TANIZAWA, Yasuhiro
Assistant Professor

Genome Informatics Laboratory.

Newly developed DFAST, a pipeline for prokaryotic genome annotation enabling anyone to make precise annotations of genes

- equipped with an additional function for auto-creating a DDBJ registration format

Assistant Prof. Tanizawa specialized in inorganic chemistry up to his Master's course and shifted to bioinformatics in his doctoral program after working at his family's company. He originally developed DDBJ Fast Annotation and Submission Tool (DFAST), a flexible pipeline enabling anyone from beginners to researchers familiar with information analysis to analyze microbial genomes simply and with high accuracy. He has been continuing his work on comparative genome analysis of a wide variety of organisms including lactic acid bacteria (*Lactobacillus*), liverworts (*Marchantia polymorpha* L.), and novel coronavirus (severe acute respiratory syndrome coronavirus 2).

Here is a link to DFAST

After many twists and turns, I decided to follow a path toward informatics…

I specialized in inorganic chemistry during the period when I was enrolled at the Faculty of Science, The University of Tokyo, and then in the Master's program at the Graduate School, and performed research on the surface structures of catalysts, among others. After that, I joined my family's business and engaged in work including the building of internal databases, taking advantage of self-taught knowledge. Being fond of playing games and operating computers, I had experience of making programs to develop my own analytical tools during the Master's program. In 2012, I was interested to learn that Prof. Yasukazu Nakamura at the Genome Informatics Laboratory, Department of Informatics, NIG, who was looking for technical staff. I applied for a job at the laboratory and was employed to work on building websites and databases within the Nakamura Group. With this as a starting point, I started to pave my way toward a role of performing research on bioinformatics. In 2013, I entered the doctoral program in Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, and obtained my doctorate.

In the doctoral program, I performed research on high-accuracy analysis of the whole genomes of lactic acid bacteria as the topic for my doctoral dissertation under the direction of Masanori Arita, former professor at The University of Tokyo and current professor at NIG (Biological Networks Laboratory, Department of Informatics) and director of the DNA Data Bank of Japan (DDBJ). The lactic acid bacteria, microorganisms belonging to the Lactobacillales that are extremely diverse, can be classified into as many as 200 species. I performed whole-genome sequencing of individual species of these bacteria and compared the results. I also performed an analysis to identify what types of characteristic genes are found in which groups of these bacteria. At that time, a variety of genome analysis tools was available, but they had a number of problems, such as being inconvenient and previously identified genes turning out to not actually be genes. Besides, we were required to register analyzed genome data in public nucleotide sequence databases including DDBJ, but the protocols for registration were complicated, making the registration of our data difficult.

For this reason, I developed a pipeline, DFAST, aiming to create an easy-to-use, high-accuracy analytical pipeline specifically for lactic acid bacteria. The first version of DFAST had a simple design produced by combining an existing tool with reference data on these bacteria (such as known gene sequences and functions) and a file-creating function for DDBJ database registration. Then, I added the species of analyzable microorganisms to the tool and developed my original elemental technology. In this way, I extended the capabilities of the tool step by step. As the number of users increased, the tool gained a growing reputation because it enabled even beginners to easily register their data in DDBJ. It was also held in high esteem by DDBJ. In this way, my own tool has been adopted by DDBJ as its official tool, being effectively used at present.

