Research Article

The Journal of Molecular Biology (ISSN 2995-8601)

Protein 3D Structure Identification by AlphaFold: a Physics-Based Prediction or Recognition Using Huge Databases?

Alexei V. Finkelstein,[1,2,*] Dmitry N. Ivankov[3]

1. Laboratory of Protein Physics, Institute of Protein Research of the Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia
2. Biology Department, the Lomonosov Moscow State University, Moscow, 119991, Russia
3. Center of Molecular and Cellular Biology, Skolkovo Institute of Science and Technology, Moscow, 121205, Russia

Corresponding author

Alexei V. Finkelstein , Laboratory of Protein Physics, Institute of Protein Research of the Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia.
Tel : +7-903-257-6694,
Email : afinkel@vega.protres.ru

Received Date : Feb 23, 2024
Accepted Date : Feb 25, 2024
Published Date : March 20 2024

ABSTRACT

The great success of AlphaFold programs poses the questions: (i) What is the main reason for this success? (ii) What AlphaFolds does: physics-based prediction of the spatial structure of a protein from its amino acid sequence or recognition of this structure from similarity of the target sequence to some parts of sequences with already known spatial structures? The answers given here are: (i) the main reason for the AlphaFold's success is the usage of huge databases which already cover virtually all protein superfamilies existing in Nature; (ii) using these databases, multiple sequence alignments, and coevolutionary information – like correlations of amino acid residues of the contacting chain regions – AlphaFold recognizes a spatial structure by similarity of the target sequence (or its parts) to related sequence(s) with already known spatial structures. We emphasize that this does not diminish the merit and utility of AlphaFold but only explains the basis of its success.

Keywords

bioinformatics | similarity of 3D structures | sequence identity| databases | AlphaFold.

INTRODUCTION

The great success of the AlphaFold, AlphaFold2 and then OpenFold programs [1 3] in identifying three-dimensional (3D) protein structures from their amino acid (a.a.) sequences is obvious [4], but it raises two questions [5]: What is the main reason for this success? What exactly does AlphaFold do: prediction of the 3D protein structure from its a.a. sequence and the protein chain physics or recognition of this structure from the similarity between the target sequence and some of sequences with already known 3D structures?
According to bioinformatics [6 10], 20 25% or higher identity of a.a. sequences is usually sufficient to ensure a close similarity of 3D protein folds with a small, <2Å, average difference between their coordinates. More specifically [7, 8], the residue identity below 20% in pairwise sequence alignments often does not provide reliable alignments of 3D structures; the identity of 20-25% is the "twilight" zone [7]; and only an identity of >25% ensures correct alignments of 3D structures. Here we show that the sequence similarity between the domain-, half-domain- and 1/3-domain-size fragments of a random protein sequence and the like ones from modern databases exceeds this threshold.

METHODS

Derivation of the expected sequence identity.
According to the Poisson distribution, the probability that the random a.a. sequence matches another random a.a. sequence of the same length in positions is

$$P_{m,pn} = \frac{(pn)^m}{m!} e^{-pn} \qquad (1)$$

when each type of a.a. falls out with probability $\boldsymbol{p}$.
If the random sequence is compared not with one but with random sequences of the same length n (forming the set Σ_N), $P_{m,pn}$ N is the expected number of the set Σ_N members matching S_n in m positions. Thus, the equation $P_{m,pn}$ N= *1*

The Journal of Molecular Biology (ISSN 2995-8601)

determines the maximal expected number M of matches of the sequence S_n with the closest in similarity sequence from the set Σ_N.

Assuming $p << 1$, long enough sequences ($1 << pn$), and significant sequence identity ($1 << pn << m$), one can use the Stirling's approximation ($m! \approx (m/e)^m$, where $e \approx 2.72$) and get

$$P_{m,pn} = \left(\frac{ep}{m/n}\right)^m e^{-pn} \quad (2)$$

The maximal expected value of m/n (denoted as M/n) follows from the equation $1/P_{M,pn}$,

or

$$\left(\frac{M/n}{pe}\right)\ln\left(\frac{M/n}{pe}\right) = \left(\frac{1}{npe}\right)\ln N - \frac{1}{e} \quad (3)$$

A "novel fold" coverage with parts of known protein structures.

Here, we considered the smallest "novel fold" from the CASP 14 experiment [12], the 102-residue long target T1035 (PDB code: 6VR4, chain A, a domain of a.a. residues 235-336). To preserve the protein secondary structure, we chose splitting points in the T1035 loops between the secondary structure elements. As a result, we split T1035 into three parts: a central part of 50 residues, a C-terminal part of 43 residues, and an N-terminal part of 9 residues, see Figure 2. We ran the TM align program [13] separately for each of these parts of T1035 against all PDB structures published by the May 2020 (because AlphaFold2 did not "see" any PDB structure published after this date [2] during its training).

