Computational Breakthroughs in Aquatic Taxonomy: The Role of Deep Learning and DNA Barcoding

Nadiia Kasianchuk [1,2], Sofiia Harkava[3], Sofiia Onishchenko[4], Olesia Solodka[5], Daria Shyshko[6,7], Eduard Siemens[8], Halina Falfushynska[8,9] and Taras Ustyianovych[10]

[1]*Faculty of Biology, Adam Mickiewicz University in Poznan, Uniwersytetu Poznańskiego Str. 6, 61712 Poznań, Poland*
2Faculty of Pharmacy, Bogomolets National University, Taras Shevchenko Str. 13, 01601 Kyiv, Ukraine
[3] *Dnipro regional branch of the Ukrainian Junior Academy of Sciences, Gagarin Avenue 26, 49000 Dnipro, Ukraine*
[4] *Kharkiv regional branch of the Ukrainian Junior Academy of Sciences, Skrypnyka Str. 14, 61000 Kharkiv, Ukraine*
[5] *Rivne regional branch of the Ukrainian Junior Academy of Sciences, Symon Petliura Str. 17, 33028 Rivne, Ukraine*
[6] *Faculty of Biology and Ecology, Oles Honchar Dnipro National University, Gagarin Avenue 72,49000 Dnipro, Ukraine*
[7] *Klinik für Psychiatrie und Psychotherapie, Charite - Universitätsmedizin Berlin, Hindenburgdamm Str. 30, 12203 Berlin, Germany*
[8] *Anhalt University of Applied Sciences, Bernburger Str. 57, 06366 Köthen, Germany*
[9] *University of Rostock, Albert Einstein Str. 3, 19059 Rostock, Germany*
[10] *Department of Artificial Intelligence Systems, Lviv Polytechnic National University, Kniazia Romana Str. 5, 79005 Lviv, Ukraine*
nadkas2@st.amu.edu.pl, sofiiaharkava@gmail.com, sonyastudy982@gmail.com, olesiasolodka@gmail.com, daria.shyshko@charite.de, eduard.siemens@hs-anhalt.de, halina.falfushynska@uni-rostock.de, taras.o.ustyianovych@lpnu.ua

Keywords: Aquatic Ecosystems; Deep Learning; Taxonomic Identification; Metagenomic Sequences; Environment Modeling; Machine Learning.

Abstract: Aquatic ecosystems are crucial in maintaining environmental equilibrium and sustaining human well-being. However, the traditional manual methods used in hydrobiological research have limitations in providing a far-reaching understanding of these intricate ecosystems. Data science, machine learning, and deep learning techniques offer a variety of opportunities to overcome these limitations and unlock new insights into aquatic environments. This study highlights the impact of computational tools in areas such as taxonomic identification, metagenomic sequence analysis, and water quality prediction. Deep learning techniques have demonstrated superior accuracy in classifying organisms, including those previously unidentified by conventional methods. In metagenomic sequence analysis, machine learning aids in effectively assembling DNA sequences, aligning them with known databases, and addressing challenges related to sequence repeats, errors, and missing data. Furthermore, predictive models have been developed to provide insights into water quality parameters, such as eutrophication events and heavy metal concentrations. These advancements lead to informed conservation measures and a deep understanding of the intricate relationships within aquatic ecosystems. However, challenges persist, including data quality issues, model interpretability, and the need for robust training datasets. Thus, data integration strategies designed specifically for environmental and genomic studies are necessary. Data fusion and imputation can help address data scarcity and provide a comprehensive view of hydrobiological processes. As the study of aquatic ecosystems continues to evolve, the synergy between computational methods and traditional hydrobiological techniques holds immense potential. By leveraging the power of data science and cutting-edge technologies, researchers can gain a deep understanding of aquatic environments, monitor changes in biodiversity, and develop informed strategies for sustainable management amidst global environmental shifts.

1 INTRODUCTION

Hydrobiological research stands as a cornerstone in the comprehensive understanding of aquatic ecosystems, which are critical to both environmental equilibrium and human sustenance [1]. Historically, hydrobiological inquiries were primarily anchored in manual sampling techniques and observational methodologies, offering insights that were often limited by the scope and resolution of available tools.

With the advent of modern technologies, there has been an exponential surge in the volume and granularity of data obtainable from aquatic environments, necessitating the integration of more sophisticated analytical tools [2].

The taxonomic identification of organisms within aquatic ecosystems is of paramount importance in hydrobiological research [3]. It serves as a fundamental pillar in assessing the health and biodiversity of these ecosystems, which in turn impacts their ecological functions and resilience to environmental changes. Accurate taxonomic identification allows scientists to monitor the presence of indicator species, assess shifts in community composition, and detect potential invasive species that can disrupt the balance of aquatic ecosystems [4]. However, the field of taxonomic identification faces common challenges, including the vast diversity of aquatic organisms, taxonomic ambiguities, and the need for rapid and cost-effective identification methods [5]. Additionally, environmental stressors such as climate change and pollution further underscore the importance of robust taxonomic identification to monitor and mitigate their effects on aquatic ecosystems [6].

