Application of Machine Learning and Swarm Intelligence to Wastewater Treatment

Charles C. Zhou[1], Shuo Han[2], Weiwei Liu[3], Zhoulin Zhang[4]

[1] Cascade Clean Energy, Inc. Cupertino, CA 95014, charles.zhou@ccleanenergy.com
[2] Kansas State University, Manhattan, Kansas 66506, shuohan@ksu.edu
[3] Kunming Research Institute of Metallurgy and Environment, China, [4] South China University of Agriculture

ABSTRACT

In this paper, we show a self-awareness concept and theory of swarm intelligence (SI) that can be used to discover authoritative and popular information as well as emerging and anomalous information when traditional connections among information nodes (e.g., hyperlinks or citations) are not available. The different categories of information can be all high-value depending on the application requirements. A self-awareness of swarm intelligence is a data-driven framework, modeled and measured using a recursive distributed infrastructure of machine learning. The combination of the machine learning and swarm intelligence to be extended and enhanced in a completely new perspective.

We used the technology described above, built a data model from USPTO database, NCBI database, JGI (Joint Genomic Database) and KEGG database, as well as our own bio-database. We used the machine learning method on these data models to select microbial consortia for wastewater treatment using the swarm intelligence of microbes. The collective behaviors of the selected microbes are used for cleaning wastewater and convert bio-wastes to usable energy.

Keywords: artificial intelligence, big data, swarm intelligence, microbes, wastewater

1 INTRODUCTION

Swarm intelligence (SI) is a branch of artificial intelligence, which has been existing in nature among grouping and activities of animals, birds, ants, fish or even microbes. Swarm Intelligence is the collective behavior of decentralized, self-organized systems, natural or artificial. The concept was originally used by Beni and Wang in the context of cellular robotic systems.

In this paper, we show a self-awareness concept and theory of swarm intelligence that can be used to discover authoritative and popular information as well as emerging and anomalous information when traditional connections among information nodes (e.g., hyperlinks or citations) are not available. The different categories of information can be all high-value depending on the application requirements. A self-awareness of swarm intelligence is a data-driven framework, modeled and measured using a recursive distributed infrastructure of machine learning. The combination of the machine learning and swarm intelligence to be extended and enhanced in a completely new perspective.

Since swarm intelligence systems consist typically of a population of simple agents interacting locally with one another and with their environment. The inspiration often comes from nature, especially biological systems. The agents follow very simple rules, and although there is no centralized control structure dictating how individual agents should behave, local, and to a certain degree random, interactions between such agents lead to the emergence of "intelligent" global behavior, unknown to the individual agents[2, 8].

We used the system self-awareness and swarm intelligence to built a data model from USPTO database, NCBI database, JGI (Joint Genomic Database) and KEGG (Kyoto Encyclopedia of Genes and Genomes) database, as well as our own CASCADE (computer assisted strain construction and development engineering) database. We used the machine learning method on these data models to select microbial consortia for wastewater treatment using the swarm intelligence of microbes. The collective behaviors of the selected microbes are used for cleaning wastewater and convert bio-wastes to usable energy.

2 METHODS

The swarm intelligence (SI) concept in this paper has an analogue in nature. As human being often ponder: what is the mechanism behind flocking swarms seem successfully achieve collective goals， such as looking for food or going to places in an optimized fashion even Pareto optimal using only local and simple communications as shown in Figure 1 [1]. Often a swarm can apply some SI to maximize a total value, e.g., get to a food target in a shortest distance. A swarm finds an optimal solution using pheromone. A pheromone is a chemical substance produced and released into the environment by a mammal or an insect which affects the behavior or physiology of others. Microbes behaves in a similar way as a form of special swarm on microscales.

Fig. 1. The collective behaviors of flocking swarm such ants, ants, fish, birds, buffalo .

In this paper, genetic information is added into the core of Computer-Assisted Strain Construction and Development Engineering （CASCADE）. The first type of information, is generic gene functions or the percentage of gene usages in 23 general function categories. According to Monica Riley's classification, there are 23 gene role categories in the genome of a microorganism[1]. The number and percentage of genes for a given microorganism in every gene function category are taken from "The Comprehensive Microbial Resource (CMR) ."

Second, codon usages or more granular measures for recording the amino acid composition of protein are also input with 64 codons for a given organism. Last, metabolic gene functions or the unique gene numbers in 137 metabolic function categories for a given organism are inserted. All three of these data sets are treated as input or known information to the system. More detailed functional categories can be found in the enhanced databases such as The Institute for Genetic Research (TIGR). Here one finds private experimental data can be also used for enhancement.

