Technical Report

Development of Statistical Method for Identification of Microorganisms Responsible for Wastewater Treatment

Toshikazu FUKUSHIMA* Junichi NAKAGAWA
Shuichi KAWANO Mamoru OSHIKI

Abstract

Activated sludge process, which is a biological wastewater treatment process, has a history of over 100 years and has also been applied to a part of the wastewater from the steelmaking industry. However, the microorganisms responsible for the wastewater treatment have been regarded as a black box, which is one of the reasons for the instability of the process. Next generation sequencing technology has revealed that the microbial community in the process is very complex. However, the relationship between wastewater treatment and microorganisms is still unknown due to the complex community, so called "big data". Therefore, we developed novel statistical methods by transdisciplinary approaches, namely integrating microbiology and mathematics, and successfully identified important microorganisms responsible for wastewater treatment. This method is expected to contribute to the further stabilization and efficiency of biological wastewater treatment processes.

1. Introduction

Activated sludge process is a representative biological wastewater treatment process and was commercialized more than 100 years ago. It has been widely applied all over the world to the treatment of wastewater, such as sewage and industrial wastewater.[1] **Figure 1** shows the treatment flow of a conventional activated sludge process schematically. Wastewater is inflowed into an aerobic (oxygen-dissolved) reaction tank, called an aeration tank, where microorganisms are acclimatized beforehand. The microorganisms degrade pollutants, mainly organic matter. In the next settling tank, the treated water is separated from the microorganisms and is discharged. Microorganisms that perform this treatment are designated as activated sludge. In the steel industry as well, activated sludge process has been applied to the treatment of coke oven wastewater. Activated sludge process for coke oven wastewater was reported more than 40 years ago.[2] Thus, activated sludge process in steel industry also has a long history.

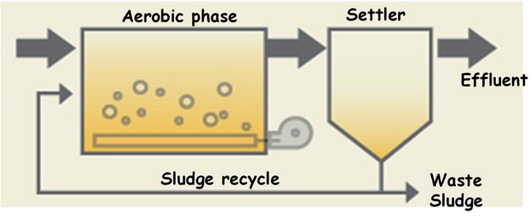

Fig. 1 Schematic of conventional activated sludge process

As described above, activated sludge process has a long history and is one of the biological processes that has been widely applied for practical use, whereas the physiology and ecology of the microorganisms responsible for the treatment have only been partially elucidated. As shown in **Fig. 2**, the quality of treated wastewater depends on the configuration of activated sludge process and the characteristics of wastewater. Since the mechanism of the process is regarded as a black box, further improvement in the treatment efficiency is often difficult. Specifically, the change of operational conditions results in the transition of dominant microbial species. As a consequence, the treatment efficiency also changes. Better understanding of the ecology of microorganisms responsible for the wastewater treatment enables the operation of the wastewater treatment process with high efficiency and stability.

Advancements in next-generation sequencing technology have made it possible to comprehensively analyze microbial communities in the biological wastewater treatment processes. In recent years, the

* Senior Researcher, Ph.D. (Environmental Studies), Environment Research Lab., Advanced Technology Research Laboratories
20-1 Shintomi, Futtsu City, Chiba Pref. 293-8511

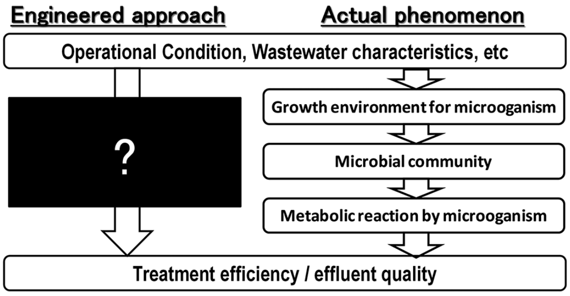
Engineered approach
Actual phenomenon
Operational Condition, Wastewater characteristics, etc
?
Growth environment for microoganism
Microbial community
Metabolic reaction by microoganism
Treatment efficiency / effluent quality

Fig. 2 Conceptual flow diagrams of biological wastewater treatment

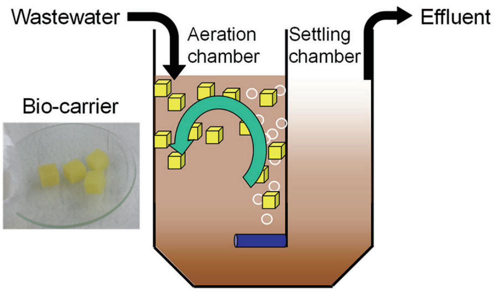

Fig. 3 Schematic diagram of laboratory-scale moving bed biofilm reactor (MBBR)