Furthermore, global environmental challenges, including climate change, pose significant threats to aquatic ecosystems. Climate-related shifts in temperature, precipitation, and water chemistry impact species distributions, increase the prevalence of cyanobacterial blooms with toxic effects, disrupt hydrological patterns, and lead to ocean acidification and glacial retreat [7, 8]. These global issues, while vast in their implications, are intricately linked to the microcosmic interactions occurring within water bodies.

The integration of data science, machine learning, and deep learning into hydrobiological research offers an unprecedented opportunity to decipher these complex interactions, predict future trends, and formulate strategies for sustainable management and conservation [9]. The sheer volume, complexity, and multidimensionality of data sourced from aquatic ecosystems have transcended the analytical capabilities of conventional methods. Data science, machine learning, and deep learning offer sophisticated frameworks that can seamlessly process, analyze, and interpret this deluge of information, enabling researchers to glean deeper insights and uncover subtle patterns that were previously elusive. These computational tools not only enhance the precision and accuracy of analyses but also empower researchers to model intricate ecological interactions, forecast environmental shifts, and optimize interventions for aquatic conservation [10]. In a realm as dynamic and intricate as hydrobiology, the fusion of computational prowess with biological inquiry heralds a new era of informed decision-making and robust ecosystem management.

2 ADDRESSING COMMON CHALLENGES IN METAGENOMIC ANALYZES

In recent years, the biotechnological landscape has undergone a huge transformation. Traditionally, DNA sequencing was predominantly executed on either cultivated cells or genetic material derived from a specific organism with known taxonomic classification. That is to say, homology-based approaches for sequence analysis have emerged as a popular solution for taxonomic classification [11]. They are grounded in the principle that sequences sharing a common ancestry will exhibit similarities. These approaches compare a query sequence against a database of known sequences, seeking matches or alignments that indicate a common typology. These methods are characterized by the very high strength and precision, especially when the genome is already cataloged in the database.

However, a significant challenge arises from the vast number of sequences that remain unclassified. Estimates suggest that at most, only 9% of ocean species have been described [12]. Furthermore, the effectiveness of taxonomic classification varies depending on factors like the sample origin, desired taxonomic level, and database specifics. Such limitation became even more evident with the rise of DNA sequence assembly aims at reconstructing the original structure of the DNA in question, by aligning [13] and merging fragments of a DNA sequence.

Sequence assembly, the process of reconstructing full DNA sequences from fragmented reads, relies heavily on advanced computational algorithms and specialized software [14]. While these tools offer precision, technical challenges, such as missing data and genomic intricacies like sequence repeats and heterozygosity, can impede accurate reconstruction. Such oversights can compromise the assembly's integrity [15]. As the demand for computational efficiency and optimal resource utilization grows [16], the integration of data science techniques becomes increasingly crucial, promising improved accuracy and streamlined assembly workflows (Figure 1).

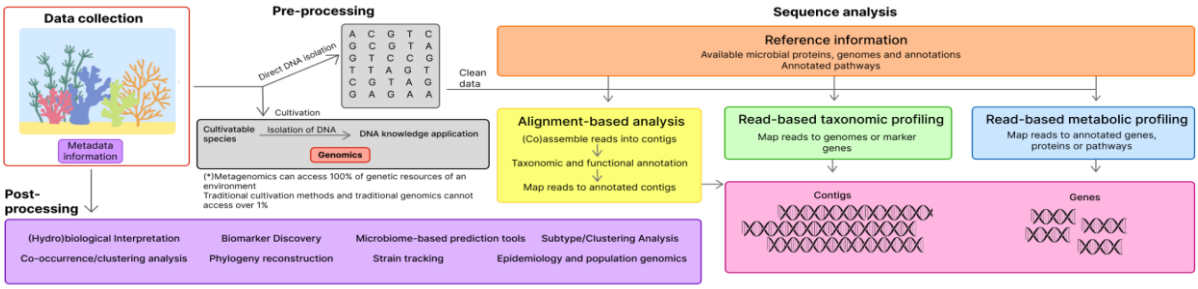

Figure 1: Schematic workflow for metagenomic sequences analysis. The workflow depicts the journey from raw sequence data acquisition, through preprocessing and quality control, to assembly and annotation.

