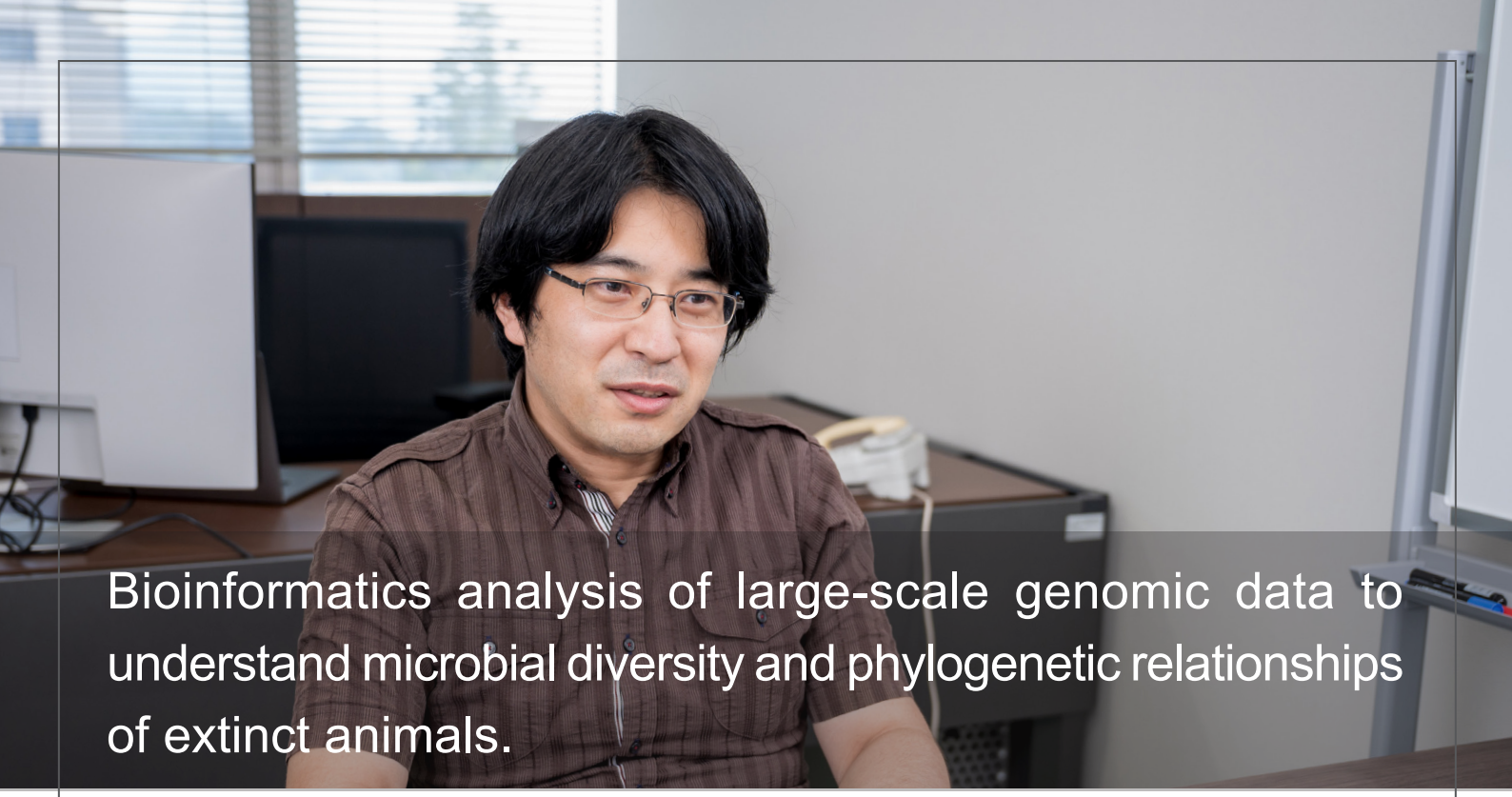


Bioinformatics analysis of large-scale genomic data to understand microbial diversity and phylogenetic relationships of extinct animals.

MORI, Hiroshi Associate Professor

Genome Diversity Laboratory
Department of Informatics

Graduated from the Department of Biology, Faculty of Science, Shinshu University. Completed his master's degree at the Nara Institute of Science and Technology (NAIST). Ph.D. in Science from the Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology. Assistant Professor, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, during 2011-2016. Assistant Professor, Laboratory of Genome Evolution, Division of Informatics, National Institute of Genetics, since 2016. Associate Professor, Laboratory of Genome Diversity, since 2021.

