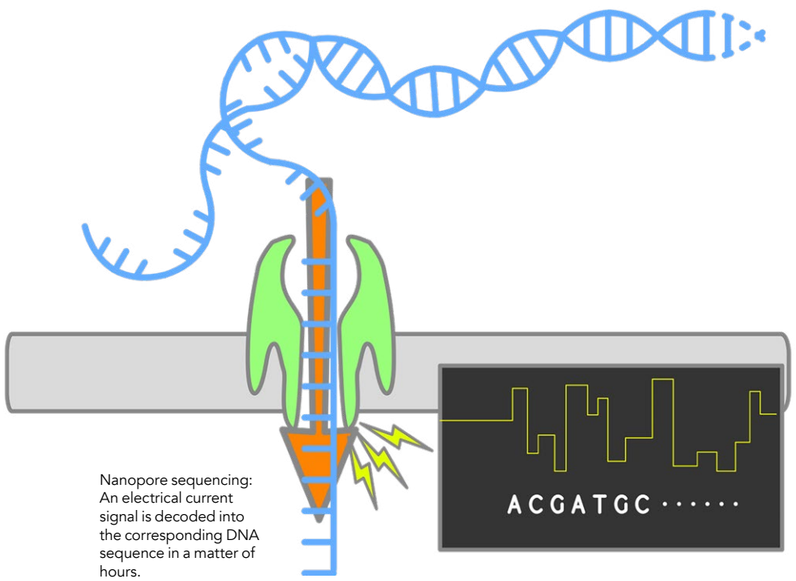
Machine learning paves the way to advances in genome sequencing

Genome sequencing platforms are transforming the field of genetic disease research as they offer a closer look at human genes and DNA for clinical diagnostics. Based on machine learning methods, DeepSimulator provides simulated datasets to train and test sequencing analytical tools while WaveNano has innovated the process of translating a raw signal sequence into a DNA read. Dr Xin Gao is a professor of computer science in CEMSE Division at King Abdullah University of Science and Technology (KAUST), Associate Director of Computational Bioscience Research Center, and Deputy Director of Smart Health Initiative at KAUST. With his team he studies the intersections between computer science and biology/biomedicine to develop genome sequencing methods based on machine learning.



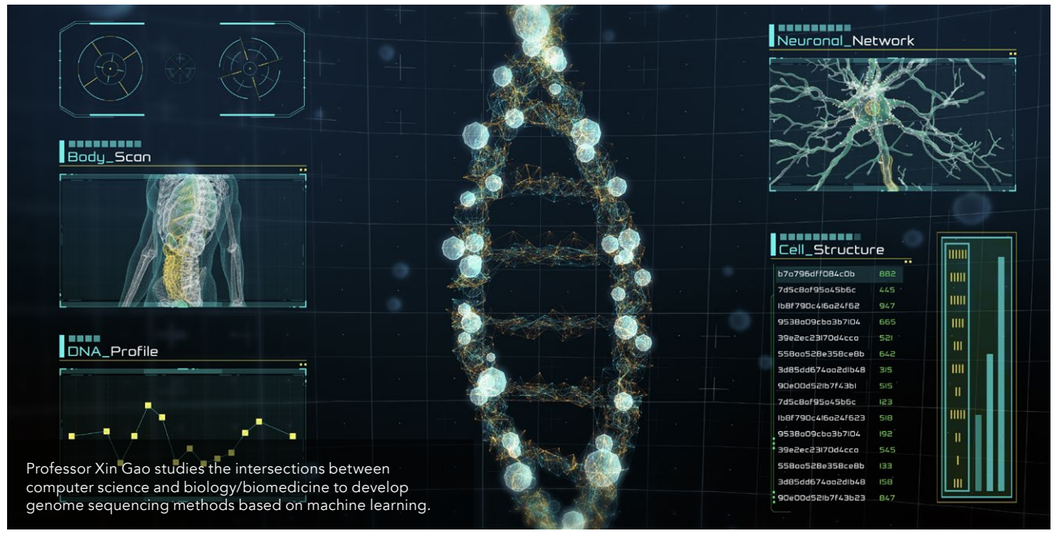
Nanopore sequencing: An electrical current signal is decoded into the corresponding DNA sequence in a matter of hours.

Some diseases such as cystic fibrosis and sickle cell anaemia occur as a result of an error in a single gene classifying them as 'single-gene disorders'. These so-called Mendelian Disorders are passed down through generations which presents an opportunity to researchers: if they can fix the gene error, they will eliminate the disease.

Genome sequencing is the process of determining a DNA sequence of an organism's genes usually with the purpose of identifying mutations, or errors, in the DNA. Over the past decade, the technology has revolutionised diagnostics of genetic diseases such as the Mendelian Disorders as researchers are able to detect genetic changes in a sample with great precision.

