Biomass Conversion and Biorefinery
https://doi.org/10.1007/s13399-023-04766-w

ORIGINAL ARTICLE

Bioinformatics approaches in upgrading microalgal oil for advanced biofuel production through hybrid ORF protein construction

Ihtesham Arshad[1] · Muhammad Ahsan[2] · Imran Zafar[3] · Muhammad Sajid[1] · Sheikh Arslan Sehgal[4] · Waqas Yousaf[5] · Amna Noor[1] · Summya Rashid[6] · Somenath Garai[7] · Meivelu Moovendhan[8] · Rohit Sharma[9]

Received: 13 February 2023 / Revised: 27 June 2023 / Accepted: 15 August 2023

Abstract

Microalgae are promising for biofuel production due to their high oil content and fast biomass growth, but increasing their oil content is essential for economic viability. In this study, we conducted in silico investigations to identify oil-producing genes in various microalgal species. We selected six genes from different species: ACCD and F751_4275 from *Chlorella protothecides*, C2E21_7193 and C2E21_2849 from *Chlorella sorokiniana*, and COO60DRAFT_1295191 and COO60DRAFT_1481410 from *Scenedesmus* sp. We utilized the NCBI genome database and performed BLASTp analysis to identify these genes' superfamilies (PLN02349, DUF212, BKR SDR, PRK08591, ACCD, and SET LSMT). The open reading frames (ORFs) of the selected genes were analyzed using the ORF Finder tool to determine their lengths and the locations of their start and stop codons. Based on this analysis, we constructed two hybrid ORFs by combining the ORFs from different genes. Hybrid ORF 1 had a length of 5166 base pairs, while hybrid ORF 2 was 3516 base pairs long. The thermodynamic evaluation was performed on these hybrid ORFs to assess their stability and GC content. We translated the hybrid ORF sequences into protein sequences using the Translate feature of Expasy. Tertiary structure predictions and bioinformatics approaches were employed to analyze the permissible regions for amino acid dihedral angles, providing insights into the potential functionality of these hybrid ORF proteins. The results of this study indicated that both hybrid ORFs have the potential to produce high lipid contents, making them promising candidates for biofuel production. However, it is essential to conduct further in vitro experiments to validate the functionality of these hybrid proteins. Our study contributes to understanding oil-producing genes in microalgae and their potential applications in the biofuel and pharmaceutical industries. The identified genes and hybrid ORFs provide valuable insights into microalgae species' genetic manipulation and biology, paving the way for advancements in renewable energy and other biotechnological applications.

Keywords Microalgae · Biofuel production · Open reading frames · Hybrid ORFs · Clone designing · Bioinformatics · In silico analysis · Protein structure prediction

1 Introduction

Microalgae have emerged as a promising and sustainable source for producing advanced biofuels. Their ability to efficiently convert solar energy and atmospheric carbon dioxide into lipids, particularly triacylglycerides (TAGs), makes them an attractive alternative to traditional fossil fuels. However, to harness the full potential of microalgal oil for biofuel production, there is a need for advanced techniques to enhance lipid content and quality. One promising approach is the construction of hybrid open reading frame (ORF) proteins, which involves the fusion of functional domains from different proteins to create novel and optimized enzymes. By leveraging bioinformatics tools and techniques, researchers can analyze genomic and transcriptomic data, predict protein structures, and design modifications that improve the efficiency of microalgal oil production.

The escalating global energy crisis, fueled by the exponential increase in energy demands and the depletion of fossil fuel reserves, has become a pressing issue in today's world [1]. The over-reliance on petroleum-based fuels, driven by rapid population growth, urbanization, and industrial development, not only leads to the exhaustion of finite resources but also contributes significantly to releasing CO2 into the atmosphere [2]. With CO2 levels reaching alarming heights of 400 parts per million, the urgent need to address

Extended author information available on the last page of the article

Published online: 25 August 2023

this environmental challenge has become more apparent [3]. Moreover, industrial effluents further compound the issue by intensifying the accumulation of greenhouse gases, exacerbating global warming and climate change [3, 4]. Consequently, scientists worldwide actively seek innovative solutions to mitigate the environmental impact and find sustainable alternatives to combat these challenges [5]. Biofuels have emerged as a compelling area of research, offering a promising avenue for addressing the energy crisis while minimizing CO2 emissions [6, 7]. Derived from plant biomass through chemical and physical processes, biofuels such as biodiesel, bioethanol, and biogas present viable options that exhibit reduced environmental footprints [6, 7]. Plant biomass encompasses any living material, primarily plant-based, that has stored energy through photosynthesis, including wood, plant remnants, and agronomic waste products [8]. This vast array of potential feedstocks provides renewable and sustainable energy generation opportunities. Biodiesel, produced from algal oil extracted from microorganisms such as *Chlorella vulgaris*, represents a promising avenue for biofuel production [9]. Microalgae possess unique characteristics, including producing storage starches comparable to those found in photosynthetic organisms, accumulating large amounts of phosphatidylcholine, and effectively converting CO2 through photosynthetic electrostatic interactions [3]. These features make microalgae an excellent candidate for efficient biodiesel production.

The conversion of Pongamia (Karanja) oil into bioethanol presents another viable option in pursuing sustainable energy sources [10]. Pongamia oil exhibits favorable properties as a feedstock for bioethanol production, and its conversion into ethanol can contribute to the diversification of the biofuel market. Biofuels offer several advantages over traditional fossil fuels. Firstly, they help reduce greenhouse gas emissions, mitigating the adverse effects of climate change. By utilizing plant biomass, biofuels promote carbon neutrality, as the CO2 released during combustion is recaptured by plants during photosynthesis [8]. Biofuels provide opportunities for sustainable agricultural practices and reduce reliance on non-renewable resources, thereby supporting long-term energy security [11, 12]. In addition to experimental approaches, in silico methods have gained prominence in biofuel research due to their cost-effectiveness and ability to provide valuable insights. These computational techniques offer unique guidelines for optimizing biofuel production processes, enhancing efficiency, and reducing costs. Microalgae have garnered significant attention from scientists due to their exceptional lipid content and rapid growth rate [13]. These microorganisms, encompassing both autotrophic eukaryotic and prokaryotic species, thrive in diverse environments and possess chlorophyll pigments that enhance their photosynthetic efficiency compared to terrestrial plants [14]. When compared to macroalgae, microalgae offer numerous advantages, including a simple cellular structure, fast reproduction rates, and a high oil yield, making them an ideal choice for biofuel production by industrial firms [15]. Selecting an appropriate biofuel feedstock is crucial for cost-effective production, and microalgae excel in this regard. They offer an abundant supply of oil suitable for biodiesel production. Biodiesel, a renewable biofuel composed of methyl esters and fatty acids derived from plant oils, microalgal oils, and animal fats, holds immense potential as an alternative fuel source [16]. Among various bioenergy prospects, microalgae stand out due to their high oil content, rapid growth rate, and ease of cultivation. Under optimal conditions of temperature, light, and nutrient availability, microalgae can double their biomass within approximately 24 h [16]. During the exponential growth phase, microalgae can double their biomass every 120 min, showcasing their remarkable productivity potential.

The oil output of microalgae for biofuel production is expected to range between twenty thousand to eighty thousand liters per acre, surpassing other oil-producing crops by approximately 7 to 31 times [16]. Microalgae can produce a diverse range of biofuels, including bio-hydrogen, biogas, bio-oil, bioethanol, and biodiesel, further emphasizing their versatility as a renewable energy resource [17]. The production of biodiesel from microalgae involves two key steps. In the first step, lipids are extracted from microalgal cells, utilizing various extraction methods, to recover the valuable oil content. In the second step, the extracted oil is transformed into biodiesel through a process known as transesterification, where the oil is reacted with an alcohol, typically methanol, in the presence of a catalyst [18]. In the main context, biofuels hold immense potential as sustainable alternatives to address the energy crisis. Biodiesel, bioethanol, and biogas offer environmentally friendly and commercially viable solutions to the increasing demand for energy while minimizing the detrimental effects of CO2 emissions. Further research and development efforts are necessary to optimize biofuel production techniques, improve scalability, and realize the full potential of biofuels in achieving a more sustainable energy future. The microalgae hold immense promise as a biofuel feedstock due to their high oil content, rapid growth rate, and ability to produce various biofuels [19]. Continued research and development efforts are necessary to optimize cultivation techniques, enhance lipid extraction processes, and improve overall efficiency to make microalgae-based biofuels a commercially viable and environmentally sustainable energy option.

