–3342.
Lemos, B., Bettencourt, B.R., Meiklejohn, C.D., Hartl, D.L., 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol. Biol. Evol. 22, 1345–1354.
Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552–558.
Loos, M.S., Ramakrishnan, R., Vranken, W., Tsirigotaki, A., Tsare, E.-P., Zorzini, V., Geyter, J.D., Yuan, B., Tsamardinos, I., Klappa, M., et al., 2019. Structural basis of the subcellular topology landscape of Escherichia coli. Front. Microbiol. 10, 1670.
Mansfield, M.L., 1994. Are there knots in proteins? Nat. Struct. Biol. 1, 213–214.
Marino-Buslje, C., Monzon, A.M., Zea, D.J., Fornasari, M.S., Parisi, G., 2019. On the dynamical incompleteness of the protein data bank. Briefings Bioinf. 20, 356–359.
Mathews, C.K., Van Holde, K.E., Appling, D.R., Anthony-cahill, S.J., 2012. Biochemistry (4th Edition), fourth ed. Pearson.
Mirsky, A.E., Pauling, L., 1936. On the structure of native, denatured, and coagulated proteins. Proc. Natl. Acad. Sci. U.S.A. 22, 439–447.
Mitchell, A.L., Attwood, T.K., Babbitt, P.C., Blum, M., Bork, P., Bridge, A., Brown, S.D., Chang, H.-Y., El-Gebali, S., Fraser, M.I., et al., 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 47, D351–D360.
Monod, J., Wyman, J., Changeux, J.P., 1965. On the nature of allosteric transitions: a plausible model. J. Mol. Biol. 12, 88–118.
Monzon, A.M., Zea, D.J., Fornasari, M.S., Saldaño, T.E., Fernandez-Alberti, S., Tosatto, S.C.E., Parisi, G., 2017. Conformational diversity analysis reveals three functional mechanisms in proteins. PLoS Comput. Biol. 13, e1005398.
Orengo, C.A., Pearl, F.M., Bray, J.E., Todd, A.E., Martin, A.C., Lo Conte, L., Thornton, J.M., 1999. The CATH Database provides insights into protein structure/function relationships. Nucleic Acids Res. 27, 275–279.
Orr, M.W., Mao, Y., Storz, G., Qian, S.-B., 2020. Alternative ORFs and small ORFs: shedding light on the dark proteome. Nucleic Acids Res. 48, 1029–1042.
Ortlund, E.A., Bridgham, J.T., Redinbo, M.R., Thornton, J.W., 2007. Crystal structure of an ancient protein: evolution by conformational epistasis. Science 317, 1544–1548.
Pancsa, R., Schad, E., Tantos, A., Tompa, P., 2019. Emergent functions of proteins in non-stoichiometric supramolecular assemblies. Biochim. Biophys. Acta Protein Proteonomics 1867, 970–979.
Pancsa, R., Tompa, P., 2012. Structural disorder in eukaryotes. PloS One 7, e34687.
Pascal, G., Médigue, C., Danchin, A., 2005. Universal biases in protein composition of model prokaryotes. Proteins 60, 27–35.
Perdigão, N., Heinrich, J., Stolte, C., Sabir, K.S., Buckley, M.J., Tabor, B., Signal, B., Gloss, B.S., Hammang, C.J., Rost, B., et al., 2015. Unexpected features of the dark proteome. Proc. Natl. Acad. Sci. U.S.A. 112, 15898–15903.
Perdigão, N., Pina, P.M.C., Rocha, C., Tavares, J.M.R.S., Rosa, A., 2020. Dark proteome database: studies on disorder. High-Throughput 9.
Philpot, J.S.T.L., Eriksson-Quensel, I.-B., 1933. An ultracentrifugal study of crystalline pepsin. Nature 132, 932–933.
Pieters, B.J.G.E., van Eldijk, M.B., Nolte, R.J.M., Mecinović, J., 2016. Natural supramolecular protein assemblies. Chem. Soc. Rev. 45, 24–39.
Porter, L.L., Looger, L.L., 2018. Extant fold-switching proteins are widespread. Proc. Natl. Acad. Sci. U.S.A. 115, 5968–5973.
