Quantitative Biology 2021, 9(3): 242–250
https://doi.org/10.15302/J-QB-021-0255

FEATURE

Early bioinformatics research in China

Runsheng Chen*

Chinese Academy of Sciences Key Laboratory of Nucleic Acid Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China
* Correspondence: chenrs@ibp.ac.cn

Received June 11, 2021; Revised June 13, 2021; Accepted June 15, 2021

This article records the author's experience in participating in the early human genome and bioinformatics research in China, especially the non-coding sequence of the genome. It also introduced the beginning of human genome research in china, including the experts and teams involved in the International Human Genome Project. All the progress of bioinformatics originates from the inheritance of theoretical biology and the layout of philosophers in china.

THE ORIGIN AND CONTENT OF "BIOINFORMATICS"

Bioinformatics, a new scientific discipline, emerged in the late 1980s and developed with the human genome research in the early 1990s. Dr. Hwa A. Lim (aka HAL), a Malaysian [1], coined the term "bioinformatics" in 1988 while working in the U.S. He defined bioinformatics as a new subject of genetic data collection, analysis, and distribution to the research community. Bioinformatics achieved significant progress and growth thanks to the implementation of the U.S. Human Genome Project (HGP): the First Five Years FY 1991–1995, jointly launched by the U.S. National Institutes of Health (NIH) and the Department of Energy (DOE). According to HGP, "Genome informatics is a scientific discipline that encompasses all aspects of genome information acquisition, processing, storage, distribution, analysis, and interpretation." In [2], genome informatics is a relatively narrow definition of bioinformatics. It means that bioinformatics takes the analysis of genomic DNA sequence information as the source to decipher the genetic language hidden in the DNA sequence, especially the essence of the non-coding region; and after discovering all kinds of gene information, it carries out the simulation and prediction of the spatial protein structure of translation. From the perspective of natural philosophy, tahe research goal of bioinformatics is to reveal the complexity of genome information structure and the fundamental laws of genetic language. It is an organic combination of the three major scientific issues of "genome", "information structure" and "complexity" in the field of natural science and technological science. As rising of multiple omics such as transcriptome, proteome, epitome, metabolome, the connotation of bioinformatics has become richer, and software and databases have become more diverse. Bioinformatics has also become indispensable when biological science enters the era of big data.

THE BEGINNING OF HUMAN GENOME RESEARCH IN CHINA

International research on the human genome was brewed in the late 1980s and started in the early 1990s. In 1991, Chinese scientists had discussed how to participate in genome research. In his article titled "Science Foundation and I," Prof. Min Wu (an academician of the Chinese Academy of Science) said, "I was appointed Director of the Life Science Department of the National Natural Science Foundation of China (NSFC) in 1991. So certainly I would work hard to make the Chinese HGP a success during my tenure. At the end of 1991, I submitted to the NSFC a significant proposal on the Chinese Human Genome Project. Two days before National Day in 1992, I slipped, fell, and sustained multiple injuries when putting on clothes after swimming. The injuries, consisted of two bloody nasal fractures, and

This article is dedicated to the feature in 20th anniversary of Human Genome (Eds. Michael Q. Zhang and Xuegong Zhang).

left eyebrow bruises were so severe that I had five stitches and received a tetanus shot. This incident had happened at a very poor timing, as the defense on the proposal of Chinese Human Genome was to be held at the NSFC on Oct 6, 1992. Prof. Zongliang Zhao, deputy director of the NSFC Life Science Department, was very anxious as any of my replacement would not be able to defend the proposal adequately. Fortunately, I did not suffer any inflammations. On the day of the defense, I wrapped my head in gauze and reported the importance and far-reaching significance of implementing the HGP in China. Fortunately, my defense was accepted"[3]. In this article, Prof. Wu also wrote, "In March 1993, a symposium on Chinese human genome review was held in Wuxi, Jiangsu Province. After a detailed discussion, the whole committee of the NSFC approved my proposal. After multiple reviews in more than a year, I eventually could get the project on the list eligible for the NSFC fund. At that time, I was 67 years old. Having considered this human genome research project would take many years to complete, and possibly well into the 21st century, I realized the necessity of younger scientists to lead this important task. Thus, Academicians Zhu Chen and Boqin Qiang, relatively younger at the time, were selected as the project leaders, and a group of young and middle-aged scholars was recruited to take the primary task of this project. After the NFSC approve the fund to support the project entitled "Comparative Study on the Genetic Structure of Several Loci in Chinese Genome", the research on the human genome in China was carried out on a large scale [3]. On September 28, 1993, in Shanghai, an expert group representing the NSFC Life Science Department and led by Prof. Jiazhen Tan, deliberated and endorsed "Research on the Genetic Structure of Several Loci in Chinese Genome", a paper submitted by Prof. Boqin Qiang and Prof. Zhu Chen. This endorsement marked the official beginning of the Chinese HGP research. In 1998, this project was successfully completed thanks to the joint efforts of 19 research groups from 16 research institutions. It obtained several significant results and some breakthroughs.

