broadening and shifting of the spectral lines.

A very narrow spectral line was observed at a temperature just below the phase-transition temperature to superfluid helium. This line was four times narrower than that observed for the inner-shell excitation of dysprosium[8]. It was even narrow enough to reveal a splitting of the atoms' energy levels, known as hyperfine splitting, which arises as a result of interactions between the electron and the antiproton. The measured resolution (one part per million of the transition frequency) is remarkable in liquid helium.

If this resolution can be matched for other exotic atoms, it might be possible to test theories proposing that dark matter decays or is annihilated in the Milky Way[9]. Current estimates[5] of the spectral linewidth of pionic helium suggest that it is up to 100 times that of the antiprotonic helium measured by Sótér and colleagues. However, if a transition with a similarly narrow linewidth also exists in pionic helium immersed in liquid helium, it might be a good candidate for such investigations. Finding the optimal combinations of energy levels and liquid-helium conditions will undoubtedly require guidance from theoretical calculations.

The fact that the linewidth measured by Sótér *et al.* narrowed suddenly at a temperature close to that at which normal fluid helium transitions to superfluid helium, and broadened again at a lower temperature, is intriguing from the viewpoint of chemical physics. Although not discussed in depth by the authors, the reason for this seems to be unrelated to the phase transition itself, but instead to be linked to the characteristics of bulk helium. More research is needed to reveal the relevance of this temperature, its relationship to the properties of helium, and the physics behind this connection.

Yukari Matsuo is in the Department of Advanced Sciences, Hosei University, Tokyo 184-8584, Japan.
e-mail: yukari.matsuo@hosei.ac.jp

1. Bothwell, T. *et al. Metrologia* **56**, 065004 (2019).
2. Sótér, A. *et al. Nature* **603**, 411–415 (2022).
3. Hori, M. *et al. Science* **354**, 610–614 (2016).
4. Ulmer, S. *et al. Nature* **524**, 196–199 (2015).
5. Hori, M., Aghai-Khozani, H., Sótér, A., Dax, A. & Barna, D. *Nature* **581**, 37–41 (2020).
6. Hori, M. *et al. Phys. Rev. Lett.* **87**, 093401 (2001).
7. Toennies, J. P. & Vilesov. A. F. *Annu. Rev. Phys. Chem.* **49**, 1–41 (1998).
8. Moroshkin, P., Borel, A. & Kono, K. *Phys. Rev. B* **97**, 094504 (2018).
9. Cuoco, A., Krämer, M. & Korsmeier, M. *Phys. Rev. Lett.* **118**, 191102 (2017).

The author declares no competing interests.

Molecular biology

An oracle for gene regulation

Andreas Wagner

A long-standing goal of biology is the ability to predict gene expression from DNA sequence. A type of artificial intelligence known as a neural network, combined with high-throughput experiments, now brings this goal a step closer. **See p.455**

Gene expression affects every aspect of life, from the survival of bacteria in specific environments to the anatomy and physiology of the human body. The ability to accurately predict how strongly a gene is expressed on the basis of the DNA sequences that regulate such expression would transform how researchers study biology. But the biochemical machinery that regulates gene expression is tremendously complex, and this goal has eluded biologists' best efforts for more than 50 years. On page 455, Vaishnav *et al.*[1] take advantage of two key technologies to produce a successful 'oracle' for gene expression in the yeast *Saccharomyces cerevisiae*.

The first technology used by the authors is a means of measuring the expression of a gene that encodes yellow fluorescent protein (YFP) in every cell of a large population of yeast cells[2]. In this population, different cells carry different regulatory DNA sequences, called promoters, that are located close to the *yfp* gene on a small piece of circular DNA — their proximity to *yfp* enables them to drive the gene's expression. Specifically, the authors used a collection of more than 30 million different promoters, each 80 base pairs long, and quantified the production of YFP by each cell containing one of these promoters.

Vaishnav *et al.* fed the resulting expression data into the second technology, an artificial intelligence (AI) called a convolutional neural network, and trained the network to predict gene expression from the data. They then validated the network's ability to predict gene expression on an impressive scale (Fig. 1).

For example, the authors synthesized thousands more promoter sequences not used for training, measured their ability to drive gene expression, and showed that the neural network very accurately predicts how well each will drive gene expression. In addition, the authors presented the network with random starting sequences, and showed that its ability to predict gene expression from sequence could be used to transform these starting sequences, through ten rounds of computer-simulated evolution, into promoter sequences predicted to drive extreme (very high or very low) YFP expression. The group then synthesized 500 of these sequences and measured their ability to drive YFP expression.