The development of graph algorithms has brought significant advancements to the field, with three primary categories traditionally employed: Overlap/Layout/Consensus (OLC), de Bruijn Graph, and greedy graph methods [13]. In classic computer science terminology, a graph is an abstraction comprising nodes interconnected by edges. Within the overlap graph approach, sequencing reads are depicted as nodes, while their overlaps are represented by edges. This graph can also be equipped with additional attributes to differentiate between the 5′ and 3′ ends of reads, forward and reverse complement sequences, read lengths, overlap lengths, and overlap types (either suffix-to-prefix containment [17]). Researchers at Leiden University have explored the use of overlap graphs for assembling genome sequences from Ciliates found in water bodies. While their findings underscored the potential of this approach for genome assembly, they also suggested refinements to enhance its efficiency and accuracy. One such refinement was the introduction of a 'partial' model, characterized by specific forbidden induced subgraphs. Notably, this model does not have a counterpart for the simple double string rule in graph rules [18]. Furthermore, the team introduced a method to directly construct the reduction graph from its overlap graph, emphasizing the capability to recover structural information seemingly lost in the overlap graph [19] . While these results are promising, further investigations are needed to determine the consistency of this method's effectiveness across various overlap graphs and organism genera.

De Bruijn approach for sequence assembly introduced great advancements in the field and underlied the development of a number of modern sequence assembly approaches [20]. In this approach, the nodes represent all possible fixed-length strings and the edges represent suffix-to-prefix perfect overlaps. One of the important forms of de Bruijn graphs is the K-mer graph (Figure 2a). Its edges represent all the fixed-length overlaps between subsequences that were consecutive in the larger sequence. According to one approach, each K-mer starting at a base corresponds to an edge, with nodes representing overlaps of K-1 bases [21]. In contrast, each K-mer starting at a base is depicted as a node, while edges signify overlaps of K-1 bases [22]. In the realm of WGS assembly, K-mer graphs can depict multiple sequences, with each read represented as a distinct path. When reads perfectly overlap, they share a common path, allowing the implicit detection of these overlaps without the necessity for pair-wise sequence alignment calculations (Figure 2b). While overlap graphs merge paths at longer repeats within a read, K-mer graphs do so at perfect repeats with a length of K or more, given that K is shorter than the read length. This makes K-mer graphs particularly susceptible to sequencing errors and repeats. A singular sequencing error can generate up to K incorrect nodes in the K-mer graph. Consequently, these erroneous nodes might align with other nodes, leading to unintended path convergences [13].

A number of approaches incorporate de Bru graphs to the sequence assembly with positive outcomes. For instance, Velvet stands out as a widely-used de novo assembler, tailored specifically for short-read sequencing data from next-generation sequencing platforms [22]. Thanks to its mechanism, Velvet is capable of managing exceptionally short reads and read pairs, facilitating the construction of meaningful genomes. Its versatility is evident from its widespread application across diverse genomic investigations, often juxtaposed with other assembly tools to gauge its competence [23]. On the other hand, Velvet's single-threaded design restricts its operation to a single processor, potentially curtailing its scalability with expansive datasets that are common in environmental and water body research [24]. That is to say, the application relies on an in-memory representation of the de Bruijn graph, which can be

memory-intensive for larger genomes [25]. Other common de Bruijn graph-based assemblers include, ALLPATHS, HaVec, ABySS, MEGAHIT, SOAPdenovo , and YAGA.

The challenge of handling high-dimensional data is recurrent across various methods. Consequently, several strategies have been developed to address this concern. Among these, DBGPS, which employs the de Bruijn graph combined with a greedy path search, stands out. This approach not only effectively manages high-dimensional data, demonstrated by its successful handling of 6.8 MB of data, but also adeptly addresses issues related to DNA breaks, rearrangements, and indels [26].

Next important step of the analysis of sequenced data is the alignment. Alignment involves matching the sequenced data to a reference genome or other sequences, a process fraught with challenges. Common difficulties [27] include handling mismatches due to genetic variations, dealing with gaps or insertions, and navigating through repetitive regions that can confound traditional alignment algorithms [28]. Metagenomic samples add layers of complexity to the analysis, amplifying the inherent challenges. Due to the vast diversity of microbial species in a single sample, distinguishing between closely related organisms can be problematic [29]. The presence of rare species means reference genomes might be absent, making traditional alignment methods inadequate [30]. Moreover, horizontal gene transfer events, common in microbial communities, can create chimeric sequences, complicating alignment [31].

There were several approaches that tried to solve the issues of conventional alignment methods. One of those is the greedy x-drop algorithm that performs sequence alignment by making locally optimal choices at each stage with the hope of finding a global optimum. While it is generally quicker than other conventional alignment algorithms, its primary disadvantage is that it doesn't always guarantee an optimal global alignment. However, due to its efficiency, it remains a popular choice in scenarios where approximate alignments are acceptable or when dealing with shorter DNA sequences [32]. Several other attempts existed to incorporate machine learning algorithms to optimise the alignment process.

