Cell Systems

CellPress

Voices

What can recent methodological advances help us understand about protein and genome evolution?

Christine Orengo
University College London

A structural lens on evolution

Dramatic advances in protein structure prediction (e.g., AlphaFold) have expanded structural data massively. A large proportion is good enough for meaningful and substantial evolutionary analyses. This vastly increased catalog will bring deep insights into biochemical and biophysical constraints and reveal evolutionary mechanisms to aid protein design and interpretation of variants.

The quantum leap in accuracy was powered by advances in AI/deep learning (DL), validated by independent assessment (CASP) and enabled by steady, exponential accumulation of structure and sequence data in standardized formats (PDB, UniProt). A landmark in scientific culture saw the public release of commercially developed algorithms together with hundreds of millions of models.

Proteins comprise one or more globular "domains"—building blocks of life. Early work exploiting DL-methods detected >300 million domain models. Powerful approaches (FoldSeek), accelerating structure comparison by >3 orders of magnitude, rapidly assigned these to known evolutionary families and revealed structural novelty. Remarkably few new families are found (<0.01% of models). Instead, much of the treasure is the amazing structural innovation across families during evolution, likely bringing functional novelty. The challenge now is to understand when and how function changes. To aid this, protein language models have enhanced the prediction of protein functional sites and surfaces. Comprehensive data on diverse domain combinations will illuminate evolutionary trajectories and help to rationalize modified functions. Prediction of protein-protein interactions remains challenging but the vast data now available on predicted domain-domain packings will accelerate progress. Mystery still surrounds the disordered regions of proteins ,but the new DL tools and much expanded structural data will shine a strong light on protein evolution.

Ian M. Ehrenreich
University of Southern California

Building networks to study evolution

Key questions about how genetic network properties impact adaptive evolution, mutational and environmental robustness, evolvability, and the diversification of organisms have yet to be conclusively answered. Addressing these questions in a definitive way requires experiments in which we design, build, and test the entire genetic networks of living cells. While such experiments were not historically possible, recent work on chromosome-scale synthetic biology (or "synthetic genomics") suggests they will be feasible in the not-distant future.

Whole chromosomes can now be built by progressively assembling small DNA pieces into increasingly large molecules. Currently, chromosomes with sizes from hundreds of kilobases to around ten megabases can be synthesized, and this upper limit will only continue to increase. Once synthesized, these chromosomes can be activated inside living cells to study their biological characteristics.

By enabling the specification of entire genetic networks in living cells, synthetic genomics should make it possible to directly probe how network features impact evolution. For example, genetic network designs can be built and tested that explore how changes in connectivity and redundancy within networks impact phenotypic expression and adaptive evolution across environments. As another example, genetic networks can be rationally or randomly reconfigured to map the relationship between network topology and the adaptive landscape or the expression of novel traits. Experiments like these will advance understanding of the mechanisms that

produce life's diversity and improve our ability to predictively engineer living systems.

Edward M. Marcotte
University of Texas at Austin

Seeing farther into deep time

By far, the biggest recent breakthrough for studying protein evolution has been the advent of new computational methods for predicting protein structures and assemblies. This development is proving transformational for evolutionary studies because proteins' amino acid sequences evolve faster than their 3D structures. Predicting 3D structures with accuracy lets us see farther back into evolutionary time and find more distant homologs by structure-structure matching.

For perhaps the first time, our ability to study protein structures has nearly caught up with our ability to sequence DNA, thanks to tools like AlphaFold, RosettaFold, ESMfold, and others. Within just 3 years, we've moved from having detailed structures for hundreds of thousands of proteins to predicting hundreds of millions with reasonable accuracy. This leap has been enhanced by new, fast structure comparison algorithms, such as FoldSeek, that let us use these vast datasets effectively. With these tools in hand, we can explore deep time, uncovering entirely new protein families, 3D folds, and higher-order complexes.

However, I'm most excited by the fact that these same tools open up studies of the least understood organisms on the planet: think of all of the strange fungi, uncultivatable microbes from deep sea core samples, extremophiles, and vast collections of phage and viruses, known only by DNA sequencing. We've essentially just been given entirely new computational lenses to peer into their proteomes and begin understanding something of how they are put together, molecularly speaking.