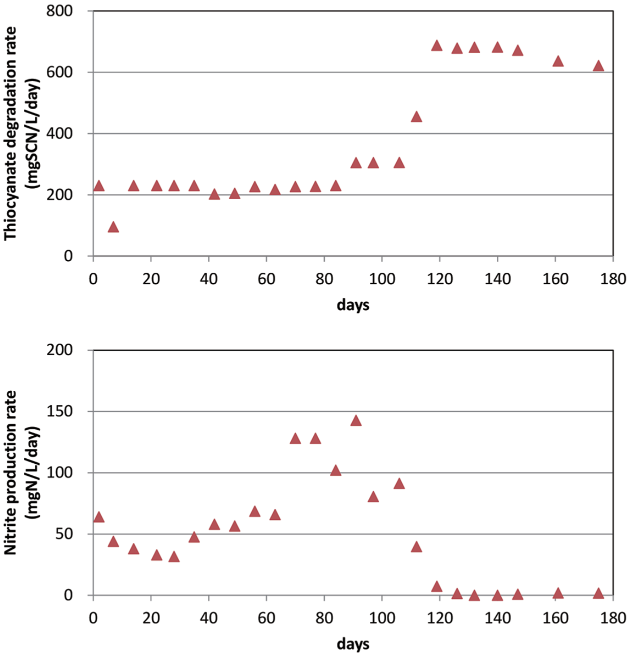

Fig. 4 Examples of estimated degradation/production rate in MBBR

sequencing technology has become more accessible as cost has dropped. The sequencing technology has revealed that activated sludge is composed of the members of several thousands to several tens of thousands of species of microorganisms. It is still difficult to elucidate the causal relationship with wastewater treatment from such vast and complicated microbial information. For example, the microorganisms present in the wastewater treatment process are thought to contribute significantly to wastewater treatment. What is the minimum relative abundance required for microorganisms to be called dominant species? Which microorganisms contribute to the treatment of which pollutants? Researchers can only answer these questions according to their own experience and judgment.

In view of the above-mentioned situation, we came to develop new statistical methods[3)] by integrating biology and mathematics, aiming at objectively identifying the important microorganisms engaged in wastewater treatment from "big data" obtained by next-generation sequencing. The details are reported below.

2. Main Subjects

To develop the new statistical methods,[3)] we first constructed experimental data. To control environmental conditions such as water temperature, we operated a laboratory-scale wastewater treatment process and obtained water quality and microbe data. The treated water quality and microbial communities were fluctuated by changing the operational conditions. We examined the validity of the statistical analysis methods by comparing the experimental data and predicted data from the statistical analysis results.

2.1 Construction of experimental data

2.1.1 Wastewater treatment data

A laboratory-scale moving bed biofilm reactor (MBBR) was operated, and water qualities of influent and effluent were examined. In activated sludge process, it is necessary to return activated sludge settled in the settling tank. In the MBBR process, most of the microorganisms are attached to carriers and thus can be retained in the aeration chamber. In this report, we adopted the MBBR process instead of activated sludge process to simplify the operation. As shown in **Fig. 3**, the MBBR consists of a 3.4 L aeration chamber and a 2.1 L settling chamber. The 1 cm^3 -sponge carriers with a volume ratio of about 33% of the aeration chamber and activated sludge collected from a coke oven wastewater treatment process were inoculated into the aeration chamber. The synthetic coke oven wastewater was prepared by mixing natural seawater and industrial water at a ratio of 60:40 and adding phenol, thiocyanate, and thiosulfate. Ammonia, carbonate, and phosphorus were also added. Oxygen was supplied by aerating the aeration chamber at an air flow rate of 4 L/min with an air pump. The sponge carriers were fluidized by aeration. Water temperature was kept at 30°C by a temperature controller and a submerged heater. The pH was constantly measured and adjusted to 7.5 or higher by adding a sodium hydroxide solution as needed.

During the operation period, the inflow rate of wastewater was controlled to vary the hydraulic retention time (HRT) and the treatment efficiency. The quality of the treated water was periodically analyzed. The degradation or production rate (mg/L/day) of each pollutant was calculated. **Figure 4** shows an example of the degradation or production rates calculated from the water quality data. In the MBBR process, thiocyanate was used as the pollutant. The degradation rate was increased by shortening the HRT, namely increasing the loading rate. Some of the ammonia is oxidized to nitrite by ammonia-oxidizing bacteria. For this reason, the microbial reactivity of ammonia was calculated as the nitrite production rate.