RESULTS

The expected similarity of a target sequence with the closest in identity chain from protein banks.

First, we estimated the expected maximal similarity of a continuous random sequence S_n of n a.a. residues (where each a.a. type falls out with probability p) with the closest in identity chain from a set Σ_N of N other continuous random sequences of the same size and a.a. content. This estimate is given by Eq. 3 in Methods.

We found that for p≈1/20, typical of proteins with 20 a.a. types, n≈100, typical of protein domain size, and N≈1.5×10^5, the number of protein structures in the Protein Data Bank (PDB) in 2020 (https://www.rcsb.org/stats/growth/growth-protein), the expected maximal fraction M/n of identical residues in the target sequence and the closest in similarity chain from the sequence set Σ_N is 0.19. For p≈1/20, n≈100, and N≈1.9×10^8, the number of protein sequences in the sequence database UniProtKB (https://academic.oup.com/nar/article/49/D1/D480/6006196) in 2020, M/n≈0.24. The sequence identity of 19-24% (Fig. 1, blue bar) is already sufficient to usually result in a small, 1.7 ± 0.5 Å [6], root mean square difference (RMSD) between coordinates of compared proteins.

For a "half-domain" size (≈ 50-residue) random chain with p≈1/20, and N≈1.5×10^5-1.9×10^8, M/n rises from 19-24% to 27-35%. This provides an RMSD of 1.4 ± 0.2 Å between coordinates of such compared protein fragments. For a 1/3-domain chain, the RMSD can be expected to be 1.3 ± 0.2Å. However, the actual level of identity between the randomly taken protein sequence and its closest database analogue is higher than 19 24% because the above comparison ignored possible insertions or deletions. The structural alignment of related protein sequences of n≈100 residues usually requires 2 4 insertions or deletions of 1 20 residues each [8, 11]. These insertions/deletions, together with possible shifts of our 100-residue sequence relatively to the database-stored one, increase the number N of independent sequence comparisons by ≈ 4 8 orders of magnitude. Thus, the best expected (by Eq. 3) identity between a ≈ 100-residue random sequence and its closest database analogue rises from 19 24% to a more realistic M/n≈25-32% (Fig. 1, yellow bar).

Figure 1

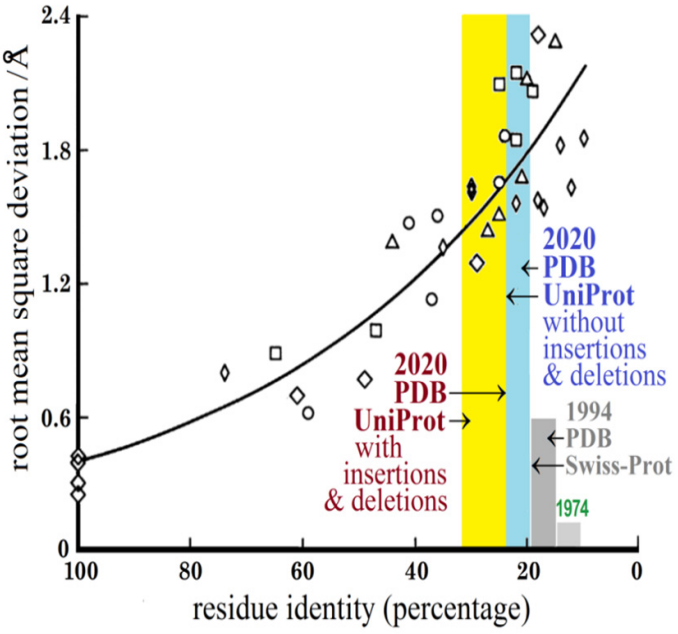

Figure 1. Structural divergence of homologous proteins plotted against the sequence identity (different black symbols referring to different protein families; adapted from [4]). Colored bars show the expected ranges of the residue identity of a "domain-size" (n≈100)

The Journal of Molecular Biology (ISSN 2995-8601)

random sequence to the closest in similarity sequence from UniProtKB and PDB databases of different years. The structural difference is measured by the root mean square deviations (Å) of the main-chain C_α positions for residues of the optimally superimposed "protein cores" (comprising the main secondary structures and covering >90% of the chains with a <50% residue identity, and ≈ 50% of the chains with a ≈ 20% residue identity) [4]. The sequence similarity is measured in the percentage of residues that are identical in the superimposed cores.