Recent research underscores the promise of reinforcement learning algorithms in tackling key challenges associated with sequence alignment, demonstrating encouraging results. Specifically, these algorithms have demonstrated superior performance compared to conventional matrix-based tools, such as ClustalW and MAFFT, especially when applied to multiple sequence alignments of several benchmark datasets [33].

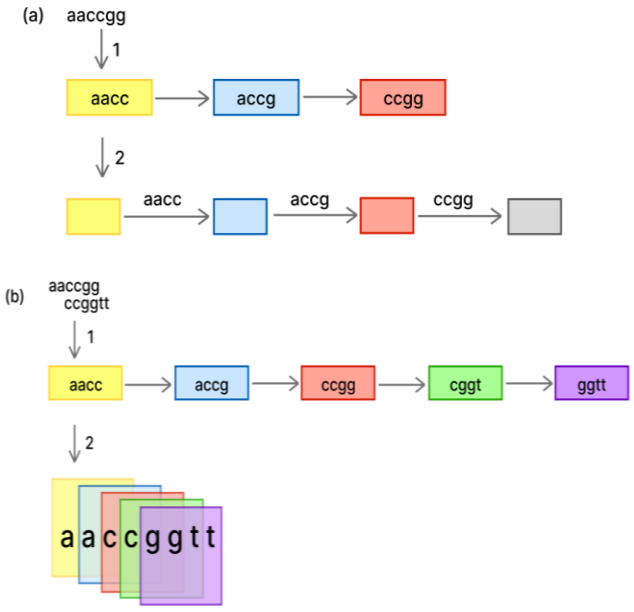

Figure 2: De Bruijn approach: a) a single read is mapped using two K-mer graph models to highlight the node-edge relationship for overlapping K-mers. Ideal for depicting simple paths due to minimal repeats. b) Pair-wise Overlap via K-mer Graph demonstrates error-free overlap of two reads and their unified representation in a K=4 K-mer graph, facilitating straightforward consensus sequence reconstruction through a simple path.

However, a prevalent constraint among these methods is their tendency to align sequences of a particular length, a limitation stemming from fixed input and network dimensions. To surmount this obstacle, Song and Cho (2021) integrated the DQNalign approach with the x-drop algorithm [34]. This amalgamation enhanced alignment performance, reduced complexity, and minimized computational time. Notably, when employed in the alignment of the E. coli genome, this method's accuracy paralleled that of conventional techniques. Another notable reinforcement learning-based solution is EdgeAlign, strategically tailored for the effective alignment of DNA sequences on edge devices. This innovative approach showcases the integration of a highly compact yet robust deep Q-network (DQN) agent, akin to the previously described method, ensuring a uniform hardware resource footprint regardless of sequence lengths [35].

In summary, recent advancements in genomics have ushered in a new era of metagenomics and DNA sequence analyses, challenging traditional methodologies. The integration of data science approaches, exemplified by use of machine learning methods, promises to enhance sequence alignment efficiency and accuracy. Altogether, it will speed-up advances in the hydrobiology due to high processing throughput of these systems. The automation of time-consuming and routine activities will contribute to high-quality data-driven approaches to tackle urgent climate change issues and its impact on water ecosystems and resources.

3 COMPUTATIONAL ADVANCES IN THE AQUATIC TAXONOMY

The taxonomic classification of organisms within water bodies plays a crucial role in ecological research and environmental monitoring [36]. Over time, this classification has progressed significantly, from traditional homology-based methods to the adoption of advanced deep learning techniques. Understanding the composition and dynamics of aquatic ecosystems is essential for assessing water quality, tracking changes in biodiversity, and studying the impact of environmental factors, such as climate change and pollution, on aquatic habitats [37].

From a technical standpoint, most classification tools hinge on the similar methodologies as described above, namely local alignments, k-mers, Burrow–Wheeler transformations, minimizers, or hybrid methodologies [38]. For instance, the discriminative k-mers method has been applied to classify metagenomic sequences, demonstrating remarkable accuracy, especially for short metagenomic reads. Notably, this algorithm exhibited impressive speed, capable of processing up to 32 million metagenomic short reads per minute [38]. Its performance aligns with other k-mer-based tools known for their accuracy, speed, and minimal memory requirements [39]. However, it's important to note that these methods primarily focus on the genus and/or species levels, with limited evaluation at higher taxonomic ranks. Overall, k-mer-based approaches excel in terms of speed, but their recall and precision may exhibit variability. On the contrary, local alignments, while highly precise, can be computationally intensive and may yield restricted recall rates.