This study explores the application of bioinformatics approaches in upgrading microalgal oil for advanced biofuel production through hybrid ORF protein construction. We delve into various aspects, including genomic and transcriptomic analysis to identify candidate ORFs involved in lipid biosynthesis. We uncover conserved domains and

functional motifs that can be harnessed for protein engineering by employing sequence alignment and homology analysis. We investigate the use of protein structure prediction methods to gain insights into the three-dimensional structure of hybrid proteins. This structural information aids in understanding their function, stability, and potential interactions with other molecules involved in lipid metabolism pathways. Furthermore, in silico mutagenesis and rational design techniques are employed to optimize the hybrid protein's properties, such as catalytic activity and substrate specificity, for enhanced microalgal oil production. Through systems biology and pathway analysis, we integrate experimental data with computational models to construct metabolic and regulatory networks associated with microalgal oil production. This holistic approach enables the identification of key genes, pathways, and regulatory elements that can be targeted for maximizing biofuel production efficiency. We explore the potential of data mining and machine learning algorithms to extract meaningful patterns and predictive models from large-scale omics datasets. This data-driven approach allows us to discover novel insights and optimize conditions for improved microalgal oil production and biofuel synthesis. By combining the power of bioinformatics with microalgal oil production, we aim to contribute to developing sustainable and efficient strategies for advanced biofuel production. Applying bioinformatics in hybrid ORF protein construction offers a pathway towards upgrading microalgal oil and unlocking its full potential as a renewable energy resource.

2 Materials and methods

2.1 Datasets

A literature analysis was conducted to obtain comprehensive datasets for essential microalgal superspecies involved in biofuel production. The aim was to identify datasets containing two expression patterns and gene products from the same developmental stage, which would provide valuable insights for research and enhance biofuel production. The analysis focused on *Chlorella sorokiniana*, *Scenedesmus* species, and *Chlorella*, as significant research efforts have been dedicated to these species, resulting in valuable findings, as mentioned in Table 1. Open-access databases like the NCBI Sequence Read Archive (NCBI-SRA) were utilized to access the necessary data, which included ORF and transcriptome datasets. These datasets have proven highly useful for conducting in silico investigations and supporting recent research in the biofuel production field.

Our research utilized the Bligh and Dyer method (1959) for lipid extraction. This method is widely recognized and relies on combining chloroform and methanol to effectively extract lipids from biological samples. The general procedure involved combining chloroform and methanol in a specific ratio, typically ranging from 2:1 to 3:1, to dissolve the lipids present in the sample. Our extraction mixture did not require additional reagents beyond the standard chloroform and methanol combination. We also explored alternative lipid extraction methods employed in recent studies. One such method is the Carpio method (2015), which utilizes a combination of chloroform, methanol, and NaCl as key reagents for lipid extraction. The precise proportions of these reagents may vary depending on the specific methodology employed. Another method we considered is the Long and Abdelkader method (2011), which also employs chloroform, methanol, and NaCl for lipid extraction.

Furthermore, we investigated the Doan method (2011), which differs from the techniques mentioned earlier in terms of its approach and reagents. The Doan method utilizes methanolic-HCl and nitrogen gas for lipid extraction from biological samples. Lastly, we examined the accelerated solvent extraction method with the ASE 350 system, which has emerged as a newer lipid extraction technique. This method involves using a combination of solvents, including methanol, dimethyl sulfoxide, hexane, and ethyl ether, for efficient lipid extraction from biological samples. The accelerated solvent extraction method offers potential advantages over traditional techniques, such as increased speed and efficiency.

2.2 Data retrieval techniques

To identify microalgal species with high oil or lipid content, a comprehensive literature review was conducted. The search was performed using keywords such as "microalgal biofuel," "biofuel," "microalgae as a biofuel feedstock," and "lipid content of microalgal species." The literature was sourced from reputable platforms, including Google Scholar, PUBMED, and NCBI, following the methodology employed by earlier researchers [34–36]. From the collected articles, those directly associated with the specified keywords were selected for analysis. After careful evaluation, three microalgal species exhibiting significant lipid content were identified as promising candidates for further research. These species will be the focus of subsequent investigations to explore their potential as valuable feedstock for biofuel production.

2.3 Identification of microalgal genes and functional proteins

This study conducted a meticulous screening process to select specific microalgal species. Comprehensive data mining was performed to identify genes responsible for oil production utilizing various online resources and the NCBI protein and nucleotide database [37, 38]. Six genes

Table 1 A comprehensive literature analysis for the most significant microalgal superspecies from years 2011 to 2022, which are implicitly and explicitly involved in biofuel production

No.	Microalgal strain	Lipid extraction method	Reagents	Year	References
1	*Scenedesmus obliqnus*	Bligh and Dyer's (1959) method was used for lipid extraction.	Chloroform, H_2SO_4, methanol, C17:0-TAG, and 0.9% NaCl	2021	[20]
2	*Chlorella luteorividis*	An accelerated solvent extraction method with the ASE 350 system was used for lipid extraction.	Methanol, dimethyl sulfoxide, hexane, and ethyl ether	2020	[21]
3	*Chlorella pyrenoidosa*	Bligh and Dyer's (1959) method was used for lipid extraction.	HCL, chloroform, and methanol	2020	[22]
4	*Chlorella sorokiniana*	Carpio's method (2015) was used for lipid extraction.	Chloroform, methanol, and 0.9% NaCl	2019	[23]
5	*Chlorella protothecides*	Bligh and Dyer's (1959) method was used for lipid extraction.	Chloroform and methanol	2017	[24]
6	*Chlorella regularis*	Bligh and Dyer's (1959) method was used for lipid extraction.	Chloroform, methanol, and 0.1% NaCl	2017	[25]
7	*Chlorella vulgaris*	Bligh and Dyer's (1959) method was used for lipid extraction.	Chloroform and methanol	2015	[26]
8	*Scenedesmus bijuga*	Long and Abdelkader's method (2011) was used for the lipid extraction.	Chloroform, methanol, and 0.9% NaCl	2014	[27]
9	*Scenedesmus* sp.	Bligh and Dyer's (1959) method was used for lipid extraction.	Chloroform and methanol	2014	[28]
10	*Chlorococcum infusionum*	Doan's method (2011) was used for the lipid extraction.	Methanolic-HCl and nitrogen gas	2013	[29]
11	*Chlorella saccharophila*	Bligh and Dyer's (1959) method was used for lipid extraction.	Chloroform, methanol, nonadecanoic acid, and butylated hydroxytoluene	2012	[30]
12	*Scenedesmus quadricauda*	Bligh and Dyer's (1959) method was used for lipid extraction.	Chloroform and methanol	2012	[31]
13	*Scenedesmus dimorphus*	Bligh and Dyer's (1959) method was used for lipid extraction.	Chloroform, methanol, 0.9% NaCl, and sodium sulfate anhydrous	2011	[32]
14	*Chlorella ellipsoidea*	Bligh and Dyer's (1959) method was used for lipid extraction.	Chloroform and methanol	2011	[33]

were carefully chosen from three diverse microalgal species for further analysis. The nucleotide sequences of these genes were acquired in FASTA format from the NCBI Nucleotide database as per the methods of earlier researchers [39–42]. To supplement our research, valuable information regarding functional proteins was collected from the NCBI Protein database, enhancing the depth of our investigation.

