Perspective: the opportunities and possibilities unleashed by clustered regularly interspaced short palindromic repeats and artificial intelligence

Simeng Yin[1,2*], Juehua Yu[3*], Ziwei Li[2,3], Haiqiu Huang[2,3]

[1]The Middle School Affiliated to Beijing Jiaotong University, Beijing 100081, China; [2]Shanghai Biotecan Pharmaceuticals Co., Ltd., Shanghai 201204, China; [3]Cancer Research Institute, Hangzhou Cancer Hospital, Hangzhou 320000, China

Contributions: (I) Conception and design: All authors; (II) Administrative support: None; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: S Yin, J Yu; (V) Data analysis and interpretation: S Yin, J Yu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*These authors contributed equally to this work.

Correspondence to: Ziwei Li, PhD. Shanghai Biotecan Pharmaceuticals Co. Ltd., First Shanghai Center, 180 Zhangheng Rd., Pudong New District, Shanghai 201204, China. Email: zwli@biotecan.com; Haiqiu Huang, PhD. Cancer Research Institute, Hangzhou Cancer Hospital, 34 Yanguan Ln, Shangcheng District, Hangzhou 320000, China. Email: tennisqiu@gmail.com.

Abstract: Clustered regularly interspaced short palindromic repeats (CRISPR) was discovered in the 1980s in *E. coli* and its function was elucidated in 2007 in *S. thermophilus*. Coupled with the CRISPR-associated protein nuclease (Cas), CRISPR/Cas forms a defense system enabling the organisms to respond to and eliminate invading exogenous organisms and foreign nucleic acids. The discovery and development of CRISPR/CRISPR-associated protein-9 nuclease (Cas9) system offered superior precision and simplicity in genome editing, which can greatly benefit and has enormous potential in both biological and medical science research. Artificial intelligence (AI) was conceptualized in the 1950s and the recent resurgence of machine learning research assisted a rapid advancement in AI capability and functionality. Combining these two technologies, equipped with the machine learning and pattern recognition capability, AI has the potential to lift the biological and medical science research to a new level by (I) improving CRISPR efficacy and precision with AI-assisted analysis and design of targets; (II) performing more efficient analysis of molecular pathways and deepening understanding of their interactions through pattern recognition, and (III) incorporating CRISPR and AI to better understand multifactorial disorders, identify critical mutations, and design therapeutic targets.

Keywords: Clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein nuclease (Cas); artificial intelligence (AI)

Received: 28 November 2017; Accepted: 10 December 2017; Published: 04 January 2018.
doi: 10.21037/amj.2017.12.05
View this article at: http://dx.doi.org/10.21037/amj.2017.12.05

It does not happen very often that articles published in peer-reviewed journals are picked up by the mass media so quickly that one could be reading the headlines side by side. It also does not happen very often that almost everyone agrees that a technology is so fundamental and transcending that the major concerns about the technology are not the feasibility or efficacy but the ethics and implications. This is exactly what is happening now right in front of us, not in one but two occasions, namely, CRISPR/Cas9 and AI. Individually, CRISPR/Cas9 is promised to be the master gene-editing tool that makes everything between correcting and designing somatic or germinal mutation possible; and the eventual destination of AI is expected to be the omnipotent solution for all the problems facing the material world. Either of these promises may still be years or decades away, nevertheless, biological and medical science may benefit greatly from CRISPR/Cas9 and AI technology in very near future, if not already. In this article, the state of CRISPR/Cas9 and AI research in the field of biological and medical science will be reviewed, and the potential directions and new opportunities will be discussed.

What is clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein-9 nuclease (Cas9)