HOW DID I "SQUEEZE" INTO THE HUMAN GENOME RESEARCH?

My interest in human genome research began in the late 1980s when many articles elaborated on measuring and interpreting human genetic codes. I gained more interest in this subject when reading Prof. James Watson's paper, "The Human Genome Project: Past, Present, and Future," which appeared in *Science* in 1990 [4]. As Prof. Watson was so committed to the direction and promotion of human genome research at NIH from 1988 to 1992 and served as the first director of the Office for Human Genome Research in the USA, I read through his paper on genome research with the increased interest. At the beginning of that paper, Prof. Watson mentioned, "Although the final monies required to determine the human DNA sequence of some 3 billion base pairs (bp) will be an order of magnitude smaller than the monies needed to let men explore the moon, the implications of the Human Genome Project for human life are likely to be far greater…When finally interpreted, the genetic messages encoded within our DNA molecules will provide the ultimate answers to the chemical underpinnings of human existence. They will not only help us understand how we function as healthy human beings but will also explain, at the chemical level, the role of genetic factors in a multitude of diseases, such as cancer, Alzheimer's disease, and schizophrenia, that diminish the individual lives of so many millions of people." These words made me so excited that I just wanted to speak out what had long been in my mind. As I could not find any colleagues in China to exchange ideas on human genome research, I took the liberty of writing a letter to Prof. Watson, sharing my understanding and support for this remarkable scientific research. At that time, I had never connected with Prof. Watson, and I was just an ordinary, unknown Chinese researcher. I knew that such a letter was only an expression of my observations about this event; I did not expect any feedback or reply from Prof. Watson. To my surprise, about a month later, I received an email from NIH, a reply from another professor on behalf of Watson. The letter had two pages, and the first page was words of appreciation thanking me for my understanding and support of HGP. On this page, he explained once again the profound impact and significance of HGP. On the second page, he said that he sent two documents to me. One was the official document of the First Five-year Plan for HGP in the U.S., and the other was a brochure about institutes at NIH. He invited me to visit these institutes. I read these document of U.S. HGP word by word, and I became more determined to devote myself to genome research.

I also realized that bioinformatics analysis, including sequence assembly and functional element identification, was the key to HGP. During that period, I was trying to search for some clues about human genome research in China. In 1992, I heard that Prof. Min Wu was organizing a human genome research project in China. How to participate in this project became my only pursuit at the time. Prof. Wu did not know me. How could I tell him about my ideas? I thought for a long time and was afraid that he might turn me down.

Moreover, Prof. Wu picked up his team members primarily from medical research institutes, not from the Chinese Academy of Sciences (CAS) institutes. I was a researcher at the Institute of Biophysics of CAS at the moment. If I went to his Institute directly, I feared that

Prof. Wu would refuse to meet me or even refuse to give me the chance of working with him. After much consideration, I sent Jun Xu (who graduated from Tsinghua University and now works in the U.S.), one of the outstanding graduate students in my lab, to visit Prof. Wu on my behalf. Jun Xu was smart and quick sensed. I told him when he met Prof. Wu, and he just talked about two points. One was telling the story about my letter to Watson and his reply to emphasize my urgent desire to participate in Chinese HGP and telling him about my interpretations of the U.S. HGP official documents. Making this point to Prof. Wu, I would like Prof. Wu to understand that the theoretical biology work that my research group had been doing was necessary and complimentary for genome research in China and that we wanted to participate in the research process of sequence assembly and analysis. The other point was that we knew their funding was tight, we could only work without their funding. To my surprise, Prof. Wu, a well-known, influential scientist in China, met my student in person and listened to his introduction. I always recalled that if Prof. Wu did not meet Jun Xu, there certainly would be no opportunity for me to engage in bioinformatics research at the earliest research of HGP in China, nor would I become an expert in genomics and bioinformatics.