2.4 Superfamily-based gene screening and ORF analysis

To identify proteins corresponding to the selected genes, the BLASTp software was employed, leveraging superfamilies as the basis for screening. Subsequently, the proteins were further classified based on their respective superfamilies, providing a comprehensive categorization. Within a gene, the ORF encompasses a continuous sequence of nucleotides that can be translated into a protein, encompassing both the start and stop codons [43]. Typically, the largest ORF within a gene is utilized for protein synthesis. To locate the ORFs within the selected genes, the ORF Finder tool—a graphical analytical resource offered by NCBI—was utilized. This enabled the identification of ORF lengths, conserved region sequences, and the positions of start and stop codons within the genes [44]. This detailed analysis facilitated a deeper understanding of the chosen microalgal genes' genetic components and protein-coding potential.

2.5 Creation of hybrid ORFs via superfamily conserved regions

The ORF retrieval from the selected genes was accomplished using the ORF Finder tool, utilizing the accession numbers as identifiers. The most substantial ORFs, identified within the nucleotide sequence format of each gene, were extracted. Subsequently, an amalgamation of these ORFs was performed to generate hybrid ORFs, integrating conserved regions derived from all of the selected genes [45].

2.6 Restriction enzyme analysis of hybrid ORF

To analyze the hybrid ORF molecules, we utilized the SnapGene software [46]. This software allowed us to visualize the molecule and identify the locations of restriction sites where specific restriction enzymes can cut. The information about these restriction sites is crucial for constructing hybrid ORFs for expanded cloning purposes.

2.7 Analysis of hybrid ORF

For confirmation of the hybrid ORFs, we employed the Vector NTI® Express Designer software [47]. This software provided an ORF Finder tool, which allowed us to verify the presence of open reading frames within the hybrid DNA molecule. This step ensured that the hybrid ORFs were adequately constructed and suitable for further analysis.

2.8 Thermodynamic analysis of hybrid ORF

We analyzed thermodynamics using the Vector NTI® Express Designer software [48]. This analysis evaluated various parameters to determine the stability of the hybrid ORFs, providing valuable insights into their structural integrity and potential functional behavior.

2.9 Clone designing of hybrid ORF

To design the clones incorporating the hybrid ORFs of selected genes, we employed the SnapGene software [49]. This software facilitated the construction of the clones by seamlessly fusing the desired gene or gene fragment with a vector. In-fusion cloning, a reliable method for combining genetic elements, was utilized to ensure the efficient generation of the desired clones for downstream applications.

2.10 Protein primary structure prediction

We utilized the hybrid ORF sequence to predict the protein's primary structure. For this purpose, we accessed Expasy, a SIB bioinformatics service site that offers a range of scientific tools and access to scientific data [50]. Specifically, we used the Translate tool provided by Expasy, which enabled us to convert the nucleotide sequence of the hybrid ORF into its corresponding protein primary sequence. This step allowed us to obtain insights into the protein's amino acid composition and sequence.

2.11 Hybrid ORF protein 3D structure prediction

To predict the three-dimensional structures of the proteins encoded by the hybrid ORFs, we employed the I-TASSER server [51]. The primary sequences of both hybrid ORFs were individually submitted to the I-TASSER server, which utilizes advanced algorithms and methods to generate accurate 3D models of proteins. Subsequently, we evaluated the quality and reliability of the generated models. To perform structure validation, we used the Ramachandran plot analysis provided by the PROCHECK website, which assessed the stereochemical quality of the predicted structures [52]. Additionally, we utilized the ERRAT tool within PROCHECK to further validate the three-dimensional structures of the hybrid proteins, ensuring their overall accuracy and reliability as per the investigation of earlier researchers [53–56].

3 Results

3.1 Analysis of microalgal species for oil production

In our study, we analyzed the lipid content in three prominent species: *Chlorella sorokiniana*, *Scenedesmus* sp., and *Chlorella protothecides*. These species are renowned for their high lipid content, which makes them particularly interesting for industries that rely on lipid-rich biomass [57]. Our findings, summarized in Table 2, revealed intriguing results regarding the lipid content of these species. These findings underscore the considerable lipid content found in these three species, making them promising candidates for further exploration and utilization in various industries.

Table 2 presents the lipid oil content and lipid productivity of three microalgal species: *Chlorella sorokiniana*, *Scenedesmus* sp., and *Chlorella protothecides*. *Chlorella sorokiniana* exhibits a lipid oil content of 28.91 ± 2.28% dry weight, indicating its high oil richness. Its lipid productivity is 85.77 ± 21.06 mg/L/day, reflecting the amount of oil produced over a given time. *Scenedesmus* sp. has a higher lipid oil content (38.36 ± 0.80% dry weight) and lipid productivity (131.47 ± 2.40 mg/L/day), making it a potentially superior candidate for oil production compared

Table 2 Microalgal species selected for the study, along with their oil content

No.	Microalgal species	Lipid oil content (%dry weight)	Lipid productivity/ ($mg{\cdot}L^{-1}{\cdot}d^{-1}$)	References
1	*Chlorella sorokiniana*	28.91 ± 2.28	85.77 ± 21.06	[58]
2	*Scenedesmus* sp.	38.36 ± 0.80	131.47 ± 2.40	[59]
3	*Chlorella protothecides*	31.23 ± 1.09	39.60 ± 5.45	[60]

to *Chlorella sorokiniana*. *Chlorella protothecides* shows a lipid oil content of 31.23 ± 1.09% dry weight, similar to *Chlorella sorokiniana*, but has lower lipid productivity (39.60 ± 5.45 mg/L/day). These variations emphasize the importance of selecting the appropriate microalgal species for oil production, as different species possess distinct potentials. *Scenedesmus* sp. is the most promising option due to its high lipid oil content and lipid productivity. However, it is crucial to note that these findings are based on limited studies, and further research is necessary to evaluate each species' oil production potential. Additional considerations such as growth conditions, production costs, and environmental impact should also be considered when selecting the most suitable species for oil production.

3.2 Oil-producing genes and their functional proteins

Genes responsible for microalgal oil production and their corresponding functional proteins were identified using the NCBI genome database. A total of six genes were chosen from different microalgal species, and their respective nucleotide sequences were downloaded in FASTA format for further analysis. The NCBI Protein database was utilized to gather information about the functional proteins associated with these genes. The details of the six selected genes, their corresponding proteins, and their accession numbers are presented in Table 3. In our investigation, a total of six genes were carefully chosen to study their involvement in microalgal oil production. Two of these genes were identified in *Chlorella protothecides*: ACCD and F751 4275.

Additionally, two genes, namely C2E21 7193 and C2E21 2849, were found in *Chlorella sorokiniana*. The remaining two genes, COO60DRAFT 1295191 and COO60DRAFT 1481410, were discovered in *Scenedesmus* sp. Specifically, the gene C2E21 7193 from *Chlorella sorokiniana* was responsible for encoding the glycerol-3-phosphate acyltransferase protein, while the gene C2E21 2849 from the same species encoded the phosphatidic acid phosphatase protein. Moving on to *Scenedesmus* sp., the gene COO60DRAFT 1295191 was identified as the encoding source for the acyl carrier protein. Similarly, the gene COO60DRAFT 1481410 from *Scenedesmus* sp. encoded the biotin carboxylase protein. In the case of *Chlorella protothecides*, the gene ACCD was associated with the mRNA of the acetyl-CoA carboxylase beta subunit (ACCD) protein. Additionally, the gene sequence F751 4275 from *Chlorella protothecides* encoded the ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit N-methyltransferase protein.