A NOVEL SEQUENCING TOOL

One of these genome sequencing tools is the Nanopore third generation sequencing platform, the MinION, which is a portable device the size of an iPhone. Without the need for PCR amplification, the MinION is simple to use and particularly good at reading very long stretches of DNA. The MinION consists of a membrane of wells made of proteins, or nanopores, with an electrical current running through the system. The nanopores are small enough to fit a single strand of nucleic acids (DNA or RNA) disrupting



Professor Xin Gao studies the intersections between computer science and biology/biomedicine to develop genome sequencing methods based on machine learning.

Iokanan VFX Studios/Shutterstock.com

the current as it passes through the well. The two-step process, known as base-calling, takes the electrical current signal and outputs a segmented curve which is then decoded into the corresponding DNA sequence in a matter of hours.

Despite the MinION offering genome sequencing both fast and accessibly, certain DNA errors, so-called indels, are frequently not identified in the final sequence output. This is where the research of Professor Gao and his team at The Structural and Functional Bioinformatics group, King Abdullah University of Science and Technology (KAUST) comes in. The researchers show how this shortcoming is largely due to the divided base-calling process and present a novel base-calling method, WaveNano, borrowing techniques from speech recognition deep machine learning (a type of artificial intelligence able to learn from large datasets without human interference). By considering the nanopore signal as a speech signal, the base-calling process can then effectively be viewed as speech recognition. WaveNano jumps over the traditional segmentation step and directly decodes the raw signal sequence substantially reducing the amount of indel errors.

SIMULATED VALIDATION

The large amount of output data then needs to be analysed with downstream analytical tools comparing the DNA sequence from the MinION to a reference genome in order to recognise important differences. Due to the rapid growth of the field, validating the efficiency of new analytical tools is not always possible due to lack of empirical data (e.g. annotated data or patient samples), paving the way for novel approaches.

Prof Gao and his team have developed the first signal-level simulator for Nanopore: DeepSimulator, which imitates the entire physical process of Nanopore sequencing using deep learning. With DeepSimulator, a researcher is able to generate sample datasets to test and validate analytical tools from a known starting sequence, or ground truth. Starting from a given reference genome, the DeepSimulator creates simulated electrical current signals by a context-dependent deep learning model, followed by a base-calling process outputting the final simulated reads. There are only a few simulators on the market for the Nanopore MinION technology, however, none of them simulates the crucial step of generating a read from electrical current signals, which is novel to the MinION. Unlike the previous simulators which only generate the reads from the statistical patterns of the real data, DeepSimulator simulates both the raw electrical current signals and nucleotide reads with up to 97% accuracy.

As well as being a tool to generate standard datasets to evaluate new methods for Nanopore sequencing data analysis, DeepSimulator also helps researchers better understand empirical datasets. Due to the high demand from the community on using DeepSimulator and numerous requests on customizing the

DeepSimulator provides simulated datasets to train and test sequencing analytical tools, cwDTW provides a highly efficient solution to align raw signals with DNA reads, while WaveNano has innovated the process of translating a sample into a DNA sequence.

Genome sequencing platforms are transforming the field of genetic disease research as they offer a closer look at human genes and DNA for clinical diagnostics.

SynthEx/Shutterstock.com

simulator, Prof Gao recently developed DeepSimulator1.5.

Another technical challenge is to align the raw signal sequences with the DNA reads, both of which are extremely long: the DNA reads of Nanopore range between 10K to 100K, whereas the raw signal sequence is even ten times longer. To this end, they developed a highly efficient alignment algorithm, cwDTW, which is 3000 times faster than the original dynamic time warping (DTW) algorithm while achieving almost 100% accuracy in alignment.

Pushing the methods further into applications in various scientific domains, Prof Gao tested the in-house end-to-end pipeline on clinical diagnosis of genetic diseases, antibiotic resistance gene detection, and the effect of genomic editing. In some cases, rare genetic variants are missed by the genome sequencing systems which has devastating effects on patients if mutations are not discovered in time. By labelling and sequencing individual DNA molecules, Prof Gao and a team of researchers identified large structural variations induced by the unique gene editing tool CRISPR-Cas9 in human stem cells. This could both help the safe testing and improvement of the gene editing technology while serving as an accurate assessment of rare genomic variations in patients. Additionally, Prof Gao's team was able to diagnose genetic RNA variants in 13.5% more patients who had previously been declared 'negative' by other sequencing methods. In doing so, the study paved the way to combine clinical RNA diagnostics with genome sequencing and analysis.

Prof Gao's team diagnosed genetic RNA variants in 13.5% more patients who had previously been declared 'negative' by other sequencing methods.