2.1.2 Microbe data

Microbial community analysis was performed by Nippon Steel Eco-Tech Corporation. Sponge carrier samples were collected from the MBBR process. The sampled sponge carriers were cut into four pieces. DNA was extracted from the biomass attached to the carrier by using Extrap Soil DNA Plus ver.2. Using the obtained DNA, the 16S rRNA gene V4-5 regions of bacteria and archaea were amplified by the polymerase chain reaction (PCR). Subsequently, the sequencing of the PCR amplicon was performed by using the MiSeq

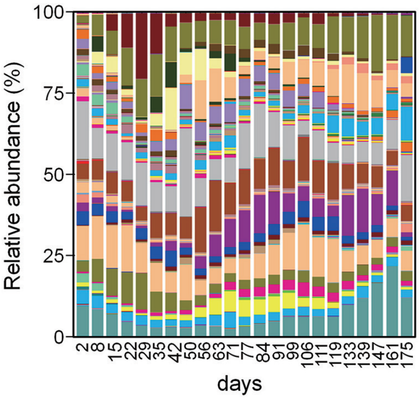

Fig. 5 Example of microbial community in MBBR

System (Illumina). Sequences with a distance-based similarity of 97% or greater were grouped into operational taxonomic units (OTUs) by Quantitative Insights Into Microbial Ecology (QIIME).[4] For convenience, OTUs were regarded as the microbial species. The microbial communities were determined from the obtained OTUs.

An example of the microbial community analysis results is shown in **Fig. 5**. The relative abundances of the microbial species (OTUs) are shown. There are at most several dozens of the microorganism species that can be visually recognized. In practice, however, there were several thousand microbial species. It is virtually impossible to study the several thousand microbial species present in the wastewater treatment process. On the other hand, the minimum abundance ratio of microorganisms to be regarded as important microorganisms depends on the wastewater treatment process and the concentration of pollutants. The identification of important microorganisms can only be done subjectively. Microbial community analysis is the first step to understanding the microorganisms that have been regarded as a black box. The change of relative abundances shown in Fig. 5 is considered to have a causal relationship with wastewater treatment.

2.2 Statistical analysis

As mentioned above, even in the laboratory-scale MBBR with controlled wastewater composition and operating conditions, several thousands of microbial species existed. It is difficult to identify important microorganisms responsible for wastewater treatment.

Regression analysis can be considered for analyzing the relationship between wastewater treatment and microorganisms. The relationship between the wastewater treatment performance and the microorganisms is analyzed, for example, with the degradation (production) rate of pollutants as the response variable y and the microorganism species responsible for the treatment as the explanatory variable x. In a conventional regression analysis, coefficients are applied to all the microbial species or explanatory variables. This is a problem. In the case of thiocyanate degradation, all the microorganisms that exist in the laboratory-scale MBBR do not have the ability to degrade thiocyanate. There are also many microorganisms involved in the degradation of other pollutants. With regression analysis, however, some coefficients are also applied to microorganisms that are not involved in any decomposition. Therefore, the decomposing microorganisms are underestimated and the non-decomposing microorganisms are overestimated. In addition, the number of microbial species is 100 times or more than the number of wastewater treatment data, namely the number of samples collected. This makes it all the more difficult to apply accurate coefficients. In other words, it is difficult to apply the regression analysis to a wastewater treatment process in which there are multiple types of pollutants and a huge number of microbial species exist. For this reason, we developed new statistical methods for identifying important wastewater treatment microorganisms. The developed Lasso+bootstrap method and its improved method will be explained below.

2.2.1 Lasso+bootstrap method

Based on the Lasso (least absolute selection and shrinkage operator) estimation that is a sparse estimation,[5] we attempted to identify the "important microorganisms responsible for wastewater treatment" that degrade major parts of the pollutants. Sparse is an English word that means "scant" or "scarce". The number of existing species is enormous as with the microbe data in this report, but the number of microbial species that actually contribute significantly to the wastewater treatment is small. When these conditions are considered, this approach applies here. The sparse estimation is a method that performs regression coefficient estimation and variable selection at the same time. Here, we select important microorganisms. In addition to the Lasso estimation, a stability analysis incorporating the Bootstrap method[6] was performed to consider the stochastic variation (variability) of the microbe data. Since the Bootstrap method is often used for the stability analysis of phylogenetic trees that show the relationship between microbial species, it is a method familiar to microbiology experts.