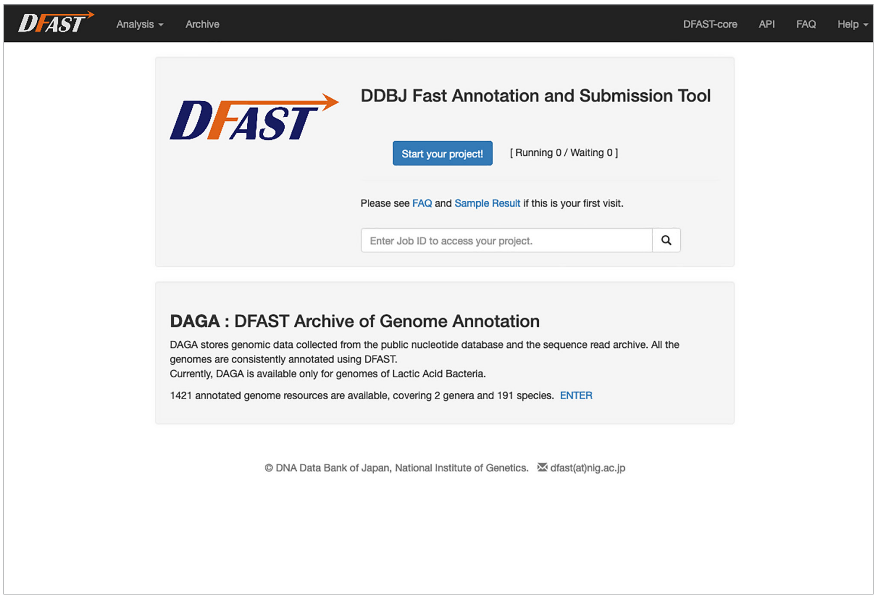
Aiming to enable annotations of genetic structures and functions to be performed easily with high accuracy

The first step of the genome analysis process is the nucleotide sequencing of DNA. In a typical bacterial genome analysis, the second step is to assemble the sequenced DNA fragments together to reconstruct the original whole-genome sequence (genome assembly). With the wide spread of next-generation sequencers capable of performing rapid sequencing and improvements in assembly technology, we became able to acquire highly accurate genome sequences easily; however, the genome sequences at this stage are no more than string data consisting four letters each representing a nucleotide base. The next very important step is to extract biological information (annotation) from the data. Annotation is largely classified into two types: structural annotation for predicting regions encoding genes and functional annotation for predicting their biological roles and functions.

Organisms are classified into eukaryotes, having a nuclear structure in the cell, and prokaryotes, having no such structure. Animals including humans and plants are eukaryotes, while bacteria such as lactic acid bacteria and archaea typified by thermophiles/methanogens are prokaryotes. Structural annotation of prokaryotes is relatively easy because of their small genome size and simple gene structure. On the other hand, functional annotation is based on the comparison with known sequences available from public sequence databases such as DDBJ and UniProtKB, using them as references for picking up genes with similar functions to predict the function of the target potential gene by analogy. However, the results obtained by functional annotation cannot be used without manual modification due to noise and variations such as spelling errors in reference data and genes with the same function being redundantly registered with different gene names. To address these problems, in developing DFAST my colleagues and I started with making highly reliable reference data that can generate reliable results ready to submit to DDBJ without any further modification.

The initial reference data for lactic acid bacteria were developed by collecting data on the genomes of these bacteria from databases and extracting the gene sequences already identified. Among these data, similar genes were collected and classified into groups (clustering). The names of proteins (gene products) produced based on genetic information were checked for each group and erroneous descriptions were corrected. This work was so unsophisticated, requiring over 2 months to compile a huge amount of data into satisfactory reference data. At that time, reference data for *Escherichia coli*, *Lactobacillus bifidus*, *Helicobacter pylori*, and others were prepared in the same manner. Later, I planned to expand the number of species that could be analyzed by DFAST to all prokaryotes. I collected sequence data of representative genomes from well-studied microorganism species and compiled them into a reference database. Now DFAST can be used to a wider variety of prokaryotes.

The DFAST pipeline is composed of two main elements: one is the web service, which is operated by the user through a web browser (web version), and the other is an annotation engine element (DFAST_core), working inside the Web Version. The web version was designed with the aim of enabling even beginners to operate it intuitively without being confused. In particular, by simply filling in data in the web form according to the given instructions,



DDBJ Fast Annotation and Submission Tool

[Running 0 / Waiting 0]

Please see FAQ and Sample Result if this is your first visit.

