The AlphaFold Database Ages

Authors

Ifigenia Tsitsa (*), Anja Conev(*), Alessia David, Suhail A Islam (%) & Michael J E Sternberg (%$)

(*) Joint first authors

(%) Joint last authors

($) Corresponding author

Affiliation

[1] Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, London SW7 2AZ, UK

Abstract

The AlphaFold database provides 200M protein structures predicted by AlphaFold2 in 2022 from sequences in UniProt. However, of the 20,504 full-length human structures in the AlphaFold database, 575 entries conflict with UniProt (version 2024_05); and there is a similar discrepancy for other species. This highlights how bioinformatics resources as exemplified by the AlphaFold database can rapidly age.

Text

The AlphaFold database (AlphaFoldDB)[1] is a major bioinformatics resource that provides predicted three-dimensional models for over 200 million protein sequences obtained from UniProt[2]. These models were generated using AlphaFold2[3], the deep learning algorithm developed by DeepMind. The importance of this database is underscored by its subsequent incorporation into UniProt[3], one of the most widely used protein databases. These models are widely used by the community for tasks including molecular replacement (e.g. Phaser[4] and MolRep[5]), missense variant prediction (e.g. FoldX[6] and Missense3D[7]) and as templates in structure prediction (e.g. SwissModel[8] and Phyre2.2[9]). The AlphaFoldDB was last updated in October 2022 using UniProt sequences from that time, contains models that may become outdated as UniProt sequences are regularly updated, a discrepancy researchers should be aware of when studying specific proteins. Moreover, this is a major challenge for developers of general bioinformatics resources that require the maintenance of integrative and robust data management. Here we have compared

AlphaFold models with their respective UniProt sequences across six species and identified these discrepancies. It is vital for the community to be aware of these challenges to preserve the value of such resources.

In November 2024 we performed a comparison of AlphaFoldDB models with UniProt (version 2024_05) for the human proteome revealed that, among 20,504 unique UniProt accessions with AlphaFold2-predicted structures, 575 sequences differed, corresponding to a 2.80% discrepancy between AlphaFold and UniProt sequences (Figure 1). Out of the conflicting 575 models, 93 correspond to the UniProt sequence entries which were made redundant since the release of AlphaFoldDB. The redundant UniProt sequences are generally the result of obsolete or reassigned genes, updated taxonomic data, changes in reference databases, annotation errors, or periodic data consolidation to improve consistency and accuracy (*example: Q5TG92*). Sequence length mismatches were observed for 273 models, with UniProt sequences being longer than AlphaFold2 for 147 models and shorter for 126 models. For example, ZNT2 (Q9BRI3), a zinc transporter protein, implicated in the transient neonatal zinc deficiency, has a region of approximately 50 amino acids inserted in the new UniProt sequence around the position 100 of the sequence of the outdated AlphaFold. There were also 209 AlphaFoldDB models where the sequence length was the same, but the sequence differs. In the latter category is Serpin B11, a non-inhibitory intracellular serpin, with six amino acid substitutions between the two sequences.

To assess the importance of proteins with differing sequences, we found that 384 out of the 575 had high annotation scores (4 and 5), 473 were "reviewed" proteins, and 125 were associated with diseases (as annotated by UniProt). Furthermore, 386 of these proteins were associated with Gene Ontology biological process terms and 382 with molecular function terms, highlighting their potential biological and functional relevance.

We expanded this analysis to include five additional model organisms (Table 1), with mouse showing a very similar trend to human, with 2.4% differing sequences. Interestingly, human and mouse sequences showed greater variation compared to other organisms, probably reflecting the increased depth of research and updates associated with these species. *D. melanogaster* and *A. thaliana* both had sequence differences around 0.1%, with *S. cerevisiae* showing the least difference between AlphaFoldDB and UniProt sequences from all species (0.07%). Remarkedly *D. rerio* shows a 43.4% difference in its sequences, as nearly half of its proteome appears to be absent from the current UniProt version. This is due to a major curation effort that removed thousands of obsolete or redundant entries, reducing the reviewed *D. rerio* proteome to 3,343 proteins, of which 90 are not present in the AlphaFold database. Notably, some entries, such as Q802X2 and B7ZC96 have AlphaFold models but are excluded from the downloadable version of AlphaFoldDB, likely due to their annotation as "unreviewed" structures. Reviewing their UniProt history reveals they were only recently marked as reviewed (Q802X2 in July 2024 and B7ZC96 in June

2023) well after the last update of AlphaFoldDB in 2022, illustrating both sequence and annotation aging in the AlphaFoldDB. Excluding all the AlphaFoldDB models no longer corresponding to the updated UniProt database, the total number of differing protein sequences is reduced to 14, resulting in a UniProt-AlphaFoldDB difference that aligns with trends observed in other organisms.