Associate Professor Hiroshi Mori specializes in metagenomic analysis of microbial communities and the analysis of ancient DNA of extinct animals using bioinformatics. He has also developed analytical tools that can easily estimate the phylogenetic composition of microbes, and a high-speed metagenome sequence similarity search tool. At the root of his research is his childhood curiosity about the ecology of organisms.

From Animal Behavioral Ecology to Bioinformatics

I studied in behavioral ecology as an undergraduate, in part because I loved catching and observing insects and fish as a child. Because snails are slow-moving and cannot fly, their populations are often isolated by terrain, such as a large river. This can cause speciation, making them very diverse animals. An important feature of snails is that most are hermaphrodites. My undergraduate thesis focused on the differences in reproductive behavior between closely related species living in the same area. At that time (2000-2003), the results of the Human Genome Sequencing Project were just coming out, which had a huge impact on biology. Perhaps because of this, I was also interested in genome science, and read genome-related books and papers in between my undergraduate research.

My research on the reproductive behavior of snails was successfully published(*1), and I decided to move into genomics as a graduate student. At that time, sequencing the genome of an organism was very expensive and laborious. However, bioinformatics, which uses computers to analyze genomic data, could be studied without high budgets using genome sequence data in public databases. This seemed very interesting to me and ultimately led to my decision to enter the Nara Institute of Science and Technology (NAIST). I had also considered the National Institute of Genetics (Graduate University for Advanced Studies) and attended a graduate school tour, but after much deliberation, I decided to go to Nara. During my master's study, I studied with Prof. Ken Kurokawa (who is now also at the National Institute of Genetics) on the comparative analysis of bacterial genomes. I also studied metagenomics, the analysis of the genomes of microbial communities, as the metagenomic sequence data were just becoming available at that time.

When Prof. Kurokawa moved to the Tokyo Institute of Technology, I also entered the Tokyo Institute of Technology as a graduate student. In my master's study, I analyzed the genome sequence data of certain bacteria sequenced by using Sanger sequencers. However, around 2006, so-called “next-generation sequencers” (NGS) were developed, in which long DNA fragments are cut into short fragments and

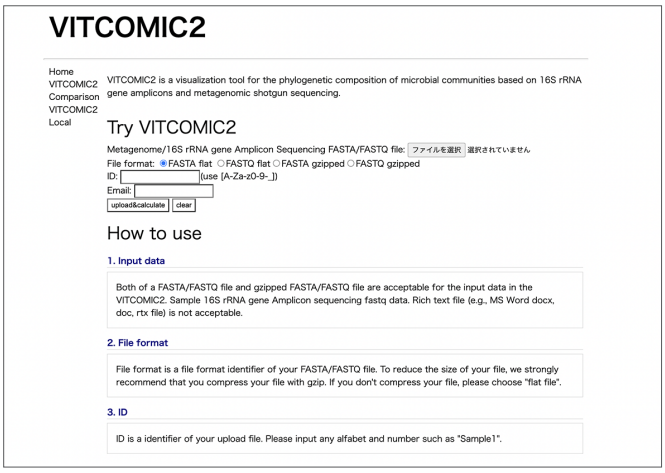
sequenced in parallel without electrophoresis. By sequencing more than millions of DNA fragments simultaneously with a single NGS, sequencing speed and cost are dramatically reduced. NGS has quickly become the standard for DNA sequencing. NGS has made it possible to read short genome fragments from microbial communities in large numbers at low cost, and the field of metagenomic analysis has rapidly become popular. Metagenomic analysis using NGS provides a large number of very short sequences. To estimate the member composition (phylogenetic composition) and gene function composition of microbial communities from these new data, it was essential to develop new bioinformatics analysis methods and software. Therefore, I developed analysis methods and web applications that can use metagenomic sequencing data. I published a paper and received my Ph.D. After that, I stayed at Tokyo Institute of Technology as an assistant professor and started my career as a researcher.

