Nir Ben-Tal
Tel Aviv University

AI-based view of protein space

In proteins, sequence determines structure and function, and yet, in spite of decades of intensive research, we still lack full understanding of the interplay between sequence, structure, function, mechanism, and dynamics. Recent advances in AI may aid in this. AlphaFold, which effectively solved protein structure prediction from sequence, has demonstrated its power in capturing sequence-structure relationship. Further demonstration of the power of AI is the success of protein language models (PLMs). PLMs capture the essence of proteins, in a form called "embeddings." In other disciplines, co-embeddings, i.e., concerted embeddings of multiple facets in the same space, revealed non-trivial connections, e.g., between images and text describing them. By analogy, it should be possible to co-embed protein sequence, structure, and function, as well as other information like mechanism. Embeddings describe objects as vectors, such that the vectors of similar objects are near each other.

In PLMs, vectors corresponding to evolutionarily linked sequences are near each other in sequence latent space (see Kolodny's Voice). Similarly, embeddings of proteins sharing structural similarity would cluster together in structure latent space and those of proteins of similar function (e.g., a shared ligand) in function latent space. Co-embedding the three types of vectors, would provide a unified view of protein space. Such a learned co-embedding of sequence, structure, and function would offer the most holistic and comprehensive view that one can hope for of protein space, with multiple practical and conceptual implications.

Carl G. de Boer
University of British Columbia

DNA synthesis writes the next chapter

While sequencing DNA has provided us snapshots into the products of evolution, our ability to synthesize DNA is driving our ability to understand its functional consequences. DNA synthesis has been around for decades, but the last few years has seen the cost of DNA synthesis go down to the point where new types of experiments are becoming feasible. For instance, one can now purchase libraries of millions of short (150–400 nt) single-stranded oligos or even assemble desired >100 kb sequences from 3 kb gene synthesis clones. And the pace of DNA synthesis technological improvement is poised to continue. Soon, our progress will be limited more by our abilities to design the best experiments.

DNA synthesis technology has already enabled us to test the effects of genetic variation, in both proteins and *cis*-regulatory DNA (e.g., enhancer/promoter), and to test the functions of distant orthologs. But the genetic variation within extant organisms reflects only an infinitesimal proportion of the variation that ever existed or could exist. Exploration of these unseen possibilities will enable us to learn better sequence-function maps across cellular systems, including *cis*-regulatory DNA and proteins, and their interactions that result in cellular and ultimately organismal phenotypes. In the longer term, computational models trained on these synthetic DNA sequences will enable us to design sequences for our benefit, enabling us to bypass evolution entirely.

Claire D. McWhite
Princeton University

Protein interactions decoded

The recent advent of protein language models has opened fresh pathways for understanding proteins and their interaction networks. These models are generated through processing millions of protein sequences, capturing the inherent rules and patterns that define what sequences are possible in the language of proteins. Language model-based protein structure predictors, such as AlphaFold2, have enabled the prediction of interactions across the entire human proteome, revealing numerous candidate novel stable protein complexes.

Furthermore, protein language models have granted unprecedented control over designing protein interactions. This includes creating entirely new peptides and antibodies that bind specific target proteins.

Looking ahead, there will be potential to engineer variants of interacting proteins with altered binding affinities. The ability to modify or destroy a specific interaction with minor sequence changes will pave the way for deeper inquiries into the functional significance of protein proximity within cells. As we harness language models to reveal amino