Patterns

CellPress
OPEN ACCESS

Preview

How deep can we decipher protein evolution with deep learning models

Xiaozhi Fu[1,*]
[1]Department of Life Sciences, Chalmers University of Technology, Kemivägen 10, SE-412 96 Gothenburg, Sweden
*Correspondence: xiaozhi.fu@chalmers.se
https://doi.org/10.1016/j.patter.2024.101043

Evolutionary-based machine learning models have emerged as a fascinating approach to mapping the landscape for protein evolution. Lian et al. demonstrated that evolution-based deep generative models, specifically variational autoencoders, can organize SH3 homologs in a hierarchical latent space, effectively distinguishing the specific Sho1^{SH3} domains.

Proteins are the building blocks of life. Variations in their sequences, and secondary and tertiary structures, contribute to the diversity of life forms. The boom in biological big data, driven by rapid sequencing technologies and atom-level structure prediction, has allowed us a glimpse into some of the evolutionary trajectories of proteins over millions of years using both sequence-based and structure-based models.[1] With machine learning (ML), sequential and structural information could be converted into distributed vector representations, linked to functional properties via transfer learning, and further used for evolutionary fitness landscape mapping.[2]

Although protein structures can also be clustered with approaches such as Foldseek cluster,[3] sequence-based ML models advance faster in the area of protein evolution due to the lower computational consumption and setup complexity compared with structure prediction and alignment. Latent space models can theoretically model high-order epistasis of sequences without exponentially increasing the number of parameters. Thanks to recent advances in stochastic variational inference, such as the variational autoencoder (VAE) approach, continuous latent space models can be readily learned for hundreds of thousands of sequences. Ding et al.[4] learned latent space models with the VAE and captured phylogenetic relationships between sequences from the alignment of multiple sequences of the fibronectin type III domain and the cytochrome P450 family. Ziegler et al.[5] further constructed a latent generative landscape with the VAE that could capture the phylogenetic, functional, and fitness properties of various protein families. These and other similar approaches allow researchers to capture snapshots of the static evolutionary state of functional proteins. Additionally, when it comes to understanding why certain proteins within the same family are favored and endowed with unique functions across species, Ziegler et al.[5] provided some evidence in the local regions of the landscape, which showed the distance between cold-sensitive transmembrane protein 8 and non-cold-sensitive ones. However, overall speaking, we still lack a systematic approach to unveil the full evolutionary mystery of proteins.

To untangle some of the mysteries, Lian et al. described in their paper published recently in *Cell Systems*[6] a hierarchical latent space mapping approach with the VAE to reflect hierarchical relationships, distinguishing between Sho1^{SH3} orthologous and paralogous sequences within the Src homology 3 (SH3) family. Sho1^{SH3} is the only SH3 domain among 26 other paralogous domains in the genome that can support osmosensing in the Sho1 pathway. The authors compared the Boltzmann machine direct-coupling analysis model and two VAE models (a generic form called vanillin VAE and a variant form, InfoVAE, known for information maximizing) in learning the sequential information of the SH3 family. They found that InfoVAE outperformed the other models in capturing epistatic information and showed higher accuracy in producing a hierarchical organization of SH3 homologs, where functional distinctions are primary, and phylogeny is secondary. The Sho1^{SH3} group seemed to fall into a constrained space in the 3D latent space embedding mapping of all 5,299 natural SH3 homologs.

More significantly, Lian et al. presented a wet-lab assay to calibrate the Sho1 functionality of all 5,299 natural SH3 homologs. They demonstrated that the experimentally validated functional sequences were localized in the vicinity of the Sho1^{SH3} paralog group in the 3D latent space embedding mapping. The annotations of these 132 validated sequences revealed that they were all orthologs of Sho1^{SH3} across the fungal kingdom, including Sho1^{SH3} domains from distant *Basidiomycota* and even non-Dikarya species.

The authors further defined a minimal polygon in the latent space ("convex hull") that bounds the natural sequences displaying full function in the *S. cerevisiae* Sho1 pathway, and the majority of Sho1^{SH3} orthologs in the fungal kingdom (155/172) lie within the hull, and only very few sequences within the hull are not functional. After sequence generation by InfoVAE, Lian et al. provided an approach for synthetic sequence selection by sampling locally within the convex hull from the 3D latent space mapping. Remarkably, 78.3% of the synthetic sequences sampled from the convex hull in the InfoVAE 3D latent space mapping were proven to be functional in the osmosensing assay, which offers a novel approach for functional synthetic protein sequence design.

