Research Article

Fodil Azzaz, Jacques Fantini*

The epigenetic dimension of protein structure

<https://doi.org/10.1515/bmc-2022-0006>

received December 17, 2021; accepted February 8, 2022

Abstract: Accurate prediction of protein structure is one of the most challenging goals of biology. The most recent achievement is AlphaFold, a machine learning method that has claimed to have solved the structure of almost all human proteins. This technological breakthrough has been compared to the sequencing of the human genome. However, this triumphal statement should be treated with caution, as we identified serious flaws in some AlphaFold models. Disordered regions are often represented by large loops that clash with the overall protein geometry, leading to unrealistic structures, especially for membrane proteins. In fact, AlphaFold comes up against the notion that protein folding is not solely determined by genomic information. We suggest that all parameters controlling the structure of a protein without being strictly encoded in its amino acid sequence should be coined “epigenetic dimension of protein structure.” Such parameters include for instance protein solvation by membrane lipids, or the structuration of disordered proteins upon ligand binding, but exclude sequence-encoded sites of post-translational modifications such as glycosylation. In our view, this paradigm is necessary to reconcile two opposite properties of living systems: beyond rigorous biological coding, evolution has given way to a certain level of uncertainty and anarchy.

Keywords: AlphaFold, Robetta, protein structure, intrinsically disordered proteins, membrane, lipid

Introduction

The prediction of 3D protein structures by the AlphaFold project is clearly a breakthrough for scientific and medical

research [1]. These structures, which now cover almost the entire human proteome, can be downloaded freely from the European Bioinformatics Institute database (<https://www.ebi.ac.uk>) and from the Uniprot server (<https://www.uniprot.org>). This long awaited technological breakthrough has been compared to the sequencing of the human genome, and the benefits are expected to be just as spectacular. However, without lessening this biotechnological achievement in any way, certain questions and potential limitations of this strategy should be considered.

While discussing with several colleagues about the publication recounting the success of AlphaFold, we realized that most of them believed that the problem of the protein structure had been completely solved. Indeed, this is what media and social networks have extensively relayed and commented on. Thus, many scientists consider browsing the AlphaFold database to download the structure of proteins of their interest, convinced that the AlphaFold structure is equivalent to the pdb structures obtained by crystallization and X-ray diffraction, cryo-electron microscopy, or nuclear magnetic resonance (NMR).

Therefore, it seemed interesting to us to wonder about the possible limitations of the AlphaFold algorithm. We present several representative examples of such proteins whose 3D structure is not accurately predicted by AlphaFold.

Based on this analysis, we propose a new concept which complements current models for 3D protein structure. This new paradigm, coined “epigenetic dimension of protein structure,” takes into account any parameter that is not encoded in the amino acid sequence of the protein. This concept excludes classical post-translational modifications of proteins such as acylation or glycosylation, which still depend on a perfectly determined code.

Methods

AlphaFold models were retrieved from Uniprot (<https://www.uniprot.org>), a high-quality resource of protein sequence and functional information. The tridimensional structures of proteins were predicted using the web-based services Robetta, an online software with a graphical interface user

* **Corresponding author: Jacques Fantini**, Department of Biology, Aix-Marseille Université and INSERM UMR_S 1072, Marseille, France, e-mail: jacques.fantini@univ-amu.fr, tel: +33-679-444-584
Fodil Azzaz: Department of Biology, Aix-Marseille Université and INSERM UMR_S 1072, Marseille, France
ORCID: Jacques Fantini 0000-0001-8653-5521

for ab-initio or homology modeling (<https://robetta.bakerlab.org/>). The amino acid sequence of the proteins of interest were submitted to Robetta in order to generate models obtained by the ab-initio modeling method. Two membrane proteins (the sodium-dependent serotonin transporter, Uniprot #P31645, and the synaptic vesicle glycoprotein SV2A, Uniprot #Q7L0J3), and a soluble protein (the cytosolic protein Mid1-interacting protein 1, Uniprot #Q9NPA3) were selected to be modeled for their high degree of structural versatility and divergence. The final models were visualized using Pymol (<https://pymol.org/2/>) and compared to the structural prediction proposed by AlphaFold.

