Editorial Commentary

Setting patterns and predicting: the role of artificial intelligence in synthetic and natural promoter screening

Ananda Sanches-Medeiros[#], Leonardo Martins-Santana[#], Rafael Silva-Rocha

Systems and Synthetic Biology Lab, FMRP - University of São Paulo, Ribeirão Preto, SP, Brazil

[#]These authors contributed equally to this work.

Correspondence to: Rafael Silva-Rocha. Ribeirao Preto Medical School, University of Sao Paulo, Av. Bandeirantes, 3.900, CEP: 14049-900, Ribeirão Preto, São Paulo, Brazil. Email: silvarochar@usp.br.

Comment on: Wu MR, Nissim L, Stupp D, *et al*. A high-throughput screening and computation platform for identifying synthetic promoters with enhanced cell-state specificity (SPECS). Nat Commun 2019;10:2880.

Received: 16 October 2019; Accepted: 30 October 2019; Published: 20 December 2019.

doi: 10.21037/jmai.2019.11.01

View this article at: http://dx.doi.org/10.21037/jmai.2019.11.01

The emergence of the 5th Industrial Revolution, also referred as Synthetic Biology Revolution, arose because of the necessity of generating biological parts to be used for biological circuits design and building (1). The purpose of this new era was established on a hinged possibility of engineering life similarly to computational circuits. From this moment on, there were a growing and constant search for new parts that enable new organism manipulation, metabolism pathway design, gene expression based on cell-type specificity and many other biological processes which depend on cellular behavior.

One of the most relevant parts for biological engineering in the biotechnological or biomedical fields are *cis*-regulatory elements, like promoters. Promoters are sequences of DNA responsible for driving the transcription of a new messenger RNA upstream to a gene-coding region. These sequences are very suitable for biological engineering because they can respond to distinct cellular stimuli, cell type and state, moment or spatiotemporal conditions for the circuit actuation. However, it is not usually simple to find new biological regulators that attend to our necessities, not only because of our small knowledge about cell functionality and screening difficulties, but also due to emergent properties of genetic parts (like promoter complexity) (2).

In order to circumvent such bottlenecks, currently it is possible to associate experimental and computational approaches to design genetic parts in a faster and more accurate manner. This is particularly feasible because of 4th Industrial Revolution, which brought us the possibility of using artificial intelligence (AI) as a source to better understand biological patterns, often generated from big data analysis, then allowing biological modeling predictions to supply biotechnological and biomedical parts demand (3). In the biotechnological field, the search for parts is mainly focused on the engineering of unusual strains as well as on parts that respond or confer resistance to specific environments, aiming at bioprocesses improvement. Recent studies used machine learning associated with experimental validation to characterize biotechnological microbial promoters of interest (4,5). These approaches resulted from computational progress rely on the improvement of *in silico* parameters, which culminates in promoter accuracy, better capacity to analyze big data and to identify patterns, as well as allows for predicting new putative promoters from a synthetic library or even genomes.

Applications of computational modeling for promoter identification is already a target of research in some organisms, mainly for those easily grown and already genetically characterized. In this sense, Liu *et al*. (4) integrated machine learning approaches to try to elucidate how the same transcription factor (TF) could confer different levels of gene expression in different operons in *Escherichia coli*. In order to understand the transcriptional dynamics responsible for this system control, the authors performed mutations on known operators for some repressor TFs to observe a range in gene expression levels, which resulted in a mutated operators library used as input to train the

algorithm. In a similar study, Gilman *et al*. (5) also applied machine learning to find putative promoters in a bacterial thermoresistant strain (*Geobacillus thermoglucosidasius*). In their study, authors analyzed genomes strains from the same genus and screened for putative promoters, selecting some of them to further *in vivo* characterization. The results generated fed a machine learning approach to better understand the pattern of promoter structure in these strains. In that case, it was not possible to predict new promoters based on the machine learning approach because the number of promoters used as input data was small, but an increase in high-throughput experiments could provide enough data for promoter prediction.