Around that time (2010 or so), NGS had become widely available and the usefulness of metagenomic analysis was better understood, so a large amount of metagenomic sequencing data was being produced around the world. However, the sequence data and metadata generated from metagenomic studies were not effectively used due to the “difficulty of reproducing the same community”. Therefore, we decided to collect and curate the public metagenomic data and develop an integrated database that would facilitate microbial omics research. In parallel, we developed analysis methods for metagenomic sequencing data. In 2016, Prof. Kurokawa moved to the National Institute of Genetics (NIG), and I also moved to the institute as an assistant professor. Since coming to NIG, I have continued my research on metagenomics and also started “ancient DNA research”, to infer the phylogenetic relationships of organisms from several thousand years ago using fossil-derived DNA (ancient DNA). In 2021, I established my own lab as an associate professor, and I now supervise graduate students while conducting several research projects of my own.

Development of VITCOMIC and VITCOMIC2, Tools for Easy Phylogenetic Composition Analysis

I developed VITCOMIC (Visualization Tool for Taxonomic Compositions of Microbial Communities) [*2], a tool to quickly and accurately estimate the phylogenetic composition of microbial communities from large metagenomic NGS data, and an improved version, VITCOMIC2[*3]. VITCOMIC2 is an online web application that performs a rapid sequence similarity search against reference DNA sequence data to estimate the phylogenetic position of each sequence and visualizes the aggregated phylogenetic composition for all metagenomic sequences.

The advantage of VITCOMIC2 is its ease of use, even for beginners. By uploading metagenomic sequence data of a microbial community, VITCOMIC2 automatically performs the analysis and provides the phylogenetic composition in just a few minutes. All operations are performed in a web browser, so even users who are not familiar with bioinformatics analysis can use VITCOMIC2. Our server is used for computation, averaging over a dozen analyses per day. Of course, I also use it for my own research. Now that metagenomic analysis has become widespread, there are many other good analysis tools that complement VITCOMIC2. Since phylogenetic composition data are the basic data for metagenomic analysis, researchers now use different tools depending on the purpose of the analysis.

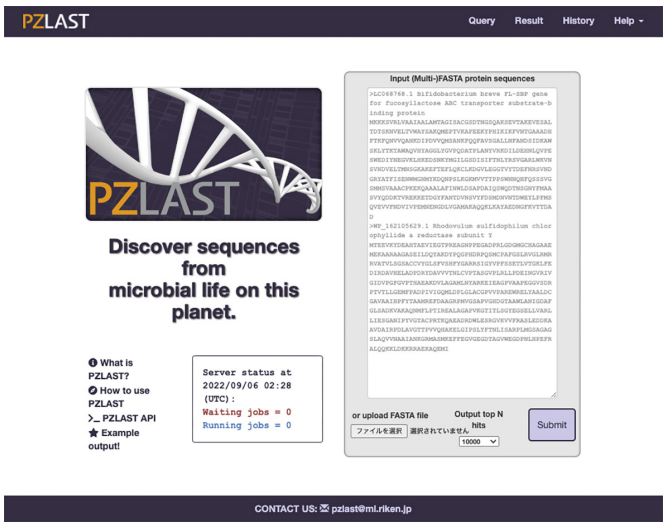


PZLAST Enables Similarity Searches and Integrated Databases of Metagenomes

In metagenomic analysis, the amount of data collected from several hundred samples is enormous, exceeding a terabyte. Public nucleotide sequence databases already contain metagenomic sequence data from several hundred thousand samples. Even if you want to perform a sequence similarity search against terabytes of reference sequence data, existing commonly used tools such as BLAST cannot be used for this purpose. Therefore, we have developed a web application called “PZLAST”. This tool searches for sequence similarity in approximately 3.6 terabytes of amino acid sequence data derived from existing metagenome samples[*4]. These 3.6 terabytes consist of 63.6 billion predicted proteins and approximately 2.4 trillion amino acid residues. There is almost no other sequence similarity search tool besides PZLAST that can accurately search such a large amount of metagenomic sequence data in a few minutes.

The metagenomic sequence data being produced and accumulated daily is a treasure trove of diverse microbial genes and proteins. However, due to the huge amount of data, there has been no tool to search this huge amount of metagenomic sequence data. PZLAST, which is available online, has a very simple metagenomic sequence similarity search function. It can be used to search for protein sequences similar to the enzyme that performs similar functions to existing enzymes but differs in its active temperature, pH, etc.