Results

The AlphaFold algorithm produces a per-residue confidence score (pLDDT) between 0 and 100 [1]. It is acknowledged by the authors that “some regions with low pLDDT may be unstructured in isolation” [1]. As a matter of fact, most structural models are concerned by this limitation, as shown in Figure 1 for the cytosolic protein Mid1-interacting protein 1 (Uniprot #Q9NPA3). Indeed, regions with very high confidence (pLDDT > 90) represent at best only one third of the 3D structure, corresponding to α -helical domains. The other parts of the model are mostly represented as unresolved loops with low ($70 > \text{pLDDT} > 50$) and very low scores (pLDDT < 50). In this case, submitting the amino acid sequence to Robetta allowed to obtain an improved model by resolving all the loops without major perturbation of the core α -helix organization (Figure 2). Note that both modeling strategies converged for the localization of the core α -helix domains of this protein.

Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions with low pLDDT may be unstructured in isolation.

Figure 1: Example of AlphaFold structure. AlphaFold model of Mid1-interacting protein 1 downloaded from Uniprot #Q9NPA3. The confidence of the pLDDT is indicated by a color code.

Figure 2: Comparison of AlphaFold and Robetta for Mid1-interacting protein 1. The arrow shows the same domain in the AlphaFold and Robetta models.

The case of membrane proteins is more problematic. The difficulties are illustrated by three distinct proteins each displaying a typical and representative membrane topology. The first example is given by the Amyloid Protein Precursor (APP; Uniprot #P05067), a 770 amino acid protein expressed by neurons which, upon proteolytic cleavage, releases Alzheimer’s β -amyloid peptide in the extracellular space [2]. The precursor is a type 1 membrane glycoprotein with its N-terminus extracellular and its C-terminus cytosolic [3]. A single transmembrane (TM) helix links both the extra- and intra-cellular domains of the protein (Figure 3). Yet in the AlphaFold model, this topology is not respected at all, and the proposed structure is totally unrealistic.

For the second example we chose the sodium-dependent serotonin transporter [4] (Uniprot #P31645), a typical protein displaying 12 α -helical TM domains (Figure 4). In the AlphaFold model, a large unresolved loop corresponding to

Figure 3: Nonaccurate prediction of a membrane protein by AlphaFold. Amyloid protein precursor (APP) is a type 1 membrane glycoprotein with a single TM domain (left panel). The AlphaFold model (right panel) is totally unrealistic.

Figure 4: Comparison of AlphaFold, Robetta, and X-ray (pdb 5i6x) models for sodium-dependent serotonin transporter. The N- and C-terminus is indicated for all models.

the intracellular N-terminus crosses the plasma membrane and reaches the extracellular space. Interestingly, Robetta solved this problem by constructing a realistic intracellular domain from the disordered N-terminal region of the protein. This structural refinement did not affect the global organization of the 12 α -helical TM domains of the transporter, which remained almost identical in both AlphaFold and Robetta. Moreover, both AlphaFold and Robetta models are consistent with the X-ray diffraction structure of the transporter obtained by X-ray diffraction (pdb file #5i6x) [4]. However, this pdb model lacks the 75 first (N-terminus) and the 12 last amino acid residues (C-terminus) of the transporter (Figure 4). Hence, the prediction of these missing regions is critical to assess the confidence of the algorithms.

Finally, our third example is the synaptic vesicle glycoprotein 2A [5] (SV2A; Uniprot #Q7L0J3). This protein has a complex topology, with a series of perpendicular α -helix domains and a large β -structure extracellular domain (Figure 5). This protein is recognized by botulinum neurotoxins which use it as a receptor to gain entry into host cells [6].