The provided information in Table 3 offers a comprehensive overview of the genes present in various microalgae species, including their accession numbers, gene names, species genomic sizes, species median protein counts, species median GC% values, and associated proteins. These genes are involved in key metabolic pathways and enzymes related to carbon metabolism, fatty acid synthesis, and biotic carboxylation. Understanding the genetic makeup of microalgae is crucial for unravelling the intricate mechanisms that regulate carbon metabolism and lipid synthesis in these organisms. By examining the genes involved in these pathways, researchers can gain valuable insights into the underlying biological processes that contribute to biofuel production from microalgae. The data includes lipid oil content and

Table 3 Selection of high-profile species used in biofuel production based on a diverse range of functionalities

No.	Organism	Accession	Gene	Genomic size	Median protein count	Median GC%	Protein	References
1	*Chlorella sorokiniana*	LHPG02000015.1	C2E21_7193	58.6108 Mb	10,384	64.0482	Glycerol-3-phosphate acyltransferase	[58]
2	*Chlorella sorokiniana*	LHPG02000005.1	C2E21_2849				Phosphatidic acid phosphatase	[61]
3	*Scenedesmus* sp.	JACERP010000075.1	COO60DRAFT_1295191	93.2383 Mb	12,172	57.05	Acyl carrier protein	[59]
4	*Scenedesmus* sp.	JACERP010000009.1	COO60DRAFT_1481410				Biotin carboxylase	[62]
5	*Chlorella protothecides*	JN831941.1	ACCD	22.9246 Mb	6433	62.5803	Acetyl-CoA carboxylase beta subunit (ACCD) mRNA	[60]
6	*Chlorella protothecides*	KL662078.1	F751_4275				Ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit N-methyltransferase	[63]

lipid productivity information for three specific microalgal species. *Chlorella sorokiniana* exhibits a lipid oil content of 28.91 ± 2.28% and lipid productivity of 85.77 ± 21.06 mg/L/day. *Scenedesmus* sp. demonstrates a higher lipid oil content of 38.36 ± 0.80% and a productivity of 131.47 ± 2.40 mg/L/day. *Chlorella protothecides*, on the other hand, possess a lipid oil content of 31.23 ± 1.09% and a lipid productivity of 39.60 ± 5.45 mg/L/day. The specific genes associated with these microalgal species have been identified, including glycerol-3-phosphate acyltransferase, phosphatidic acid phosphatase, acyl carrier protein, biotin carboxylase, acetyl-CoA carboxylase beta subunit, and ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit N-methyltransferase. The knowledge of these genes provides valuable insights into the biology of these microalgae species and their potential application in advanced biofuel production.

3.3 Mapping of ORFs and superfamilies

The identification of superfamilies for specific genes was carried out using the BLAST program. Through this analysis, six proteins were found to belong to distinct superfamily: PLN02349, DUF212, BKR SDR, PRK08591, ACCD, and SET LSMT. In this context, it is essential to note that an ORF refers to a specific gene section that includes both the start and stop codons. To locate the ORFs of the selected genes, the ORF Finder program was utilized. This process involved extracting information such as the length and position of the start/stop codons and the nucleotide and amino acid sequences of the ORFs. To facilitate further analysis, hybrid ORFs were created by selecting the largest ORFs from each gene.

3.4 Construction of hybrid ORFs

The ORF Finder tool was employed to extract the sequences of the largest ORFs, while the SnapGene online tool was utilized for building the hybrid ORFs. In total, two hybrid ORFs were generated through this process. The first hybrid ORF, called hybrid ORF 1, was constructed by combining the ORFs from all six genes. As depicted in Fig. 1, hybrid ORF 1 had a length of 5166 bp. The ORFs from the six genes were represented in various colors within the hybrid ORF. The ORF of glycerol-3-phosphate acyltransferase was highlighted in red, phosphatidic acid phosphatase in blue, acyl carrier protein in green, biotin carboxylase in accent orange, acetyl-CoA carboxylase beta subunit (ACCD) mRNA in black, and acyl carrier protein in yellow. Furthermore, the large subunit N-methyltransferase for ribulose-1,5-bisphosphate carboxylase/oxygenase was distinguished in purple.

In this investigation, four distinct ORFs were combined from four different genes, namely F751_4275, C2E21_7193, COO60DRAFT_1295191, and COO60DRAFT_1481410. These ORFs were responsible for encoding the big subunit N-methyltransferase (HNMT, HMT) of various proteins, including ribulose-1,5-bisphosphate carboxylase/oxygenase, glycerol-3-phosphate acyltransferase, acyl carrier protein, and biotin carboxylase. The resulting hybrid ORF, referred to as hybrid ORF 2, possessed a length of 3516 base pairs, as illustrated in Fig. 2. Each of the four ORFs used in the construction was visually differentiated using a distinct color scheme. The ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit N-methyltransferase was represented in purple, glycerol-3-phosphate acyltransferase in red, acyl carrier protein in green, and biotin carboxylase in accent orange.

3.5 Hybrid ORF restriction enzyme analysis

The examination of hybrid ORFs using restriction enzymes was conducted using the SnapGene software. This software facilitated the visualization of restriction sites within the hybrid ORFs, which could be targeted and cleaved by specific restriction enzymes. The analysis revealed the potential obstruction points of various restriction enzymes

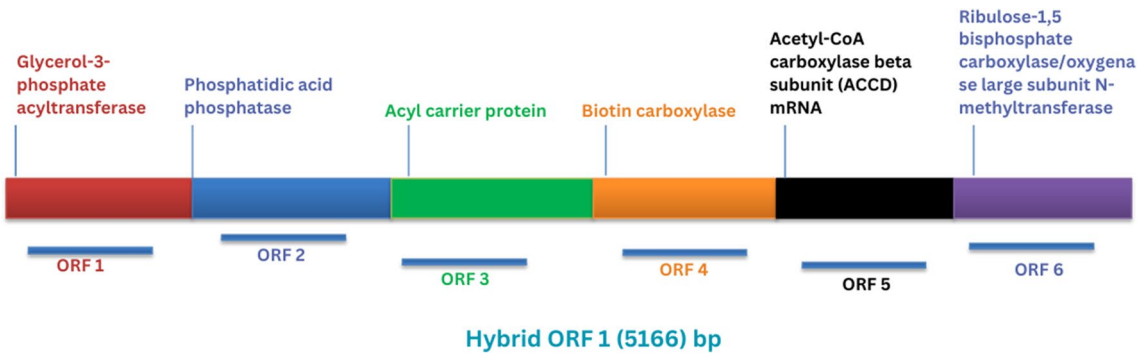

Fig. 1 A hybrid ORF 1 designed using SnapGene

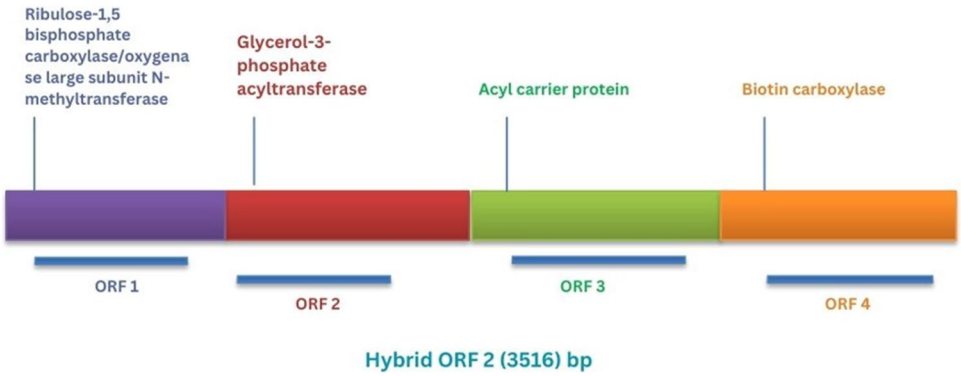

Fig. 2 Hybrid ORF 2 constructed by using SnapGene

along the hybrid DNA sequences in hybrid ORF 1 and hybrid ORF 2.

We performed restriction enzyme analysis of hybrid ORF 1 demonstrating that different restriction enzymes could impede or cleave the sequence at specific locations. Each restriction enzyme used in the study was assigned a distinct color to aid visual differentiation. The corresponding results of the restriction enzyme analysis of hybrid ORF 1 are depicted in Fig. 3. Similarly, Fig. 4 showcases the restriction enzyme analysis of hybrid ORF 2, illustrating that different restriction enzymes could obstruct or cleave the sequence at various positions. A different color represented each restriction enzyme utilized in the construction process.