To address these challenges, deep neural network (DNN) approaches have emerged as a promising solution. These deep learning techniques, rather than merely relying on database similarities, model intricate relationships between DNA sequences and their taxonomic classes. An example of such an algorithm is a recently developed model called BERTax, that is based on the state-of-the-art natural language processing architecture BERT (bidirectional encoder representations from transformers) updated with additional layers. BERTax is able to classify the sequences in question on 3 taxonomic levels (superkingdom, phylum and genus). A notable advantage of such methods compared to the conventional approaches is that it does not focus on local similarity, but on the overall image and, therefore, is not subjected to the common restrictions of comparable tools. As a result of such novelty, the tool was able to deal with novel organisms, not existing in the initial databases, which still remained a challenge for similar tools [40].

DNA barcoding, another widely-adopted genomic approach for taxonomic classification, owes its popularity to several key attributes [41]. This method utilizes a standardized region of DNA, enabling efficient species identification, insights into molecular lineage, and applications in conservation biology [42]. Traditionally, DNA barcoding relied on similarity-, character-, and tree-based methodologies. However, as computational capabilities have advanced, the integration of machine and deep learning techniques has emerged as a transformative approach. The evaluation of various algorithms using both empirical and synthetic datasets has demonstrated the remarkable efficacy of the k-nearest neighbors algorithm in addressing these tasks, outperforming other techniques such as Naive Bayes, Random Tree, and SVM [43]. While these algorithms

exhibit high performance on synthetic datasets, real-world data introduces additional challenges, including the high dimensionality of DNA barcode sequences, limited interspecific sequence variation, and numerical constraints due to the diversity of species. Consequently, a recent study has employed a sophisticated deep learning model to tackle these complexities in classifying fish from different families. This novel approach combines an Elastic Net-Stacked Autoencoder (EN-SAE) with Kernel Density Estimation (KDE), effectively mitigating the aforementioned challenges and enhancing classification accuracy [44].

Computer vision techniques have also been effectively employed in the taxonomic detection and classification of aquatic organisms. A noteworthy study utilised the Faster Region-based Convolutional Neural Network (R-CNN) extended with a supplementary classification branch. Remarkably, the model posted mean average precision scores of 74.64% at the genus level and 81.17% at the class level source, however the dataset's uneven distribution, with varying instance percentages of specific genera and biological classes, influences the model's efficacy. This limitation underscores the impending need for well-balanced, high-quality datasets for the algorithm training, as this step directly influences the model's performance [45]. Another successful application of taxonomy in aquatic habitats is exemplified by the work of Memmolo et al. (2020), where they conducted algorithm training using diatom test slides. Leveraging a substantial dataset comprising 8,731,800 elements, an average of 174.636 augmented phase-contrast images were generated from a single hologram record. This extensive dataset contributed to training a model that achieved an impressive classification accuracy of 98% [46]. These findings underscore the practicality and cost-effectiveness of utilizing species test slides as a valuable approach for training classification models in taxonomic studies.

In conclusion, the adoption of advanced computational methods, including deep learning and DNA barcoding, holds significant promise for enhancing aquatic taxonomic classification. These approaches offer notable advantages, such as improved accuracy and scalability, enabling more precise species identification and ecological analysis. However, challenges related to data quality, model interpretability, and the need for robust training datasets remain. Despite these obstacles, the potential benefits for biodiversity assessment, environmental monitoring, and conservation efforts are substantial. Further research and refinement of these methodologies are essential to unlock their full potential in advancing our understanding of aquatic ecosystems.

4 CONCLUSIONS AND FUTURE

The intersection of machine learning and hydrobiological research has opened doors to a plethora of possibilities in understanding aquatic ecosystems. These computational tools offer a sophisticated framework for handling the increasing volume and complexity of aquatic data, making previously elusive patterns more discernible. The taxonomic identification of organisms, a pivotal aspect of hydrobiological research, has seen remarkable advancements with the integration of deep learning techniques, improving precision and enabling detection of novel organisms. Furthermore, in the domain of metagenomic sequences analysis, machine learning has addressed challenges related to sequence assembly, alignment, and taxonomy, paving the way for more efficient and accurate methods.

However, despite the significant strides made, challenges persist. The quality of data, model interpretability, and the need for robust training datasets are among the hurdles faced. Therefore, there is a need for developing robust data integration and augmentation strategies, custom-tailored specifically for environmental and genomic studies. Most of the studies utilize common practices, such as image rotation and cropping, synthetic data generation, and manual modification. To overcome the data scarcity, usage of data fusion techniques can become helpful, however such approaches are not precisely studied in terms of hydrobiological research. Additionally, utilization of data imputation might fill gaps in the datasets and provide a precise view of the explored hydrobiological processes. Furthermore, while computational models offer enhanced accuracy and efficiency, their real-world application necessitates a comprehensive understanding of local aquatic nuances and characteristics.

