CellPress
OPEN ACCESS

Trends in
Plant Science

Forum

Toward learning the principles of plant gene regulation

Jan Zrimec [1,*], Aleksej Zelezniak,[2] and Kristina Gruden[1,*]

Advanced machine learning (ML) algorithms produce highly accurate models of gene expression, uncovering novel regulatory features in nucleotide sequences involving multiple *cis*-regulatory regions across whole genes and structural properties. These broaden our understanding of gene regulation and point to new principles to test and adopt in the field of plant science.

Innovations in gene expression prediction

Deciphering the properties of gene expression, the process that includes transcription, mRNA stability and translation, will have an important impact on understanding plant physiology and improving crop productivity [1,2]. To this end, advances in data-driven **ML** (see Glossary) algorithms, such as **deep neural networks (DNNs)**, have enabled significant improvements in predicting gene expression patterns across a number of model organisms [2,3] (Box 1). By further examining the 'learned' internal structure of these models, we can decipher gene regulatory features that the models infer from the data to make such accurate predictions [2,4,5] (Figure 1). The aim of the present article is to give an overview of recent ML developments and uncovered regulatory principles, indicating how they can impact future research in the field of plant science.

Gene regulatory structure jointly controls expression patterns

The regulation of gene expression is traditionally considered to be mainly driven by the promoter, a ~100–3000 bp region immediately upstream of the transcription start site that carries multiple protein-binding sites for transcription factors (TFs) and other RNA polymerase-related proteins regulating transcription initiation (Figure 1A). However, using DNNs, studies have shown that the whole DNA sequence at and around the gene is highly predictive of expression levels [2,6–8]. According to genome-scale models, ~2 kb of this DNA already contains the majority of information on average mRNA abundance across multiple conditions or tissues, explaining over 82% of their variation and revealing expression-related effects of sequence motifs and their associations [2]. In higher eukaryotes, improved performance is achieved by models with inputs of up to 100 kb, spanning distant enhancer–promoter interactions [9].

This contributes to the awareness that gene expression regulation spans different coding and noncoding regions that include the enhancer, promoter, untranslated regions (UTRs), and terminator [2,8] (Figure 1A). It is affected by the enzymatic accessibility of DNA defined by chromatin and epigenetic states [3,10]. Since mRNA abundance is a result of both mRNA synthesis and degradation, it is controlled not only by TFs and core promoters, but by a more complex set of *cis*-acting elements carried mostly by UTRs [11]. Whereas promoter regions were found to explain up to 96% of the variation of gene expression according to DNNs, coding regions can explain up to 69% and 5′ and 3′ UTRs as much as 89% [10]. The different regions also carry complementary information, with different parts coevolving and predictive of the activity of others [2]. A key characteristic of the gene regulatory structure seems to be that the initiating regions, namely the promoter and 5′ UTR, define large-scale expression properties (turning it on/off) and can overshadow the contributions of downstream terminating regions (3′ UTR and terminator) involved in smaller-scale changes of mRNA abundance, which are harder to capture in models [8].

Including the information from the whole gene regulatory structure as **explanatory variables** in DNNs (Box 1) was shown to improve predictive performance both in plants (*Arabidopsis*) [2,8] as well as other organisms from lower to higher eukaryotes [2,7]. Moreover, experimentally varying certain regions (e.g., terminator) while keeping others intact (e.g., promoter) was found to have a large effect on gene expression, with perturbations of up to two orders of its magnitude [2]. Most prediction and design approaches, however, still focus on a single regulatory region (most frequently the promoter or a part of it), tested only with a specific reporter gene and not optimized for genome-wide application [1,6]. Regulator design could thus be greatly improved by taking the whole gene regulatory structure into account.

Glossary

Deep neurals networks (DNN): ML models with multiple hidden layers, each learning an informative representation of the data, such as DNA regulatory grammar.
DNA structural properties: physicochemical and conformational numerical variables computed from the nucleotide sequence, such as DNA shape.
Engineered features: new variables derived from the initial explanatory variables with the goal of improving model accuracy.
Explanatory variables: a set of input features based on which the ML model predicts the value of the target variable.
Machine learning (ML): a set of algorithms that automatically build models from training data.
Shallow ML: classical ML algorithms that produce models with few hidden layers and that do not automatically learn informative data representations, instead relying on the modeler to provide engineered features.
Target variable: the variable or set of variables whose values the ML model learns to predict to a certain degree of accuracy.