GENOMIC DIAGNOSTICS 'ANYWHERE, ANYTIME'

Through their academic collaboration, Professors Gao and Mo Li co-founded the start-up, Peregrine Genomics, with a goal of providing real-time genomic diagnostics 'anywhere, anytime'. Peregrine Genomics – Accurate Biomedical Genomics Anywhere Anytime

The Paregrine Genomics system combines the portable long-reads Nanopore sequencer and the innovative data analysis solution, in the hope of overcoming the current technical imperfection in genetic diagnosis and accelerate its application to human healthcare. In September 2019, the start-up won the Taqadam competition in Saudi Arabia and ranked in the top 25 start-ups among all the 175,000 participating start-ups from all over the world at the Entrepreneurship World Cup in October 2020. Using an algorithm that compares the reference sequence of the gene of interest with the raw signal of patient samples from the sequencer, the machine learning method is built to detect mutations in real-time.

During the past three years, Prof Gao and his team have developed a series of methods and algorithms to provide an end-to-end pipeline for Nanopore sequencing, covering everything from basic research, to practical and clinical applications, and to technology transfer.

Behind the Research

Professor Xin Gao

E: xin.gao@kaust.edu.sa **T:** +966-12-8080323 **W:** http://sfb.kaust.edu.sa

Research Objectives

Professor Gao's research interest lies at the intersection between computer science and biology. In the field of computer science, he is interested in developing machine learning theories and methodologies related to deep learning, probabilistic graphical models, kernel methods and matrix factorisation. In the field of bioinformatics, his group works on building computational models, developing machine learning techniques, and designing efficient and effective algorithms to tackle key open problems along the path from biological sequence analysis, to 3D structure determination, function annotation, understanding and controlling molecular behaviours in complex biological networks, and, recently, to biomedicine and healthcare.

Detail

Xin Gao
Building 3, Room 4217, CBRC, KAUST, Thuwal, 23955, Saudi Arabia

Bio

Xin Gao is Professor of computer science in CEMSE Division at KAUST. He is also the Acting Associate Director of the Computational Bioscience Research Center (CBRC), Deputy Director of the Smart Health Initiative (SHI), and the lead of the Structural and Functional Bioinformatics Group at KAUST. Prior to joining KAUST, he was a Lane Fellow at Lane Center for Computational Biology in School of Computer Science at Carnegie Mellon University. He earned his BSc in Computer Science, Tsinghua University in 2004 and PhD in Computer Science, University of Waterloo, 2009. He has published more than 230 papers in the fields of bioinformatics and machine learning. He is associate editor of *Genomics, Proteomics & Bioinformatics, BMC Bioinformatics, Journal of Bioinformatics and Computational Biology, and Quantitative Biology, and the guest editor-in-chief of IEEE/ACM Transactions on Computational Biology and Bioinformatics, Methods, and Frontiers in Molecular Bioscience.*

Funding

King Abdullah University of Science and Technology (KAUST)

Collaborators

- Professor Mo Li, BESE Division at KAUST
- Dr Fowzan Alkuraya, King Faisal Specialist Hospital and Research Center, Saudi Arabia

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

References

S. Wang, Z. Li, Y. Yu, and X. Gao. (2018). WaveNano: a signal-level nanopore base-caller via simultaneous prediction of nucleotide labels and move labels through bi-directional WaveNets. *Quantitative Biology*. 6(4): 359-368.

Y. Li, R. Han, C. Bi, M. Li, S. Wang, and X. Gao. (2018). DeepSimulator: a deep simulator for Nanopore sequencing. *Bioinformatics*. 34(17): 2899-2908.

Y. Li, S. Wang, C. Bi, Z. Qiu, M. Li, and X. Gao. (2020). DeepSimulator1.5: a more powerful, quicker and lighter simulator for Nanopore sequencing. *Bioinformatics*. 36(8): 2578-2580.

R. Han, Y. Li, X. Gao, and S. Wang. (2018). An accurate and rapid continuous wavelet dynamic time warping algorithm for end-to-end mapping in ultra-long nanopore sequencing. *Bioinformatics*. 34(17): i722-i731.

R. Han, S. Wang, and X. Gao. (2019). Novel algorithms for efficient subsequence searching and mapping in nanopore raw signals towards targeted sequencing. *Bioinformatics*. 36(5): 1333-1343.

S. Maddirevula, *et al*. (2020). Analysis of transcript-deleterious variants in Mendelian disorders: implications for RNA-based diagnostics. *Genome Biology*. 21: 145.

C. Bi, L. Wang, B. Yuan, X. Zhou, Y. Li, S. Wang, Y. Pang, X. Gao, Y. Huang, and M. Li. (2020). Long-read individual-molecule sequencing reveals CRSPR-induced genetic heterogeneity in human ESCs. *Genome Biology*. 21: 213.

Personal Response

Based on your research, what do you think is the next step in genome sequencing?

"In my opinion, the mutations, insertions and deletions have been relatively well solved in genome sequencing, but large structural variations (SVs) remain a big challenge. Therefore, I expect the next step is to develop computational methods to resolve SVs. In addition, how to interpret the variations and to associate them with traits or diseases is another key challenge in the field."