Later, Prof. Wu arranged our lab to participate in Chinese HGP. It's my great honor to join the Major Project of NSFC entitled "Comparative Study on the Genetic Structure of Several Loci in Chinese Genome." This project was launched in 1993 with total funding of 3 million RMB and was led by Prof. Zhu Chen and Prof. Boqin Qiang per Prof. Wu's recommendation. What made me even more overjoyed was that the project team allocated funds to our team, although their funds were severely insufficient and that I had made it clear that I did not expect their financial support. I was assigned to study the methods of DNA sequence splicing, assembly, and identification of functional elements (mainly coding genes) for his project. Our team established various DNA sequence statistical analysis methods, fractal dimension analysis, neural network, complexity, local degeneracy, and so on. In particular, we were the first in the world to discover the cryptographic method [5]. Combined those methods for gene recognition, we could improve the success rate of prediction. It's a coincidence that we applied cryptography knowledge to analyze DNA sequences. In the second half of 1991, we heard that Prof. Kencheng Zeng and Prof. Dingyi Pei were going to organize a cryptography training class. I wondered whether genetic code could relate to passwords widely used in military or commerce. Why didn't we take part in this training class? Jun Xu, a graduate student in my lab, spent almost one year for the training, while I only attended the lectures. Although we did not understand everything, we did learn some technical methods to deal with the four characters of DNA sequence by changing some formulas. We scored excellent results and applied the learned cryptography analysis technology in genome analysis. Due to our achievements in the field of genomic informatics, I was invited to give a commemorative speech at the 15th International Committee on Science and Technology Data (CODATA) held in Tsukuba, Japan, on September 29, 1996, and was awarded the "Kotani Prize" (biological field). Although we have developed and applied various algorithms, we have only identified a few coding regions in the genome sequence. In the beginning, we always suspected that the algorithm we had developed might not be good. Having communicated with scientists doing the human genome research worldwide, we noted that they also could not find more coding regions in the human genome sequence. Together with our international counterparts, we gradually came to a consensus that only a few sequences available might be only used to encode proteins in the human genome. Our initial estimate of the sequences was about 10% and soon we realized that this 10% was also overestimated. Nowadays, we know that there is only around 2% of the human genome sequence available to encode protein.

JOINED THE "HUMAN GENOME PROJECT"

Although the exact proportion of coding sequences could not be known in the early 1990s, it was evident that coding sequences accounted for only a small part of the human genome. The sequence in the genome that did not code for a protein was called "JUNK" DNA. I work in theoretical analysis of human genome sequence, and I always feel that 98% of "JUNK" DNA is neither comforting nor illogical! I firmly believe that the "JUNK" DNA must be functional. At the end of 1993, my research group had paid more attention to genome non-coding sequences research. In some academic conferences, I repeatedly elaborated on the "JUNK" DNA. But there was little response, and my research group did not make much progress. The reason was simply lacking experimental data. I wrote about the 73rd puzzle of "What's the function of 'JUNK' DNA?" in the book "100 Scientific Puzzles of the 21st Century" [6] published in June 1998. This writing of mine represented some of my thinking about the non-coding sequences.

Although China launched the human genome research in 1993, it was not conducted on a large scale, regardless of the persistent promotion by Prof. Min Wu and serious implementation led by Prof. Zhu Chen and Prof. Boqin Qiang. It only had detected some specific loci and did not touch upon the complete genome sequence. From 1998 to 1999, three genomics research centers were established in

China, namely "National Genomics Northern Research Center" established in September 1998 (Academician Qiang Boqin was the Director, Academician Wu Min was the Honorary Director of the Academic Committee), the "National Genome Southern Research Center" (Academician Zhu Chen was the Director) established in October 1998 and the "Beijing BGI Gene Research Center" established in July 1999. They indicated that the large-scale sequencing of the human genome began in China. They performed in line with international standards. They participated in the international "Human Genome Research" and undertook the task of sequencing the 30 million bases of the short arm in human chromosome 3 (about 1% of the human genome). There have been many reports about the work done during this period of history, especially about the history of BGI. The following is only a summary based on what I have participated in and what I have known.

