Molecular Cell

CellPress

Voices

Artificial intelligence in molecular biology

In recent years, computational methods and artificial intelligence approaches have proven uniquely suited for studying patterns in molecular biology. In this focus issue, we spoke with researchers about using these tools to address various biological questions and explore both current implications and future possibilities.

Anshul Kundaje
Stanford University, USA

Unlocking the genome's regulatory code

Transcription factors bind complex *cis*-regulatory logic encoded in regulatory DNA sequences to modulate gene transcription. Decoding this "*cis*-regulatory code" is essential for understanding how genes are regulated across cellular contexts, individuals, and species as well as how non-coding genetic variations impact traits and diseases. Over the past decade, machine learning (ML), particularly deep learning, has become a powerful tool for tackling this challenge.

Fueled by advances in high-throughput molecular profiling technologies, models of regulatory DNA are typically trained in a supervised manner to explicitly map DNA sequences spanning across the genome to regulatory and transcriptional activities in one or more cellular contexts. These models have evolved to incorporate different approaches and architectures, aiming to address key biological questions such as deciphering the *cis*-regulatory code, predicting the effects of genetic variation, and designing synthetic DNA sequences with desired properties.

Local sequence models: The most mature models focus on local sequence contexts (typically < 4 kb), predicting outcomes such as transcription-factor binding and chromatin accessibility. These models have progressed from binary predictions to high-resolution, quantitative outputs, including base-pair resolution prediction. Convolutional neural networks dominate this space, often coupled with recurrent layers for additional context. Despite being "black boxes" with opaque internal parameters, advances in model interpretation techniques have shed light on motif lexicons, and their higher-order arrangements (e.g., spacing, orientation, combinatorial density). Local sequence models have also been quite effective at predicting effects of genetic variants and synthetic sequence perturbations on local regulatory activity. They have also shown success in designing cell-context-specific regulatory sequences, although mostly in cell lines. However, these models inherently do not model long-range interactions, such as the influence of distal enhancers on gene expression.

Expanding the scope with long-context models: To capture long-range regulatory interactions, new models with greater capacity—often based on transformer architectures—have been developed. These "long-context models" are typically trained to jointly predict genome-wide regulatory and transcriptional activity across thousands of diverse cellular contexts and even across species. However, these large models introduce challenges with interpretability and robustness and are more susceptible to learning spurious patterns. Recent benchmarks have revealed that these large, kitchen-sink models do not always outperform smaller, specialized, local-context models and especially struggle to learn long-range regulatory interactions despite their design, resulting in poor prediction of distal effects of regulatory elements and variants on gene expression.

Several promising strategies could improve these models without sacrificing interpretability and robustness. These include stage-wise training methods that progressively build long-context models from local context models and model optimization strategies that leverage multiple informative axes of variation of regulatory and transcriptional activity—across the genome, across cell types and states, across diverse data modalities, and across diverse genetic backgrounds. The deluge of multi-modal single-cell data and large-scale perturbation screens will power these models.

The potential of DNA language models: Inspired by advances in natural language processing, self-supervised "DNA language models" have very recently emerged as a promising complementary approach. These models are currently trained solely on large collections of genomic sequences across species to learn generalizable patterns by reconstructing artificially masked or truncated DNA sequences. It remains an open question whether these models can effectively learn representations of diverse classes of functional DNA without using any annotation or context-specific molecular data. So far, annotation-agnostic, long-context DNA language models of mammalian genomes have not delivered on their promise. Although still in their infancy, if successful, these large pre-trained models could serve as ‘‘foundations’’ to fine-tune more specialized models on smaller datasets, potentially reducing data requirements and broadening applications.

Toward a unified future: Looking ahead, we will likely witness the emergence of effective long-context, supervised, and self-supervised "foundation models" for regulatory genomics. However, smaller, interpretable task-specific models will remain indispensable for addressing focused biological questions and deriving robust mechanistic insights. But we must learn from the lessons of the past to ensure rapid progress by avoiding premature hype. Success will hinge not only on innovative model design and training strategies but also on meticulous data curation, exhaustive model interpretation, and rigorous, adversarial benchmarking. Regardless of the approach, predictive models of gene regulation will continue to drive hypothesis generation at unprecedented resolution and scale, offering exciting opportunities for model-driven discovery and iterative experimental design.