The above estimates do not take into account the fact that some proteins have many more homologues than others. The existence of multiple homologues reduces the number of "independent" comparisons. However, even if this number is reduced tenfold, the ranges of 19 24% and 25 32% change only slightly, to 18 23% and 24 31%, respectively. For a "half-domain" (50-residue) random chain, M/n rises from 27 35% to more realistic 36 44%, which ensures an RMSD of 1.2 ± 0.2 Å between coordinates of superimposed protein fragments. For a 1/3 domain chain, the RMSD can be expected to be 1.1 ± 0.2 Å. As Eq. 3 shows, with given number N of independent sequence comparisons, the expected M/n value decreases with the increasing chain length n. However, one part of a target protein chain can match some part of one known protein, while its another part can match some part of another known protein. AlphaFold is able to dock [1] such separate known parts, and the result of this docking may form a "novel fold" that has not yet been found among known proteins.

A "novel fold" coverage with large parts of known protein structures.

To demonstrate that a "novel fold" can be presented as a combination of parts of known structures, we show this for the target T1035 from the CASP 14 experiment [12], where T1035 was the smallest "novel fold". We split T1035 into a few parts and ran TM-align program [13] against all PDB structures known in the training phase of AlphaFold (see Methods). Figure 2 features the T1035 structure as a combination of three fragments with RMSD of about 0.4 2.0 Å for each of them. It should be noted that, in fact, the TM align program gave a lot of nearly equally good options for covering the T1035 "novel fold" with parts of known protein structures that have almost the same RMSD; and we have presented only one of them in Figure 2.

Figure 2

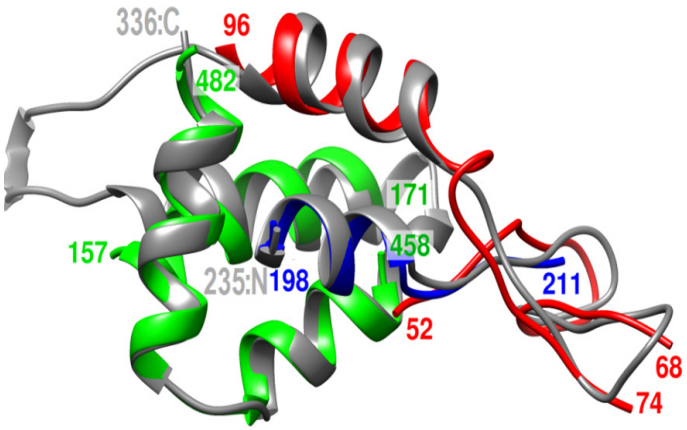

Figure 2. A "novel fold" (target T1035 from CASP 14) as a combination of fragments of already known structures. Superposition of this "novel fold", shown in gray (PDB code: 6VR4, chain A, residues 235 (N terminus) - 336 (C terminus)) and fragments of three structures which were available to AlphaFold during the training. A fragment shown in blue, which comprises residues 198 211 of a protein with PDB code 1GB3, chain A, is superimposed on the N terminal part of T1035; the fragments shown in in red, which comprise residues 52 68 and 74 96 of a protein with PDB code 5A29, chain A, are superimposed on the central part of T1035; and the fragments shown in green, which comprise residues 157 171 and 458 482 of a protein with PDB code: 5W40, chain B, are superimposed on the C terminal part of T1035.

DISCUSSION

With huge databases currently available, a "new" sequence whose 3D structure is to be identified typically either has ≈ 25 32% identity to some protein whose 3D structure (or that of its homolog) is already known or can be divided into a few domain-size or approximately half-domain-size parts that have ≈ 25 44% sequence identity to some parts of known protein structures and can be subjected to docking. The above sequence identities correspond to proteins whose structural divergence is about 1.4 ± 0.4 Å (Fig. 1).

It should be noted that although the assumption that similar sequences have similar folds [6, 10] is practically 100% correct for natural proteins, some specially designed proteins show that mutation of only one special a.a. residue can drastically change their structure and function [14]; the sequence-based recognition of structure can be problematic in such a case.

It is also worth noting that AlphaFold contains ~21×10^6 adjustable parameters ([2], and J. Jumper, private communication) - 1000 times more than the number of parameters needed to describe the physics of protein chains, including all the pairwise, triple, and even quadruple [15] interactions of all their atoms. This 1000 fold excess

The Journal of Molecular Biology (ISSN 2995-8601)

shows the ratio between the AlphaFolds' similarity-based (bioinformatics) and physics-based efforts.