3.6 Hybrid ORF analysis

The sequences of both hybrid ORFs were analyzed using Vector NTI® Express Designer software to verify their ORFs. The ORF Finder tool in Vector NTI® Express Designer was used to confirm the hybrid ORF 1, and it identified ORFs on both the forward and reverse strands. The forward-strand ORFs were represented by arrows pointing from left to right, while the reverse-strand ORFs were represented by arrows from right to left. For hybrid ORF 2 verification, the ORF Finder in Vector NTI® Express Designer was employed, which revealed ORFs on both the forward and reverse strands. The forward-strand ORFs were indicated by arrows pointing from left to right, and the reverse-strand ORFs were represented by arrows from right to left. Figures 5 and 6 display the results of constructing two hybrid ORFs, hybrid ORF 1 and hybrid ORF 2. The directional arrows in the figures indicate the locations of the ORFs on both the forward and reverse strands. Different colors were used to represent the ORFs of various genes involved in the construction. The ORF of ribulose-1,5-bisphosphate carboxylase/large oxygenase subunit N-methyltransferase was shown in purple, glycerol-3-phosphate acyltransferase in red, acyl carrier protein in green, and biotin carboxylase in accent orange. These figures visually illustrate the study's findings and provide a clear depiction of the ORF locations within both hybrid ORFs.

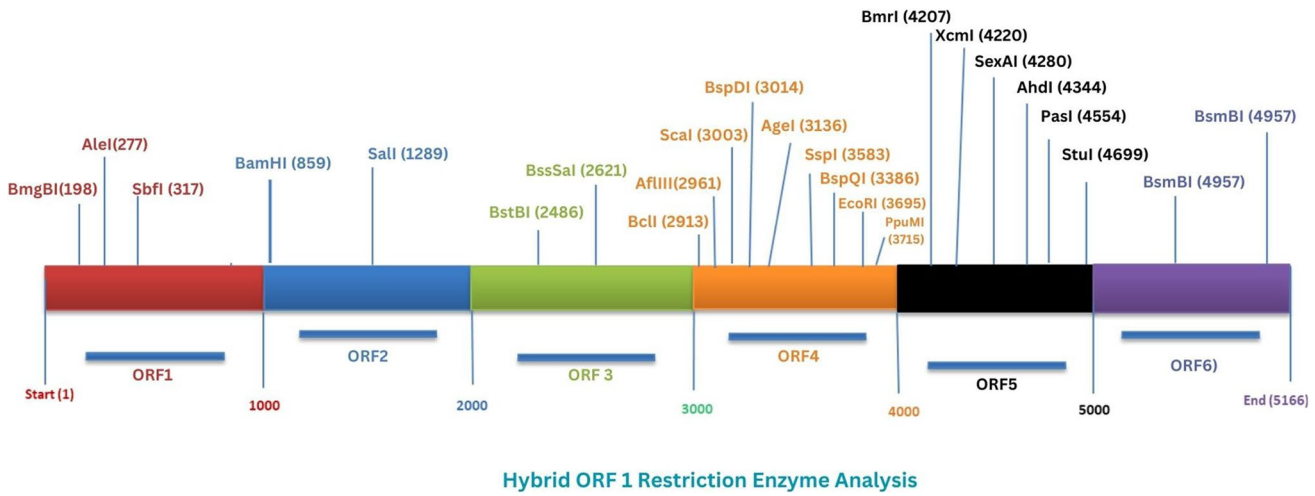

Fig. 3 Restriction enzyme analysis of hybrid ORF 1

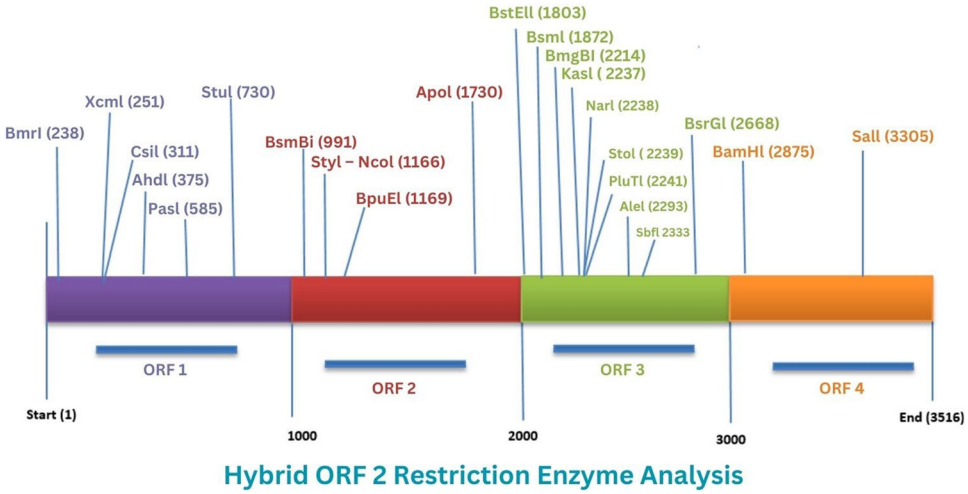

Fig. 4 Restriction enzyme analysis of hybrid ORF 2

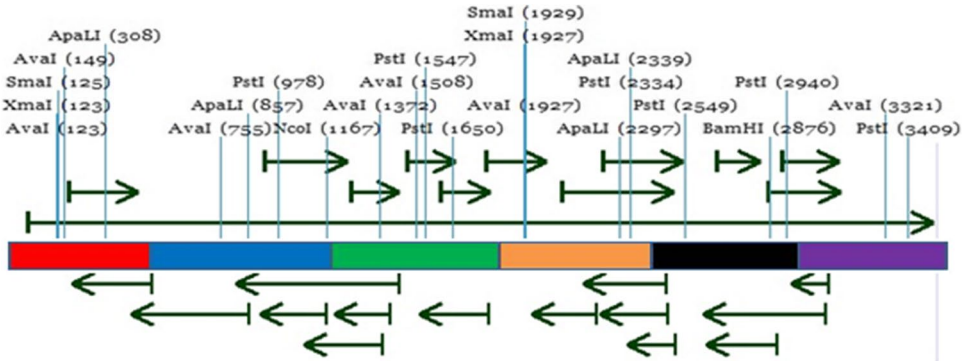

Fig. 5 Hybrid ORF 1

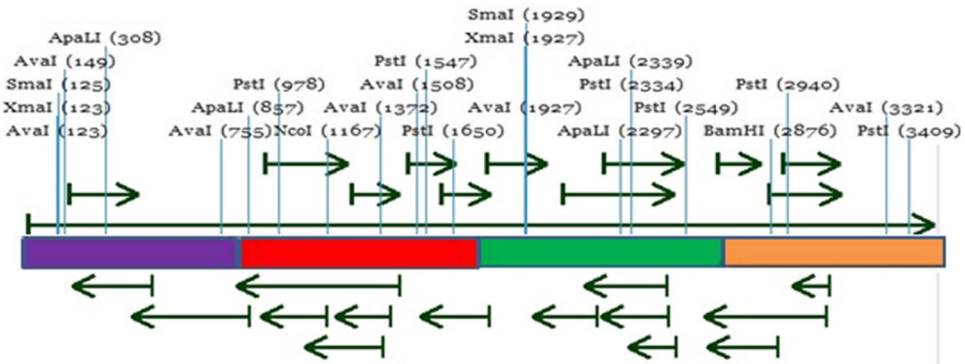

Fig. 6 Hybrid ORF

3.7 Hybrid ORF thermodynamic analysis

The stability of the hybrid ORFs was assessed using thermodynamic evaluation in the Vector NTI® Express Designer software. Tables 4 and 5 present the results of the thermodynamics analysis for hybrid ORF 1 and hybrid ORF 2. The analysis considered several parameters, including dG temperature (°C), probe concentration, and salt concentration, as the baseline analytical settings. The study focused on molecular weight, GC%, thermodynamic temperature, and GC% temperature. The Vector NTI® Express Designer predicted two melting temperatures for DNA/RNA oligonucleotides: thermodynamic and GC% temperature. The

Table 4 Hybrid ORF 1 sequence thermodynamic analysis through Vector NTI® Express Designer