Figure 2 illustrates the CASCADE method which uses predictive targets such as average metabolic efficiency (AME), a measure of the average frequency that a gene appears in an organism's metabolic pathway, to compute electrogenicity, a microorganisms ability to generate electricity, a key component in Microbial Fuel Cell applications while cleaning the wastewaters.

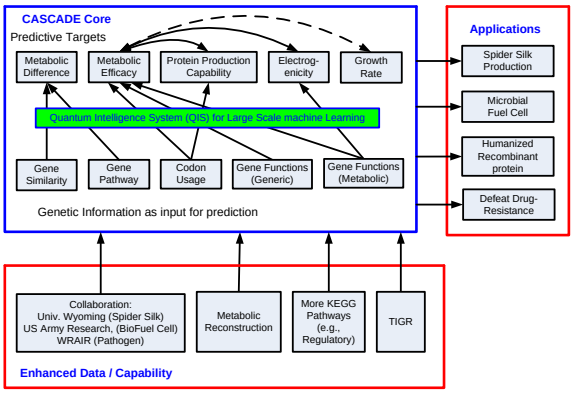

Fig. 2. Data model for selecting bacteria to use their collective intelligence to clean and reuse wastewater, and convert wastes to energy.

Microbs are one of the most prolific organisms on Earth. Harnessing the power of "smart" microbs for energy generation while cleaning the wastewater and a host of other life-improving applications such as wastewater treatment has become more and more crucial to a sustainable world.

"Smart" microbs are aptly named for their extraordinary ability to generate energy and materials like electricity, hydrogen, methane and proteins from organic sources. Enhancing how we use the unique capabilities of these microbes is important. Unlocking predictive patterns between a microorganisms genetic fingerprints and their possible "smart" metabolic capabilities opens the door to improving the interpretation of information in compiled databases of existing research which could lead to revolutionary new screening methods. For scientists, it means that microbe analysis for specific bioenergy and biotechnological uses becomes more efficient.

Machine learning, which focuses on prediction based on known properties, is the basis for the technology that we termed CASADE. This in silico methodology was able to uncover predictive relationships between a microb's genetic information and its metabolic behavior.

We applied metabolic pathways from public databases, such as KEGG and investigated metabolic reconstructions for the organisms with only genomic information. Our selection included 327 bacteria in 13 groups.

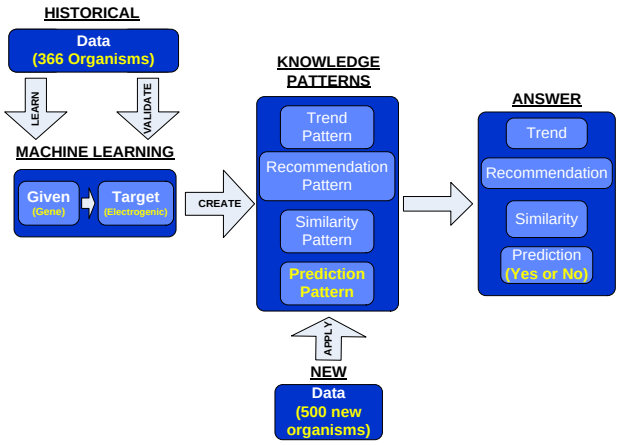

Fig. 3: Supervised leaning models of how to select microbes for our wastewater problem by using the complimentary and collective behaviors of microbes

We applied CASCADE technology to microb populations using a defined measure termed as average metabolic efficiency (AME), a gauge that highly correlated with metabolic capabilities that occur in real life. This measurement allowed us to explore electrogenicity for improving microbial fuel cells (MFC), methanogenicity for methane generation in anaerobic digester and protein production for spider silk.

One notable result in our work occurred with methane experiments. Here, CASCADE-selected microbs were not

only consistent (5/7 overlap) with current scientific selection, but also allowed the prediction of two additional microorganisms not previously selected by conventional methods.

This machine leaning method promises to help researchers find a cocktail of mixed microorganisms that could work more efficiently and more powerfully than a single microbs. In silico research in predictive metabolomics and computational biology has the potential to speed the rate of discovery process and prediction and lower the expense of lab work and experimental trials.

3 SCREENING MICROBS

Recovering methane from cleaning wastewater processing is important for two reasons: 1) methane emitted into the air after sludge digestion has an impact 23 times worse than CO_2 on the greenhouse effect 2) yet, in an opposite perspective, wastewater is an unexpected and surprisingly rich resource for electricity, hydrogen, and clean water.