The above 21×10^6 adjustable parameters "trained" on protein databases form the "own", that is, the internal memory of AlphaFold. In addition, AlphaFold can use external memory contained in protein databases, i.e., in PDB (now containing hundreds of thousands of proteins with about hundred of millions of amino acid residues with billions of atomic coordinates) and UniProtKB (now containing hundreds of millions of sequences with about hundred of billions of amino acid residues). AlphaFold can work without this external memory, but the predictions of protein structures made using the external memory are better and more confident [16].

It is widely debated how well AlphaFold has learned the physics of proteins. Yes, AlphaFold knows the stability of protein structure elements, since it knows the statistics of PDB-stored protein structures, which is directly related to their stability [10, 17] due to evolutionary preservation of their stable features; but it (yet?) knows nothing about the process of protein folding. Besides, Figure 5b of [2] clearly demonstrates that bioinformatics is much more important than physics for AlphaFold predictions. This Figure shows a correct but contradicting to physics prediction of a non-compact structure of some protein chain. This non-compact structure lacks interactions that can support it – it is fixed by surrounding protein chains not introduced to AlphaFold. However, knowledge of similar complexes is sufficient to AlphaFold to make correct bioinformatics recognition, though contradicting to physics of this separate target chain. The same is demonstrated by Figure 3.

Figure 3

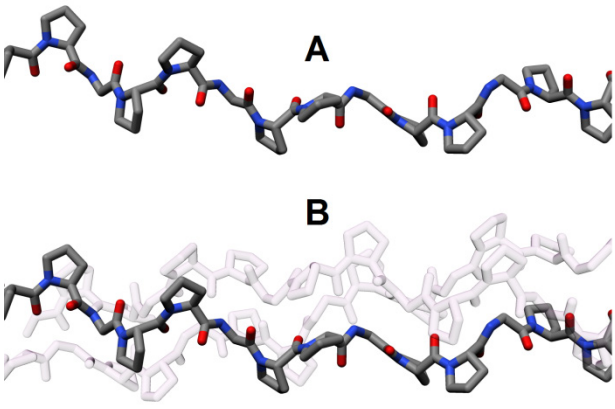

Figure 3. (A) Structure of a 13 residue fragment of the "collagen-like" sequence (Gly-Pro-Pro)13 predicted by AlphaFold program [2, 19]. This chain is not supported by any "extraneous" interactions, and thus has to have no definite structure, because it is much longer than a persistent length of a polypeptide chain (especially since it contains a lot of glycines, and this is the most flexible amino acid residue [10]). Nevertheless, its conformation is just the same as in collagen. (B) The same chain fragment (colored) in the context of triple helix of collagen, PDB: 5CTD [20], where this chain is supported by the remaining (shown in light-gray) two chains of the collagen triple helix.

Now we can answer the questions posed at the beginning of this paper. The main reason for the tremendous success of AlphaFolds is, apart from great programming, the usage of huge protein databases which, as Cyrus Chothia predicted 30 years ago [21], now seem to cover all or almost all protein superfamilies. Using these databases, multiple sequence alignments, and the resulting coevolutionary information like correlations in contacting pairs, triplets, etc. of a.a. residues, and evolutionary conservation of their stable features, Open- and AlphaFolds outline the sought-for 3D structure by similarity of the target to related sequence(s) with known 3D structure(s). It is worth noting that AlphaFold, if trained on databases-1994 (the year of the first CASP - Critical Assessment of Protein Structure experiment) would work significantly worse than now because databases-1994 were much smaller: they contained ≈ 30000 protein sequences [22] and ≈1000 structures (https://www.rcsb.org/stats/growth/growth-released-structures). With databases-1994, the highest Eq. 3 expected sequence identity for the best continuous alignment of a 100-residue chain would be not 19 24% as it is now, but only 15 19% (Fig. 1, dark gray bar). This is below the "twilight zone" where sequence- and structure alignments are often different. As for the alignments with insertions and deletions, in 1994 we would have a "twilight" recognition with an expected sequence identity of 19 26% (instead of the present 25 32%). And it goes without saying that in 1974, when the first international assessments of protein structure prediction [23] started and only ~10 of protein structures and ~1000 sequences were known, AlphaFolds would not be able to determine protein structures from a.a. sequences (Fig. 1, light gray bar).

CONCLUSIONS

In this article, our attention is paid not to the operation of AlphaFold programs, which is well described in [1 4] and some other articles. Our attention has been paid to AlphaFold's source of wisdom. We see that the basis of AlphaFold's great success is a skillful usage of huge protein databases collected during 60 years and clearly presenting evolutionary conservation of stable features of 3D protein structures. Now they give a possibility to predict, or rather recognize stable protein structures from their a.a. sequences without considering the process of protein folding [10, 24] that

The Journal of Molecular Biology (ISSN 2995-8601)

creates these structures. We emphasize that the presented paper does not diminish the merit and utility of AlphaFolds; it only explains the basis of their success.