Katherine S. Pollard
Gladstone Institute of Data Science & Biotechnology, USA

Investigating chromatin organization *in silico*

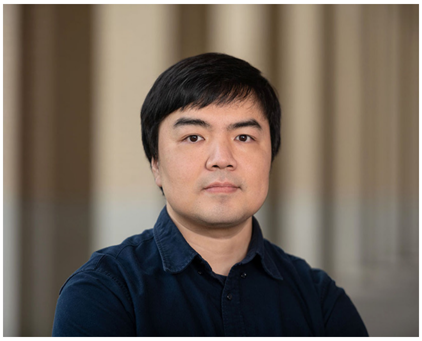
Chromatin organization is integral to how eukaryotic cells work, evolve, and break in disease. One grand challenge is to identify the epigenetic changes in multicellular organisms that enable a single genome sequence to generate many unique cell types. Another important problem is discovering causal genetic variants in non-coding loci associated with disease. Increasingly, these and other questions are being addressed using computational models that predict local chromatin states or three-dimensional (3D) genome folding from DNA sequences.

Why is ML so important in the study of chromatin organization? Several years ago, the field viewed these questions as data problems. This led to the development of imaging and genomic assays that measure 3D chromatin interactions and 1D chromatin states with ever-increasing throughput and resolution. But these observational datasets have several limitations. First, many tissues and cell states are difficult or impossible to access. Second, understanding how chromatin organization is encoded in DNA requires experiments that can establish causality. Machine learning addresses both of these gaps.

In terms of bolstering observed data with model predictions, the key breakthrough was the development of sequence-based models that are highly accurate (https://doi.org/10.1038/s41588-021-00782-6,https://doi.org/10.1038/s41592-020-0958-x). With this level of model performance, researchers can impute chromatin data for conditions that lack measurements and design sequences with desired epigenetic activities. Because DNA is cheaper and easier to collect than cell-type-specific epigenetic data, whole-genome sequences now exist for about one million people. Using these as inputs, the ML models can generate personalized chromatin organization data across cell types at biobank scale. Such predictions are particularly useful for the least inaccessible tissues.

Beyond the predictions themselves, ML can also help to establish mechanistic understanding of how sequence variation affects chromatin organization. Experimental genome editing and synthetic DNA assays enable candidate causal variants to be tested in cells. But current editing technologies are not high throughput enough to explore trillions of variants and variant combinations. On the other hand, models can be queried with huge numbers of sequence modifications (*in silico* mutagenesis). Furthermore, interpretation tools can distill what sequence-function relationships a model has learned. This toolkit reveals many surprising sequences underlying chromatin organization, such as the importance of repetitive elements in genome folding and transcription factors that switch from activators to repressors in different contexts. Thus, models are helping scientists to generate novel hypotheses and to prioritize sequences for experimental testing.

Looking ahead, there are several challenges and emerging opportunities. First, it will be important to increase the cell-type specificity of existing models through increased training data and changes to the methodology. Second, we need to link predicted epigenetic changes to function so that their importance to cell biology can be investigated. For example, current models can predict how a genetic variant alters chromatin accessibility but struggle to predict downstream changes to RNA or protein levels (https://doi.org/10.1038/s41592-024-02331-5). It will be exciting to see the next generation of sequence-based models that integrate multi-modal data from large cohorts to solve these problems!



Jian Ma
Carnegie Mellon University, USA

Decoding single-cell genome structure and function

While we can now map the entire genome from telomere to telomere, understanding how it is organized in 3D space and how this relates to cellular function remains a major challenge. Advances in genomic mapping and imaging technologies have probed different aspects of 3D genome architecture, but the principles governing nuclear structure and its role in cell function are largely unclear. A key challenge is deciphering the spatial and temporal dynamics of multiscale 3D genome features at single-cell resolution and understanding how these variations influence gene expression and cellular processes.

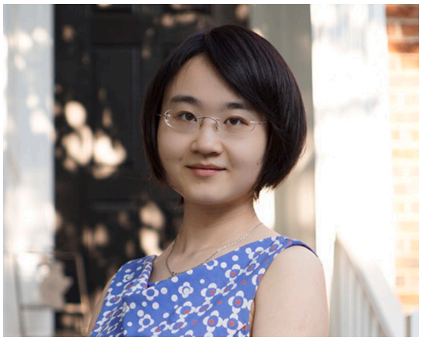
Technologies like single-cell Hi-C and genome imaging have provided new insights into how 3D genome organization contributes to cell identity. Artificial intelligence/machine learning (AI/ML) models are beginning to reveal deeper connections between spatial genome features and their biological significance. However, we still lack comprehensive models that predict how 3D genome organization shapes gene expression at the single-cell level. To address this, predictive models must integrate DNA sequences, chromatin architecture, and epigenomic data to uncover the complex relationships driving cell function. Beyond chromatin interactions, genome function is modulated by other biomolecules, such as RNA, proteins, and subnuclear structures that participate in 3D interaction patterns within the nucleus. Developing multimodal AI/ML models that capture these interactions alongside chromatin will be essential. Such models could answer critical questions such as how chromatin is organized into multiscale 3D structures, how functional elements collectively regulate genes, and which nuclear features are crucial for processes like DNA replication and repair. Additionally, cell-to-cell variability in 3D genome organization may play a crucial role in shaping gene expression patterns and cellular behavior, yet capturing and interpreting this variability remain a significant challenge.