CRISPR is the abbreviation of clustered regularly interspaced short palindromic repeats. CRISPR was discovered in the 1980s in *E. coli* and later found in 84% of sequenced archaeal genomes and approximately 45% of bacterial genomes (1,2). Coupled with the CRISPR-associated protein nuclease (Cas), CRISPR/Cas forms a defense system enabling the organisms to respond to and eliminate invading viruses, plasmids, and other foreign nucleic acids. However, the function and mechanism of action of CRISPR weren't reported until 2007, in which *S. thermophilus* was shown to acquire resistance against a bacteriophage by integrating a fragment of the genome of an infectious virus into its CRISPR locus (3). Since then, comparative genomic analysis of bacterial and archaeal genomes had identified more than 45 cas gene families (4,5). Currently, CRISPR-Cas systems were categorized into three major types (type I, II, and III) and ten different subtypes, among which type II is the best understood. The type II CRISPR-Cas system cut invading DNA from viruses or plasmids into small fragments, which are incorporated into a CRISPR locus amidst a series of short repeats (~20 bps). The transcripts of these loci are then processed to generate CRISPR RNAs (crRNAs), which are incorporated into effector complexes, where the crRNA guides the complex to the invading nucleic acids and the Cas proteins degrade the nucleic acids (6,7). A unique feature, as well as a critical advantage, of the type II CRISPR system is that only one Cas protein (Cas9) is required for gene silencing, in which Cas9 participates in both the processing of crRNAs and the destruction of the target DNA (6,8).

The human genome project has sequenced and published over 99% of the euchromatic human genome, which greatly elevated our understanding of many genetic diseases, including mutations linked to different forms of cancer (9). The hype and promises of genome editing to correct genetic mutation was hampered by a number of technical difficulties and unsatisfactory efficacy of gene editing tools, such as transcription activator-like effector nucleases (TALENs) and zinc-finger nucleases (ZFNs). The discovery and development of CRISPR/Cas9 system offered superior precision and simplicity in genome editing. The adaptation of CRISPR/Cas9 for genome editing involves only three key components: Cas9 and the corresponding crRNA and trRNA and was first reported in 2012 by the Doudna and Charpentier labs (6), who further simplified the system to two components by combining crRNA and trRNA into a single synthetic single guide RNA (sgRNA). Currently, three different variants of the Cas9 nuclease have been adopted in genome-editing protocols with different specificities and enzymatic activities (10-16).

What is artificial intelligence (AI)

The concept of AI is generally considered to date back to the 1950s and initially described by Turing (17) and the term AI was coined in 1955 by McCarthy, a math professor at Dartmouth (18). Narrowly defined, AI is the science and engineering of making intelligent machines, especially intelligent computer programs. A more expansive definition includes using computers to understand human intelligence and mimic human-like "cognitive" functions, such as "learning" and "problem-solving" (19). The present stage of AI is considered as narrow AI (or weak AI), which can perform narrow or defined tasks (such as math calculation, web search, or self-driving). Looking forward, the majority of researchers agree that general AI (AGI or strong AI) can be expected, but the timeline is heavily disputed.

Recent improvements in computing power have brought about a resurgence of machine learning research, which is fueled by breakthroughs in deep learning algorithms (20). From voice assistants on your phone (such as Siri and Google assistant) to self-driving cars to AI-assisted cardiac imaging and diagnosis, deep learning powered machine learning technology is progressing rapidly in all fronts and making achieving better AI a possibility.

The most significant advances in AI over the past decade are in two key areas: perception and cognition, in other words, machine learned to see (or hear) and understand. Image and voice recognition, with their dramatical improvement assisted by big data and deep learning, can now achieve human-level recognition, if not better. The second part of the major improvement is in solving the problem using existing data or data generated through the learning process with or without human supervision. Machine learning has generated chess and Go algorithms that can beat human champions (Beep Blue and AlphaGo), as well as more practical applications such as improved malware detection for cybersecurity (Deep Instinct) and the cooling efficiency at Google's data centers.

Equipped with improved perceptive and cognitive functions, AI is not only being implemented in tasks already delegated to machines and computers but are now working its way into many tasks that were once done best and only by humans. It may be years or decades before an AI system with 100% accuracy can be achieved, but once the human-level performance is reached, such as less than 5% error rate, wide implementation of AI can vastly expand human capabilities. For example, Enlitic, a medical deep learning company, is training AI to scan medical images to help diagnose cancer; and Verily, a subsidiary of Alphabet's life sciences research organization, is developing algorithms to detect causes of diabetics associated blindness.

Combining CRISPR with AI

The opportunities and possibilities afforded by CRISPR and AI are potentially limitless, and the combination of the two technologies can greatly advance biological and medical science and bring about drugs and treatments that are previously out of the question.