The mystery of evolution is akin to an iceberg; while various approaches can describe and unravel the visible part above water, the vast majority of the

latent space remains largely unexplored beneath the surface. Intriguingly, the authors introduced the concept of protein design and engineering by attempting to define locality within a representation space. With advancements in fast sequencing and structural biology, researchers can now apply various models to perform multiple sequence alignments and structural analyses, aiming to identify essential amino acid locations either sequentially in 1D or in the physical 3D structure. Lian et al. contributed to this endeavor by defining a constrained region in latent space, offering a new perspective to explore the depths of evolutionary mysteries beneath the surface. It would be fascinating to investigate whether similar constrained regions exist for orthologs in other protein families, how these regions have changed over millions of years of evolution, and whether it is possible to predict future changes. Additionally, efforts in encoding and decoding 3D structural coordinates using VAEs have been successful,[7] which, together with this work, makes us wonder how latent space embeddings will look like when decoding and encoding structural information and whether this will provide additional dimensions for inferring orthologous relationships.

Inferring orthology is important for clarifying the evolutionary history of genes and reconstructing phylogenetic trees; however, it remains challenging, which is still like mysteries hidden under the surface. High computational demand is generally needed for conventional computational tools of orthology analysis to compare hundreds or thousands of genomes or proteomes with each other.[8] Lian et al.'s approach could possibly infer orthology by capturing the epistatic relationships from just the amino acid sequences of proteins, which makes it more accessible and scalable for orthology analysis.

Practically, it will also be fascinating to investigate whether the approach developed by Lian et al. has broader applicability for ortholog analysis and design in other protein kinds, such as antigens and enzymes. For instance, polyethylene terephthalate (PET)-degrading enzymes, primarily from the hydrolase family, have garnered significant interest in recent years. Although most hydrolases cannot catalyze PET, certain hydrolases across different prokaryotic[9] and eukaryotic[10] species can degrade PET. It would be compelling to determine if similar constrained regions in the latent space landscape could also be defined for enzymes such as PET-degrading hydrolases and further apply the results for protein design and engineering.

DECLARATION OF INTERESTS

The author declares no competing interests.

REFERENCES

1. Malbranke, C., Bikard, D., Cocco, S., Monasson, R., and Tubiana, J. (2023). Machine learning for evolutionary-based and physics-inspired protein design: Current and future synergies. Curr. Opin. Struct. Biol. *80*, 102571.
2. Bepler, T., and Berger, B. (2021). Learning the protein language: Evolution, structure, and function. Cell Syst. *12*, 654–669.e3.
3. Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C.L.M., Wein, T., Varadi, M., Velankar, S., Beltrao, P., and Steinegger, M. (2023). Clustering predicted structures at the scale of the known protein universe. Nature *622*, 637–645.
4. Ding, X., Zou, Z., and Brooks, C.L., III (2019). Deciphering protein evolution and fitness landscapes with latent space models. Nat. Commun. *10*, 5644.
5. Ziegler, C., Martin, J., Sinner, C., and Morcos, F. (2023). Latent generative landscapes as maps of functional diversity in protein sequence space. Nat. Commun. *14*, 2222.
6. Lian, X., Praljak, N., Subramanian, S.K., Wasinger, S., Ranganathan, R., and Ferguson, A. (2024). Deep learning-based design of synthetic orthologs of SH3 signaling domains. Cell Syst. Published online August 5, 2024. https://doi.org/10.1016/j.cels.2024.07.005.
7. Eguchi, R.R., Choe, C.A., and Huang, P.-S. (2022). Ig-VAE: Generative modeling of protein structure by direct 3D coordinate generation. PLoS Comput. Biol. *18*, e1010271.
8. Glover, N., Dessimoz, C., Ebersberger, I., Forslund, S.K., Gabaldón, T., Huerta-Cepas, J., Martin, M.-J., Muffato, M., Patricio, M., Pereira, C., et al. (2019). Advances and Applications in the Quest for Orthologs. Mol. Biol. Evol. *36*, 2157–2164.
9. Erickson, E., Gado, J.E., Avilán, L., Bratti, F., Brizendine, R.K., Cox, P.A., Gill, R., Graham, R., Kim, D.-J., König, G., et al. (2022). Sourcing thermotolerant poly(ethylene terephthalate) hydrolase scaffolds from natural diversity. Nat. Commun. *13*, 7850.
10. Brinch-Pedersen, W., Keller, M.B., Dorau, R., Paul, B., Jensen, K., Borch, K., and Westh, P. (2024). Discovery and surface charge engineering of fungal cutinases for enhanced activity on poly(ethylene terephthalate). ACS Sustain. Chem. Eng. *12*, 7329–7337.