Figure 5: Comparison of AlphaFold and Robetta for the synaptic vesicle protein SV2A. The arrows indicate the main differences between AlphaFold and Robetta models.

In the AlphaFold structure, the N-terminal region, normally cytosolic, forms a long α -helix rod which crosses the plasma membrane and reaches the extracellular space (this α -helix is located on the right side of the AlphaFold model shown in Figure 5). Robetta allowed to reconcile the model with the actual topology of SV2A. The N-terminal region still forms a α -helix but this helical domain is now correctly located in the intracellular region.

Discussion

The goal of the present study was to draw attention on the current limitations of the AlphaFold structure prediction algorithm. These limitations have been illustrated by four distinct proteins, one cytosolic and three with a membrane localization. In all cases, AlphaFold gave unrealistic structures. Thus, the promise that “AlphaFold is able to produce confident predictions” even for membrane proteins [1] is an overstatement. Our analysis supports the notion that any effort to model the 3D structure of proteins based on their amino acid sequence is intrinsically limited by the fact that other parameters, apart from the sequence, can significantly influence these structures. In this respect, the fact that Robetta appeared to be more respectful of the topology of these proteins does not mean that these particular models are totally accurate. In fact, the same limitations that apply to AlphaFold also apply to Robetta. However, these data suggest that, in some cases, it will be easier to work with protein structures obtained by Robetta instead of AlphaFold. To which extent this can be established as a general rule will require hundreds of structure comparisons, and this task is clearly out of the scope of the present study. And in any case the models proposed by Robetta should be further refined by energy minimization [7] and molecular dynamics simulations [8–10]. To sum up, it is not our intention to criticize the AlphaFold approach, but rather to take advantage of this recent advance to bring up to date the problem of protein folding and its biological coding.

The problem of predicting the 3D structure of proteins is a very old one. The concept started with the formulation of the central dogma of molecular biology [11,12]. This dogma sums up in a simple way the central notion according to which biological information is a universal code that is stored in a nucleic acid (DNA in most cases but also RNA if we include those viruses whose genome consists of RNA) and flows unidirectionally from nucleic acids to proteins [13]. We can thus summarize it in one line: DNA $\rightarrow$ RNA $\rightarrow$ Protein.

Figure 6: The epigenetic dimension of protein structure: a new paradigm. Left panel. Epigenetics modifications of DNA alters which genes are on or off, in response to environmental parameters, with no alteration of DNA sequence. This biological information is translated into a go no go decision of gene expression. Post-translational modifications of histones and DNA methylation play a key role in these mechanisms. Right panel. Epigenetics modifications of proteins alter their conformation, in response to environmental parameters, with no alteration of amino acid sequence. This biological information is translated into conformational guidelines. Biological membranes and ligand binding for IDPs play a critical role in these mechanisms. In both cases (DNA and protein epigenetics), the biological information can be transmitted from biomolecule to biomolecule.

An important issue was to determine whether the genetic information is sufficient to unambiguously determine the folding of a protein into a stable 3D structure on which its function is based. Anfinsen's experience (Nobel Prize in 1972) seemed to answer "yes" to this question [14]. Starting from the functional 3D structure of an enzyme, ribonuclease, Anfinsen set out to destroy this 3D structure by exposing it to denaturing agents. Having obtained in this way an inactive ribonuclease, he succeeded in renaturing the ribonuclease which recovered its initial 3D structure and its activity. The central dogma of molecular biology then acquired its corollary: 1 gene $\rightarrow$ 1 protein $\rightarrow$ 1 structure $\rightarrow$ 1 function. In other words, the 3D structure of proteins is entirely encoded by its amino acid sequence which is itself stored in a gene.

This fully deterministic view has dominated biology for decades. But since the deciphering of the human genome, deviations from this model have followed one another.