Box 1. Data-driven machine learning (ML) for deciphering gene regulatory principles

ML algorithms use training data to automatically build predictive models, which can be examined to interpret new regulatory principles (see Figure 1 in main text). Continuing the success of classical, shallow ML architectures, deep neural networks (DNNs) have brought improvements to the prediction of transcription factor binding sites [4] and gene expression patterns [2,3], such as those obtained via high-throughput sequencing, including ChIP-seq, DAP-seq, RNA-seq, and ATAC-seq [4]. This is due to their ability to extract information directly from raw input nucleotide sequence instead of relying on **engineered features** as with ML, where the sequence is encoded with numerical variables [10]. Since DNNs automatically learn predictive motif representations [4], cooperative binding interactions [2,5] and genotypic variation effects [3], they represent a powerful approach to uncover the detailed *cis*-regulatory grammar of genomic sequences (see Figure 1C in main text). Shallow ML approaches remain a highly useful alternative for assessing the performance of numerically encoded DNA sequence features, such as DNA structural (shape) properties [12].

DNA shape guides protein–DNA interactions to define gene expression

The specificities of protein–DNA interactions are defined both by direct protein–DNA readout, facilitated by the major groove of the DNA helix, and indirect readout, mediated through DNA backbone and minor groove contacts. The latter comprises 'weak' protein–DNA interactions, with base pairs that are not directly contacted by the protein but defined by DNA conformational and physicochemical (shape) properties at or around the specific binding sites [5,12].

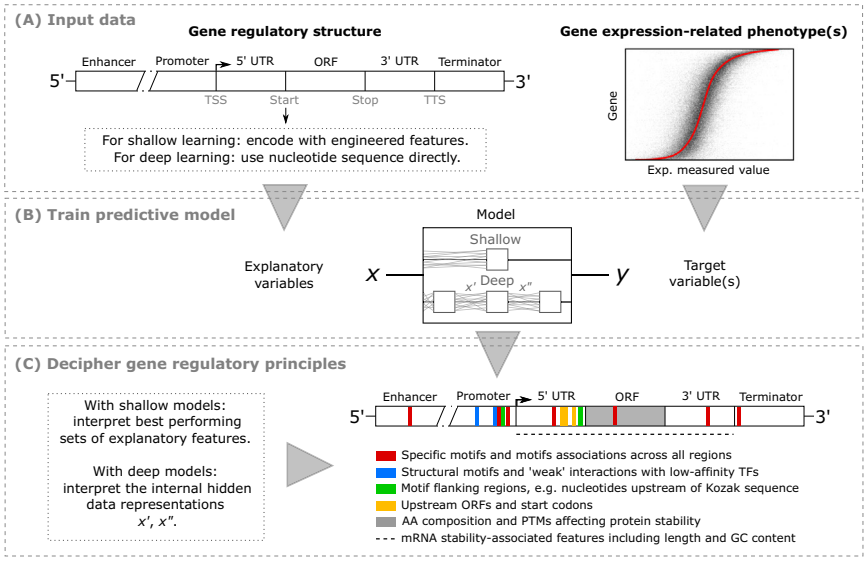

Figure 1. Overview of data-driven machine learning to uncover the principles of plant gene regulation. In supervised learning a model is trained to predict *y* (**target variable**) from the set of features *x* (explanatory variables) present in the training dataset. (A) *x* is a set of nucleotide sequences or numerical engineered features (e.g., position weight matrices, k-mer frequencies, DNA shape) and *y* describes some property related to gene expression [e.g., transcription factor (TF) binding, mRNA abundance, chromatin accessibility]. (B) Differing from classical shallow architectures, deep neural networks (DNNs) are abstracted by multiple hidden layers between *x* and *y*, each learning a new and informative representation of the data. For instance, the most frequently applied DNNs in genomics, 1D convolutional networks, scan DNA and learn to recognize different patterns such as partial and full motifs in early layers, combining these into associated sets in later layers. (C) Regulatory knowledge can be interpreted by either: (i) evaluating the performance of models trained on different subsets of explanatory variables, or (ii) inferring the representations learned by the hidden DNN layers (e.g., by occluding portions of the input sequence and measuring their effects on the output to reconstruct the most important motifs). Finally, the gene regulatory structure with recently uncovered regulatory elements is depicted (bottom right). Abbreviations: AA, amino acid; ORF, open reading frame; PTM, post-translational modification; TSS, transcription start site; TTS, transcription termination site; UTR, untranslated region.

Using **shallow ML** models, it was recently shown in multiple eukaryotic organisms, including *Arabidopsis*, that most TFs likely combine both types of readout to recognize their binding sites, as the integration of these features improves binding site predictions [12]. In line with this, local **DNA structural properties** can be more highly conserved than nucleotide sequence and were similarly found to contribute to TF cooperativity, a mechanism where TFs bind DNA cooperatively and thereby strengthen their affinity [5].