Rachel Kolodny
University of Haifa

A revolution of protein comparison

Recent methodological developments offer new and improved ways to compare proteins—a foundational tool used in bioinformatics to study molecular evolution. The classical example of a comparison tool is the BLAST implementation for sequence alignment, which was designed with biological intuition and few learned parameters to search for homologues in large databases. Using AI, we can instead compare the so-called protein embeddings: learned fixed-size vector representations for proteins. Learned embeddings for sequences are called protein language models. Trained to predict masked, or hidden, parts (in a self-supervised manner), an embedding learns a distilled form of the amino acid sequence.

Intuitively, that embeddings learn an ''internal grammar'' allows them to model proximities in protein space in a meaningful manner. Thus, embeddings can be used to compare pairs, or larger sets, of proteins and to efficiently search in large databases. Even embeddings that were trained only on sequences capture information beyond sequence, as evidenced by their utility in predicting structural and even functional properties. However, embeddings can be further improved by co-embedding alongside the sequences, structural and functional information (see Ben-Tal's Voice). This will not only render ways to compare the different facets of proteins but also lead to a better holistic comparison. Contemporary AI tools can revolutionize comparison, and just imagine the effect this will have on our understanding of the world.

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

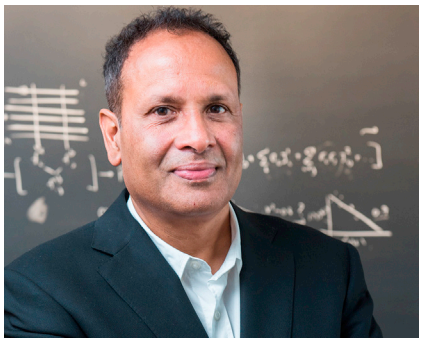
Protein interactions decoded

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

acids critical to interactions, we can begin to follow the evolutionary paths that shape interaction specificity and functionality.

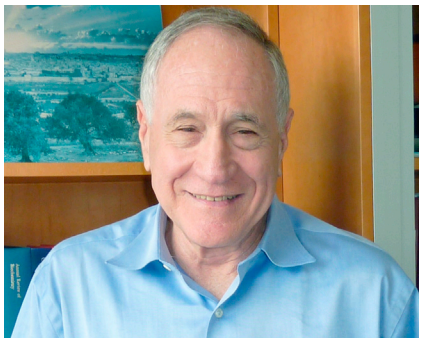


Rama Ranganathan
University of Chicago

"Statistics before physics" in biology

A basic goal in biology is to understand how evolution works as a design process to build machines from the scale of atoms to ecosystems that operate in living systems. A remarkable characteristic of such "natural machines" is that they can approach theoretical limits of performance while maintaining two key properties—robustness and adaptability—that ensure fitness as environments change. The co-existence of high-performance, robustness, and adaptability is not obvious in man-made systems and seems to originate from two unique characteristics of evolved systems: (1) strong heterogeneity, such that parts make unequal or even distinct contributions, and (2) strong non-linearity, such that global behaviors emerge from the cooperative actions of the parts.

How can we learn the design principles of the kind of large, nonlinear dynamical systems that occur in living things? A direct mechanistic approach often fails us because it is hard to have intuition about the right effective variables that control global behaviors. We end up learning local details but not fundamental design principles, and potential simplifications go unnoticed. A major advance at the current time is to take a "statistics before physics" approach to biology, using machine learning tools to learn low-dimensional representations of seemingly complex biological systems from sequence and experimental data. From the scale of proteins to microbial communities, this approach is producing models that are both interpretive (providing new mechanistic insights) and generative (capable of designing synthetic systems). Most importantly, the low dimensionality of these models inspires confidence that we can now make real progress to learn how these systems work and why they are built the way they are through the process of evolution.



Barry Honig
Columbia University

The centrality of 3D structure

As someone who has, for many decades, used computational biophysics to study protein structure and function, the past few years have been both exciting and a bit overwhelming. In particular, the AlphaFold revolution in protein structure prediction and the continuing explosion of sequence information offer great opportunity but also a new set of challenges: how best to exploit the new technologies and the massive amount of data they produce; how best to gain conceptual understanding from AI methods whose insights are buried in black boxes; do we need to "understand" for progress to be made? Obviously, there are many areas where deep understanding may not be necessary, but scientific progress is often fueled by conceptual advances, and if this is to continue, we need to leverage the dramatic developments we are witnessing without sacrificing the quest for concepts, global principles and deep understanding of biological phenomena.