- ❖ AI-assisted improvement of CRISPR efficacy and precision. Though CRISPR has been shown to be superior to its predecessor technologies (TALENs or ZFNs) in many aspects, including precision, efficacy, and cost, improvements are still needed in targeting efficiency and off-target mutation for CRISPR to be successful in medical and therapeutic applications. Currently, CRISPR/Cas9 can achieve 70% efficiency in zebrafish and plants, and up to 78% efficiency in one-cell mouse embryos (21-23). Off-target mutations tend to occur in sites that have highly similar sequences with differences in only a few nucleotides, because Cas9 can tolerate up to 5 base mismatches within the protospacer region or a single base difference in the PAM sequence (24). A 2017 study reported widespread off-target mutations in vivo using CRISPR/Cas9 (25), while this particular study is debated, the risk and possibility of off-target mutations are far from uncertain. Computer programs developed to help identify potential CRISPR target sites and assess the off-target cleavage potential are already available, such as CRISPR design tool (26). A more powerful tool can be designed by employing the supervised approach of deep learning so that the design and off-target detection algorithm can be improved after each iteration. CRISPR design and the corresponding experimental data can be fed back to the system, preferably with off-target mutations annotated, with which the design system can analyze and detect the pattern of mismatch. Such training requires a great amount of data and annotations, which can be pulled from published work in journals and crowd-sourced from researchers around the world.
- ❖ AI- and CRISPR-assisted analysis of molecular pathways. The past decade witnessed the vast expansion of sequencing capacity and a precipitous decrease of cost compared to that of the Human Genome Project. A detailed map of genes and disease-related mutations is being put together piece by piece. Using such information, the current generation of molecular pathway analysis tools is developed and largely curated by human scientists, such as MetaCore by Thomson Reuters and Ingenuity Pathway Analysis by QIAGEN Bioinformatics. Molecular pathway analysis involves a large amount of pattern recognition and analysis to reveal the correlations and causal relations between genes. Machine learning algorithm is particularly apt in pattern recognition, which is essentially the mathematical basis of the deep learning algorithm. As has been shown in the case of AlphaGo and AlphaGo Zero (27), machine learning, supervised or unsupervised, can achieve better than the human level of recognition in the complex game of Go. If introduced to analyze molecular pathways, similar superior results may be achieved. Machine learning will need an exceedingly large amount of data, which can be accumulated with relatively high precision and low cost with CRISPR. The ease of inducing point mutations using CRISPR and analyses of resulting changes in expressions of a wide range of genes using high-throughput sequencing can generate the data needed for training the pattern recognition algorithm. In turn, the trained algorithm can identify interesting nods in the pathway for further mutation and analyses. Such iterations can quickly help the machine learning algorithm to move beyond the supervised learning approach, and, as indicated by AlphaGo Zero, the unsupervised

learning may lead to new pathways or targets that have evaded human scientists.

- Multifactorial disorders and beyond. Current development of the CRISPR and other genome editing technologies mainly focus on monogenic disorders. For example, the first clinical trial involving CRISPR technology is aimed to disable the PD-1 gene (28,29) and another trial was designed to remove the viral genes in the HPV infected cells (30,31). Monogenic disorders are relatively better understood and can benefit a great deal from the CRISPR-enabled therapeutic approaches. On the other hand, the majority of human diseases are not stemmed from single gene mutation. To start tackling multifactorial disorders, their mechanisms, including the molecular and environmental mechanisms, need to be better elucidated. If a common and critical mutation can be identified in a cohort and is proven to account for a significant portion of the disease outcome, such disorders can be treated as a pseudo-monogenic disorder, and the corresponding mutation can be selected as the primary target and corrected using CRISPR. Again, the pattern recognition capability of machine learning can be helpful in identifying the critical mutations. Isolated, such analyses of critical mutation may be difficult; but when fuzzy data can be pooled, and *in vitro* analyses and animal models can be employed, instead of being intimidated by the amount of data, machine learning can benefit from it.

Conclusions

AI and CRISPR are both widely believed to be the transcending technologies of our time, and yet both are in their infancy. Thanks to the expansion of computational power and sequencing capability, machine learning and CRISPR can be employed with relatively low cost. Taken together, the decades of biomedical data archive and the ever-increasing ability to generate new date will make the combination of CRISPR and AI a useful tool and intriguing guide in researching the mechanisms of various diseases and developing new therapeutic treatments.

Acknowledgements

Funding: None.