One day in 1992, Maynard V. Olson, one of the planners and leaders of the U.S. "Human Genome Project," came to New York University and personally invited Prof. Jun Yu to work in his lab and to co-develop the key technologies needed for the "Human Genome Project" in his lab. Consequently in 1993, Prof. Jun Yu resigned as an assistant professor from New York University and joined the Maynard V. Olson laboratory, working at the genomics research center of Washington University, where he completed the most accurate physical map. At the same time, Prof. Jun Yu discussed with Prof. Jian Wang (who worked at the University of Washington) on how to extend human genome research to China so that China could catch up with the international level in this field. They visited Prof. Huanming Yang in Denmark and Prof. Siqi Liu in Texas. These four Chinese scientists decided to come back to China for genome research in China. In 1994, Prof. Jian Wang first returned to China and established the Beijing Hua Da Ji Bi Ai Biotechnology Co., Ltd. (GBI), focusing on promoting domestic genome research. Prof. Huanming Yang returned to China on the same year and served as a professor in the Institute of Basic Medical Science, Chinese Academy of Medical Sciences. In November 1997, Prof. Huanming Yang, a director of the Youth Committee of the Chinese Genetic Society at the time, organized a seminar in Zhangjiajie, Hunan Province, and invited Prof. Jun Yu to introduce the history and development trend of the human genome project in the U.S. After careful discussion, Prof. Huanming Yang, Prof. Jian Wang, Prof. Jun Yu and others put forward the strategic concept of China's Human Genome Project, which became the starting point of their shared career. In 1998, Prof. Shouyi Chen (Director of the Institute of Genetics and Developmental Biology, CAS) and Prof. Lihuang Zhu (Deputy Director of Institute of Genetics and Developmental Biology, CAS) invited Prof. Jun Yu and Prof. Huanming Yang to set up a human genome center at the Institute of Genetics and Developmental Biology, CAS. Prof. Jun Yu returned to China in 1998 and established the human genome center on August 12, 1998, at the Institute of Genetics and Developmental Biology. Apart from Prof. Jun Yu, Prof. Huanming Yang, Prof. Jian Wang, and Prof. Siqi Liu, all participated in this center. Prof. Huanming Yang served as Director, Prof. Jian Wang as Executive Director, Prof. Jun Yu and Prof. Siqi Liu as Deputy Directors. I attended this center-founding conference. I remember that this meeting was held in the courtyard in front of the main building of the Institute of Genetics and Developmental Biology. After the conference, I saw a small building on the side, which was especially used for this human genome research center. There were only desks and chairs in the room, and some of them were damaged. They had not got any research equipment and instruments. It is fair to say that they started from scratch. On June 26, 1999, this center applied for the international HGP of the NIH. HGP announced on their webpage that China had registered to join the international sequencing organization. This announcement marks that China became the sixth country to join the organization after the United States, Britain, Japan, Germany, and France. In the article "Science Foundation and I", Prof. Min Wu recalled, "In the summer of 1999, Prof. Huanming Yang came to my office and told me that he decided to participate in the international genome conference to be held in Cambridge, UK, and would be striving for participating in the 1% of human genome sequencing research at the conference. He would like to hear my opinions. I immediately gave him my support and encouragement [3]. Thanks to the kind support of Maynard V. Olson and other experts, Chinese scientists made it to the Fifth Conference on large-scale sequencing strategy of human genome held in Cambridge, UK, in September 1999. Prof. Huanming Yang requested to join the human genome project and undertake the 1% of the sequencing task on behalf of China. The Chinese government at the time had not approved this project yet. On November 10, 1999, the 1% of the human genome sequencing project was made on the national project list, and it was determined that BGI would take charge of the whole project, National Genomics Northern Research Center and National Genome Southern Research Center would participate in it.

BGI was officially established on September 9, 1999, in a factory building close to the Beijing Capital Airport. I remember that in 2000, I spent most of my time working in BGI. I commuted 4 times a week from the Institute of Biophysics (located by Datun Road, Chaoyang District) and BGI. There were about 20 members in the Bioinformatics team at the time, each with a cubicle

(about 2 m × 2 m), and the cubicles were the same for everyone. Jun Wang was in charge of the Bioinformatics team, and my student Wei Li was his assistant. The key task was to integrate and develop algorithms for large-scale genome splicing, assembly, and gene identification.

DECIPHERING THE GENOME OF THERMOPHILUS TENCHONG–TRAINING BEFORE THE STUDY OF THE HUMAN GENOME

Before the large-scale human genome sequencing, Prof. Huarong Tan's team (Institute of Microbiology, CAS), Prof. Huanming Yang's team (Institute of Genetics and Developmental Biology, CAS), and my team (Institute of Biophysics, CAS) agreed to sequence the whole microbial genome in 1998 so as to prepare and train the teams. A seminar on choosing microorganisms to be the target was held in the Institute of Microbiology. The Institute of Microbiology prepared four or five candidate microorganisms, the large one was estimated to have six million bases, and the small one was about three million bases. After discussing all factors, we finally selected the hot spring bacteria B4 from Tengchong Hot Spring in Yunnan Province as the research material on the following considerations: (1) This strain is heat-resistant, and it is possible to find new functional thermostable enzymes (similar to the thermostable enzymes in the famous PCR reaction); (2) This strain lives in water, and it may be a relatively old species, which is helpful to explain the evolution of life; (3) The Institute of Microbiology discovered this strain, and they had the right to name such a new species of microorganisms. Its patent should belong to China; (4) This strain is the smallest of all the candidate microorganisms, with only 3 million bases. At that time, the cost of sequencing in China was estimated to be 1 RMB per base, excluding the cost of manpower and so on. Thus, it would take 3 million RMB to complete the sequencing of hot spring bacteria B4, which was an ideal material for this project. However, even the wise are not always free from error. We ignored the fact that the AT content of this bacterial genome is very high (62.4%), and the GC content is very low (only 37.6%). This significantly increased the amount of sequencing and the difficulty of splicing and assembly. In the end, the total amount of sequencing was no less than that of a six million base genome with high GC content. In the beginning of this project, there was no ready-made software tool. We had to manually complete all the algorithms and programs for splicing, assembly, and gene identification. I'm glad that this is the first complete genome of an organism in China, after all. My students Zhenyu Xuan, Wei Li, and Jian Yang were the primary participants in bioinformatics analysis. Then they took part in the bioinformatics work of the human genome. On June 26, 2000, Bill Clinton and Tony Blair jointly announced the completion of the "working framework map" of the human genome jointly undertaken by the six countries.