No.	Thermodynamic analysis parameters	Resulted value
1	Probe concentration (pMol)	250.0
2	dG temperature (Celsius)	25.0
3	Formamide percentage	0.0
4	Salt concentration (mMol)	50.0
5	GC content	61.2%
6	Thermodynamic temperature	100.0
7	Percentage GC temperature	84.4
8	Palindromes (base pairs)	6
9	Stem length (base pairs)	3
10	3′ end dG	−14.7
11	Nucl. repeats (base pairs)	4
12	3′ end length (base pairs)	7
13	dH	−8232.3
14	Molecular weight	308,313.8
15	dS	−20,324.7
16	dG	−2170.7

Table 5 Hybrid ORF 2 sequence thermodynamic analysis through Vector NTI® Express Designer

No.	Thermodynamic analysis parameters	Resulted value
1	Probe concentration (pMol)	250.0
2	dG temperature (Celsius)	25.0
3	Formamide percentage	0.0
4	Salt concentration (mMol)	50.0
5	GC content	73.2%
6	Thermodynamic temperature	100.0
7	Percentage GC temperature	89.3
8	Palindromes (base pairs)	6
9	Stem length (base pairs)	3
10	3′ end dG	−12.8
11	Nucl. repeats (base pairs)	4
12	3′ end length (base pairs)	7
13	dH	−9063.1
14	Molecular weight	309,114.3
15	dS	−22,137.4
16	dG	−2461.0

experiment was conducted with default values of 250 pM for the probe concentration and 50 mM for the salt concentration.

Hybrid ORF 1 was constructed with a GC content of 61.2% and exhibited a higher GC% temperature of 84.4, as indicated in Table 4. Similarly, hybrid ORF 2 had a GC content of 73.2% and a GC% temperature of 89.3, as shown in Table 4. GC concentrations exceeding 60% are generally considered favorable for gene design, protein expression, and primer designing in polymerase chain reactions. The strength of hydrogen bonding between GC base pairs significantly influences the stability of DNA. Moreover, the GC concentration affects the secondary structure of mRNA and the annealing temperature of DNA templates in polymerase chain reactions. The stability of the molecules depicted in the tables was assessed based on the analysis mentioned above.

Table 4 provides the results of a thermodynamic analysis. The parameters measured include probe concentration, temperature, formamide percentage, salt concentration, GC content, thermodynamic temperature, percentage GC temperature, palindromes, stem length, 3′ end dG, nucleic repeats, 3′ end length, dH, molecular weight, dS, and dG. The values in the "Resulted value" column represent the results of the analysis for each parameter. The probe concentration (250.0 pMol) represents the amount of probe used in the analysis. The thermodynamic Gibbs free energy (dG) at a temperature of 25.0 degrees Celsius and the concentration of formamide (0.0%) both influence the stability of the probe. The salt concentration (50.0 mMol) represents the salt present in the solution and can also impact probe stability. The GC content (61.2%) is the percentage of G and C nucleotides in the sample. The thermodynamic temperature (100.0) and the percentage GC temperature (84.4) are related to the stability of the probe and can impact its ability to bind to target molecules. The palindromes (6 base pairs), stem length (3 base pairs), and 3′ end dG (−14.7) are parameters related to the structure of the probe. The nucleic repeats (4 base pairs) and 3′ end length (7) can also impact probe stability and specificity. The dH (−8232.3), molecular weight (308,313.8), dS (−20,324.7), and dG (−2170.7) are thermodynamic parameters that describe the stability of the probe and the strength of its binding to target molecules.

Table 5 presents the results of a thermodynamic analysis performed on a sample. The parameters analyzed include probe concentration, temperature, formamide percentage, salt concentration, GC content, thermodynamic temperature, GC temperature, palindromes, stem length, 3′ end Gibbs free energy, nucleotide repeats, 3′ end length, enthalpy, molecular weight, entropy, and Gibbs free energy. The results showed a probe concentration of 250.0 pMol, a dG temperature of 25.0 Celsius, 0.0% formamide, a salt concentration of 50.0 mMol, a GC content of 73.2%, a thermodynamic temperature of 100.0, a GC temperature of 89.3, 6 palindromic base pairs, a stem length of 3 base pairs, a 3′ end Gibbs free energy of −12.8, 4 nucleotide repeats, a 3′ end length of 7 base pairs, an enthalpy of −9063.1, a molecular weight

of 309,114.3, an entropy of −22,137.4, and a Gibbs free energy of −2461.0.

3.8 Designing of clone

The pDE-Cas9 cloning vector, a plasmid-based tool, was utilized to insert hybrid ORF 1 at a specific location within the plasmid. SnapGene software, known for its genetic construct design, visualization, and annotation capabilities, facilitated the cloning process. By employing SnapGene, hybrid ORF 1 was successfully cloned into the pDE-Cas9 cloning vector, generating a modified plasmid harboring the hybrid ORF 1. This modified plasmid can then express the hybrid protein in a suitable host cell.

The cloning process was also carried out for hybrid ORF 2, generating clones of both hybrid ORFs using SnapGene software. The sequence of hybrid ORF 1 had a length of 5166 bp, while the sequence of hybrid ORF 2 was 3516 bp long. Figures 7 and 8 display the in-fusion cloning vectors, namely pDE-Cas9 and pGL-basic. The pDE-Cas9 vector, with a size of 15,758 bp, was employed for the in-fusion cloning of hybrid ORF 1. On the other hand, the pGL-basic vector, with a size of 4818 bp, was used for the in-fusion cloning of hybrid ORF 2.

The in-fusion cloning technique was employed to generate clones of both hybrid ORFs. In-fusion cloning is a versatile method that enables the seamless fusion of a desired gene or gene fragment with a vector. The process involves mixing the relevant overlapping-end DNA fragments with a linearized vector for the infusion procedure. SnapGene software automatically adds appropriate primers for the vector and the fragment using PCR. After adding the fragment to the linearized vector, SnapGene reveals the fused result, confirming the successful cloning.

3.9 Primary structure prediction of protein

The protein sequences of all hybrid ORFs were generated by translating their respective nucleotide sequences using the Translate feature provided by Expasy. Each hybrid ORF produced six frames, three in the 5′ to 3′ direction and three in the 3′ to 5′ direction. The Translate tool in Expasy can convert amino acid query sequences to nucleotide query sequences, which can be utilized for predicting tertiary structures.

Careful selection was made to identify the frames with the most ORFs. Figures 9 and 10 visually depict the primary protein sequences of hybrid ORF 1 and hybrid ORF 2, respectively. The primary protein sequences are highlighted