In an anaerobic system [3-7] the majority of the chemical energy within a starting organic material is released as methane through the conversion of complex organic molecules to intermediate molecules. These end products also include sugars, hydrogen and acetic acid. This conversion is divided into four key biological and chemical stages, namely i) Hydrolysis; ii) Acidogenesis (Fermentation); iii) Acetogenesis; iv) Methanogenesis, and four physiologically distinct groups of microorganisms, including acetic acid-forming bacteria (acetogens) and methane-forming bacteria (methanogens), are involved in anaerobic digestion. Alternative to this scientific method, we used the CASACADE model to select smart microorganisms capable of converting sludge in wastewater to methane according to the constituents of the water.

In Figure 4, we applied the characteristics discovered from CASCADE to query and identify microorganisms that were not previously identified as methanogenic by KEGG, literature or industries. The blue boxes highlight the microorganisms linked to the pathway characteristics (in red circles). For instance, Pyruvate/Oxoglutarate_ oxidoreductases_ High denotes higher gene activities than average are observed in the pathway Pyruvate/Oxoglutarate oxidoreductases linked to a cluster of microorganisms that are methanogens – represented in three letter abbreviations, e.g. *MAC(Methanosarcina acetivorans)*, *MJA (Methanococcus jannaschii)* , *MTH (Methanothermobacter thermautotrophicus)*. It is also linked to the microorganism *BJA*(*Bradyrhizobium japonicum*) in a blue box. We found three microorganisms in the population possessing both "Pyruvate/Oxoglutarate _oxidoreductases_ High" and "Tetrachloroethene_ degradation_High" which are not identified as methanogenic from the current understanding. The three microorganisms are *BJA* (*Bradyrhizobium japonicum*), *CTE*(*Chlorobium tepidum*) and *CAC* (*Clostridium acetobutylicum*). We also found *ECO* (*E. coli K-12 MG1655*) links to the group via the "ATPASE_HIGH" characteristics as shown in Figure 4.

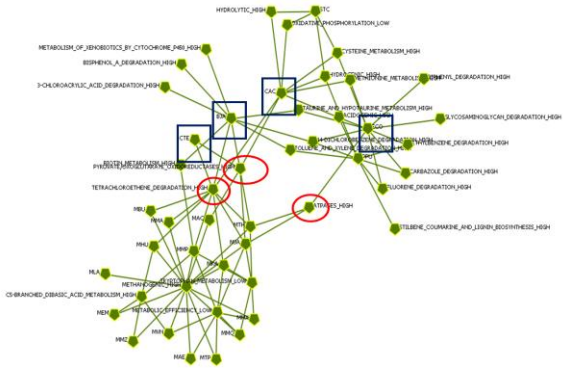

Fig. 4. Search for New Microbs for Methanogenic Using Pathway Characteristics

Although the four microbs marked in Figure 4 are not directly identified as methanogens in literature, our CASCADE model shows that they can produce methane and can function as anaerobic methanogens. It is suggested that the conversion of pyruvate to acetyl-CoA is catalyzed by pyruvate oxidoreductase in all archaebacteria. A similar pyruvate ferredoxin oxidoreductase is found in anaerobic eubacteria, therefore, it might be possible here to use a consortium to simulate anaerobic methanogens [9]. It is also suggested that the 'Pyruvate/oxoglutarate oxidoreductases' and 'C5-branched dibasic acid metabolism' are related to energy generation including valine and isoleucine biosynthesis from pyruvate (i.e. acetolactate synthase) [10]. Metabolic capacity is defined based on microbial community gene content.

After testing several anaerobic bacteria, including four strains of acetate-using methanogens, we found that *Methanosarcria* sp., *Methanosarcria mazei* cultures, and DCB-1 could degrade tetrachloroethene to trichloroethene. The process by which methanogens dechlorinate tetrachloroethene is a co-metabolic process and appears to be dependent on forming methane from the carbon source.

Figure 5 shows a graph similar to 4 that some substrates and products, which are critically linked to the methanogens in three letter nodes starting with 'M' in Figure 5 , are also linked to non-methanogens in three letter nodes in blue boxes. For example, C15489_OUT_HIGH in a red circle in Figure 3 has four links. They represent a higher than average product C15489 (acetyl-CoA), in connection with the methanogens.

Based on this graph, we selected four microorganisms, CAC (*Clostridium acetobutylicum)*, LAC(*Lactobacillus acidophilus)*, *ECO* (*Escherichia coli K-12 MG1655*), and *PPU* (*Pseudomonas putida*) in blue boxes, which are connected with the substrates and products in red circles. These substrates and products are C15489 (acetyl-CoA), C00331(pyruvate), C00138(reduced ferredoxin), C00139(oxidized ferredoxin),C02565(N-Methylhydantoin), C00043(UDP-N-acetylglucosamine), C00238(potassium cation) and C00070(copper). The four microbs could form a consortium to

enhance the functions of methanogens and clean the wastewater[11].