Furthermore, due to the ability of DNNs to learn and uncover multiple informative data representations in their internal hidden layers (Figure 1B), they can automatically learn the intrinsic protein–DNA binding properties underlying gene expression, which is likely the reason for their high performance [2,9]. The learned regulatory grammar contained in the models' hidden representations can be interpreted by a number of methods and collectively includes both known *cis*-regulatory elements as well as multiple novel elements and features [10] (Figure 1C). These include weakly interacting motifs and motif-flanking regions, which are known to underlie weak interactions with low-affinity TFs and can have highly conserved structural profiles that affect coregulator or TF binding around or at the binding site [5,10]. Also uncovered are motif associations across the whole gene regulatory structure, which were found to explain practically the entire dynamic range of gene expression, further indicating how all regions contribute to the joint regulation of gene expression levels [2].

Synthetic regulator design and future challenges

Although the majority of current regulatory insights were obtained in non-plant organisms, such as the eukaryotes yeast and human as well as bacteria, the innovative ML approaches represent an invaluable resource for the plant science field, both for improving predictions as well as for designing novel synthetic regulators. A crucial point is

that the whole gene regulatory structure should be considered in regulator designs, so that when changing one regulatory region also the adjacent regions and coding region properties are taken into account.

Future challenges thus include: (i) building and evaluating sequence-to-expression models to capture regulatory fitness landscapes across different plant organisms, including tissue and condition-specific models; (ii) quantifying the amount of regulatory information in different regions and deciphering the regulatory grammar of both transcription and translation; (iii) development of population-scale models based on multiple genotypes to increase the resolution and accuracy of evaluating sequence variant effects; (iv) analyzing and cataloging the natural regulatory elements found in plant model organisms and their joint effects, including the use of different coding regions, on gene expression; (v) further unraveling the mechanisms and contributions of DNA structural properties to protein–DNA binding and gene regulation; and (vi) development of approaches to generate *de novo* regulatory elements, resulting in libraries of both natural and synthetic elements with documented combinatorial effects.

To conclude, the application of DNNs in deciphering gene regulation can substantially reduce the extent of required wet-lab experimentation and validation. Once built, models capturing detailed regulatory fitness landscapes enable pure *in silico* investigation of gene regulatory principles and designs, addressing fundamental research questions and facilitating multiple development objectives in the plant sciences. This can greatly impact plant breeding approaches, protein production *in planta*, and biosensor design.

Acknowledgments

This study was supported by the Slovenian Research Agency (ARRS) grant no. J2-3060 and grant no. P4-0165, Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia grant no. 11013-9/2021-2, Swedish Research council (Vetenskapsrådet) starting grant no. 2019-05356, and SciLifeLab funding.

Declaration of interests

No interests are declared.

[1]Department of Biotechnology and Systems Biology, National Institute of Biology, Večna pot 111, 1000 Ljubljana, Slovenia
[2]Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, 412 96, Gothenburg, Sweden

*Correspondence:
jan.zrimec@nib.si (J. Zrimec) and
kristina.gruden@nib.si (K. Gruden).

https://doi.org/10.1016/j.tplants.2022.08.010

References

1. Jores, T. *et al.* (2021) Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nat. Plants* 7, 842–855
2. Zrimec, J. *et al.* (2020) Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.* 11, 6141
3. Zhao, H. *et al.* (2021) PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. *Nucleic Acids Res.* 49, W523–W529
4. Lai, X. *et al.* (2019) Building transcription factor binding site models to understand gene regulation in plants. *Mol. Plant* 12, 743–763
5. Ibarra, I.L. *et al.* (2020) Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. *Nat. Commun.* 11, 124
6. Vaishnav, E.D. *et al.* (2022) The evolution, evolvability and engineering of gene regulatory DNA. *Nature* 603, 455–463
7. Agarwal, V. and Shendure, J. (2020) Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep.* 31, 107663
8. Washburn, J.D. *et al.* (2019) Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proc. Natl. Acad. Sci. U. S. A.* 116, 5542–5549
9. Avsec, Ž. *et al.* (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* 18, 1196–1203
10. Zrimec, J. *et al.* (2021) Learning the regulatory code of gene expression. *Front. Mol. Biosci.* 8, 673363
11. Srivastava, A.K. *et al.* (2018) UTR-dependent control of gene expression in plants. *Trends Plant Sci.* 23, 248–259
12. Sielemann, J. *et al.* (2021) Local DNA shape is a general principle of transcription factor binding specificity in *Arabidopsis thaliana*. *Nat. Commun.* 12, 6549