However, the main challenges and opportunities for future studies in this area include high complexity and non-linearity of the data, noise, lack of covariates, and compositionality. This particular problem can be solved with a large model well-versed in the specifics of the domain area, similarly the way large language models are trained.

In essence, the fusion of computational methodologies with traditional hydrobiological techniques holds immense potential. As research continues and technologies evolve, the synergy of

these domains will undoubtedly lead to more informed strategies for the preservation and understanding of aquatic ecosystems. The emerging developments of data science in hydrobiology are mainly in their early stages and are becoming a powerful assistance tool for regular research activities. Their applications are vital for water quality monitoring, environment and climate change processes modeling. We expect the number of data science models and tools to grow in the incoming years with improvements in predictive performance and data quality. It is worth preparing solid data resources to make such studies possible and consolidate the available knowledge into a high-quality model.

The leveraged data and machine learning technologies are offering a significant boost to the development of specific areas of hydrobiology and aquatic environment modeling. The reviewed studies explore usage models and algorithms with a varying level of complexity depending on the input requirements. We consider that despite the exponential growth and usage of deep learning technologies, classical linear- and tree-based machine learning algorithms provide enough efficiency, flexibility, and accuracy to assist with modeling of hydrobiological properties and characteristics.

We observe a rising number of deep learning applications for unstructured data classification. It is worth mentioning that such an approach is very robust when dealing with image, sensor or text data formats. Nevertheless, the usage of deep learning networks requires large amounts of data in order to be efficient, which is the primary issue of the reviewed studies. Data collection for these research areas is still a major challenge due to a number of factors: sensor design and technologies, their deployment and high costs associated with them, a need for appropriate equipment and computational resources. From a hydrobiological perspective the following difficulties arise during data assemblage: vast diversity of unclassified microbes; heterogeneity, repeats, and duplications; quick genetic changes complicate the clarity of taxonomic categorization; external impact (environment or reagents contamination).

ACKNOWLEDGMENTS

We acknowledge support by the German Research Foundation (Deutsche Forschungsgemeinschaft DFG) - and the Open Access Publishing Fund of Anhalt University of Applied Sciences.

REFERENCES

[1] D. B. Oerther, L. Gautham, and N. Folbre, “Environmental engineering as care for human welfare and planetary health,” Journal of Environmental Engineering, vol. 148, no. 6, Jun. 2022, doi: 10.1061/(asce)ee.1943-7870.0002013.

[2] D. Y. Kwon, J. Kim, S. Park, and S. Hong, “Advancements of remote data acquisition and processing in unmanned vehicle technologies for water quality monitoring: An extensive review,” Chemosphere, vol. 343, p. 140198, Dec. 2023, doi: 10.1016/j.chemosphere.2023.140198.

[3] K. I. Suh, J. M. Hwang, Y. J. Bae, and J. H. Kang, “Comprehensive DNA barcodes for species identification and discovery of cryptic diversity in mayfly larvae from South Korea: Implications for freshwater ecosystem biomonitoring,” Entomological Research, vol. 49, no. 1, pp. 46-54, Jan. 2019, doi: 10.1111/1748-5967.12334.

[4] V. Gomez-Alvarez, H. Liu, J. G. Pressman, and D. G. Wahman, “Metagenomic Profile of Microbial Communities in a Drinking Water Storage Tank Sediment after Sequential Exposure to Monochloramine, Free Chlorine, and Monochloramine,” ACS ES&T Water, vol. 1, no. 5, pp. 1283-1294, Mar. 2021, doi: 10.1021/acsestwater.1c00016.

[5] C. O. Coleman and A. Radulovici, “Challenges for the future of taxonomy: talents, databases and knowledge growth,” Megataxa, vol. 1, no. 1, Jan. 2020, doi: 10.11646/megataxa.1.1.5.

[6] A. C. Staudt, et al., “The added complications of climate change: understanding and managing biodiversity and ecosystems,” Frontiers in Ecology and the Environment, vol. 11, no. 9, pp. 494-501, Nov. 2013, doi: 10.1890/120275.

[7] H. Falfushynska, N. Kasianchuk, E. Siemens, E. Henao, and P. Rzymski, “A review of common cyanotoxins and their effects on fish,” Toxics, vol. 11, no. 2, p. 118, Jan. 2023, doi: 10.3390/toxics11020118.

[8] R. C. Allen, B. E. Rittmann, and R. Curtiss, “Axenic Biofilm Formation and Aggregation by Synechocystis sp. Strain PCC 6803 Are Induced by Changes in Nutrient Concentration and Require Cell Surface Structures,” Applied and Environmental Microbiology, vol. 85, no. 7, Apr. 2019, doi: 10.1128/aem.02192-18.

