AlphaFold, AI and ontologies

Barry Smith[1]

[1] *University at Buffalo, Buffalo, NY 14051, USA*

Abstract

What follows is a comment on the 2022 ICBO paper "What AlphaFold teaches us about deep learning with prior knowledge" by Jobst Landgrebe. It seeks to throw light on the sense in which the prior knowledge used by AlphaFold is to be understood in ontological terms.

Keywords

AlphaFold, AI, ontology, Protein Ontology, UniProtKB, protein structure

1. Introduction

The AI models developed in the life sciences have a much lower predictive power than the models developed in domains such as engineering or physics. Why is this so? In a paper for this conference, Jobst Landgrebe analyzes AlphaFold [1], one of the few examples of applying AI to biology that is predictively successful. Landgrebe shows that this success turns on the fact that AlphaFold is able to use prior knowledge about protein folding that has already been assembled through experimental efforts invested in the decoding of protein sequences.

For a cluster of such decoded sequences, Alphafold can be applied to identify certain patterns in each protein homologous to the cluster, which then allow it to make highly successful predictions about its structure. This is remarkable given that, before AlphaFold, a very low degree of success had been attained in making protein folding predictions.

As Landgrebe explains, the prior knowledge about protein folding ingested as input into the AlphaFold machine learning algorithm takes two forms: 1) as protein structure data (CIF files); and 2; as knowledge about protein homology groups. These form the decisive factors which enable the predictive success of the algorithm, which uses only the protein's amino acid sequence as input and the heavy atom angle information as output. Like other prediction algorithms in the field, the ability of the AlphaFold model to predict protein structure can be applied only for proteins homologous to those with established structures. This ability depends on a part of the implicit model capturing the relationship between these known folding structures and sequence clusters. This is why the model succeeds; but also why it fails to predict structures for those molecules which are not homologous to proteins for which the structures had been already determined using classical protein crystallography. The model can therefore not create new protein folding knowledge – this must still be obtained from experiments, which can take several years per protein or fail altogether after unsuccessful efforts (as is often the case, for example, for transmembrane domains of proteins). On the other hand AlphFold stands out as compared to other structure prediction algorithms because it achieves high accuracy even for sequences with fewer homologous sequences [2].

2. The role of ontology

There are two meanings of the term "ontology": (i) as a branch of classical metaphysics dealing with the fundamental structure of the world, and (ii) as a scientific discipline that developed over the last 30 years, and which deals with organizing data and information about the world in a structured form to enable various sorts of data exploitation, for

ICBO 2022, September 25–28, 2022, Ann Arbor, USA
EMAIL: ifomis@gmail.com.
ORCID: 0000-0003-1384-116X

example in what is called ‘data science’. A significant fraction of the work carried out today under the label of ontology in sense (ii) is influenced by our understanding of ontology in the more traditional sense (i), above all in its use of the distinction between universals, organized in taxonomical hierarchies, and their respective instances – typically entities such as cells and molecules, which exist in time and space.

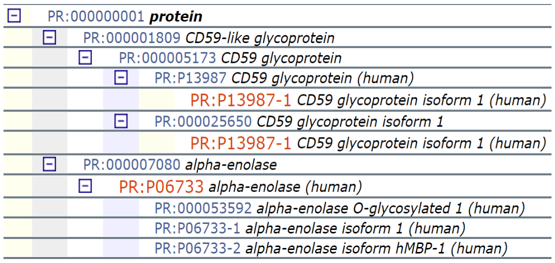
When Landgrebe claims that the prior knowledge that was used by AlphaFold is a form of ontology, he is referring to the CIF files, each of which represents the structure of the protein it describes. In what sense, then, is a CIF file an ontology, or a part of an ontology? This is a deep question, which takes us to the very foundations of sense (ii) ontology, namely to the distinction between universals and instances. The terms protein, amino acid chain, histone, chordin, and so forth, are unquestionably ontological – they are all terms from the Protein Ontology [3]. But so also, and for the same reason, is the term:

> PR:P06733-2: alpha-enolase isoform hMBP-1 (human)

whose position in the PRO hierarchy is illustrated in Figure 1. This term is defined in PRO as:

> An alpha-enolase (human) that is a translation product of some mRNA whose exon structure and start site selection renders it capable of giving rise to a protein with the amino acid sequence represented by UniProtKB:P06733-2.

