Deep learning for regulatory genomics

Yongjin Park & Manolis Kellis

Computational modeling of DNA and RNA targets of regulatory proteins is improved by a deep-learning approach.

A fundamental unit of gene-regulatory control is the contact between a regulatory protein and its target DNA or RNA molecule. Biophysical models that directly predict these interactions are incomplete and confined to specific types of structures, but computational analysis of large-scale experimental datasets allows regulatory motifs to be identified by their over-representation in target sequences. In this issue, Alipanahi *et al.*[1] describe the use of a deep learning strategy to calculate protein–nucleic acid interactions from diverse experimental data sets. They show that their algorithm, called DeepBind, is broadly applicable and results in increased predictive power compared to traditional single-domain methods, and they use its predictions to discover regulatory motifs, to predict RNA editing and alternative splicing, and to interpret genetic variants.

Diverse statistical models have been proposed for regulatory motif discovery[2], but current models still have considerable limitations[3], especially for RNA-binding proteins that recognize both sequence components and secondary (or tertiary) structural components. Moreover, regulatory proteins bind in the context of dozens of other proteins that compete for occupancy or exert synergistic effects, by binding nearby or partly overlapping positions. This results in higher-order structures and motif combinations that are not easily recognizable by traditional methods.

Deep learning, a recent modification of multi-layered artificial neural networks, is a particularly powerful approach for learning complex patterns at multiple layers. Originally inspired by the layers of neurons that receive and combine information in the human brain, neural networks have been remarkably adept at learning complex tasks with relatively simple building blocks arranged in complex networks. However, their internal representations have generally been difficult to interpret, and training deeply layered models has been algorithmically intractable and statistically prone to overfitting.

So-called 'belief networks' have recently combined the learning architectures of neural networks with generative models framed in the context of an internal 'representation' of the world, from which samples are drawn with varying probabilities. Model parameters that match observed data are fit with Bayesian statistics and are used to classify elements in complex datasets by means of nonlinear decision boundaries, similar to traditional neural networks. The generative nature of belief networks results in explicit representations of the world that reduce overfitting and may yield new insights about a problem domain at multiple levels of resolution[4,5]. At each layer, these internal representations provide lower-dimensional views of high-dimensional data, and can be accessible to visualization and human interpretation. These features are refined during the learning process, akin to manual feature engineering, which traditionally required extensive domain knowledge, and leave a record of concept learning that can be of great interest. The resulting weights associated with each concept reflect its contribution to overall classification accuracy, and are akin to feature selection.

Training deep models poses far greater challenges than training shallow models, both for defining model parameters and model structures. Back-propagation algorithms used for traditional neural networks propagate the difference between observed and predicted output to adjust parameters[5], but their focus on classification accuracy can miss general properties common to multiple classes and overfit small datasets. By contrast, training of belief networks seeks also to optimize the overall fit to the data, either through optimization functions that explicitly combine data likelihood and classification error or through sequential learning algorithms[6] that start with generative pre-training to capture representations agnostic to output labels and continue with discriminative optimization to fine-tune classification of pre-trained models. Training only two layers at a time facilitates rapid convergence and reduces overfitting. Model depth and breadth can be automatically tuned, thus adjusting model complexity.

Deep learning models have helped revolutionize the field of machine intelligence. Somewhat counterintuitively, classical artificial intelligence was highly successful for tasks difficult for humans (e.g., equations, symbolic reasoning, chess), but seemingly simple human tasks (e.g., scene understanding) proved much more difficult. Deep learning led to multiple breakthroughs in cognition, surpassing expert systems for scene recognition, producing increasingly accurate automated translations, and making speech recognition commonplace in most smart phones.

Deep learning is well suited to genomics. The multiple learning layers can capture multiple levels of information processing and abstraction within cells. The explicit representations at each layer can reveal insights about the biological structures inferred at multiple levels of resolution. They enable data summarization into lower-dimensional representations at each layer, and their modular structure allows integration of diverse input data types at higher layers.

For protein–nucleic acid binding prediction, meaningful features can be individual sequence motifs or *k*-mers at the lowest layers combinations of motifs at intermediate layers and complex motif 'grammars' at the highest layers. Traditional motif scanning algorithms can discover lower-level features but would only capture combinations if they co-occurred frequently, and they would miss higher-level

Yongjin Park & Manolis Kellis are at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.
e-mail: manoli@mit.edu