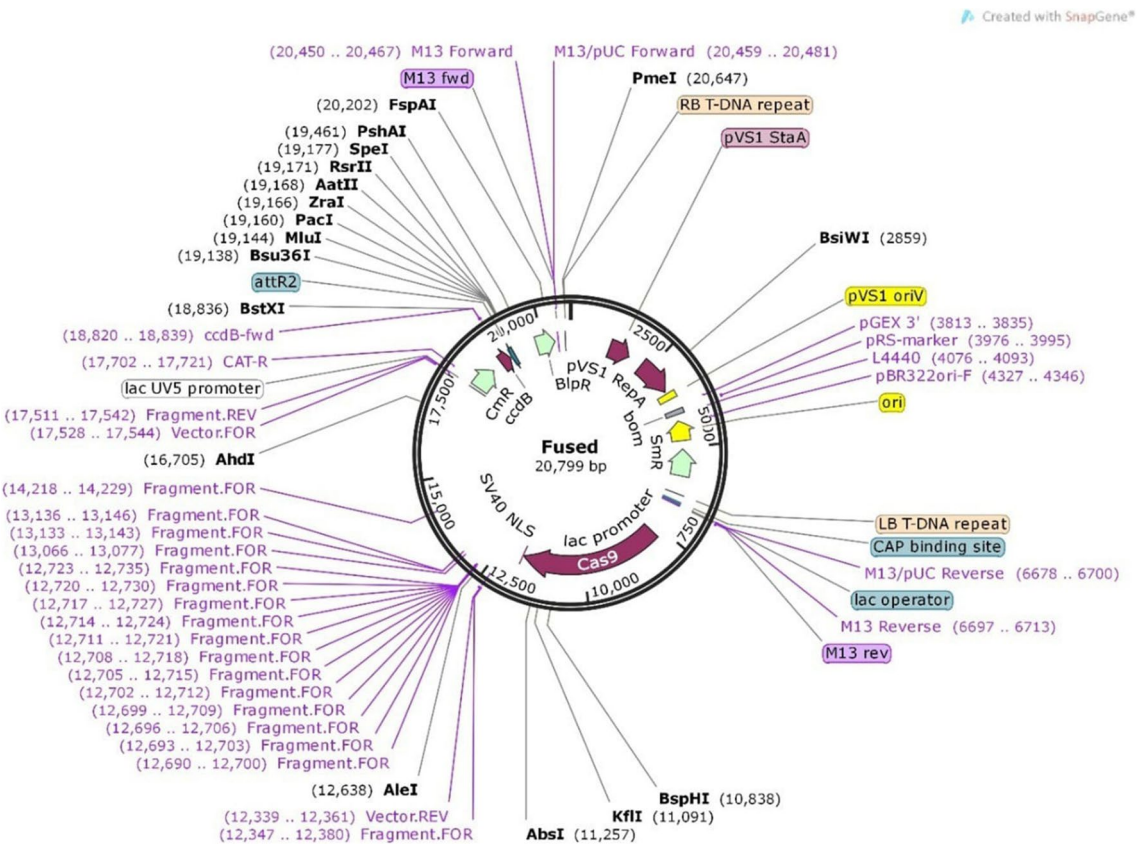

Fig. 7 *pDE-Cas9* cloning vector for the in-fusion cloning of hybrid ORF 1 by using SnapGene

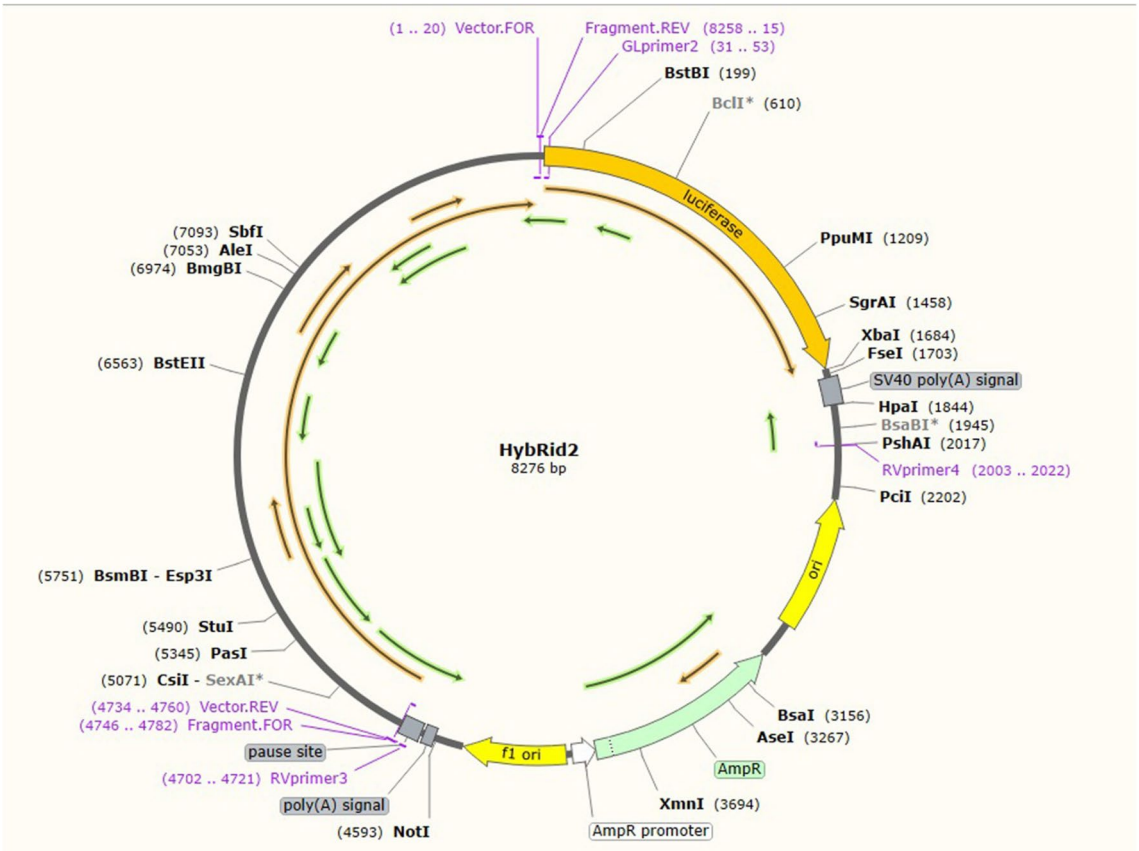

Fig. 8 pGL-basic cloning vector for the in-fusion cloning of hybrid ORF 2 by using SnapGene. The area with the gray color indicates the sequence of hybrid ORF 2

```
MNPCCCCCCQPMLPTVRHAVPLQGWPPVPDAALLSLCCTQTPRSSAGHTGQAAGPDMRSWCVCSTHVLCECGAGRLLQQLLVAALDGAVALTQVHHVAKLVSQHLQGVHQQQQQQQQQQQQQQQG
GEPEASKPAAQTAGAACCRVSKQAQEQQQEHNTSKVDCQLTPAHTTTTKLSCSSSPQAAAAMILEHNPAQRYAHTSITPGTRCGVGSARSARCTRCRCQRLPLPPAAPPSAAAGSPPAWWPGACH
AHHHLMRSSSSSQCHAVPAAGVEPVNNSKVSDATSKGSPVGRPEGAPFKRMCSAALLQLHSSTSLLHRRHCCMPGSGMRHTTCWTRCGWPSSAHQCRLRFGCAWLLLFGMRQPHLLWPQQAKQLY
SAVLKVTLLLCKQDSSCRAGLVLQPSLACRTATARLLSAVHCRFHSQCQGSQHPGRRTWPRGQPEGNARKRAQSAQGLLQVNNQRGAAAAPRRYLLDLSAQLLRNDGAHISQHLCLVLRSKIGAP
CPWPTLLPSALLNRSAGDCARARGHLRLPGAGCGGGAGGHWCADCCLLLRLLGCDLGPSCPQLGLEAVLAGTGELHASFAVLHLTSAQMLCLCNQQQLGVWQLERYSPPLASCPCRGQGAAAGQA
GGALLRGGLLGLLCLLGLLGLLLLQWKKGGRASGCKLCARAVPACRLLSCAQIRVTPFFIDFVCRRVCKEYNRMQAVTLRFWAVVGNLSLQLLLAPMNAGGVRGVRPHAGGDSGARVPGGRGLQP
ALAGIRSRAPGVIGLALCGLKRLLPPALAAAPLPGTGAAAAAAYSSRLPSCCRLCSSLRRGKRATSGVSGTKQGSAGQFEPGCACTRHRPRRLPPLTWPAPHPPPLAAPALPPLPPAPPLLAAAR
VGAAAAPAAAARRRGCPPPSCCTAARLSDAPGLLLQAAALLPCHSRAPRRTAAAGGRSPCVSILSWIENQRKLKLLNAPKYNQPESDGSQGLWTRCDHCGVILYIKHLKENQRVCFGCGYHLQMS
STERIDYLIDTGTWRPFDETVSPCDPLEFRDQKAYTERLKDAQERTGLQDAVQTGTGLLDGIPVALGVMDFNFMGGSMGSVVGEKITRLIEYATQEGLPVILVCASGGARMQEGILSLMQMAKIS
AALHVHQNCAKLLYISVLTSPTTGGVTASFAMLGDPIFAEPKALIGFAGRRVIEQTLQEQLPDDFQTAEYLLHHGLLDLIVPRSFLKQALSETITLYKEAPLKMQGRIPYGERGPLTKIREEQLR
RFAKAPKNPQYSNLVAEFEQLLELLTSDKNMLSSVDVFAPETVNKAFELACGSQTRLDWLNDKTNQFRLRPVFMQLRESWKALTGLASRLAAEAGAPSPPTESDFLWAHAVFWSRAIAFPCPPEA
LPGGRSTTAGPSPQEGVVPGLDFCNHAMRPVARWTVYGAELDGAGGGRRHVHLVPGVAPPAPGAELCISYGDDKSNEELLFLYGFALPEASGLARLMLPLPLPAAAEDWTPRLHARVALLAARGL
PPQVFLPRAALAGGGPGPLGAVPDAALRTLEVFVLDDAELARGLEAAERGDGAAGAEGAPGAEGPEACAARARLERGRVRETRVRDSPFQLGRSGCACRVRHAVFWHVHLLMGEDCHSNCPQPPL
PLPACPDPGLRLALTGALVGLLQLMVQGLETGDGCQTLEADAALSSTPTLMAGLSANQRAALMYRMEQKSLARDWLAHAKSLRAMAMEELRAQGAG-
```