The flow of the analysis is as follows:

1. Generate Bootstrap samples for the obtained experimental data (the number of samples is arbitrary but is 1 000 sets here).
2. Apply the Lasso estimation to each sample and count the occurrence frequency of the microorganism species with a selected frequency of 1 or more.
3. Extract microorganism species whose occurrence exceeds any occurrence frequency (600 times in this case).
4. Construct a linear regression model by using the extracted microorganism species and calculate the R^2 value by the cross-validation method.
5. Compare the R^2 value and the measured water quality data with the linear regression model and consider the results.

An example of the results is shown in **Fig. 6**. Although 3 752 microbial species were detected by the microbial community analysis, the Lasso+bootstrap method found that only 3 to 6 types of microorganisms were considered to be significantly involved in the decomposition of the pollutants. The linear regression model was cross-validated by using the identified important microorganisms as explanatory variables. As shown in Fig. 6, the predicted degradation/production rates were similar to those observed. The Lasso +bootstrap method was applied to the removal of a total of 6 types of pollutants in two laboratory-scale MBBRs operated under different conditions. The R^2 values were large at 0.61 to 0.77. An excellent linear regression model was considered to have been constructed. We surveyed the literature on the ecophysiology of the species phylogenetically close to the identified important microorganisms. We could not obtain information on many microorganisms, but we found that some of the species identified as important microorganisms for nitrite production were ammonia oxidizing bacteria that oxidize ammonia to nitrite. These were considered appropriate analysis results from a microbiological point of view. It should be noted however that the identified microbial species are related to the variation in the degradation rates and do not directly indicate the degradation of pollutants. Such microorganisms may have been identified as important microorganisms because they are in a symbiotic rela-

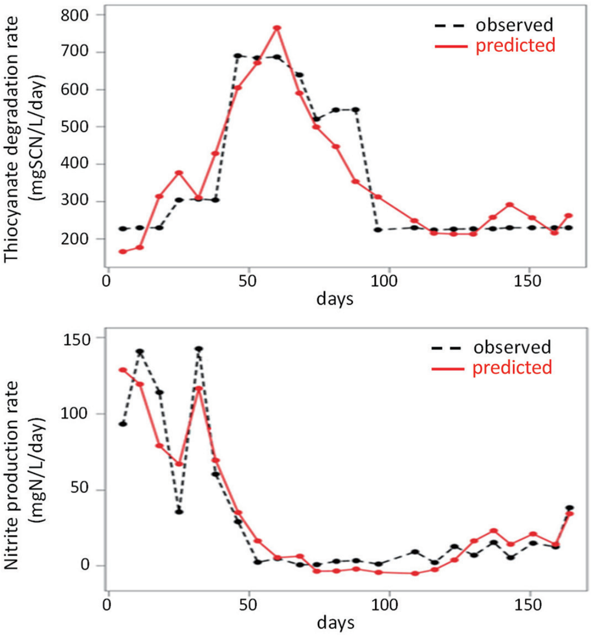

Fig. 6 Analysis results by Lasso+Bootstrap method

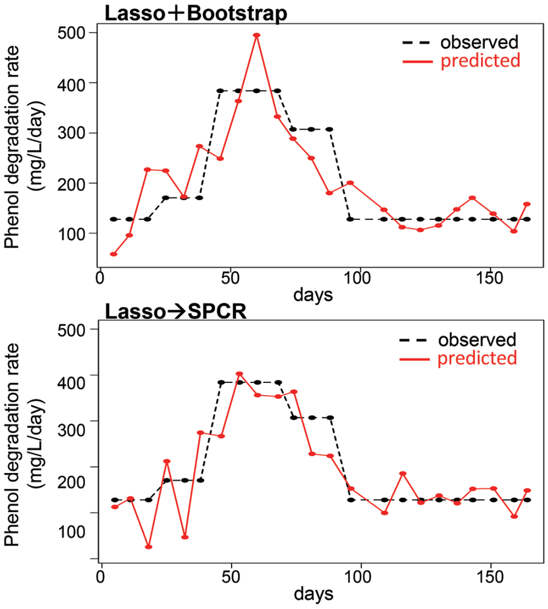

Fig. 7 Comparison of analysis results by Lasso+Bootstrap method and SPCR method

tionship with non-degrading microorganisms, for example. Despite these issues, when only a few species of microorganisms are identified, subsequent biological investigations will be practically feasible, and new discoveries can be expected.