The mentioned UniProt sequence is an example sequence for a certain class of molecules (briefly: molecules having the same translation site and exon structure, where ‘same’ means ‘belong to the same class’).



```
⊟ PR:000000001 protein
    ⊟ PR:000001809 CD59-like glycoprotein
        ⊟ PR:000005173 CD59 glycoprotein
            ⊟ PR:P13987 CD59 glycoprotein (human)
                PR:P13987-1 CD59 glycoprotein isoform 1 (human)
            ⊟ PR:000025650 CD59 glycoprotein isoform 1
                PR:P13987-1 CD59 glycoprotein isoform 1 (human)
    ⊟ PR:000007080 alpha-enolase
        ⊟ PR:P06733 alpha-enolase (human)
            PR:000053592 alpha-enolase O-glycosylated 1 (human)
            PR:P06733-1 alpha-enolase isoform 1 (human)
            PR:P06733-2 alpha-enolase isoform hMBP-1 (human)
```

Figure 1: Fragment of the PRO hierarchy

Why, now, do we regard PRO as an ontology, and UniProtKB as a database? There are a number of answers to this question. Most importantly, as Figure 1 makes clear, the content of PRO is organized in terms of hierarchies of representations of universals of greater and lesser generality; something which is not the case for the content of UniProtKB.

Indeed, there is a sense in which UniProtKB comprises instance data as its content: almost all of the protein sequences it provides are derived through translation of the coding sequences (CDS) submitted to public nucleic acid databases on the basis of analysis of biological samples, i.e. of instances. In this sense each sequence is, given its provenance, itself a specific instance (it is the sequence of a corresponding specific sample).

Yet at the same time each such sequence is found in (is the sequence of) many trillions of molecules. It is for this reason that UniProtKB is useful to biological and biomedical research. Each UniProt sequence in fact represents a universal with these many corresponding instances. UniProtKB, it is true, lacks an explicit hierarchy for these universals, though one could infer an implicit hierarchy from the information in the entry. (We know, for example, that all information about the sequences is derived from the indicated gene.) UniProt, as contrasted with PRO and other ontologies, also lacks explicit definitions – though, again, these are implied. PRO is explicit in its representation of the molecules themselves and--for those cases that derive from UniProtKB--makes explicit those implied hierarchies and definitions.On the other hand, all the entries in UniProtKB (and in practically all other putative databases maintained by biological scientists), consist of representations of universals. Each of the protein sequences contained in UniProtKB, for example, almost certainly exists in some trillions of instances.

In the same way, each of the mmCIF files in the Protein Data Bank (PDB) represents a protein structure that, again, almost certainly exists in trillions of instances. [4] The collection of mmCIF files is already structured into protein homology families, and as UniProt and PDB develop we can expect that more and more of the hierarchical ontology structure incorporated into PRO will become explicit in these resources, too. We can also expect that more and more of this ontological knowledge – which means knowledge that is organized in such a way as to make explicit the relations between universals – as it is made available in computable form, will in the future help to drive progress in applying AI to the life sciences.

Why, then, are there not already more predictive models in biology? Because organisms are complex systems and it is only certain aspects of such systems that can be modelled using mathematics [5]. Due to evolutionary pressure and the high costs of evolutionary change when it occurs in biological systems, nature has conserved protein homology families to a high extent. That is a pattern of regularity in a complex system that is amenable to mathematical modelling. But many other aspects of biological systems are not conserved. The task of applied mathematics in biology is to find the patterns of regularity that can be modelled using implicit models such as those exploited by AlphaFold.

Acknowledgements

Thanks are due to Jobst Landgrebe, Darren Natale and Asiyah Lin for helpful comments.

References

[1] J. Landgrebe, What AlphaFold teaches us about deep learning with prior knowledge. ICBO 2022.

[2] J. Jumper, R. Evans, A. Pritzel, et al. Highly accurate protein structure prediction with AlphaFold. Nature 596 (2021) 583–589.

[3] D. A. Natale, C. N. Arighi CN, J. A. Blake et al. Protein Ontology: a controlled structured network of protein entities. Nucleic acids res. 42 (2014) D415-21.

[4] P. D. Adams, P. V. Afonine, K. Baskaran, et al. Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB). Acta Crystallographica Sect D: Struct Biol. 75 (2019) 451–454.

[5] J. Landgrebe and B. Smith. Why Machines Will Never Rule the World: Artificial Intelligence Without Fear. Routledge, 2022.