1. The notion that each protein has a single function has been challenged by the discovery of moonlighting proteins [15]. These proteins exert different functions according to their state of oligomerization, their subcellular location, or physicochemical parameters. For instance, neuropilin is a membrane receptor expressed by endothelial cells (vascular endothelial growth factor receptor) [16]. In nervous axons, it is also a receptor but here it interacts with a different ligand, semaphorin [17].
2. The notion that each protein has a unique 3D structure has been contradicted twice. First, with the discovery of amyloid proteins which may adopt two distinct types of secondary structures, α or β , the $\alpha \rightarrow \beta$ transition

being associated with a pathological process [18,19]. Second, with the discovery of a set of proteins that do not have any stable structure in solution because they lack a sufficient proportion of apolar amino acid residues to form an apolar core in water. These proteins, referred to as intrinsically disordered proteins (IDPs) do not have any stable structure in solution, for the greatest benefit of regulatory ligands [20]. As much as 40% of the human proteome corresponds to IDPs or proteins with disordered domains.

3. In addition to these exceptions, we can also cite the case of single nucleotide polymorphisms (SNPs). Because many SNPs do not induce any change in amino acid in the protein, it is generally assumed that they do not affect protein folding. However, changing a frequent codon to a rare one for the same amino acid may alter the rate of ribosome traffic on mRNA due to decreased availability of aminoacyl transfer RNAs (tRNAs). Then, abnormal translation kinetics in some cases can produce a different protein conformation [21]. In this specific case, structure prediction algorithms are totally inefficient.
4. Finally, membrane proteins may be the group of protein for which AlphaFold gives the less realistic results. The solvation of membrane proteins with lipids has a major impact of the 3D structure of membrane proteins which need to adjust their topology to membrane constraints (including fluidity, rigidity, lipid exchange, and membrane domains). It is not just a matter of hydrophobic effect [22,23] since lipid solvation in a membrane environment is totally distinct from water solvation [3]. In soluble proteins, nonpolar amino acid residues form an apolar core stabilized by van der Waals interactions,

whereas water molecules are stuck on the protein surface [23]. In the case of membrane proteins, lipid-protein interactions control the association of nonpolar residues with selected lipids, among which cholesterol and sphingolipids play a central role [24,25]. Correspondingly, membrane proteins may display several lipid-specific binding sites that will be operative only when the protein meets the adequate lipid, and this dynamic process may induce significant conformational fluctuations of the protein. In summary, it might be possible to use the amino acid sequence to predict the folding of a soluble protein in water, but rather difficult to anticipate how a protein will be shaped by membrane lipids by considering only its amino acid sequence.

Conclusion and perspectives

Although the AlphaFold project has opened a new era for the determination of protein structures, it cannot overcome the basic principle that the amino acid sequence alone does not contain all the information required for a unique protein folding formula. Protein structure is not completely cast in the stone of genes. Indeed, beyond biological coding, evolution has given way to a certain level of uncertainty and anarchy [26].

In this respect, claiming to have solved the problem of protein structure using an algorithm, no matter how sophisticated, is in contradiction with the laws of biology, only because a large number of proteins do not have a stable structure. In order to relativize the impact and promote the real usefulness of AlphaFold, we suggest recapitulating a series of well-established notions of protein structure into a global concept, coined “epigenetic dimension of protein structure.” We consider in this paradigm, all parameters that affect the structure of a protein without being strictly encoded in the amino acid (genetic) sequence (Figure 6). The comparison is intentionally made with epigenetics, considered as an “epi-information beyond the DNA sequence,” chiefly regulation of gene expression [27]. Thus, the epigenetic dimension of protein structure can be defined as an “epi-information beyond the amino acid sequence.” Its field of application is protein folding. DNA epigenetics can be transmitted from one generation to the next [28]. The conformational information acquired by a protein in response to specific environmental conditions can also be transmitted from protein to protein as shown for prion propagation mechanisms [29–31]. However, post-translational modifications such as glycosylation or acylation, which are determined by consensus amino acid motifs [32–34], are obviously not concerned by the protein epigenetics concept.