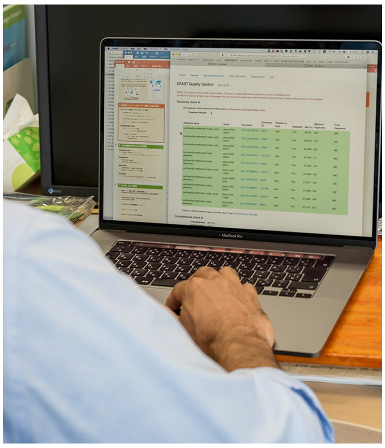
Enter Job ID to access your project.

DAGA : DFAST Archive of Genome Annotation

DAGA stores genomic data collected from the public nucleotide database and the sequence read archive. All the genomes are consistently annotated using DFAST.
Currently, DAGA is available only for genomes of Lactic Acid Bacteria.

1421 annotated genome resources are available, covering 2 genera and 191 species. ENTER

© DNA Data Bank of Japan, National Institute of Genetics. ✉ dfast(at)nig.ac.jp



Microbial genome analysis pipeline, DFAST, enabling the user to perform a series of tasks of assembling, annotating, and registering in DDBJ.
[DDBJ Fast Annotation and Submission Tool (DFAST)]

Yasuhiro Tanizawa

Graduated from Department of Chemistry, Faculty of Science, The University of Tokyo. Completed the doctoral course at The University of Tokyo Graduate School of Computational Biology and Medical Sciences. Received a Ph.D. in Science. Since 2012, he has been working at Department of Informatics, National Institute of Genetics (NIG). In 2018, he was appointed as Assistant Professor after working as a technical staff member and a project researcher.

2. Web版とStand-alone版が利用可能
Web version (Graphical user interface face)
Stand-alone
--config sample.cfg
Bioconda からもインストール可能
conda install -c bioconda dfast
https://github.com/nigyta/dfast_core
MacBook Pro

Assistant Prof. Tanizawa conducts all daily research activities on a computer.

users can automatically create a necessary file for registering data in DDBJ. An additional function to manage many jobs executed by the user has been integrated. My experience in developing web service and databases at the Nakamura Group was very helpful in this work.

Another element, DFAST_core acts as an internally working annotation engine. DFAST_core uses MetaGeneAnnotator, a gene region prediction tool originally developed by Hideki Noguchi, a Project Professor at Advanced Genomics Center, NIG, and GHOSTX, a fast search tool for amino acid sequences, developed at Tokyo Institute of Technology. GHOSTX is capable of performing computations faster than the commonly used BLAST by a factor of approximately 10, contributing to a reduction in DFAST execution time. One of the features of DFAST_core is leveraging these softwares developed by Japanese researchers. DFAST_core's flexible integration of existing tools and reference databases enables the user to easily perform customization, providing one of the main advantages of DFAST.

Microbial genome annotation pipeline DFAST was made openly accessible to the public.

In 2016, we reported the web version in our first research paper entitled "Microbial genome annotation pipeline DFAST" (*1). The initial web version had been intended to analyze lactic acid bacteria. Then, an original annotation engine, DFAST_core, was developed to expand the targets for searches to all prokaryotes, resulting in the prototype of the current DFAST (*2). Initially, DFAST was operated using my own server, but, in 2021, the web service moved to the supercomputer at NIG. Consequently, the analytical performance dramatically improved enough to provide the result of analysis of a bacterial genome with a typical size of millions of base pairs in 2 or 3 minutes. Simply by uploading a file containing DNA sequences, analysis can be performed and the results can be provided at high speed, as indicated by the name DFAST; therefore, I also recommend beginners use DFAST. A wide variety of pipelines other than DFAST have been developed. I recommend DFAST as the first step of a series of analysis just to find out what information come out.