Acknowledgments

We are grateful to N.V. Dovidchenko, S.O. Garbuzynskiy, G. Vriend and especially J. Jumper for discussions and E.V. Serebrova for editing the manuscript. We acknowledge funding from the Russian Science Foundation (grant № 21-14-00268).

Competing Interest:

The authors declare no competing interests.

REFERENCES

1. A.W. Senior et al., Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710 (2020).

2. J. Jumper et al., Highly accurate protein structure prediction with AlphaFold. Nature 596, 583-589 (2021).

3. G. Ahdritz et al., OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. bioRxiv preprint (2022) https://doi.org/10.1101/2022.11.20.517210.

4. J.P. Roney, S. Ovchinnikov, State-of-the-art estimation of protein model accuracy using AlphaFold. Phys. Rev. Lett. 129, 238101 (2022).

5. A.V. Finkelstein, Does AlphaFold predict the spatial structure of a protein from physics or recognize it (its main parts and their association) using databases? bioRxiv [Preprint], 2022. https://doi.org/10.1101/2022.11.21.517308

6. A.M. Lesk, C. Chothia, The response of protein structures to amino-acid sequence changes. Phil. Trans. R. Soc. Lond. A 317, 345–356 (1986).

7. S.Y. Chung, S. Subbiah, A structural explanation for the twilight zone of protein sequence homology. Structure l4, 1123–1127 (1996).

8. S.R. Sunyaev et al., From analysis of protein structural alignments toward a novel approach to align protein sequences. Proteins 54, 569–582 (2004).

9. E. Krissinel, K. Henrick. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Cryst. D60, 2256-2268 (2004). doi:10.1107/S0907444904026460.

10. A.V. Finkelstein, O.B. Ptitsyn, "Protein Physics. A Course of Lectures". 2-nd Ed. (Academic Press, 2016), lectures 3–7, 16, 22.

11. S.K. Chan, M. Hsing, F. Hormozdiari, A. Cherkasov, Relationship between insertion/deletion (indel) frequency of proteins and essentiality. BMC Bioinformatics 8, 227 (2007).

12. L.N. Kinch, R.D. Schaeffer, A. Kryshtafovych, N.V. Grishin, Target classification in the 14th round of the critical assessment of protein structure prediction (CASP14). Proteins: Structure, Function, and Bioinformatics 89, 1618–1632 (2021).

13. Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research 33, 2302-2309 (2005).

14. P.A. Alexander, Y. He, Y. Chen, J. Orban, P.N. Bryan, A minimal sequence code for switching protein structure and function. Proc. Natl. Acad. Sci. U.S.A. 106, 21149–21154 (2009).

15. L.B. Pereyaslavets, A.V. Finkelstein, Development and testing of PFFsol.1, a new polarizable atomic force field for calculation of molecular interactions in implicit water environment. J. Phys. Chem. B 116, 4646–4654 (2012).

16. D. Ivankov, M. Pak, A. Finkelstein, AlphaFold: predicts or recognizes the protein structure? PROGRAM of the XXVIII Symposium on Bioinformatics and Computer-Aided Drug Discovery, May 24-26, 2022, Moscow (2022). http://www.way2drug.com/dr/bcadd2022_program.php

17. A.V. Finkelstein, A.Ya. Badretdinov, A.M. Gutin, Why do protein architectures have a Boltzmann-like statistics? Proteins 23: 142–150 (1995).

18. E. Krieger, T. Darden, S.B. Nabuurs, A. Finkelstein, G. Vriend, Making optimal use of empirical energy functions: force-field parameterization in crystal space. Proteins 57: 678-683 (2004).

19. AlphaFold, standalone version 2.0. https://github.com/google-deepmind/alphafold.

The Journal of Molecular Biology (ISSN 2995-8601)

20. S.P. Boudko, H.P. Bachinger. Structural insight for chain selection and stagger control in collagen. Sci. Rep. 6, 37831-37831 (2016).

21. C. Chothia, One thousand families for the molecular biologist. Nature 357, 543–544 (1992).

22. A. Bairoch, B. Boeckmann, The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Res. 22, 3578–3580 (1994).

23. G.E. Schulz et al., Comparison of predicted and experimentally determined secondary structure of adenilate kinase. Nature 250, 140–142 (1974).

24. S.-J. Chen et al., Opinion: Protein folds vs. protein folding: Differing questions, different challenges. Proc. Natl. Acad. Sci. U.S.A. 120, e2214423119 (2023).