MY RESEARCH ON NON-CODING RNA

The more I participated in genome analysis, the more I believed that non-coding sequences had biological functions. We had little progress in this field because there were too little experimental data. There were not many researchers who engaged in non-coding research in the world at the time. To understand the function of non-coding from the analysis of biological information was like making bricks without straw. At the end of 1999, we saw the vigorous development of domestic human genome sequencing experiments and the process of realizing large-scale sequencing of BGI from scratch, which touched our desire to establish a wet laboratory and obtain non-coding research data. But it was not easy for theorists to do biological experiments at the molecular level. Where were the talents? Where were the funds? What was the device? Do what? How to do? All needed to be solved. If there is a will, everything will come true! This was a test of my will and determination. I invited Wei Deng, who was good at the experiment from our institute. He graduated from the University of Science and Technology of China, where I graduated, too. He was much younger than me. In the future, he would be in charge of doing experiments on discovering new non-coding nucleic acids. Fortunately, it was not that hard to get the fund. After I came back from studying in Germany at the end of 1987, I secured financial support from the "National High-tech Program of China (863 Program) Protein Project" funded by the Ministry of Science and Technology. In the early 1990s, I also took part in the Chinese human genome project. Most of these funds obtained were not spent because it cost nothing to do the bioinformatics analysis as long as we could use any computers free of charge. Although 0.6 million RMB was not that much, it could enable us to do lots of experiments. What project to do was a test for me. If I could not make a right choice, I would certainly fail. I was very clear that we needed to find transcripts from non-coding sequences of the genome. If we found the transcripts, it meant that non-coding sequences had information dissemination. These transcripts were needed to perform biological functions, and there would be substantial progress in non-coding research. What kind of species should we take as samples? Microorganisms cannot because the proportion of non-coding sequences in the genome is too small (generally 10%–20%), and higher organisms are too complex. *C. elegans* is the most suitable. Although it is simple, it is a multicellular organism. The proportion of

non-coding sequences in its genome has reached 70%. Moreover, as a model organism, its physiology, development, cell, and genome are very clear. But how many transcripts were there, and how about their distribution? We had no idea. For this reason, we made an unscientific artificial regulation before the experiment, limited the length of transcripts to 50–500 bases. Later experiments proved that this artificial limitation was essential to achieve faster results. The upper limit of 500 bases was the desired transcript, which came from a single reading frame and did not contain introns. Such a transcript was a functional unit without splicing and assembly, and the experiment was much simpler. The set lower limitation of 50 bases was because microRNAs' length was 21–24 bases, and 50 was more than twice that of 24. Therefore, it could ensure that our discovery is new and different from the family of microRNAs. How could we systematically find these non-coding RNAs? There was no precedent! We should design the experimental process by ourselves and explore the experimental conditions and parameters constantly. We made a large of clones, but we could not do more sequencing because it was too expensive. Wei Deng had recorded more than 400 documents of various exploration related to this experiment for more than four years. In theory, we also established our own non-coding gene prediction method. We found 161 new non-coding genes in *c. elegans*. Thus, two non-coding gene families were identified, and three specific non-coding gene promoters were found. The results showed that non-coding genes and coding genes had their own transcriptional regulation systems. The paper was published in *Genome Research* on January 6, 2006 [7]. Later, EurekAlert, a scientific review journal of AAAs, published a long article on January 9 to introduce these research results. This introduction confirmed the above findings and pointed out that the efficiency of the experimental technology was 10 times higher than that in the world. All the non-coding genes were included in GenBank (NCBI access number: ay948555 – ay948719). Then, using the whole set of non-coding gene identification methods established in *C. elegans* research, we independently identified non-coding genes in human chromosome 3 and found nearly 900 non-coding genes of various types. As a co-author, this article was published in *Nature* in 2006 [8].