In conclusion, the AlphaFoldDB has provided an unprecedented resource for structural biology, its reliance on outdated sequence and annotation data poses challenges across its widespread applications including structure-based drug design, variant interpretation and functional analysis. These discrepancies highlight the need for regular updates to maintain alignment with UniProt, especially as other groups, such as SwissModel[10], continue to develop their own structure databases. Addressing these issues will be critical for maximizing the utility of AlphaFold models and ensuring their integration into downstream tools and resources. Similarly, static bioinformatics databases, including AlphaMissense[11] that presents precomputed missense variant effect predictions, are also aging. Time flies in bioinformatics.

Species	**Common Name**	**Reference Proteome**	**Full length AlphaFoldDB Predicted Structures**	**Unique UniProt IDs**	**Total Different Sequences**	**Percentage of Different Sequences (%)**
H. sapiens	Human	UP000005640	20504	20504	575	2.8
M. musculus	Mouse	UP000000589	21615	21615	518	2.4
D. melanogaster	Fruit fly	UP000000803	13458	13458	13	0.097
S. cerevisiae	Budding yeast	UP000002311	6039	6039	4	0.07
D. rerio	Zebrafish	UP000000437	24664	24664	10704	43.4
A. thaliana	Arabidopsis	UP000006548	27434	27434	31	0.11

***Table 1**: Summary table of the sequence comparison between UniProt and AlphaFold2DB. The percentage of different sequences is calculated on the unique UniProt IDs.*

References

1. Varadi, M. *et al. NAR* 50, D439–D444. (2022).
2. UniProt Consortium. *NAR* 51, D523–D531. (2023).
3. Jumper, J. *et al. Nature* 596, 583–589. (2021).
4. McCoy, A. J. *et al. J. Appl. Cryst.* 40, 658-674. (2007).
5. Vagin, A. and A. Teplyakov, A. *Acta Cryst.* D66, 22-25. (2010).
6. Delgado, J. *et al. Bioinformatics*, btaf064. (2025).
7. Ittisoponpisan, S. *et al. JMB 431*, 2197-2212. (2019).
8. Waterhouse A, *et al. NAR* 46, W296-W303. (2018).
9. Powell, H. R. *et al. JMB* 168960. (2025).
10. Bienert S., *et al. NAR* 45, D313-D319. (2017).
11. Cheng, J. *et al. Science*, *381*(6664). (2023).

Funding

IT and AC were supported by Wellcome Trust grant 218242/Z/19/z . AD was supported by MRC grant MR/Y031091/1. SAI and MJES were supported by Biotechnology and Biological Sciences Research Council (BBSRC) grants BB/T010487/1 and BB/V018558/1. This research was funded in part by the Wellcome Trust [Grant number 218242/Z/19/z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission

Acknowledgements

We thank Harold Powell for helpful discussions.

Figure

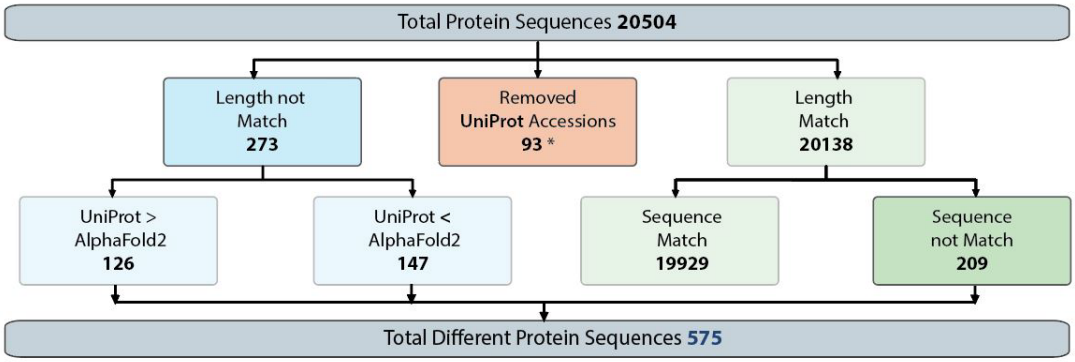

Figure 1: Sequence comparison of human AlphaFold2DB models with corresponding UniProt sequences. Blue rectangles indicate cases where the UniProt sequence length has changed, red rectangles mark accessions that have become redundant, and green rectangles denote sequences where UniProt and AlphaFoldDB match in length.

Supplementary Table 1

Supplementary Table 1: Table of the sequence analysis between UniProt and AlphaFold2 for *H. sapiens*, *M. musculus*, *D. rerio*, *D. melanogaster*, *S. cerevisiae* and *A. thaliana*.