DFAST_core has been disclosed for researchers familiar with information analysis to execute directly by command operation. It has advantages such as the pipeline being flexibly customizable, as well as batch data being processible at high speed by a command operation when substantial data are processed in a relatively large-scale analysis project. As an advantage in common with the web version, a file for registration in DDBJ may be automatically created.

Auto-creation of registration format and fast processing appeal to users, achieving the number of usages of 80,000

I have been satisfied to hear that DFAST has been used by many users in Japan and elsewhere. My paper on DFAST is also ranked first in total citations by field among the papers recently published (2016–2021) by researchers at NIG. Web Version has been used approximately 80,000 times and, at present, approximately 2,000 times a month. Overall, 30% to 40% of the users are based outside Japan. DFAST is also used in the Global Catalogue of Microorganisms (GCM) 10K type strain sequencing project, which is an international genome sequencing project aiming to determine the genomes of type strains (strains used as references in species classification) of 10,000 species of microorganism deposited in bioresource preservation/collection centers in various countries. Japan Collection of Microorganisms (JCM), RIKEN BioResorce Center (BRC), and NBRC of National Institute of Technology and Evaluation in Japan are participating in this project. The results obtained at the initial state of this project were reported in papers. This 5-year project is ongoing and some of the genomes determined by the project have been disclosed through DDBJ.

Examples of the use in large-scale genome analysis research include comparative analysis of *E. coli* genomes, which we conducted in joint research with Kyushu University. In this joint research, to acquire detailed information on gene functions, reference gene data on *E. coli* were prepared and DFAST was used to annotate approximately 800 *E. coli* genomes isolated from the bovine intestinal tract. The findings revealed that enterohemorrhagic *E. coli*, which is highly pathogenic to humans, is derived from bovine-residing *E. coli* and that it produces specific virulence factors enabling it to survive in the bovine intestinal tract. During the course of developing DFAST, I found that *Lactobacillus gasseri*, one of the lactic acid bacteria widely used in yogurt products, is largely classified into two groups and proposed a new species named *Lactobacillus paragasseri*. Generally, the name of new species is proposed based on the strain isolated by a microbiologist in a field survey; in contrast, this research has a unique advantage that it focused on information analysis using openly available data obtained from a database. The efforts made together with Dr. Ipputa Tada, an ex-student at Dept. of Genetics of The Graduate University for Advanced Studies, SOKENDAI, bore fruit in the discovery of the new species (*3).

As I mentioned earlier, DFAST's strength lies in the fact that annotation and registration in DDBJ can be easily performed. To register the genomes of bacteria in GenBank, an international genetic sequence database in the U.S., the PGAP pipeline is available. However, no standard tool for registering data in DDBJ was available; I thus take pride in considerably reducing the burden on the user in registering genome sequences via DFAST. Moreover, the annotations made by using DFAST satisfy a given quality standard; accordingly, the burden on the curators, who check the content of data to be registered, on the side of DDBJ may be reduced. Actually, in the year of 2020, the share has been so increased that DFAST was used in 90% of the genomes of bacteria with annotations registered in DDBJ.

At present, to publish papers concerning newly determined DNA sequences in many academic journals, it is necessary to register the data in a public sequence database. I think this requirement is not only a simply obligation, but also very important for researchers. It provides researchers with the advantages of ensuring the reliability and reproducibility of their studies and enabling their own research to influence other studies when researchers refer to their data or reuse it. In that sense, data registration may be considered as an essential part of research publication.

DFAST being improved and further enhanced with the aim of conducting mutation analysis of novel coronavirus

Last year, a function (DFAST_QC) to check the names of species of genomes and data integrity was additionally integrated into DFAST. In sequence databases including DDBJ, assessments are conducted to prevent inappropriate data from being registered; however, data describing an erroneous species name or poor quality genome data may be missed and registered in some cases. We aimed to solve this problem via a new function, DFAST_QC.