The future of the field lies in developing more sophisticated, multimodal AI/ML models that can comprehensively decode the ‘‘language’’ of 3D nuclear structure. Generative AI/ML models can enable *in silico* characterization of interconnected nuclear components and produce testable hypotheses to reveal underlying mechanisms. Techniques such as interpretability and causal inference will further help nominate novel targets for perturbation, which are key to understanding these complex interactions. By integrating diverse data types and uncovering how spatial genome features influence gene regulation, cell function, and biological processes, these models hold the potential to open broad opportunities for exploring the intricate relationship between genome structure and function in health and disease.

Xing Chang
Westlake University, China

Optimizing gene-editing enzymes with language models

Gene-editing enzymes, such as CRISPR and its derived systems (e.g., base editors), hold the potential to cure genetic disorders through precise genomic alterations. Challenges remain in their application for therapeutic purposes, particularly in inducing certain types of mutations, achieving high *in vivo* editing efficiency at disease-relevant genomic loci and minimizing unintended "bystander" and "off-target" edits. While traditional optimization strategies such as directed evolution and structure-based design have proven to be effective, they require extensive experimental efforts and are often assessed by a single parameter (e.g., efficiencies). In contrast, protein language models (PLMs) are deep-learning models that are trained on millions of natural protein sequences that developed throughout evolution, a process analogous to the training of natural language models. The inherent protein design principles are learned by PLM, which can be exploited to predict sequence probabilities. One key advantage of PLMs is to predict global amino acid dependencies across protein sequences, including the amino acids that may be far apart in the primary sequence. Consequently, the knowledge captured in pre-trained PLMs can be applied for optimization of specific proteins without a task-specific training dataset. For instance, evolutionary scale modeling (ESM) has been used to enhance the activities of a uracil DNA glycosylase (UNG) variant with altered substrate specificities. The resulting enhanced enzyme, when coupled with spCas9 nickase, leads to transversion base editors targeting thymines, named TSBE (i.e., T to G/C base editor). Among the top 50 PLM-predicted UNG variants, >50% exhibited over 1.5-fold enhancement in enzymatic activities. Interestingly, most top-ranking variants reside outside of the catalytic domain, which may have been missed if a directed evolution approach targeting the catalytic domain had been employed (https://doi.org/10.1016/j.molcel.2024.01.021). Furthermore, models like ProMEP have successfully predicted the outcomes of adenine deaminase mutations, recapitulating the results from previous experimental data (https://doi.org/10.1038/s41422-024-00989-2). These successes highlight the capacity of PLMs to score and provide actionable guidance on candidate variants for further assessment. Beyond functional optimization of enzymes, PLMs also hold promise in designing *de novo* gene-editing enzymes, illustrated by the latest developments in open CRISPR systems (https://doi.org/10.1101/2024.04.22.590591). Future development of PLMs will include more extensive integration of protein structural information, which will enhance their predictive power for protein-protein and/or protein-nucleic acid interactions. As these models evolve, they will likely become indispensable tools in the gene-editing field, offering an efficient path to design, evaluate, and implement next-generation therapeutic enzymes.



Mengjie Chen
University of Chicago, USA

Emerging challenges in high-resolution RNA modification mapping

Antibody-based transcriptome-wide mapping methods, including MeRIP-seq, have greatly advanced our understanding of RNA modifications. Emerging technologies, including m6A-SAC-seq, eTAM-seq, and GLORI, have leveraged enzymatic and chemical reactions to improve m6A quantification and mapping resolution. However, these approaches are dependent on enzyme efficiency, relative selectivity, and RNA accessibility, which introduce challenges for statistical modeling and data analysis.

Enzyme efficiency refers to the maximum rate at which mutations are induced onto target nucleotides. Unlike bisulfite conversion, where efficiency typically exceeds 99%, achieving a conversion rate above 90% can be challenging for enzyme-based approaches. Mutation rates are also affected by enzyme sequence preferences. For example, m6A modifications in the common G-m6A-C consensus motif tend to show higher mutation rates than those in the less common A-m6A-C motif, highlighting the importance of sequence context in conversion efficiency.

Relative selectivity regards the possibility that both m6A and unmethylated adenosines can be modified due to their similar chemical properties. Whereas eTAM-seq and GLORI target unmethylated adenosines, m6A-SAC-seq targets m6A. For m6A-SAC-seq, off-target mutations at unmethylated adenosines are a major concern, particularly given that unmethylated adenosines are ∼250 times more abundant than m6A. Accurate modeling of off-target conversion efficiency is therefore essential.