2.2.2 SPCR method

As mentioned above, we succeeded in identifying important microorganisms by the Lasso+bootstrap method and in demonstrating by comparison with the measured data that the Lasso+bootstrap method is a highly accurate method. On the other hand, we could not determine by the literature survey whether many of the identified microorganisms actually have the degrading ability and what environment (such as the pH and water temperature) suits them. It is difficult to prove experimentally because of the high difficulty of isolating microorganisms. It is not realistic to clarify the characteristics of all microorganism species. In addition, the situation is more complicated because there may be relationships such as competition and symbiosis between microorganism species.

To understand the relationships between the microorganisms and environmental factors and between the microorganisms themselves and to create regression modelling by using these relationships, we studied the possibility of applying a sparse principal component regression (SPCR) model.[7] Principal component regression involves principal component analysis in the first stage and regression analysis in the second stage. But there is the problem that the principal component scores that become the explanatory variables in the second stage are not designed to match the objective variables at all. On the other hand, the SPCR method can be regarded as single-stage principal component regression modelling. We used this method to grasp the relationships between the environment and the microorganisms and between the microorganisms themselves. We then studied the construction of regression modelling by using these relationships.

To narrow down the microorganism species in the first stage, all the microbial species identified even once were selected from 1 000 bootstrap samples by the Lasso+bootstrap method. Depending on the target data, several dozens of species were selected from several thousands of microorganism data by this operation. Subsequently, SPCR was applied to the selected microbial species. At this time, the number of principal components was set from 1 to 5. The number of principal components with the highest R^2 values was determined by the cross-validation of the SPCR method using the identified microorganism species as explanatory variables.

An example of the results is shown in **Fig. 7**. In the example of Fig. 7, the R^2 value by the Lasso+bootstrap method was 0.61, while the R^2 value by the SPCR method was higher at 0.67. In this way, it found that the application of the SPCR method increases the R^2 value of many, but not all, of the data. Also, the number of principal components differed depending on the data. The relationships between the environmental factors and the microorganisms or between the microorganisms themselves may be clarified. There are such possibilities that microorganisms identified in different principal component axes prefer different environments, have some symbiotic relationship on the same axis, or have relationships with environmental factors or with other microorganisms. So far, it is difficult to consider the results obtained and it is impossible to discuss such possibilities. However, it may be possible to acquire new ecophysiological findings of microorganisms by considering and examining the analysis results by SPCR in greater depth.

2.2.3 Analysis including time series information

The Lasso+bootstrap method and SPCR method described above are not time series analysis. The water quality and microbe data at the same time points are compared. In the actual wastewater treatment process, however, as the number of microorganisms increases or decreases, the wastewater treatment improves or worsens. There may be a time lag between these changes. It has been reported that there are time lags between the changes in environmental factors such as the pH and the changes in wastewater treatment microorganisms.[8, 9] Based on this possibility, we examined the analysis including the time series information.

In particular, as shown in **Fig. 8**, we thought that the water quality at a certain time point may depend on microorganisms at the same time point and one time point before. We studied microbe data

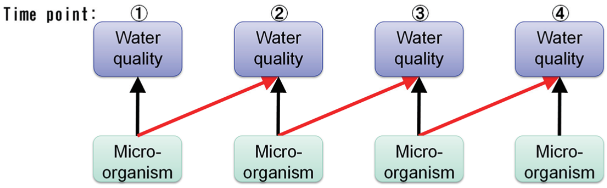

Fig. 8 Conceptual diagram of time series analysis

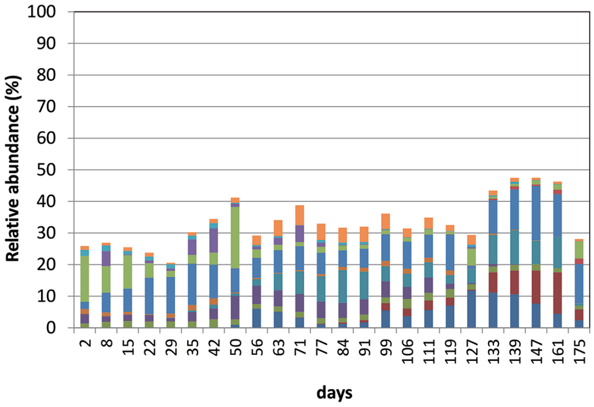

Fig. 9 Relative abundances of microorganisms identified by Lasso+ Bootstrap method

and water quality data at both time points using the developed statistical methods. For an example, the R^2 value was 0.77 in the analysis without time series but improved to 0.90 when the time series analysis was conducted. The accuracy of analysis was found to improve for many data. In addition, the microbial species identified at the same time point or one time point before were different. This is considered to represent the difference in growth rate, for example. The data in this report was collected once a week. If we can collect the data more frequently, for example, once a day, a more detailed consideration may be possible.