I believe that macromolecular structure offers a way forward. Cryo-electron microscopy has revolutionized structure determination while Cryo-electron tomography is providing increasingly detailed images of sub-cellular structures. AlphaFold has provided us with fairly accurate structure of millions of individual proteins and structural similarity methods, "Structural BLAST," allow us to leverage these structures to predict which proteins interact on a proteome-wide scale and to discover relationships not evident from sequence. Sequence relationships have been the main driver of our understanding of protein evolution, but it seems likely that cross-genome comparison among protein structures, and protein networks, will become increasingly dominant. Similarly, the ability to predict which proteins interact physically, and in multi-protein complexes, will allow a far more insightful description of biological networks than one dimensional signaling pathways that still dominate our thinking. Thus, 3D structure appears to be an increasing central component in the new high-throughput era we have entered.

Yana Bromberg
Emory University

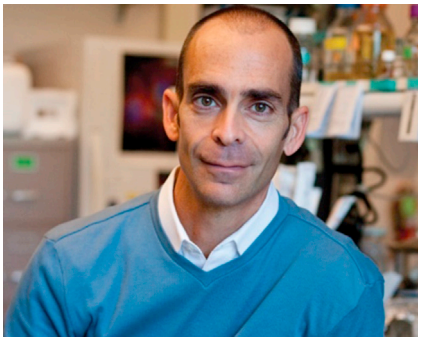
Back to the future

Organismal evolution is driven by functional changes encoded in genomic shifts. Over the years, we have come to rely on genetic or protein sequence differences as proxies of evolutionary patterns. But can we really read these "blueprint" differences accumulated over billions of years of history of life on Earth?

Recently, the field of sequence analysis has gotten a boost from the advances in machine learning and, specifically, from transformer architectures. In the 1990's, hidden Markov models (HMMs) had significantly improved our ability to recognize sequence motifs and, thus, remote homologs. Today's large language models (LLMs) hold similar promise of finding even further out homologs. For example, protein 3D-structure similarity may help infer homology. Coincidentally, sequence defines structure and can be read as such with the right LLM. Moreover, the multi-dimensional encoding of protein sequences generated by such models may itself be a good representation of the faint signatures of similarity. It's hard not to get excited—novel discoveries await!

One must be cautious, of course, of attributing magic to models. It remains to be seen if their structure predictions, or sequence embeddings, can capture evolutionary relationships well. Furthermore, as all models model what is currently known, they may not work for novel sequences. Also, what to do with disordered proteins?

LLMs also hold promise beyond describing history. For example, the current generation of models can act as agents of digital evolution to generate novel structures of putatively functional proteins. But at what resolution does predicted structure capture function? Understanding whether LLMs actually "speak" protein-ish will be important in figuring out the next breakthrough.



Joseph W. Thornton
University of Chicago

Deep and ancient

Evolutionary biochemists want to know how modern proteins evolved their structures and functions and why they took this path in history. We are now advancing on this goal by combining ancestral sequence reconstruction with deep mutational scanning.

As a protein evolves, it wanders through the vast space of all possible sequences, mostly via single-residue changes. It may also duplicate within a genome and speciate along with its host, yielding a family of proteins that follow diverging paths. Using computational phylogenetics and large sequence alignments, we can reconstruct the likely historical paths that led to modern proteins and infer the ancestral amino acid sequences that existed at each branchpoint during history—with good confidence even hundreds of millions of years into the past if sampling is dense relative to the rate of evolution. These reconstructed ancient proteins can then be synthesized and experimentally characterized to identify the precise changes in sequence, structure, and other physical properties that caused the family's diversity to evolve.

But why did evolution follow this one set of breadcrumb paths? Deep mutational scanning enables biochemists to experimentally assay huge libraries of protein variants. By applying this method to reconstructed ancestral sequences, we can now characterize the roads not taken by a protein family. Do modern proteins represent the optimal endpoints of a selection-driven process, or are they one of many possible rolls of the evolutionary dice? If "better" proteins were possible, were they even accessible, or did interactions among protein resides constrain evolution to suboptimal regions? Was history shaped by functionally inconsequential steps that opened doors to new possibilities and closed paths to others? How many different structures and functions—alternative lifeforms at the molecular level—were once accessible, and has this world of unrealized possibility become broader, narrower, or simply different from what existed in the deep past? Deep questions like these and the mechanisms that underlie the answers are now becoming experimentally tractable.

DECLARATION OF INTERESTS

E.M.M. is a co-founder, shareholder, and scientific advisory member of Erisyon, Inc., which played no role in this work. The University of Southern California has filed a non-provisional patent application (530.029WO1) with I.M.E. as named co-inventor covering the entire process of cloning natural DNA segments and assembling these segments into chromosomes inside living cells. R.R. is a board member of Evozyne Inc.