We have focused on the development of these analytical tools, but in the future, we would like to focus on the development of databases that will serve as the basis for different types of research. We are already developing the “Microbiome Datahub”, an integrated database of metagenomic data and bacterial genomes reconstructed from metagenomes. Bioinformatics techniques have improved in recent years to allow the reconstruction of genome sequences from metagenomic data, albeit imperfectly, of individual bacteria living in specific environments. A wide variety of bacterial draft genome sequences can now be obtained without cultivation, but these metagenome-derived genome data are not well organized, as the analysis methods of the data vary from paper to paper. The Microbiome Datahub we are developing will solve this problem, and in the future, we hope to integrate metagenome-derived genome data and the genome data from individual isolate bacteria to develop an integrated genome/metagenome database that covers nearly all of the bacterial diversity that exists on Earth.



Inference of Phylogenetic Relationships of Extinct Animals Using Ancient DNA

In parallel, we are also studying on ancient DNA analysis, using DNA extracted from the remains of extinct animals. Ancient DNA analysis is similar to metagenomics in that the DNA fragments obtained are short fragments derived from a large number of species. Ancient DNA analysis targets DNA derived from organisms that died more than a few decades ago, but there is no clear standard for what qualifies as “ancient”. Fossil-derived DNA has been analyzed since the 1980s. The development of NGS has accelerated this trend, and ancient genomics of large numbers of samples is now being conducted. Ancient DNA analysis allows us to predict the phylogenetic relationships of extinct organisms with living organisms, and the genetic diversity of the extinct species during different time periods. Because the phylogenetic information of many extinct animals is inferred only from fossil morphology, molecular phylogenies based on ancient DNA analysis are valuable. We have so far analyzed ancient DNA from three species: *Aepyornis* from Madagascar(*5), the brown bear from Honshu, Japan(*6), and the Japanese wolf(*7).

Aepyornis, also called the “elephant bird,” was a large flightless bird that is thought to have become extinct several hundred years ago after humans arrived in Madagascar. The largest *Aepyornis* (*Aepyornis maximus*) was more than 3 meters tall, making it one of the largest birds. Giant flightless birds such as ostriches and emus, and the flying birds of South America (Tinamous) belong to a group called *Palaeognathae*. We wanted to infer where *Aepyornis* locates within the *Palaeognathae* lineage using ancient DNA analysis.

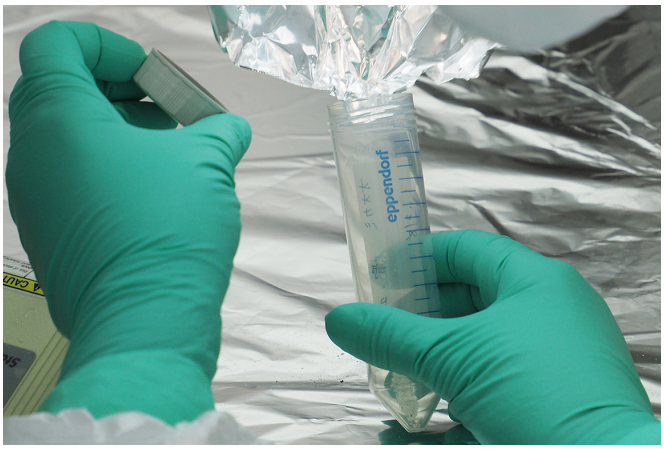
Aepyornis DNA was extracted from bones preserved by a university collaborator in Madagascar and sequenced at the NIG and other institutions. Ninety-nine percent of the DNA obtained was from microbes. However, after removing the microbial DNA from the sequencing data, partial nuclear gene sequences and complete mitochondrial genome sequences were obtained. Comparing these sequences with those of known Palaeognathae, we found that the closest relative is the kiwi, a small flightless bird that lives in New Zealand. Since Madagascar and New Zealand were connected by a continent in prehistoric times, it is thought that a common ancestor from Madagascar and New Zealand was present on both islands and evolved independently on each. We are proud of this finding, which we believe was made possible by ancient DNA analysis.

We wanted to verify whether ancient DNA analysis of an even older era was possible, so we analyzed brown bears. In modern Japan brown bears live only in Hokkaido, but tens of thousands of years ago, they also lived in Honshu. Brown bear fossils excavated from Honshu are larger in size than bones from modern Hokkaido bears, and it has remained unclear what lineage the Honshu brown bears belonged to, when they arrived, and where they came from. Therefore, we performed a similar analysis as for *Aepyornis*, using 32,500-year-old brown bear fossils from Honshu. We found that the brown bears of Honshu are an unknown lineage, distinct from the bears of Hokkaido or any other region of present-day Japan. We also found that the existing brown bear group from southern Hokkaido is a sister group to the Honshu brown bear, and that the two diverged about 160,000 years ago. These findings, together with the known fossil record, suggest that brown bears migrated from Eurasia to Honshu at least twice.