Although machine learning is a promising strategy to identify new biological parts for biological engineering, much of the efforts performed in this field are still far to be completely extended to higher organism classes, such as mammalian ones. The design of synthetic promoters for mammalian cells is a field that claims for special attention since transcriptional dynamic in these organisms requires a coordinated TFs arrangement and it is under a complex network for differentiation and cell state specificity conditions. It is worth mentioning that beyond the boundaries of gene expression, the study of promoter specificity has constantly been a target of research as metabolic and physiological regulation programs may switch depending on the health-and-disease cellular requirements (6).

In the biomedical field, there is much interest in searching promoters that are cell type or cell state-specific, like promoters that are activated only in cancer cells (7) as well as diabetic conditions-responsive promoters (8). The major focus of using this kind of regulatory sequences relies on the creation of synthetic circuits capable of diagnosis and/or the release of an effector molecule under specific disease condition (9). These features make cell-state specific responsive-promoter suitable candidates for medical research, since they can corroborate for the generation of more accurate responses as well as result in discriminative environments for cancer detection.

In order to make progress in this field, recent studies are investing in design and screening synthetic promoters allying AI approaches towards specific interests (9,10). This combination promotes a better understanding of promoter patterns, then allowing possible predictions of promoter functionality. In summary, such approaches consist of three steps. The first is the creation of a synthetic promoter library combining sequences of transcriptional factor binding sites (TFBS) associated with a known core promoter driving the expression of a reporter gene (fluorescent protein). The second is based on the screening of this library, a process mediated by fluorescence-activated cell-sorting (FACS), which consists in discriminating cells by fluorescence intensity. Finally, the third and last step is resembling the sequencing of the different generated bins to classify the promoters by activity levels. These results are processed and employed as input to machine learning. Posteriorly, the results generated after analysis could be used to identify patterns in promoter sequences as well as to correlate gene expression induction, promoter structure and cell-state, which could enable promoter predictions in other promoter libraries for another specific condition. This strategy can be used to identify eukaryotic promoters with biomedical importance or even to find microorganisms promoters of biotechnological interest (*Figure 1*).

In this sense, the recent article published by Wu *et al*. (12) reports an interesting approach to search for transcription-driven regulatory sequences regarding cell state-specificity. For this, the authors have developed a machine-learning computational method based on a library of synthetic promoters with enhanced cell state specificity (SPECS) for distinct purposes. In their study, the authors engineered a promoter library based on tandem repetitions of the same TFBS preceding a lentivirus core-promoter. Initially, TFBS were extracted from 6,107 TFBS reported at two databases. Then, this promoter library was screened by FACS in healthy breast cells and cancer breast cells, as well as in bulk glioblastoma cells and glioblastoma stem-like cells. The authors also tested the library for organoid differentiation cells, integrating FACS and confocal microscopy approaches. Eighty-one promoters tested in breast-cancer and non-breast-cancer cells obtained from experimental data were used to feed machine learning algorithms, which in turn were used to predict gene expression patterns from other 54 promoters. These promoters were also tested *in vivo*, and the results were used as input to feed the machine learning algorithm to improve its accuracy. The resultant model was used to predict the activity of promoters from the entire library.

We highlight here as one of the most relevant outcomes provided by this work the possibility of finding regulatory sequences which could oscillate in their activity strength whereas authors were also able to identify promoters with specific transcriptional activity regarding cell state. In this sense, the searching and the discovery of new regulatory sequences using AI is a reasonable strategy to explore the