Since 2000 we have collected internationally confirmed NcRNA genes and non-coding transcripts developed corresponding software and searching tools and established the NcRNA database-NONCODE [9], which is currently the most comprehensive NcRNA in the world. NONCODE has become the basic data source for many studies. The academic contribution of this work is to propose a classification system for non-coding genes. As soon as the article was published on January 21, 2005, *Science* introduced this work. Since then, we have built a non-coding RNA and protein interaction database-NPInter [10], which provides a data basis for international non-coding gene research. Due to some experimental and theoretical results, our non-coding research had some influence both internationally and domestically, and research would be easier to carry out. Recalling those periods, I have to say that it was not easy at the beginning, and there was no funding. Even the "863 protein project" could not provide funds for us. In recent years, I have obtained the "National Program on Key Basic Research Project (973 Program)" in the non-coding field, and also received the NSFC "Major Research Project" as the chief and colleagues. But I always feel that I was still in the early stage of the research. Although we worked lonely then, yet it was a period of time full of creativity, freshness, and challenges.

THE EXPERTS AND TEAMS INVOLVED IN THE EARLY RESEARCH ON BIOINFORMATICS

As early as the early 1980s, scientists in China were engaged in analyzing DNA sequence characteristics. Prof. Liaofu Luo of Inner Mongolia University had led the team shift from theoretical physics research to theoretical biology research in 1982 and put most of his resources on the study of DNA sequence. Such a study was rather pioneering and rare in China then. Throughout the whole 1980s, they published a lot of research reports on DNA sequences and genetic codes, such as base distribution, homology, and Markov properties of nucleic acid sequences [11]; nucleic acid start sequence, stop sequence, and statistical analysis of insertion sequence [12]; S-4 symmetry breaking and stop codon of mutation rate [13]; information parameters of nucleic acid molecules and molecular evolution [14]; the degeneracy rule of genetic code [15]. They even discussed why the genetic code consisted of four bases [16]. It is very important that they had researched and developed theoretical methods of DNA sequence analysis, such as fractal methods [17], especially methods of information theory, such as the average mutual information method of gene sequences [18] and the principle of maximum information [19]. I remember that as early as the early 1980s, at a theoretical biology seminar, I listened to Prof. Luo Liaofu's report on DNA sequence analysis and discussed with him the analysis of nucleic acid sequences. Only then did I know that the Inner Mongolia University team started the study of DNA sequence characteristics so early.

Prof. Chunting Zhang of Tianjin University was also engaged in DNA theoretical research after the mid-1980s. Prof. Zhang was a visiting scholar at the French National

Research Center for Theoretical Physics from 1979 to 1982. In May 1984, he was transferred to the Physics Department of Tianjin University. After that, he switched from physics to computational biology and bioinformatics. Having his first theoretical biophysics paper published in 1987 [20], Prof. Zhang made two main contributions in the field of theoretical biology research. First, he proposed using the double Sine-Gordon partial differential equations to simulate the dynamics of base movement in transcription and replicating DNA molecules in the late 1980s. Second, he proposed the Z curve theory of DNA sequences in the early 1990s [21,22], which opened up a new application to the analysis of DNA sequences by geometric methods. At present, the Z-curve theory has been widely used in genomics and bioinformatics. Due to his contributions in theoretical biology and bioinformatics, Prof. Zhang was elected as an academician of the Chinese Academy of Sciences in 1995.

It can be said that from the mid-1980s to the early 1990s, China has been prolific in the theory of nucleic acid sequence analysis. For instance, Prof. Chunting Zhang proposed the geometric theory of DNA sequence analysis, Prof. Liaofu Luo proposed the informatics theory of analysis on DNA sequence, and I proposed the cryptographic theory of DNA sequence analysis.

In 1997, two outstanding scientists in mathematical sciences, Academician Bolin Hao and Academician Yanda Li, led their team to join the information analysis of genome sequence. They strengthened the bioinformatics research team in China and inspired the scientific and technological community to study genomics. Prof. Bolin Hao is a well-known theoretical physicist in China. He graduated from Kharkiv University in Ukraine in 1959, then served as principal investigator and the Director at the Institute of Theoretical Physics, CAS. In 1980, he was elected as a member of the Chinese Academy of Sciences (academician). Prof. Hao primarily studies theoretical physics and non-linear discipline. He has made many achievements in the fields of the solid electronic energy spectrum and phonon spectrum, infrared properties of metals, polymer semiconductor theory, statistical physics, antenna theory, seismic analysis, chaotic dynamics, and symbolic dynamics. He won the first prize of the Natural Science Award of the Chinese Academy of Sciences in 1992 and 1999. He also got the second prize of the National Natural Science Award in 1993 and 2000, separately. In 1997, he entered the field of theoretical biology, adhering to the consistent belief of "being engaged in specific scientific work on the front line," and devoted himself to the just-beginning nucleic acid sequence research. I remember that it was at the BGI that I met Prof. Bolin Hao and Prof. Weimou Zheng in 2000. They were developing new algorithms to complete the assembly and information mining of the rice genome. At the same time, Prof. Hao also developed the "k-mer (k length string)" technology to reconstruct the life evolution tree of prokaryotes using whole genome data. This set of microbial relationship analysis software-CVtree had been internationally recognized. As soon as he entered the field of bioinformatics, Prof. Hao and Prof. Jixing Liu edited the book "Theoretical Physics and Life Sciences" [23]. In 2000, Prof. Hao and his wife, Prof. Shuyu Zhang, co-authored the "Handbook of Bioinformatics" [24]. In 2002, the second edition of "Handbook of Bioinformatics" was also published [25]. In 2003, Prof. Hao also wrote the book "A Brief Introduction to Bioinformatics," which specifically introduced bioinformatics. Prof. Hao has made outstanding contributions to the development of bioinformatics in China.

