Knowledge Transfer Office
香港城市大學
City University of Hong Kong

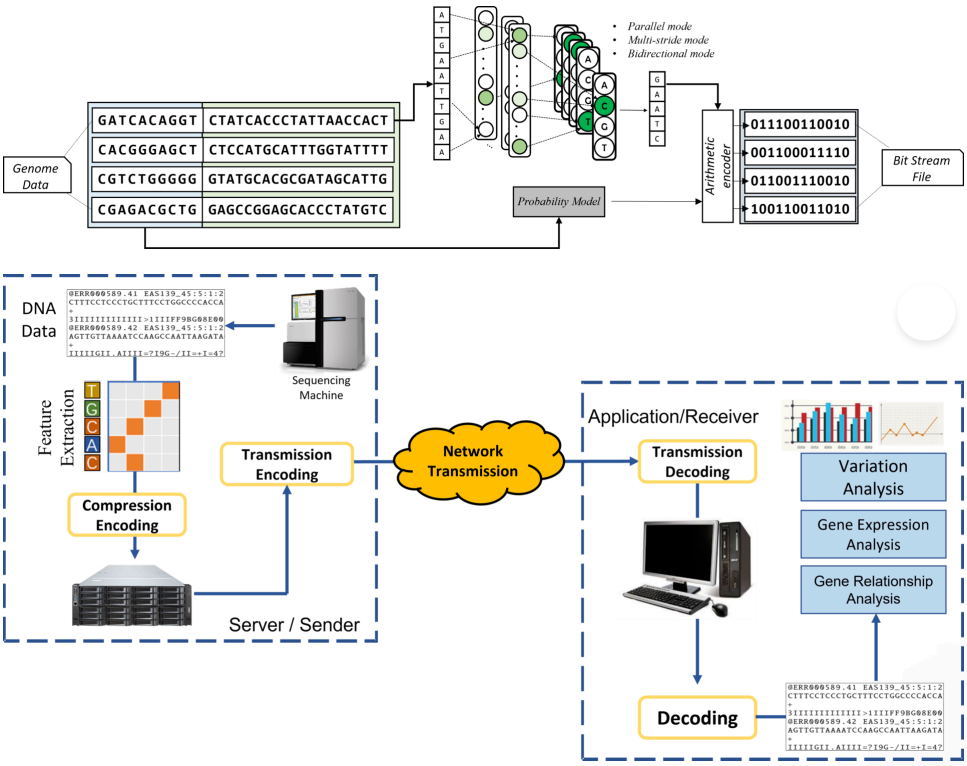
Learning-based Genome Codec

Communications & Information

Biomedical and Genetic Engineering/Chemical Products

Computer/AI/Data Processing and Information Technology

中文版本



Remarks

48th International Exhibition of Inventions Geneva (IEIG) (2023) - Silver Medal

IP Status

Patent filed

Technology Readiness Level (TRL) ? 6

Inventor(s)

Prof. KWONG Tak Wu Sam

Prof. WANG Shiqi

Dr. WANG Meng

Dr. SUN Zhenhao

Enquiry: kto@cityu.edu.hk

Opportunity

The rapid development of gene sequencing technology has fundamentally changed the way humans study the blueprint of life, and promoted the establishment and development of genomics, bioinformatics, systems biology, synthetic biology and other disciplines. With the continuous improvement of gene sequencing technology, the price of sequencing continues to decrease. However, the growth rate of sequencing data far exceeds Moore's Law, such that genome data compression and transmission have become the fundamental key technologies in the upstream of the bioinformatics industry. Relying on artificial intelligence, this invention focuses on high efficiency genome compression through machine learning and context modeling. The prototype software can achieve more than 100 compression ratio over test data, featured with configurable, low-complexity, and high compression efficiency for genome data compression. As such, storage and transmission cost can be effectively reduced for genome data.

Technology

In this invention, we use machine learning techniques to learn potential data patterns from a large amount of genetic data of the same species. By using these models, we can compress the genome data of the same specie with high efficiency. In addition, in order to improve the compression speed, we innovatively introduced three prediction methods, including parallel prediction, multi-stride prediction and bidirectional prediction. The experiment results of the prototype software show that these predictive models can effectively reduce the computational complexity of machine learning models during compression.

Advantages

- **Efficient:** The compression ratio of our invention achieves more than 100 times, which is much higher than traditional encoder (e.g., 4x)
- **Fast:** Compared with methods without codec optimization, our technologies boost the processing speeding (50x speed up), facilitating a variety of applications.
- **Intelligent:** With the compact representation, we can explore the relationship of different species, providing the fundamental support for biology research.

Applications

- Genetic cloud database
- Telemedicine
- Genetic diagnosis