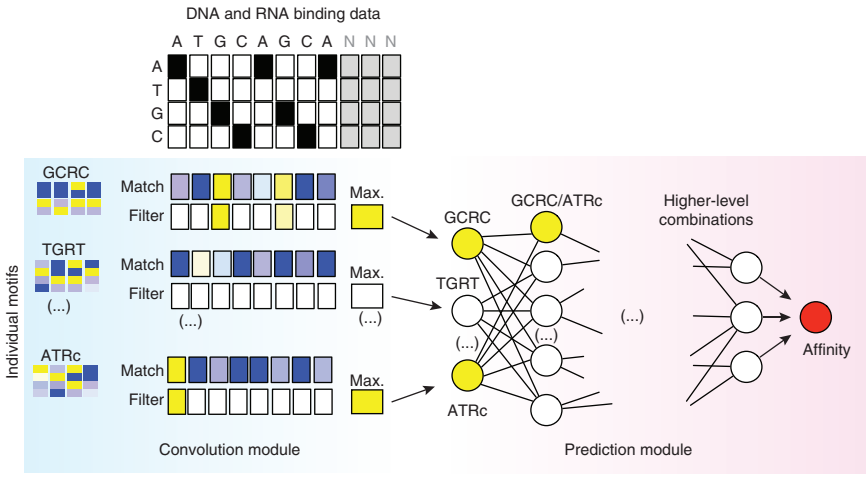

Figure 1 Illustration of the deep convolutional neural network designed by Alipanahi *et al.*[1]. The motif discovery layers apply local sequence filters and extract relatively short motifs (convolution[6]), and higher prediction layers synthesize local patterns in deep neural architecture (representational learning[4]). Raw input sequences are first converted to a sequence matrix and screened by convolution filters, which mark the location and intensity of desired sequence motifs. These filtered signals are then collected if they reach above some threshold level, pooled and fed into a deep neural network, where simplified signals, such as the presence and absence of motifs, are synthesized to capture higher-level concepts.

features that may underlie ultraconserved elements or enhancer clusters.

Alipanahi *et al.*[1] incorporated state-of-the-art techniques from deep learning in the development of DeepBind. The method predicts binding affinity of a protein to a DNA or RNA sequence in two steps, consisting of applying a convolution module for representation learning and a prediction module for feature combinations (**Fig. 1**).

The convolution module uses individual nodes akin to *de novo* motif detectors, and identifies local sequence patterns known as position-weight matrices (PWMs) that summarize nucleotide frequencies at each position. Repeatedly occurring sequence k-mers boost the weight of PWMs that increase prediction accuracy, and the shape and length of PWMs are not hard-coded but are automatically adjusted.

The prediction module synthesizes local features into higher-level structures using a nonlinear neural network that considers them in many combinations and orientations. This allows capture of longer motifs, motif pairs and combinations, and more complex patterns at multiple layers. Prediction power can only come from utilization of these features, thus forcing the ultimate classifier to use meaningful representations of the underlying biological signal.

Alipanahi *et al.*[1] evaluated the performance of DeepBind using nearly 1,000 publically available datasets, encompassing DNA binding *in vivo* (chromatin immunoprecipitation) and *in vitro* (protein binding microarrays), and RNA binding *in vitro* (high-throughput screening). DeepBind predicted protein binding microarray scores with nearly perfect accuracy, ChIP-seq with strong accuracy (area under the receiver operating curve=0.7) and RNA binding similarly to replicate experiments.

The authors also compared their approach to the best previous methods, some of which are based on extensive biological knowledge or are customized to specific biological systems. Although the ranking of competing methods varied widely depending on the types of experiments and transcription factors, DeepBind consistently outperformed all of them, even when the training and testing data sets were of different types—which means that the knowledge encapsulated in the model is truly transferrable.

The authors demonstrated diverse applications for the increased prediction accuracy, including predicting the effect of single-nucleotide variants from genome-wide association studies on regulator binding, which can help elucidate the intermediate phenotypes leading to diverse complex traits and diseases.

The high accuracy indicates that meaningful representations were learned, but the authors only make the lowest-level representations explicit, while higher-levels remain a black box. Opening this black box and interpreting higher levels presents an opportunity to gain insights about the language of gene regulation beyond the word level, a long-standing challenge in the field.

DeepBind resembles existing bioinformatics pipelines that include motif discovery and binding energy prediction, but it has the advantage that model parameters and complexity are selected automatically (see Fig. 2b in ref. 1). Recent advances in training methods[7] make DeepBind training practical, even when considering many possible combinations of structures in a probabilistic fashion[8]. It can thus be widely useful for mining increasingly large datasets, such as ENCODE[9] or Roadmap Epigenomics[10].

Looking beyond regulatory motifs, the current results illustrate the power of deep learning for biological data analysis in general. The approach can increase predictive power for specific tasks, integrate diverse datasets across data types, and provide greater generalization given the focus on representation learning and not simply classification accuracy. Systematic visualization and exploration of internal representations at each layer can yield mechanistic insights and guide new experiments and research directions.

More broadly, deep learning can serve as a guiding principle to organize both hypothesis-driven research and exploratory investigation. For this potential to be realized, statistical and biological tasks must be integrated at all levels, including study design, experiment planning, model building and refinement, and data interpretation.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Alipanahi, B., Delong, A., Weirauch, M.T. & Frey, B.J. *Nat. Biotechnol.* **33**, 831–838 (2015).
2. Stormo, G.D. *et al. Bioinformatics* **16**, 16–23 (2000).
3. Weirauch, M.T. *et al. Nat. Biotechnol.* **31**, 126–134 (2013).
4. Bengio, Y., Courville, A. & Vincent, P. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
5. LeCun, Y., Bengio, Y. & Hinton, G. *Nature* **521**, 436–444 (2015).
6. LeCun, Y. *et al. Neural Comput.* **1**, 541–551 (1989).
7. Hinton, G.E., Osindero, S. & Teh, Y.-W. *Neural Comput.* **18**, 1527–1554 (2006).
8. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
9. The ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
10. Roadmap Epigenomics Consortium *et al. Nature* **518**, 317–330 (2015).