RNA accessibility refers to the proportion of RNA molecules accessible to enzyme treatment, which can vary based on RNA secondary structures and other modifications. In eTAM-seq, RNA accessibility is estimated using *in vitro* transcription (IVT). Briefly, RNA is converted into cDNA and linearly amplified into RNA, preserving sequence and expression information without introducing any modifications. Accessibility is then modeled as a site-specific parameter shared between enzyme-treated RNA samples and IVT controls. The observed methylation is a product of the true methylation level and site accessibility, requiring statistical adjustment to estimate the true underlying methylation level.

While technologies like m6A-SAC-seq, eTAM-seq, and GLORI have advanced our understanding of the transcriptome by allowing us to map RNA modifications at single-base-pair resolution, they bring their own set of challenges that must be addressed through careful experimental design and robust statistical modeling. Statistical models must then be sophisticated enough to factor in potential off-target effects, biases linked to local sequence context, and variations in enzyme efficiency across different batches. To address these complexities, Bayesian inference approaches could be employed to meticulously model the data generation process, taking into consideration the detailed technical aspects of each method. As the field evolves, further refinement of these techniques and analytical approaches will be essential to unlock a deeper understanding of the role of RNA modifications in gene regulation and cellular function.

Remo Rohs
University of Southern California, USA

Toward a complete structural landscape

The 2024 Nobel Prize in Chemistry awarded for protein design and protein structure prediction breakthroughs highlights the fascinating possibilities in biology and medicine enabled by computational and statistical methods, particularly ML and AI.

Protein structure is the window into understanding protein function. Solving the 3D structure of a protein experimentally is laborious and not always possible. In addition, experimental structures are often limited to protein fragments and subjected to crystallization effects. Cryoelectron microscopy has overcome some of these limitations, although deriving multiple conformations or information on flexibility of a molecule is still challenging. Molecular dynamics simulations partially fill this gap but have their own limitations since they are computationally expensive, require significant run time for meaningful results, and often remain restricted to sampling of a local energy minimum.

AlphaFold dramatically changed the opportunities for structure-based research in biology, chemistry, and medicine. Structure-based approaches are no longer restricted to studies that involve previously solved structures. By taking advantage of statistical ML and AI methods, AlphaFold uses data of known structures, which are available in the Protein Data Bank (PDB), to predict the 3D structure, requiring only a protein's amino acid sequence. This, in turn, has a large impact on the acceleration of the development of structure-based methods to answer questions based on molecular interaction, binding, and recognition.

Protein-DNA readout is a field of research that benefits from the availability of the 3D structure of any protein of interest. The binding of a transcription factor to a specific site in the genome regulates genes and their expression. In the past, laborious experiments, which included protein purification and DNA sequencing, were necessary to derive the nucleotide sequence preference of DNA target sites selected by a DNA-binding protein. A recently published AI method, DeepPBS (https://doi.org/10.1038/s41592-024-02372-w), can predict the sequence specificity of any DNA-binding protein. The probability of each nucleotide occurring at a given position in the DNA target site is described as DNA-binding specificity of a protein. DeepPBS uses the 3D structure of a protein-DNA complex, which can be solved experimentally or derived computationally with AlphaFold 3 or RoseTTAFoldNA without DNA sequence information to predict the identity of nucleotides contacted by the protein.

The prediction of nucleic acid structures has experienced similar advances as the prediction of protein structures, although the number of experimentally determined DNA and RNA structures in the PDB is much smaller than that of proteins. For predicting the sequence-dependent 3D structure of the DNA double helix, known as DNA shape, a data-mining method, DNAshape, was initially developed. DNAshape uses a sliding pentamer window with structural information derived from computationally predicted DNA conformations, which cover the entire nucleotide sequence space in contrast to experimental DNA structures. A recently published AI method, Deep DNAshape (https://doi.org/10.1038/s41467-024-45191-5), expands on this approach to predict DNA shape features of *k*-mers of any length. This method enables the exploration of the role of flanking regions on conformations at the core motif of transcription factor binding sites.

Structure-based AI methods for drug design are likely to have a large impact. Current computational methods are used to virtually screen large libraries of existing drug candidates. In fact, the available chemical space is many orders of magnitude larger than drug libraries. AI methods are being developed to design small molecules. One example is a recently published method, DrugHIVE (https://doi.org/10.1021/acs.jcim.4c01193), which uses protein structure to design new and previously unknown chemical compounds based on interactions with protein cavities or surfaces. Using a density description for small molecules, DrugHIVE enables the exploration of the vast chemical space of drug-like molecules with possibly unlimited opportunities to treat diseases.

DECLARATION OF INTERESTS

K.S.P. is also affiliated with the University of California San Francisco and the Chan Zuckerberg Biohub San Francisco. R.R. and the University of Southern California have filed an international patent application (No. PCT/US2024/054511).