Prof. Yanda Li is a well-known expert in signal processing and intelligent control in China. He graduated from Tsinghua University in 1959 and was elected as a member of the Chinese Academy of Sciences (academician) in 1991. He served as the Dean of the Department of Automation of Tsinghua University, a member of the School Council of Tsinghua University, and the convener of the Disciplinary Review Group of the State Council (Degrees Committee Control Science and Engineering). He has long been engaged in the research of signal processing theory and seismic exploration data processing methods. He made the advanced international level in the research of signal reconstruction theory and algorithm. He used signal processing and pattern recognition methods for seismic exploration data analysis and achieved pioneering results. He has repeatedly won the National Natural Science Award and the National Education Commission Science and Technology Progress Award. In 1997, Academician Yanda Li and Prof. Zhirong Sun co-founded the Institute of Bioinformatics of Tsinghua University, which was developed into the Key Laboratory of Bioinformatics of the Ministry of Education in 2002. For a long time, they have cultivated a large number of backbone talents in the field of bioinformatics and have made important contributions to the promotion of bioinformatics in China and the development of bioinformatics.

As we entered the 21st century, two critical teams joined the bioinformatics research. They were the Peking University Center for Theoretical Biology and the CAS Shanghai Center for Bioinformatics. The former was initiated in 1999 under the proposal of influential Prof. Zhengdao Li and under the advocacy and support of Peking University leadership, and came into being formally on September 17, 2001. It is based on the research power of mathematics, physics, chemistry, and mechanics to carry out theoretical and systems biology research from the two directions of experiment and

theory. It has achieved significant results in the research of biological regulatory networks. Its key founding members include Prof. Luhua Lai, Prof. Zhensu She, Prof. Chao Tang, and Prof. Qi Ouyang. Prof. Chao Tang and Prof. Qi Ouyang have now been elected as academicians of the CAS. It has now been renamed as the Peking University Center for Quantitative Biology. The CAS Shanghai Bioinformatics Center was established in June 2000 and headed by Prof. Yixue Li. It is a bioinformatics support platform in the CAS Shanghai Institute for Biological Sciences. The team under the direction of Prof. Yixue Li has become the core force of the CAS Shanghai Bioinformatics Research Center, which was established in 2002. This center, affiliated with the CAS Shanghai Academy of Sciences, has integrated bioinformatics research taskforces from 11 Shanghai-based scientific research institutions, including Shanghai Institutes for Biological Sciences, National Human Genome Southern Research Center, Fudan University, Shanghai Jiaotong University, Shanghai Institute of Pharmaceutical Industry, etc. It is an independent corporate legal entity specialized in bioinformatics research and database construction, and bioinformatics software development and is a supporting institution of the Shanghai Bioinformatics Society.

Since 2001, data on the human genome and rice genome have been published successively, and data on functional genomes such as the transcriptome and proteome have continued to appear. With the rapid development of omics big data, the number of individuals or teams engaged in bioinformatics in China has increased rapidly after 2002. For example, in March 2004, Harbin Medical University officially approved the creation of the Department of Bioinformatics. Prof. Xia Li served as the head of the department. In 2007, the School of Bioinformatics and Technology was established on the basis of the department.

The rapid development of bioinformatics is also reflected in domestic academic activities. In 1998, the North China Regional Bioinformatics Seminar, the first conference on the subject of "Bioinformatics", was held in the academic lecture hall of Tsinghua University Library in China. This was only a regional academic event. And the first national bioinformatics conference was staged in April 2000 at the Academy of Military Medical Sciences in Beijing. The second National Bioinformatics Conference was conducted in Peking University in June 2002.