[9] K. Malde, N. O. Handegard, L. Eikvil, and A.-B. Salberg, “Machine intelligence and the data-driven future of marine science,” Ices Journal of Marine Science, vol. 77, no. 4, pp. 1274-1285, Apr. 2019, doi: 10.1093/icesjms/fsz057.

[10] R. H. Medina, et al., “Machine learning and deep learning applications in microbiome research,” ISME Communications, vol. 2, no. 1, Oct. 2022, doi: 10.1038/s43705-022-00182-9.

[11] R. Harr, P. Hagblom, and P. Gustafsson, “Two-dimensional graphic analysis of DNA sequence homologies,” Nucleic Acids Research, vol. 10, no. 1, pp. 365-374, Jan. 1982, doi: 10.1093/nar/10.1.365.

[12] C. Mora, D. P. Tittensor, S. M. Adl, A. G. B. Simpson, and B. Worm, “How many species are there on Earth and in the ocean?,” PLOS Biology, vol. 9,

no. 8, p. e1001127, Aug. 2011, doi: 10.1371/journal.pbio.1001127.

[13] J. R. Miller, S. Koren, and G. Sutton, “Assembly algorithms for next-generation sequencing data,” Genomics, vol. 95, no. 6, pp. 315-327, Jun. 2010, doi: 10.1016/j.ygeno.2010.03.001.

[14] A. M. Phillippy, “New advances in sequence assembly,” Genome Res., vol. 27, no. 5, pp. xi-xiii, May 2017, doi: 10.1101/gr.223057.117.

[15] K. L. Korunes and K. Samuk, “pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data,” Mol. Ecol. Resour., vol. 21, no. 4, pp. 1359-1368, May 2021, doi: 10.1111/1755-0998.13326.

[16] J. Mbatchou, et al., “Computationally efficient whole-genome regression for quantitative and binary traits,” Nat. Genet., vol. 53, no. 7, pp. 1097-1103, Jul. 2021, doi: 10.1038/s41588-021-00870-7.

[17] E. Csuhaj-Varjú, I. Petre, and G. Vaszil, “Self-assembly of strings and languages,” Theor. Comput. Sci., vol. 374, no. 1, pp. 74-81, Apr. 2007, doi: 10.1016/j.tcs.2006.12.004.

[18] R. Brijder and H. J. Hoogeboom, “Combining overlap and containment for gene assembly in ciliates,” Theor. Comput. Sci., vol. 411, no. 6, pp. 897-905, Feb. 2010, doi: 10.1016/j.tcs.2009.07.047.

[19] R. Brijder, H. J. Hoogeboom, and G. Rozenberg, “REDUCTION GRAPHS FROM OVERLAP GRAPHS FOR GENE ASSEMBLY IN CILIATES,” Internat. J. Found. Comput. Sci., vol. 20, no. 02, pp. 271-291, Apr. 2009, doi: 10.1142/S0129054109006553.

[20] B. Ekim, B. Berger, and R. Chikhi, "Minimizer-space de Bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer," vol. 12, no. 10, pp. 958-968.e6, Oct. 2021, doi: 10.1016/j.cels.2021.08.009.

[21] R. M. Idury and M. S. Waterman, “A new algorithm for DNA sequence assembly,” Journal of Computational Biology, vol. 2, no. 2, pp. 291-306, Jan. 1995, doi: 10.1089/cmb.1995.2.291.

[22] D. R. Zerbino and E. Birney, “Velvet: algorithms for de novo short read assembly using de Bruijn graphs,” Genome Res., vol. 18, no. 5, pp. 821-829, May 2008, doi: 10.1101/gr.074492.107.

[23] Y. Endo, F. Toyama, C. Chiba, H. Mori, and K. Shoji, “Memory Efficient de novo Assembly Algorithm using Disk Streaming of K-mers,” scitepress.org, Jan. 2016, doi: 10.5220/0005798302660271.

[24] E. Costa and G. Silva, “The velvet assembler using OpenACC directives,” EPiC Series in Computing, May 2023, doi: 10.29007/pzbt.

[25] R. Chikhi and G. Rizk, “Space-efficient and exact de Bruijn graph representation based on a Bloom filter,” Algorithms Mol. Biol., vol. 8, no. 1, p. 22, Sep. 2013, doi: 10.1186/1748-7188-8-22.

[26] L. Song, et al., “Robust data storage in DNA by de Bruijn graph-based de novo strand assembly,” Nature Communications, vol. 13, no. 1, Sep. 2022, doi: 10.1038/s41467-022-33046-w.

[27] J. Thompson and O. Poch, “New challenges and strategies for multiple sequence alignment in the Proteomics Era,” in Humana Press eBooks, 2005, pp. 475-492. doi: 10.1385/1-59259-890-0:475.

[28] T. N. Petersen, et al., “MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads,” PLOS ONE, vol. 12, no. 5, p. e0176469, May 2017, doi: 10.1371/journal.pone.0176469.