Pravda, L., Sehnal, D., Svobodová Vareková, R., Navrátilová, V., Toušek, D., Berka, K., Otyepka, M., Koca, J., 2018. ChannelsDB: database of biomacromolecular tunnels and pores. Nucleic Acids Res. 46, D399–D405.
Qi, G., Lee, R., Hayward, S., 2005. A comprehensive and non-redundant database of protein domain movements. Bioinformatics 21, 2832–2838.
Rao, S.T., Rossmann, M.G., 1973. Comparison of super-secondary structures in proteins. J. Mol. Biol. 76, 241–256.
Ricciardi, A.S., Quijano, E., Putman, R., Saltzman, W.M., Glazer, P.M., 2018. Peptide nucleic acids as a tool for site-specific gene editing. Molecules 23.
Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Garner, E., Guilliot, S., Dunker, A.K., 1998. Thousands of proteins likely to have long disordered regions. Pac. Symp. Biocomput. 437–448.
Shimojo, H., Kawaguchi, A., Oda, T., Hashiguchi, N., Omori, S., Moritsugu, K., Kidera, A., Hiragami-Hamada, K., Nakayama, J.-I., Sato, M., et al., 2016. Extended string-like binding of the phosphorylated HP1α N-terminal tail to the lysine 9-methylated histone H3 tail. Sci. Rep. 6, 22527.
Shin, Y., Brangwynne, C.P., 2017. Liquid phase condensation in cell physiology and disease. Science 357.
Soto, C., Estrada, L., Castilla, J., 2006. Amyloids, prions and the inherent infectious nature of misfolded protein aggregates. Trends Biochem. Sci. 31, 150–155.

Steven, A., Baumeister, W., Johnson, L.N., Perham, R.N., 2016. Molecular Biology of Assemblies and Machines, first ed. Garland Science, New York.
Storz, G., Wolf, Y.I., Ramamurthi, K.S., 2014. Small proteins can no longer be ignored. Annu. Rev. Biochem. 83, 753–777.
Su, M., Ling, Y., Yu, J., Wu, J., Xiao, J., 2013. Small proteins: untapped area of potential biological importance. Front. Genet. 4, 286.
Sułkowska, J.I., Rawdon, E.J., Millett, K.C., Onuchic, J.N., Stasiak, A., 2012. Conservation of complex knotting and slipknotting patterns in proteins. Proc. Natl. Acad. Sci. U.S.A. 109, E1715–E1723.
Tanford, C., Reynolds, J., 2001. Nature's Robots: A History of Proteins, first ed. Oxford University Press, Oxford.
Tompa, P., Szász, C., Buday, L., 2005. Structural disorder throws new light on moonlighting. Trends Biochem. Sci. 30, 484–489.
Tompa, P., 2002. Intrinsically unstructured proteins. Trends Biochem. Sci. 27, 527–533.
Trabi, M., Craik, D.J., 2002. Circular proteins–no end in sight. Trends Biochem. Sci. 27, 132–138.
Van Oss, S.B., Carvunis, A.-R., 2019. De novo gene birth. PLoS Genet. 15, e1008160.
Varadi, M., De Baets, G., Vranken, W.F., Tompa, P., Pancsa, R., 2018. AmyPro: a database of proteins with validated amyloidogenic regions. Nucleic Acids Res. 46, D387–D392.
Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A.K., Felli, I.C., Forman-Kay, J.D., Kriwacki, R.W., Pierattelli, R., et al., 2014. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. Nucleic Acids Res. 42, D326–D335.
Wei, G., Xi, W., Nussinov, R., Ma, B., 2016. Protein ensembles: how does nature harness thermodynamic fluctuations for life? the diverse functional roles of conformational ensembles in the cell. Chem. Rev. 116, 6516–6551.
Wright, P.E., Dyson, H.J., 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 293, 321–331.
Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., et al., 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. 34, D187–D191.
Ycas, M., 1976. Origin of periodic proteins. Fed. Proc. 35, 2139–2140.
Zhang, Y., Baranov, P.V., Atkins, J.F., Gladyshev, V.N., 2005. Pyrrolysine and selenocysteine use dissimilar decoding strategies. J. Biol. Chem. 280, 20740–20751.