



The Suntory Foundation Uses Fusion ioMemory™ to Process DNA Data at High Speed

Solution Focus

- Life Sciences/Bioinformatics
- Big Data analysis

Summary of Benefits

- High-speed processing of vast amounts of data
- Establishing operational infrastructure at appropriate cost
- 40X reduction in processing time

Products

- Fusion ioMemory™ ioDrive2®

“With Fusion ioMemory, we determined that we would be able to process massive amounts of data at high speeds without developing software to perform complex parallel processing.”

Mr. Satoshi Shiraishi
Function Researcher, Division
of Integrative Biomolecular
The Suntory Foundation for Life Sciences

Summary

Fusion ioMemory™ from SanDisk® was chosen by The Suntory Foundation for Life Sciences research laboratory in its efforts to uncover unknown genomes. The lab has elected to use a Fusion ioMemory PCIe application accelerator to perform high-speed processing of large amounts of DNA data that are read by a next generation sequencer.

Background

The Suntory Foundation for Life Sciences was founded in 1946, based on the vision of Mr. Keizo Saji (1919-1999), who believed that “the future Japan should contribute to the peace and prosperity of the world through academics and culture.”

Through the academic promotion of sciences related to bio-organics, the Foundation’s philosophy is to contribute to the happiness and prosperity of mankind. In 2012, the Foundation introduced a next generation sequencer, which can read nucleotide sequences at a very high speed. In order to process the massive amounts of data obtained through this, the foundation also introduced SanDisk’s Fusion ioMemory.

The Challenge

A high speed I/O was essential for analyzing massive nucleotide sequence data from a next-generation sequencer

Mr. Honoo Satake, the director and senior researcher of the Division of Integrative Biomolecular Function of the Bioorganic Research Institute, explained, “The goals of our Institute are nothing less than explaining the mechanisms of a variety of biological activities of natural organic compounds, discovering the essence of co-existence and diversity among species, contributing to humanity, and realizing a safe and secure society. Specifically, we have established two main research themes—explaining the mechanisms of the biological activities of natural organic compounds, and approaching the essence of co-existence and diversity among species.”

Among these, the Division of Integrative Biomolecular Function, which works on approaching the essence of co-existence and diversity among species, introduced a next generation sequencer in 2012, in order to perform research that makes full use of state-of-the-art bioinformatics. Bioinformatics is a new research area that fuses biology and information science.

“By adopting Fusion ioMemory, we became able to complete processing that previously took 24 hours in just 30 minutes to an hour. By making our processing time about 40 times faster, we were able to devote more time to analysis and research.”

Mr. Satoshi Shiraishi
Function Researcher, Division
of Integrative Biomolecular
The Suntory Foundation for Life Sciences

Bioinformatics makes full use of technologies for large-scale data analysis, in addition to the experimentation that was conventionally central to biological research, in order to discover new solutions. In order to obtain the bio-information of the structures of genes (DNA) and proteins, etc., which forms the basis for this research, a next generation sequencer that can analyze DNA information on organisms was essential.

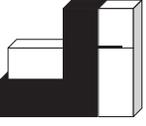
Researcher Satoshi Shiraishi at the Division of Integrative Biomolecular Function explained the relationship between the sequencer introduced by the Institute and data analysis. “When we analyze the DNA nucleotide sequences from an organism’s cells and tissue using a next generation sequencer, we receive an output of sequence data for approximately eight billion sequence pairs in a single operation. This equals several tens of gigabytes of data. We process these large-scale nucleotide sequence data with analytics software, and so we were looking for high-speed I/O functionality, in addition to high-speed calculating ability.”

The Solution

Using Fusion ioMemory to challenge large-scale data analysis in search of unknown base sequences

When analyzing nucleotide sequence data with a next-generation sequencer, information on how the four types of bases (ATGC) are arranged is read from the cells and tissue of the research subject. The next-generation sequencer that the Institute introduced has the ability to read 40 million fragments from several hundred sequence pairs simultaneously, and the data volume from a single analysis can reach several tens of gigabytes. The actual data obtained are massive quantities of alphabetical lists such as ACTACGACGTAAAC.

“In addition to the amount of data we analyze, many of the organisms that we research were yet unknown genome sequences. Therefore, we had to search for the correct sequences from what was basically a blank slate,” Mr. Shiraishi told us about some of the initial challenges. “Because of this, we were often unable to process the data just with the existing genome analysis data and software such as BLAST, and we had to develop new programs and find ways to process large-scale data in a way that would let us check nucleotide sequences by brute force.”



No. of base reads per data segment: 50 - 100b
 No. of data segments per run: Approx. 2 billion reads
 No. of base single reads per run: Approx. 75 - 100 Gb
 Pair ends: Approx. 150 - 200 Gb

Output of massive sequence data from next-generation sequencer

```

ACTACGACGTAAACNGTGAAGGCTTGATATGTGGTAACT
CGTAAACCGTGAAGGCT-GATATGTGGTAACTCAGCCA
CGTGAAGGCT-GATATGTGGTAACTCAGCCAGCACT
GCT-GATATGTGGTAACTCAGCCAGCACTATGCTC

ACTACGACGTAAACNGTGAAGGCTTGATATGTGGTAACT
CGTAAACCGTGAAGGCT-GATATGTGGTAACTCAGCCA
CGTGAAGGCT-GATATGTGGTAACTCAGCCAGCACT
GCT-GATATGTGGTAACTCAGCCAGCACTATGCTC

ACTACGACGTAAACNGTGAAGGCTTGATATGTGGTAACT
CGTAAACCGTGAAGGCT-GATATGTGGTAACTCAGCCA
CGTGAAGGCT-GATATGTGGTAACTCAGCCAGCACT
GCT-GATATGTGGTAACTCAGCCAGCACTATGCTC