This paradigm illustrates how difficult it is to make an accurate prediction of the 3D structure of a protein which may, in some cases, remain an elusive Grail. Yet this does not mean that AlphaFold models are useless, but that we should use them with caution. When downloading an AlphaFold file, a careful analysis of the pLDDT is mandatory, since only regions with pLDDT > 90 have a sufficient confident score. These regions are highlighted in blue in the Uniprot database. Regions that are colored in orange and yellow may not be resolved and are represented as large loops with poor biological significance. For membrane proteins, it is necessary to localize the TM domains to ensure that membrane topology is respected. The particular case of IDPs may be solved by merging the protein and a known ligand to obtain structural information on the folding process of these proteins. An analysis of the electrostatic surface potential of both the partners may also help to set realistic starting conditions for dynamic docking [35].

By taking the very best of algorithms such as AlphaFold or Robetta (artificial intelligence) and applying a rigorous checking and treatment of raw data (human touch with computer assistance), the problem of protein structure will not be solved (as claimed by media) but it will reach a new level of scientific excellence without infringing the laws of biology.

Acknowledgments: We thank Pr. Nouara Yahi and Dr. Henri Chahinian for helpful discussions.

Funding information: Authors state no funding involved.

Author contributions: Both authors contributed equally to this study. FA prepared the figures. JF wrote the article.

Conflict of interest: Authors state no conflict of interest.

Data availability statement: The datasets generated and analyzed during the current review are available from the corresponding author on reasonable request.

References

- [1] Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Židek A, et al. Highly accurate protein structure prediction for the human proteome. *Nature*. 2021;596:590–6.
- [2] Hardy J, Allsop D. Amyloid deposition as the central event in the aetiology of Alzheimer's disease. *Trends Pharmacol Sci*. 1991;12:383–8.
- [3] Fantini J, Yahi N. Brain lipids in synaptic function and neurological disease: clues to innovative therapeutic