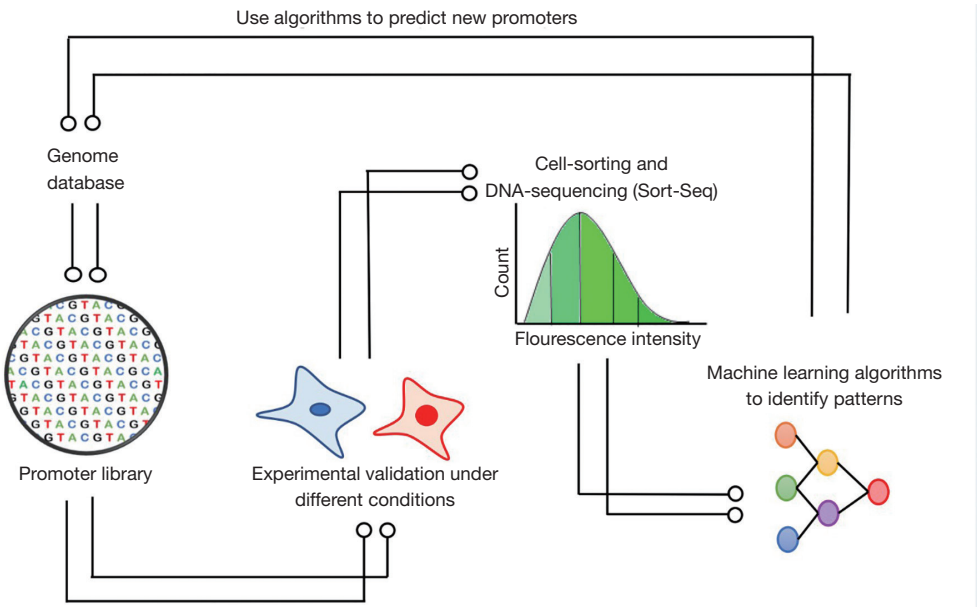

Figure 1 Schematic representation of the new approaches for screening synthetic and natural promoters using Sort-Seq and machine learning. The promoters are extracted from promoter, genomic, mutated promoters or synthetic promoter databases, to create a large library of promoters. Then, they are screened in different conditions using FACS, separating cells by fluorescence intensity in bins and each bin pool is sequenced. The screening output is analyzed, correlating promoter with fluorescence intensity. These results serve as *input* to machine learning patterns and new promoter predictions from other libraries. FACS, fluorescence-activated cell-sorting.

bottlenecks of transcriptional regulation in cells regardless of its state specificity. First, because we can search for regulatory sequences of biological interest to try to elucidate cellular mechanisms, e.g., disease conditions; secondly, using AI, a huge set of input signals could be given for machine training, and, lastly, the selection of outputs (regulatory sequences) could be performed through filtering according to desired specific conditions for further validation.

Notwithstanding, Wu *et al*. (12) showed that their approach is also a powerful way to improve the outputs concerning how cells reprogramming transcriptional networks are coordinated to ensure transcription in spite of the changing in metabolic and physiological scenarios. To validate this, the initial steps of the study consisted of the identification of responsive specific cell-state promoters in organoids under cellular differentiation processes. Since the approach could distinguish responses between cell lines, the authors constructed a library of synthetic promoters to be evaluated according to transcriptional response levels after transfection in breast cancer cells lines. The identification of promoter strength and specificity was performed using the authors' platform and resulted in the identification of sequences that were subsequently used as training inputs for the construction of a machine learning method to search for relevant regulatory promoter sequences based on the cell-state specificity studies.

Moreover, the authors sought to investigate whether their approach was able to efficiently select promoter sequences in glioblastoma cells. This is especially relevant because this is an aggressive type of brain tumor which remains resistant to therapies, making this validation choice a fair manner to demonstrate the potential of AI on medical field. For instance, this platform was able to predictively select regulatory sequences present in these glioblastoma cells as well as it could distinguish the promoters showing high levels of activity between the cell lines used by authors to validate cell-state specific condition. In the face of the present results found by Wu *et al*., we highlight the potential and relevance of coupling computational models to biomedical applications to predict cellular, physiological and/or metabolic behavior in scenarios regarding cellular differentiation.

The contributions provided by Wu *et al*. are a source of new perspectives for medical applications. However, some endeavors must be considered to improve this approach and make it reach a useful large-scale tool in a

non-distant future. We first summarize the uniqueness of specific TF binding sites in the promoter library, restricting transcriptional dynamics to only one protein. As regulator entities, TFs can interact with cellular transcriptional machinery promoting DNA bending and subsequent transcription. This is the basis for transcription initiation, and it is important to consider the occurrence of more than a simple protein binding site in DNA, mimicking what naturally happens in eukaryotic DNA. It would be exceedingly interesting to notice the same phenomena using synthetic promoters based on cell-state specificity composed of two or more TF binding sites.