[29] T. Wolf, P. Kämmer, S. Brunke, and J. Linde, “Two’s company: studying interspecies relationships with dual RNA-seq,” Current Opinion in Microbiology, vol. 42, pp. 7-12, Apr. 2018, doi: 10.1016/j.mib.2017.09.001.

[30] C. Anyansi, T. J. Straub, A. L. Manson, A. M. Earl, and T. Abeel, “Computational methods for Strain-Level microbial detection in colony and metagenome sequencing data,” Frontiers in Microbiology, vol. 11, Aug. 2020, doi: 10.3389/fmicb.2020.01925.

[31] I. L. Brito, “Examining horizontal gene transfer in microbial communities,” Nature Reviews Microbiology, vol. 19, no. 7, pp. 442-453, Apr. 2021, doi: 10.1038/s41579-021-00534-7.

[32] R. A. DeVore, G. Petrova, and P. Wojtaszczyk, “Greedy algorithms for reduced bases in banach spaces,” Constructive Approximation, vol. 37, no. 3, pp. 455-466, Feb. 2013, doi: 10.1007/s00365-013-9186-2.

[33] R. Jafari, M. M. Javidi, and M. K. Rafsanjani, “Using deep reinforcement learning approach for solving the multiple sequence alignment problem,” SN Applied Sciences, vol. 1, no. 6, May 2019, doi: 10.1007/s42452-019-0611-4.

[34] Y.-J. Song and D.-H. Cho, “Local alignment of DNA sequence based on deep reinforcement learning,” IEEE Open Journal of Engineering in Medicine and Biology, vol. 2, pp. 170-178, Jan. 2021, doi: 10.1109/ojemb.2021.3076156.

[35] A. Lall and S. Tallur, “Deep reinforcement learning-based pairwise DNA sequence alignment method compatible with embedded edge devices,” Scientific Reports, vol. 13, no. 1, Feb. 2023, doi: 10.1038/s41598-023-29277-6.

[36] M. Muthulakshmi, “A Novel Feature Extraction from Genome Sequences For Taxonomic Classification Of Living Organisms,” Turkish Journal of Computer and Mathematics Education, Apr. 2021, doi: 10.17762/turcomat.v12i2.1364.

[37] F. J. Wrona, T. D. Prowse, J. Reist, and W. F. Vincent, “Climate change effects on aquatic biota, ecosystem structure and function,” ResearchGate, Dec. 2006, doi: 10.1579/0044-7447(2006)35.

[38] V.-K. Bui and C. Wei, “CDKAM: a taxonomic classification tool using discriminative k-mers and approximate matching strategies,” BMC Bioinformatics, vol. 21, no. 1, Oct. 2020, doi: 10.1186/s12859-020-03777-y.

[39] R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi, “CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers,” BMC Genomics, vol. 16, no. 1, Mar. 2015, doi: 10.1186/s12864-015-1419-2.

[40] F. Mock, F. Kretschmer, A. Kriese, S. Böcker, and M. Marz, “Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks,” Proceedings of the National Academy of Sciences of the United States of America, vol. 119, no. 35, Aug. 2022, doi: 10.1073/pnas.2122636119.

[41] B. H. Mendoza-Ramírez, L. Páiz-Medina, T. Salvatierra-Suárez, N. Del Socorro Hernández, and J. A. Huete-Pérez, "A survey of aquatic macroinvertebrates in a river from the dry corridor of Nicaragua using biological indices and DNA barcoding," Ecology and Evolution, vol. 12, no. 11, Nov. 2022, doi: 10.1002/ece3.9487.

[42] H.-T. Vu and L. Le, "Bioinformatics Analysis on DNA Barcode Sequences for Species identification: A review," Annual Research & Review in Biology, pp. 1-12, Dec. 2019, doi: 10.9734/arrb/2019/v34i130142.

[43] M. Emu and S. Sakib, "Species Identification using DNA Barcode Sequences through Supervised Learning Methods,"2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Feb. 2019, doi: 10.1109/ecace.2019.8679166.

[44] L. Jin, J. Yu, X. Yuan, and X. Du, "Fish Classification Using DNA Barcode Sequences through Deep Learning Method," Symmetry, vol. 13, no. 9, p. 1599, Aug. 2021, doi: 10.3390/sym13091599.

[45] P. Qian, et al., "Multi-Target Deep Learning for Algal Detection and Classification," In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jul. 2020, doi: 10.1109/embc44109.2020.9176204.

[46] P. Memmolo, et al., "Learning Diatoms Classification from a Dry Test Slide by Holographic Microscopy," Sensors, vol. 20, no. 21, p. 6353, Nov. 2020, doi: 10.3390/s20216353.