2.3 Discussion

Using our developed statistical methods, we were able to screen the important microorganisms related to the wastewater treatment from more than one thousand microbial species to a few species. If there are several species, it is actually possible to study the microorganisms in greater detail and to monitor their abundances by gene quantification.[10)]

From a different point of view, we consider the relative abundance of the identified important microorganisms to all microbial species. As shown in **Fig. 9**, the identified microbial species accounted for only about 20 to 40% of the total biomass. Although more than half of the microorganisms are present, they were not involved in the wastewater treatment. Since this study examined a laboratory-scale MBBR fed with a synthetic wastewater with defined composition, it was unlikely that it contained unknown pollutants. It suggests interesting possibilities such as that there are a considerable number of microbial species that feed on and degrade dead microorganisms or that play a role other than wastewater treatment.

The extracted DNA for microbial community analysis includes not only active microorganisms, but also DNA derived from low activity or dead microorganisms. For this reason, we considered the possibility of obtaining results with higher accuracy by devising experimental methods to obtain microbe data, such as eliminating the DNA derived from dead microorganisms in advance[11)] and targeting the RNA related to activity.[12)]

In this report, we introduced the developed SPCR method and the time series analysis, in addition to the Lasso+bootstrap method. At present, the superiority or inferiority of these methods cannot be discussed. It is desirable to select the method according to the purpose or analyze with all methods and comprehensively consider the obtained results.

As discussed in this report, advanced analysis has enabled us to handle “big data” obtained from next-generation sequencing. However, more research is required to completely elucidate the “black box” pointed out at the introduction. In future, it will be important to obtain more findings from microbe data and accumulate those by approaches like those described in this report.

3. Conclusions

We developed novel statistical methods by transdisciplinary approaches, namely integrating microbiology and mathematics, and successfully identified important microorganisms responsible for wastewater treatment. In the future, microbial community data is expected to contribute to the stabilization and efficiency enhancement of biological wastewater treatment processes by the analysis of big data using statistical methods.

References

1) Seviour, R. et al.: Microbial Ecology of Activated Sludge. 1st ed. London, IWA Publishing, 2010, 688p
2) Katsumi, S. et al.: Journal of the Fuel Society of Japan. 57 (4), 227 (1978)
3) Japanese Patent Publication. 2020-036579. March 12, 2020
4) Caporaso, J. G. et al.: Nature Methods. 7, 335 (2010)
5) Kawano, S. et al.: Statistical Modeling via Sparse Estimation. 1st ed. Tokyo, Kyoritsu Shuppan Co., Ltd., 2018, 168p
6) Efron, B. et al.: An Introduction to the Bootstrap. 1st ed. New York, Chapman and Hall/CRC, 1994, 456p
7) Kawano, S. et al.: Computational Statistics & Data Analysis. 89, 192 (2015)
8) Wu, Y. J. et al.: Journal of Bioscience and Bioengineering. 115 (4), 424 (2013)
9) Fukushima, T. et al.: Water Science and Technology. 62 (6), 1432 (2010)
10) Fukushima, T. et al.: Journal of Water and Environment Technology. 5 (1), 37 (2007)
11) Vela, J. D. et al.: Water Research. 138, 241 (2018)
12) Albertsen, M. et al.: Plos One. 10 (7), e0132783 (2015)

Toshikazu FUKUSHIMA
Senior Researcher, Ph.D. (Environmental Studies)
Environment Research Lab.
Advanced Technology Research Laboratories
20-1 Shintomi, Futtsu City, Chiba Pref. 293-8511

Shuichi KAWANO
Associate Professor, Ph.D. (Functional Mathematics)
Graduate School of Informatics and Engineering
The University of Electro-Communications

Junichi NAKAGAWA
Former Chief Researcher, Ph.D. (Mathematical Science)
Mathematical Science & Technology Research Lab.
Advanced Technology Research Laboratories

Mamoru OSHIKI
Associate Professor, Ph.D. (Environmental Studies)
Division of Environmental Engineering, Faculty of Engineering
Hokkaido University