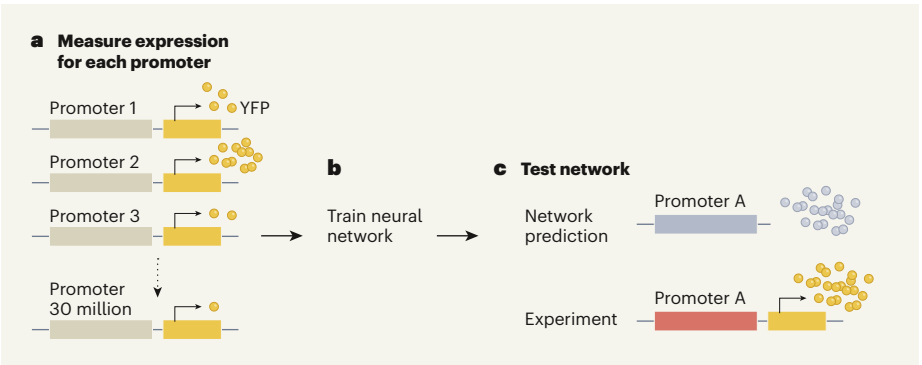

Figure 1 | Learning to predict gene expression. a, Vaishnav *et al.*[1] created a library of 30 million promoters — 80-base-pair-long DNA sequences that drive gene expression. They measured how well each could drive expression of the gene that encodes yellow fluorescent protein (YFP) in yeast cells. **b**, The group used these data to train a neural network to predict how well different promoter sequences drive gene expression. **c**, The authors then tested network's predictive ability. The group designed thousands more promoters (only one is shown here, for simplicity), and showed that the network could predict, extremely accurately, how well each promoter would drive gene expression.

The computer-simulated sequences could indeed drive very high and very low expression. This and other validation experiments showed that Vaishnav *et al.* have created a highly effective oracle to predict gene expression.

This oracle can also help to elucidate various aspects of the evolution of gene expression. For example, the authors predict computationally, and validate experimentally, that, for most starting sequences, three or four mutations are sufficient for sequences to evolve that have very high or very low expression. They also show that some 70% of yeast genes are subject to stabilizing selection on their expression (selection that favours mutations that do not cause large changes in expression). In addition, they show that genes subject to stabilizing selection have become more resistant to regulatory-DNA mutations. That is, mutations in their promoters alter gene expression to a lesser extent.

This work is important for several reasons. First, it can help to design genes that have specific expression levels. Second, it can help to clarify many aspects of the evolution of gene regulation. And, notably, like other applications of deep learning used in the past few years in biology, such as the development of a tool to predict protein folding[3], it will enable scientists to answer a broader spectrum of questions than any one group of authors could possibly address.

That said, the oracle has limitations. First, it varies only promoters — just one of several types of sequence that can affect gene expression. It does not take into account the effect of variation in the surrounding DNA, including that in protein-coding regions, which might affect gene expression. Second, it has been developed for yeast, in which gene regulation is much less complex than in humans. For example, yeast regulatory DNA is typically located within a few hundred base pairs of the regulated gene, whereas the regulatory DNA of animals can be located millions of base pairs away. As such, it is not clear whether Vaishnav and colleagues' approach will scale to more-complex gene regulation. A source of cautious optimism is that the approach is highly successful even though the 30 million sequences used for training are a tiny fraction (about 2×10^{-41}) of all 4^{80} possible 80-base-long strings that can be formed with the DNA's four nucleotides. Thus, sparse sampling of sequence space might not be a fatal obstacle for this approach.

"This oracle can also help to elucidate various aspects of the evolution of gene expression."

Finally, like the oracles of mythology, this model predicts but does not explain. It does not tell us why a promoter has high or low expression, which transcription factors bind at the promoter, or how they interact. In other words, it does little to elucidate the regulatory logic of gene expression. Overcoming this limitation requires much more work[2,4,5]. However, given the long-standing recalcitrance of the problem, it does not take an oracle to see that biologists will welcome even the ability to predict gene expression.

Andreas Wagner is in the Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich CH-8057, Switzerland, and at the Stellenbosch Institute for Advanced Study, Stellenbosch University, Stellenbosch, South Africa.
e-mail: andreas.wagner@ieu.uzh.ch

1. Vaishnav, E. D. *et al. Nature* **603**, 455–463 (2022).
2. de Boer, C. G. *et al. Nature Biotechnol.* **38**, 56–65 (2020).
3. Jumper, J. *et al. Nature* **596**, 583–589 (2021).
4. Zhou, J. & Troyanskaya, O. G. *Nature Methods* **12**, 931–934 (2015).
5. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. *Nature Biotechnol.* **33**, 831–838 (2015).

The author declares no competing interests.
This article was published online on 9 March 2022.