In addition, we are endeavoring to integrate DFAST with the DDBJ data registration system. Previously, we made annotations using DFAST and then exchanged e-mails with the person in charge of registration in DDBJ to register base sequences. Successful integration may enable us to transfer the results to DDBJ by one click for data registration. I hope that data assessment, which is currently conducted manually, becomes automated to provide a one-stop service

aimed at enabling the user to seamlessly follow a series of steps from making annotations to data publication.

NIG started to work on analysis of the genomes of new variants of novel coronavirus in cooperation with Shizuoka Prefecture. In this program, the whole-genome sequences of specimens collected in Shizuoka are determined, mutation sites are identified, and annotations are made to these sites to make the data publicly available from DDBJ. I am responsible for data registration in DDBJ in this program and have already built a customized version of DFAST for novel coronavirus using the existing pipeline. NIG is equipped with all the facilities for genome analysis, a genome sequencing center Advanced Genomics Center for genome sequencing, a supercomputer capable of large-scale analysis, and DDBJ, a portal for data registration. As a result of this, publishing correct data rapidly may contribute to tracking the route of infection of COVID-19 and identifying its new variants.

Bioinformaticians are a melting pot of a wide variety of backgrounds

Four researchers apart from me work at the laboratory. We are communicating using an online chat tool among laboratory members even before the COVID-19 pandemic, so I do not feel that it is inconvenient teleworking from home, as long as I have access to a computer. The individual researchers are proceeding with their own projects. We can report the progress of our projects and consult or discuss them with others via the chat tool at any time.

I am very grateful to my boss, Prof. Nakamura, for letting me pursue my own research. In the laboratory, a project intended to support other researchers in performing genome analysis (Grant-in-Aid for Scientific Research on Innovative Areas -“Platform for Advanced Genome Science (PAGS)”) is also ongoing, through which we perform comparative genome analysis of liverworts (*Marchantia polymorpha* L.), diatoms (Bacillariophyceae), and others in cooperation with Prof. Nakamura.

I entered the doctoral program after working at my family’s company. Since then, I have been working as a researcher in the field of bioinformatics. As my background shows, one’s path in life can change at any time. I believe that there are opportunities for anyone who seeks a change of direction to go toward their goal. Researchers in the field of bioinformatics in particular have a wide variety of backgrounds. Some researchers in this field have always been engaged in informatics from the beginning of their career. Some other researchers have shifted their field of research to informatics because they had to perform information analysis when conducting experimental work; with this as a starting point, they became engaged in informatics. The background of the members of the Nakamura Lab is also highly diverse.

Recently, various tools for genome analysis have become available, lowering the barrier to conducting information analysis, which may have contributed to an increase in the number of researchers in this field, as mentioned above. Researchers who disseminate information on how to use the tool and its usability have considerably increased. From my experience, for researchers unfamiliar with biology to enter the field of bioinformatics, it is important to acquire the necessary knowledge from one’s colleagues little by little.

Interviewer: Naoko Nishimura, a science writer
Pictures taken by: Mitsuhiko Kurusu, NIG ORD
Designed by: Naho Kanazawa, NIG ORD
July 2021

▌Cited references

*1----------

▌DFAST and DAGA: web-based integrated genome annotation tools and resources
Tanizawa, et al. 2016 BMFH.

*2----------

▌DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication
Tanizawa, et al. 2018 Bioinformatics.

*3----------

▌Revealing the genomic differences between two subgroups in Lactobacillus gasseri
Tada et al. 2017 BMFH.

▌Lactobacillus paragasseri sp. nov., a sister taxon of Lactobacillus gasseri, based on whole-genome sequence analyses
Tanizawa et al. 2018 IJSEM

▌Other related links

▌Bioinformation and DDBJ Center

▌Advanced Genomics Center

▌The Department of Genetics of The Graduate University for Advanced Studies, SOKENDAI

▌"Platform for Advanced Genome Science" (PAGS)