In the first few years of the 21st century, bioinformatics research grew fast in China, and there emerged a large number of excellent talents. As people involved in bioinformatics studies in China are quite familiar with what happened since then in this field, I will not recount them any further in this article.

COMPLIANCE WITH ETHICS GUIDELINES

The author Runsheng Chen declares that he has no conflict of interests.

OPEN ACCESS

REFERENCES

1. Timeline of Bioinformatics, http://www.dtrends.com/Bioinformatics/biotimelinecontent.html
2. U.S. Department of Health and Human Services and Department of Energy (1990) Understanding our genetic inheritance: the U.S. Human Genome Project: the first five years
3. Huang, Y. (2006) Twenty years of history, I and the Science Foundation. Scientific Chinese (in Chinese), 6, 26–30
4. Watson, J. D. (1990) The human genome project: past, present, and future. Science, 248, 44–49
5. Xu, J., Chen, R., Ling, L., Shen, R. and Sun, J. (1993) Coincident indices of exons and introns. Comput. Biol. Med., 23, 333–343
6. Chen, R. (1998) What's the function of "JUNK" DNA? In: 100 Scientific Puzzles of the 21st Century, pp. 642–647. Changchun: Jilin People's Publishing Press
7. Deng, W., Zhu, X., Skogerbø, G., Zhao, Y., Fu, Z., Wang, Y., He, H., Cai, L., Sun, H., Liu, C., *et al*. (2006) Organization of the *Caenorhabditis elegans* small non-coding transcriptome: genomic features, biogenesis, and expression. Genome Res., 16, 20–29
8. Muzny, D., Scherer, S., Kaul, R.Wang, J., Yu, J., Sudbrak, R., Buhay, C. J., Chen, R., Cree, A., Ding, Y., *et al*. (2006) The DNA sequence, annotation and analysis of human chromosome 3. Nature, 440,1194–1198
9. Liu, C., Bai, B., Skogerbø, G., Cai, L., Deng, W., Zhang, Y., Bu, D., Zhao, Y. and Chen, R. (2005) NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic Acids Res., 33, D112–D115
10. Wu, T., Wang, J., Liu, C., Zhang, Y., Shi, B., Zhu, X., Zhang, Z., Skogerbø, G., Chen, L., Lu, H., *et al*. (2006) NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. Nucleic Acids Res., 34, D150–D152
11. Luo, L., Zhou, Y. and Tsai, L. (1988) The statistical distribution of nucleic acid sequences (The homological and non-markovian analysis). Journal of Biomathematics (in Chinese), 3, 10–17
12. Luo, L., Zhou, Y.-M. and Tsai, L. (1987) Statistical analysis of 5′-CAPS, 3′-TAILS and introns of nucleic acid sequences. Acta

Biophysica Sinica (in Chinese), 3, 341–345
13. Luo, L. and Li, Q. (1985) S-4 symmetry breaking and stop codon of mutation rate. Chin. Sci. Bull., 30, 1056
14. Luo, L. F., Tsai, L. and Zhou, Y. M. (1988) Informational parameters of nucleic acid and molecular evolution. J. Theor. Biol., 130, 351–361
15. Luo, L. F. (1988) The degeneracy rule of genetic code. Orig. Life Evol. Biosph., 18, 65–70
16. Luo, L. (1986) Why are there four bases in DNA? Orig. Life Evol. Biosph., 16, 267–268
17. Luo, L. F., Tsai, L. and Zhou, Y. M. (1988) Informational parameters of nucleic acid and molecular evolution. J. Theor. Biol., 130, 351–361
18. Luo, L., Lee, W., Jia, L., Ji, F. and Tsai, L. (1998) Statistical correlation of nucleotides in a DNA sequence. Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics, 58, 861–871
19. Luo, L. and Bai, G. (1995) The maximum information principle and the evolution of nucleotide sequences. J. Theor. Biol., 174, 131–136
20. Zhang, C. (1987) A kind of equations set in B-DNA dynamics. Journal of Biomathematics (in Chinese), 2, 7–13
21. Zhang, R. and Zhang, C.-T. (1994) Z curves, an intutive tool for visualizing and analyzing the DNA sequences. J. Biomol. Struct. Dyn., 11, 767–782
22. Zhang, C.-T. and Zhang, R. (1991) Analysis of distribution of bases in the coding sequences by a diagrammatic technique. Nucleic Acids Res., 19, 6313–6317
23. Hao, B. (1997) Theoretical Physics and Life Sciences. Shanghai: Shanghai Science and Technology Press
24. Hao, B. and Zhang, S. (2000) Handbook of Bioinformatics. Shanghai: Shanghai Science and Technology Press
25. Hao, B. and Zhang, S. (2002) Handbook of Bioinformatics (Second Edition). Shanghai: Shanghai Science and Technology Press