```




Hardware: HP ProLiant DL980 G7
 Mounted Processor: Intel® Xeon®
 Processor: E7-4870 2.4GHz 8 processors x 80 cores
 Memory: 1TB; Options: Fusion ioMemory ioDrive2 1.2TB x 2 units
 OS: Red Hat Enterprise Linux 6.2

Utilizes a high-performance server and high-speed I/O to analyze large-scale output data and interpret organism mechanisms

Links sequences
 ACTACGACGTAAACNGTGAAGGNT-GATATGTGGNAACTCANCCAGCACTATGCTC

Edits
 ACTACGACGTAAACCGTGAAGGCTGATATGTGGTAACTCAGCCAGCACTATGCTC



Fusion ioMemory™ ioDrive2®

When they began their full-scale research of bioinformatics using the next-generation sequencer, Mr. Shiraishi said he drew from his previous work experience at the University of Kyoto. “At Kyoto University, we were creating an infrastructure to screen our target chemical compound from an astronomical number of chemical compounds in a short period of time for drug discovery. We determined that if we had Fusion ioMemory at this laboratory, we would be able to process massive amounts of data at high speeds without having to develop software to perform complex parallel processing,” he explained, regarding the team’s reasons for choosing Fusion ioMemory as the architecture solution.

Mr. Satake added, “We began to be concerned from 2012 that if we did not establish a life sciences research environment that made full use of data analysis, we would become completely unable to perform cutting-edge research in just five years. In order to draw life science interpretations from the massive data analyzed by the next-generation sequencer, it was essential for us to establish a system that could analyze data at high speeds.”

The Result

Bioinformatics demonstrates results even in analyzing hops genomes

“By adopting Fusion ioMemory, we became able to complete processing that previously took 24 hours in just 30 minutes to an hour,” Mr. Shiraishi said about the results of the decision. “By making our processing time about 40 times faster, we were able to devote more time to analysis and research.”

After assembling an environment where they were able to use the HP ProLiant DL980 G7 and Fusion ioMemory to analyze large-scale nucleotide sequence data obtained via the next-generation sequencer, the bioinformatics from the Institute were noticed by the bioresearch department of the Suntory Global Innovation Center (Ltd.). They jointly began an analysis of hops genomes, and published a research paper in a major international journal on plant science.

“Hops are an important plant because it is a component of beer’s aroma. At Suntory, we use the Saaz species of hops that are grown in the Czech Republic. We research the Saaz species, wild species, and domestic hops on a genetic level. If we hadn’t been equipped with a next-generation sequencer and a high-speed analytic environment, we would not have been able to obtain our research results. In the future, the genome information that we analyzed may be able to be used to create a new breed of hops.” Mr. Satake said of the significance of his analysis.

Outlook

Working to transfer skills to allow more researchers to use technology

“In the life sciences, there are ‘wet’ research areas that center around experiments, and ‘dry’ research areas in which computers are used. It used to be difficult for a researcher to become familiar with both fields. However, in the future, researchers will need to obtain results in short periods of time by combining the advantages of both areas. With this new system, we have established a processing flow in which researchers can share analytical results with each other. Also, researchers who used to only work in ‘wet’ areas will be able to perform analyses and predictions by combining with ‘dry’ areas, which will lead to big breakthroughs. It’s our goal in the future to continue to educate researchers in both ‘dry’ and ‘wet’ research areas,” explained Mr. Shiraishi.

“Currently, we are working on researching receptors, which I was also researching at the university. Receptors are like switches for living things. If you can figure out what receptors respond to what substances, it can help us understand biological mechanisms. We will continue to use next-generation sequencers and Fusion ioMemory, as well as analytical software, to continue to promote research to contribute to humanity,” Mr. Shiraishi said of his aspirations.

Contact information

fusion-sales@sandisk.com

Western Digital Technologies, Inc.

951 SanDisk Drive
Milpitas, CA 95035-7933, USA
T: 1-800-578-6007

Western Digital Technologies, Inc. is the seller of record and licensee in the Americas of SanDisk® products.

SanDisk Europe, Middle East, Africa

Unit 100, Airside Business Park
Swords, County Dublin, Ireland
T: 1-800-578-6007

SanDisk Asia Pacific

Suite C, D, E, 23/F, No. 918 Middle
Huahai Road, Jiu Shi Renaissance Building
Shanghai, 20031, P.R. China
T: 1-800-578-6007

For more information, please visit:

www.sandisk.com/enterprise

SanDisk®
a Western Digital brand

At SanDisk, we’re expanding the possibilities of data storage. For more than 25 years, SanDisk’s ideas have helped transform the industry, delivering next generation storage solutions for consumers and businesses around the globe.

The performance results and cost savings discussed herein are based on internal testing and use of Fusion ioMemory products. Results and performance may vary according to configurations and systems, including drive capacity, system architecture and applications.

©2016 Western Digital Corporation or its affiliates. All rights reserved. SanDisk is a trademark of Western Digital Corporation or its affiliates, registered in the United States and other countries. Fusion ioMemory, ioDrive, and others are trademarks of Western Digital Corporation or its affiliates. Other brand names mentioned herein are for identification purposes only and may be the trademarks of their holder(s). Suntory_CS_SanDisk_v5 06/03/16 5076EN