- strategies for brain disorders. 1st edn. New York: Elsevier; 2015.
- [4] Coleman JA, Green EA, Gouaux E. X-ray structures and mechanism of the human serotonin transporter. *Nature*. 2016;532:334–9.
 - [5] Bajjalieh SM, Peterson K, Shinghal R, Scheller RH. SV2, a brain synaptic vesicle protein homologous to bacterial transporters. *Science*. 1992;257:1271–3.
 - [6] Dong M, Yeh F, Tepp WH, Dean C, Johnson EA, Janz R, et al. SV2 is the protein receptor for botulinum neurotoxin A. *Science*. 2006;312:592–6.
 - [7] Di Scala C, Fantini J. Hybrid *in Silico/in vitro* approaches for the identification of functional cholesterol-binding domains in membrane proteins. *Methods Mol Biol*. 2017;1583:7–19.
 - [8] Flores A, Ramirez-Franco J, Desplantes R, Debreux K, Ferracci G, Wernert F, et al. Gangliosides interact with synaptotagmin to form the high-affinity receptor complex for botulinum neurotoxin B. *Proc Natl Acad Sci U S A*. 2019;116:18098–108.
 - [9] Fantini J, Chahinian H, Yahi N. Synergistic antiviral effect of hydroxychloroquine and azithromycin in combination against SARS-CoV-2: what molecular dynamics studies of virus-host interactions reveal. *Int J Antimicrob Agents*. 2020;56:106020.
 - [10] Fantini J, Carlus D, Yahi N. The fusogenic tilted peptide (67–78) of α -synuclein is a cholesterol binding domain. *Biochim Biophys Acta*. 2011;1808:2343–51.
 - [11] Crick FHC. The biological replication of macromolecules. *Symp Soc Exp Biol*. 1958;XII:138.
 - [12] Crick F. Central dogma of molecular biology. *Nature*. 1970;227:561–3.
 - [13] Monod J, Jacob F. General conclusions: teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb Symp Quant Biol*. 1961;26:389–401.
 - [14] Anfinsen CB. Principles that govern the folding of protein chains. *Science*. 1973;181:223–30.
 - [15] Singh N, Bhalla N. Moonlighting proteins. *Annu Rev Genet*. 2020;54:265–85.
 - [16] Soker S, Takashima S, Miao HQ, Neufeld G, Klagsbrun M. Neuropilin-1 is expressed by endothelial and tumor cells as an isoform-specific receptor for vascular endothelial growth factor. *Cell*. 1998;92:735–45.
 - [17] Lu D, Shang G, He X, Bai XC, Zhang X. Architecture of the Sema3A/PlexinA4/Neuropilin tripartite complex. *Nat Commun*. 2021;12:3172.
 - [18] Gasset M, Baldwin MA, Fletterick RJ, Prusiner SB. Perturbation of the secondary structure of the scrapie prion protein under conditions that alter infectivity. *Proc Natl Acad Sci U S A*. 1993;90:1–5.
 - [19] Kallberg Y, Gustafsson M, Persson B, Thyberg J, Johansson J. Prediction of amyloid fibril-forming proteins. *J Biol Chem*. 2001;276:12945–50.
 - [20] Uversky VN. Analyzing IDPs in Interactomes. *Methods Mol Biol*. 2020;2141:895–945.
 - [21] Komar AA. Genetics. SNPs, silent but not invisible. *Science*. 2007;315:466–7.
 - [22] Silverstein TP. The hydrophobic effect: is water afraid, or just not that interested? *ChemTexts*. 2020;6:26.
 - [23] Sumi T, Imamura H. Water-mediated interactions destabilize proteins. *Protein Sci*. 2021;30:2132–43.
 - [24] Fantini J, Barrantes FJ. Sphingolipid/cholesterol regulation of neurotransmitter receptor conformation and function. *Biochim Biophys Acta*. 2009;1788:2345–61.
 - [25] Fantini J, Di Scala C, Baier CJ, Barrantes FJ. Molecular mechanisms of protein-cholesterol interactions in plasma membranes: functional distinction between topological (tilted) and consensus (CARC/CRAC) domains. *Chem Phys Lipids*. 2016;199:52–60.
 - [26] Kupiec JJ. On the lack of specificity of proteins and its consequences for a theory of biological organization. *Prog Biophys Mol Biol*. 2010;102:45–52.
 - [27] Zhang Y, Sun Z, Jia J, Du T, Zhang N, Tang Y, et al. Overview of histone modification. *Adv Exp Med Biol*. 2021;1283:1–16.
 - [28] David I, Canario L, Combes S, Demars J. Intergenerational transmission of characters through genetics, epigenetics, microbiota, and learning in livestock. *Front Genet*. 2019;10:1058.
 - [29] Prusiner SB. Prions. *Proc Natl Acad Sci U S A*. 1998;95:13363–83.
 - [30] Ayers JI, Paras NA, Prusiner SB. Expanding spectrum of prion diseases. *Emerg Top Life Sci*. 2020;4:155–67.
 - [31] Yahi N, Di Scala C, Chahinian H, Fantini J. Innovative treatment targeting gangliosides aimed at blocking the formation of neurotoxic α -synuclein oligomers in Parkinson's disease. *Glycoconj J*. Jul 30, 2021 doi: 10.1007/s10719-021-10012-0.
 - [32] Gavel Y, von Heijne G. Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng*. 1990;3:433–42.
 - [33] Johnson DR, Bhatnagar RS, Knoll LJ, Gordon JI. Genetic and biochemical studies of protein N-myristoylation. *Annu Rev Biochem*. 1994;63:869–914.
 - [34] Boutin JA. Myristoylation. *Cell Signal*. 1997;9:15–35.
 - [35] Fantini J, Yahi N, Azzaz F, Chahinian H. Structural dynamics of SARS-CoV-2 variants: a health monitoring strategy for anticipating Covid-19 outbreaks. *J Infect*. 2021;83:197–206.