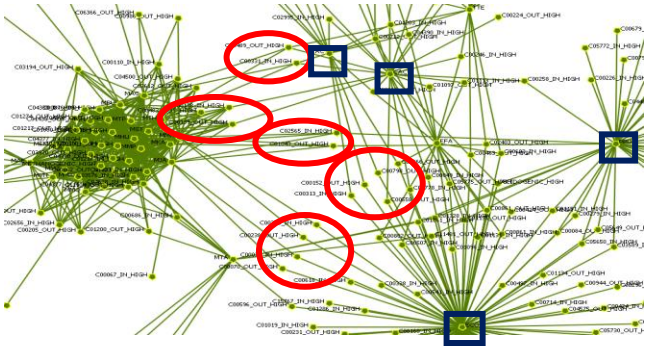

Fig. 5: Use substrates and products to link smart microorganisms for energy generation while cleaning the wastewaters

4 CONCLUSIONS

One challenge that arises is the large dimensionality from the potentially very large quantities of biology data--counting all possible combinations of features and attributes that might impact specific biological behavior and desired properties. Some objects and entities (e.g. number of microorganisms) available for analysis are relatively small compared to the dimensionality in research. Many of the traditional machine learning methods such as hierarchical clustering, neural networks, decision trees and association rules, are often not readily applicable. The CASCADE system draws the similarity between complex system analysis for biological data and text data, therefore, results in a more scalable and meaningful approaches.

The clustering method implemented in the CASCADE system associates elements, contexts, concepts, sequences, and clusters in a holistic manner to benefit semantic analysis. Data scoring favors identifying data anomalies that might be associated with novel biological behaviors. This is the competitive advantage of our method over other methods especially in the area of selecting smart microorganisms: A microorganism possessing a rare, novel and smart capability can be detected as data anomalies. Our method provides indicators of specific traits in a small population of microbs from a generic microb population where a large-scale of genetic information (e.g. number of genes) linking to a desired behavior like methanogenicity might emerge.

In the case of methane production and wsetwater treatment, we found that biologically selected microbs and CASCADE selected microbs are consistent with five out of seven. In addition, CASCADE predicted another two microbs that are not scientifically selected. The CASCADE method can be extended to a range of applications requiring screening microorganisms that have smart capabilities in addition to methanogenicity, electrogenicity and protein productivity.

REFERENCES

[1] Fleischer M., “Foundations of Swarm Intelligence: From Principles to Practice”, 2005, Retrieved from http://arxiv.org/pdf/nlin/0502003.pdf

[2] Zhao, Y. & Zhou, C. (2016). “System Self-Awareness Towards Deep Learning and Discovering High-Value Information. In the Proceedings of the 7th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference, Oct. 20-22, New York, USA. Page 109-116.

[3] Logan B.E., Hamelers, B., Rozendal, R., Schröder, U., Keller,J., Freguia, S., Aelterman, P., Verstraete,W. & Rabaey,K. (2006a). Microbial fuel cells: methodology and technology. Environ. Sci. Technol., 40(17): 5181 -5192.

[4] Rabaey, K., Lissens, G., Siciliano, S.D. and Verstraete, W. (2003). A microbial fuel cell capable of converting glucose to electricity at high rate and efficiency. Biotechnol Lett 25, 1531–1535.

[5] Niessen, J., Schro¨der, U., Rosenbaum, M. & Scholz, F. (2004a). Fluorinated polyanilines as superior materials for electrocatalytic 8 anodes in bacterial batteries. Electrochem Commun 6, 571–575.

[6] Niessen, J., Schro¨der, U. & Scholz, F. (2004b). Exploiting complexcarbohydrates for microbial electricity generation – a bacterial fuel cell operating on starch. Electrochem Commun 6, 955–958.

[7] Kim, M.J., Cho, H.S. & Kim, J. Y. (2007). Anaerobic biodegradability of plastic garbage bags based on starch polymer and aliphatic polyester. In the Proceedings Sardinia 2007, Eleventh International Waste Management and Landfill Symposium, S. Margherita di Pula, Cagliari, Italy; 1 - 5 October 2007: 517-518.

[8] Zhou, C., and Zhao Y., US patents 9,026,373 and 9,792,404, “Method and system for knowledge pattern search and analysis for selecting microorganisms based on desired metabolic property or biological behavior”

[9] Kates, M., Kushner, D. J. & Matheson, A. T. (eds), (1993). The Biochemistry of Archaea (Archaebacteria). Elsevier: Amsterdam.

[10] Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R. & Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444:1027-1031.

[11] Zhou, C., Killgrow, S. Hill，C. “Cascade Clean Energy System for Water & Wastewater Treatment”, www.energy.ca.gov/2015publications/CEC-500-2015-065