As discussed above, many efforts have been performed to compile data and train computational approaches to generate a machine-learning system capable to predict promoter behavior after chasing validation steps. In this sense, the study of Wu *et al*. used a glioblastoma-based detection system to identify promoters and after rounds of artificial intelligence only a few of them indeed presented higher activities capable to make two glioblastoma cell lines are distinguishable from each other. Despite the low rates of activities in comparison to the predictive analysis, the methodology of sorting glioblastoma cells can itself promote a decrease in cell viability, then corroborating to a deficit on efficiency in transcriptional reporter detection.

Finally, machine-learning-based predictions are relative and its successful employment is a suitable alternative to assume biological behaviors in a context cell-state specificity. The tests performed in the study of Wu *et al*. trained the algorithm to respond to a non-high number of inputs (promoter sequences), which could be improved to a wide range of sequences to amplify the obtained response pattern after validation performance using not only viable cells but also short-live clinical samples, making the merit of this promising approach still more elegant. In summary, the study published by Wu *et al*. describe an encouraging approach with so many perspectives as the multifaceted possibilities for studying regulation based on cell-state specificity. As a candidate for future applications, we highlight the potential of this method to be used in the design of synthetic biology-based circuits for the studying of many other diseases evolving cellular reprogramming as well as metabolic flux rewiring, shedding light on cancer, biomedical and biotechnological research.

Acknowledgments

Funding: The authors are thankful to lab colleagues for insightful discussion about this manuscript and FAPESP for financial support (Young Research Awards by the Sao Paulo State Foundation FAPESP, award number 2012/21922-8, and FAPESP PhD fellowships, 2018/04810-0 and 2017/17924-1).

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Journal of Medical Artificial Intelligence*. The article did not undergo external peer review.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/jmai.2019.11.01). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Di Mauro G, Dondi A, Giangreco G, et al. ENABLE 2017, the First European PhD and Post-Doc Symposium. Session 2: The OMICS Revolution. Biomolecules 2018;8. doi: 10.3390/biom8040116.
2. Barnard A, Wolfe A, Busby S. Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. Curr Opin Microbiol 2004;7:102-8.
3. Alipanahi B, Delong A, Weirauch MT, et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 2015;33:831-8.
4. Liu X, Gupta STP, Bhimsaria D, et al. De novo design of programmable inducible promoters. Nucleic Acids Res 2019;47:10452-63.

5. Gilman J, Singleton C, Tennant RK, et al. Rapid, Heuristic Discovery and Design of Promoter Collections in Non-Model Microbes for Industrial Applications. ACS Synth Biol 2019;8:1175-86.
6. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell 2013;152:1237-51.
7. Nissim L, Bar-Ziv RH. A tunable dual-promoter integrator for targeting of cancer cells. Mol Syst Biol 2010;6:444.
8. Xie M, Ye H, Wang H, et al. β-cell-mimetic designer cells provide closed-loop glycemic control. Science 2016;354:1296-301.
9. Sedlmayer F, Aubel D, Fussenegger M. Synthetic gene circuits for the detection, elimination and prevention of disease. Nat Biomed Eng 2018;2:399-415.
10. Cheng JK, Alper HS. Transcriptomics-Guided Design of Synthetic Promoters for a Mammalian System. ACS Synth Biol 2016;5:1455-65.
11. Mamoshina P, Vieira A, Putin E, et al. Applications of Deep Learning in Biomedicine. Mol Pharm 2016;13:1445-54.
12. Wu MR, Nissim L, Stupp D, et al. A high-throughput screening and computation platform for identifying synthetic promoters with enhanced cell-state specificity (SPECS). Nat Commun 2019;10:2880.

doi: 10.21037/jmai.2019.11.01
Cite this article as: Sanches-Medeiros A, Martins-Santana L, Silva-Rocha R. Setting patterns and predicting: the role of artificial intelligence in synthetic and natural promoter screening. J Med Artif Intell 2019;2:25.