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/amj.2017.12.05). The authors have no conflicts of interest 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. Ishino Y, Shinagawa H, Makino K, et al. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. J Bacteriol 1987;169:5429-33.
2. Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 2007;35:W52-7.
3. Barrangou R, Fremaux C, Deveau H, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science 2007;315:1709-12.
4. Haft DH, Selengut J, Mongodin EF, et al. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol 2005;1:e60.
5. Swarts DC, Mosterd C, Van Passel MW, et al. CRISPR interference directs strand specific spacer acquisition. PloS one 2012;7:e35888.
6. Jinek M, Chylinski K, Fonfara I, et al. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science 2012;337:816-21.
7. Terns MP, Terns RM. CRISPR-based adaptive immune systems. Current opinion in microbiology 2011;14:321-7.
8. Deltcheva E, Chylinski K, Sharma CM, et al. CRISPR

RNA maturation by trans-encoded small RNA and host factor RNase III. Nature 2011;471:602-7.

9. Cavalli-Sforza LL. The human genome diversity project: past, present and future. Nature Reviews Genetics 2005;6:333-40.
10. Gong C, Bongiorno P, Martins A, et al. Mechanism of nonhomologous end-joining in mycobacteria: a low-fidelity repair system driven by Ku, ligase D and ligase C. Nat Struct Mol Biol 2005;12:304-12.
11. Gasiunas G, Barrangou R, Horvath P, et al. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A 2012;109:E2579-86.
12. Chen B, Gilbert LA, Cimini BA, et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell 2013;155:1479-91.
13. Cong L, Ran FA, Cox D, et al. Multiplex genome engineering using CRISPR/Cas systems. Science 2013;339:819-23.
14. Gilbert LA, Larson MH, Morsut L, et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. cell 2013;154:442-51.
15. Davis L, Maizels N. Homology-directed repair of DNA nicks via pathways distinct from canonical double-strand break repair. Proc Natl Acad Sci U S A 2014;111:E924-32.
16. Cong L, Zhang F, et al. Genome engineering using the CRISPR-Cas9 system. Methods Mol Biol 2015;1239:197-217.
17. Turing AM. Computing machinery and intelligence. Mind 1950;49:433-60.
18. McCarthy J. Artificial intelligence, logic and formalizing common sense. Springer Netherlands 1989:161-90.
19. Nilsson NJ. Principles of artificial intelligence. San Francisco: Morgan Kaufmann; 2014.
20. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
21. Feng Z, Zhang B, Ding W, et al. Efficient genome editing in plants using a CRISPR/Cas system. Cell Res 2013;23:1229-32.
22. Hwang WY, Fu Y, Reyon D, et al. Heritable and precise zebrafish genome editing using a CRISPR-Cas system. PloS One 2013;8:e68708.
23. Zhou J, Wang J, Shen B, et al. Dual sgRNAs facilitate CRISPR/Cas9-mediated mouse genome targeting. FEBS J 2014;281:1717-25.
24. Fu Y. Foden JA, Khayter C, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol 2013;31:822-6.
25. Schaefer KA, Wu WH, Colgan DF, et al. Unexpected mutations after CRISPR-Cas9 editing in vivo. Nature Methods 2017;14:547-8.
26. Hsu PD, Scott DA, Weinstein JA, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol 2013;31:827-32.
27. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge. Nature 2017;550:354-9.
28. Cyranoski D. Chinese scientists to pioneer first human CRISPR trial. Nature 2016;535:476.
29. Su S, Hu B, Shao J, et al. CRISPR-Cas9 mediated efficient PD-1 disruption on human primary T cells from cancer patients. Sci Rep 2016;6:20070.
30. Zhen S, Hua L, Takahashi Y, et al. In vitro and in vivo growth suppression of human papillomavirus 16-positive cervical cancer cells by CRISPR/Cas9. Biochem Biophys Res Commun 2014;450:1422-6.
31. Page LM. Boom in gene-editing clinical trials. NewScientist 2017;234:6.

doi: 10.21037/amj.2017.12.05

Cite this article as: Yin S, Yu J, Li Z, Huang H. Perspective: the opportunities and possibilities unleashed by clustered regularly interspaced short palindromic repeats and artificial intelligence. AME Med J 2018;3:4.