Sampling bones for Ancient DNA analysis in a cave in Ueno Village, Gunma Prefecture

Having demonstrated success in analyzing ancient DNA, we next decided to investigate the origin of the Japanese wolf. The very small Japanese wolf, which is considered to have been endemic to Japan, was once widely distributed in Honshu, Shikoku, and Kyushu, but became extinct about 100 years ago. In addition, fossil evidence suggests that one of the world's largest wolves lived in Honshu more than 20,000 years ago. However, the origin of the small Japanese wolf

and its phylogenetic relationship to the giant wolf were not known. We conducted ancient DNA analysis using nuclear and mitochondrial DNA on the remains of a 35,000-year-old giant wolf and a 5,000-year-old fossil of a small Japanese wolf. We found that the small Japanese wolf originated from the interbreeding of the “old” (57,000-35,000 years ago) giant wolf lineage and the later (37,000-14,000 years ago) wolf lineage.



Sampling of wolf bones for DNA analysis.

We are now in the process of analyzing the ancient DNA of Naumann elephants, which lived tens of thousands of years ago. Large numbers of Naumann elephants have been excavated from various locations, and fossil specimens are on display in museums, but there have been no successful genome analyses until now. We have already obtained interesting results on their phylogenetic relationships and are writing a paper on this topic.

Bioinformatics Research in Japan Faces a Critical Situation

In terms of both human resources and budget, bioinformatics research in Japan is in a more difficult situation than in Europe, the USA, China and other countries. Due to the recent Machine learning boom, many students who have graduated in bioinformatics tend to choose to work in the companies that use statistical and computer science skills. There are significant differences in salaries and benefits between academic and Machine learning related companies for people with expertise in large-scale data analysis and machine learning. It is also difficult to find long-term, stable employment as an academic bioinformatics researcher.

As research budgets have also been cut, we are now facing a critical situation for the future of life science research in Japan as a whole, including bioinformatics. I have made various efforts, such as holding informatics analysis workshops and co-writing books on bioinformatics analysis, to promote young scientists, but more efforts are needed to improve the situation.

I have been involved in a wide range of projects and activities, working towards my ultimate goal - to unravel the “whole microbial diversity” through metagenomic analysis. The majority of microorganisms remain unknown and unnamed. I would like to continue to explore the relationships between microbes, their environments, their genomes and their ecology.

Of course, still there are limits to what one person can do, so I also actively promote collaborative research. I am still considered a young researcher, and I feel privileged to be in this position, because I can spend a lot of time on research and be very focused on it. However, I try not to be too closed in my own world too much, and I enjoy actively engaging with people from outside to exchange information and enjoy collaborative research.

The National Institute of Genetics has more faculty members than students, which is a good environment for learning and research. Those who want to specialize in bioinformatics are all welcome. Please feel free to knock on our door.

Interviewer:
Naoko Nishimura, Science writer

Photographer:
Mitsuhiko Kurusu, NIG ORD

August 2022

References

*1----------
Asymmetric reproductive isolation during simultaneous reciprocal mating in pulmonates
Biology Letters. 2009 Apr 23; 5(2): 240-3.

*2----------
VITCOMIC: visualization tool for taxonomic compositions of microbial communities based on 16S rRNA gene sequences
BMC Bioinformatics. 2010 Jun 18; 11:332.

*3----------
VITCOMIC2: visualization tool for the phylogenetic composition of microbial communities based on 16S rRNA gene amplicons and metagenomic shotgun sequencing
BMC Systems Biology. 2018 Mar 19; 12(Suppl 2):30.

*4----------
PZLAST: an ultra-fast amino acid sequence similarity search server against public metagenomes
Bioinformatics. 2021 Jul 8; 37(21): 3944–3946.

*5----------

▌Phylogenomics and Morphology of Extinct Paleognaths Reveal the Origin and Evolution of the Ratites

Current Biology. 2017 Jan 9; 27(1): 68-77.

*6----------

▌Ancient DNA reveals multiple origins and migration waves of extinct Japanese brown bear lineages

Royal Society Open Science. 2021 Aug 4; 8 (8):210518.

*7----------

▌Paleogenomics reveals independent and hybrid origins of two morphologically distinct wolf lineages endemic to Japan

Current Biology. 2022 Jun 6; 32(11) 2494-2504.e5