Fig. 9 Primary protein sequence of hybrid ORF 1

```
5'3' Frame 1
MQLRESWKALTGLASRLAAEAGAPSPPTESDFLWAHAVFWSRAIAFPCPPEALPGGRSTTAGPSPQEGVVPGLDFCNHAMRPVARWTVYGAELDGAGGGRRHVHLVPGVAPPAPGAELCISYGDD
KSNEELLFLYGFALPEASGLARLMLPLPLPAAAEDWTPRLHARVALLAARGLPPQVFLPRAALAGGGPGPLGAVPDAALRTLEVFVLDDAELARGLEAAERGDGAAGAEGAPGAEGPEACAATRA
RLERGRVRETRVRDSPFQLGRSGCACRVRHAVFWHVHLLMGEDCHSNCPQPPLPLPACPDPGLRLALTGALVGLLQLMVQGLETGDGCQTLEADAALSSTPTLMAGLSANQRAALMYRMEQKSLA
RDWLAHAKSLRAMAMEELRAQGAGPCPWPTLLPSALLNRSAGDCARARGHLRLPGAGCGGGAGGHWCADCCLLLRLLGCDLGPSCPQLGLEAVLAGTGELHASFAVLHLTSAQMLCLCNQQQLGV
WQLERYSPPLASCPCRGQGAAAGQAGGALLRGGLLGLLCLLGLLGLLLLQWKKGGRASGCKLCARAVPACRLLSCAQIRVTPFFIDFVCRRVCKEYNRMQAVTLRFWAVVGNLSLQLLLAPMNAG
GVRGVRPHAGGDSGARVPGGRGLQPALAGIRSRAPGVIGLALCGLKRLNPCCCCCCQPMLPTVRHAVPLQGWPPVPDAALLSLCCTQTPRSSAGHTGQAAGPDMRSWCVCSTHVLCECGAGRLLQ
QLLVAALDGAVALTQVHHVAKLVSQHLQGVHQQQQQQQQQQQQQQQGGEPEASKPAAQTAGAACCRVSKQAQEQQQEHNTSKVDCQLTPAHTTTTKLSCSSSPQAAAAMILEHNPAQRYAHTSIT
PGTRCGVGSARSARCTRCRCQRLPLPPAAPPSAAAGSPPAWWPGACHAHHHLMRSSSSSQCHAVPAAGVEPVNNSKVSDATSKGSPVGRPEGAPFKRMCSAALLQLHSSTSLLHRRHCCMPGSGM
RHTTCWTRCGWPSSAHQCRLRFGCAWLLLFGMRQPHLLWPQQAKQLYSAVLKVTLLLCKQDSSCRAGLVLQPSLACRTATARLLSAVHCRFHSQCQGSQHPGRRTWPRGQPEGNARKRAQSAQGL
LQVNNQRGAAAAPRRYLLDLSAQLLRNDGAHISQHLCLVLRSKIGA-
```

Fig. 10 Primary protein sequence of hybrid ORF 2

in red color, providing a clear visualization of the translated sequences. These protein sequences will also be valuable for predicting tertiary structures.

3.10 Tertiary structure prediction

Predicting the tertiary structure of hybrid ORF 1 protein involves determining its three-dimensional shape based on the amino acid sequence and its residues' physical and chemical properties. Computational methods, such as molecular dynamics simulations or homology modeling, are typically employed for this prediction. These methods help understand the protein's shape, interactions with other molecules, and potential functions. The prediction accuracy relies on the quality of the protein sequence data, the complexity of the protein, and the reliability of the algorithm's algorithm. However, it is essential to note that the prediction of a protein's tertiary structure may not always be entirely accurate, and experimental validation is necessary to confirm the predictions.

In this study, I-TASSER, a program known for predicting three-dimensional structures of protein molecules from their amino acid sequences, was employed. I-TASSER utilizes a fold recognition technique to search the Protein Data Bank for structural templates. The full-length structure models are constructed by reassembling structural components using threading templates, as described in previous research studies [53, 54]. I-TASSER has demonstrated high accuracy in protein structure prediction during community-wide CASP investigations. In the current study, I-TASSER was utilized to construct the three-dimensional structures of the hybrid ORF proteins, and the resulting structures were validated using the Ramachandran plot.

I-TASSER searches for threading templates similar to the query sequence based on the input amino acid sequence. It provides five structure models for each hybrid protein, which are evaluated using various methods such as ERRAT, PROCHECK, PROVE, WHAT CHECK, and VERIFY 3D. Based on these evaluation values, the best structure models are selected. Figures 11 and 12 present the three-dimensional architectures of the hybrid ORF proteins as generated by I-TASSER.

The three-dimensional structure of the hybrid ORF 1 protein represents a unique combination of the individual ORFs from different genes used to construct it. This structure is determined by the amino acid sequences within the protein and their interactions in three-dimensional space. Understanding the three-dimensional structure of hybrid ORF 1 protein is crucial for comprehending its function, stability, and interactions with other molecules. This knowledge can be valuable in developing novel applications and technologies that utilize this hybrid protein.

3.11 Structure assessment

The Ramachandran plot, or the Rama plot, is a bioinformatics approach used to illustrate permissible regions for the backbone dihedral phi angles of amino acids against psi angles. It provides insights into the stereochemical properties of a protein and is commonly generated using software like PROCHECK. This study utilized the PROCHECK software to generate a Ramachandran plot for hybrid ORF 1.

Figure 13 presents the Ramachandran plot for hybrid ORF 1, showing the distribution of residues in different regions. The plot indicates that 73.1% of the residues are located in highly favorable regions, 21.6% in extra-allowable regions, 3.0% in generously allowed regions, and 1.8% in prohibited regions. Specifically, the plot reveals that out of

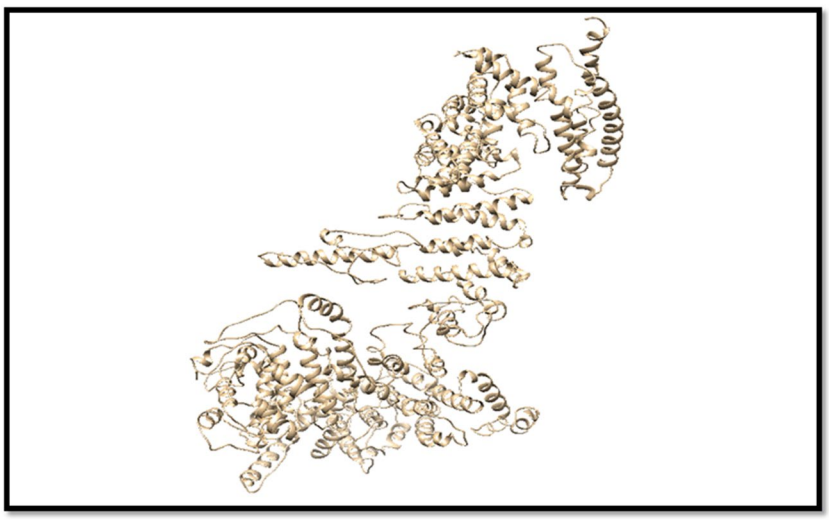

Fig. 11 Three-dimensional structure of hybrid ORF 1 protein

Fig. 12 Three-dimensional structure of hybrid ORF 2 proteins

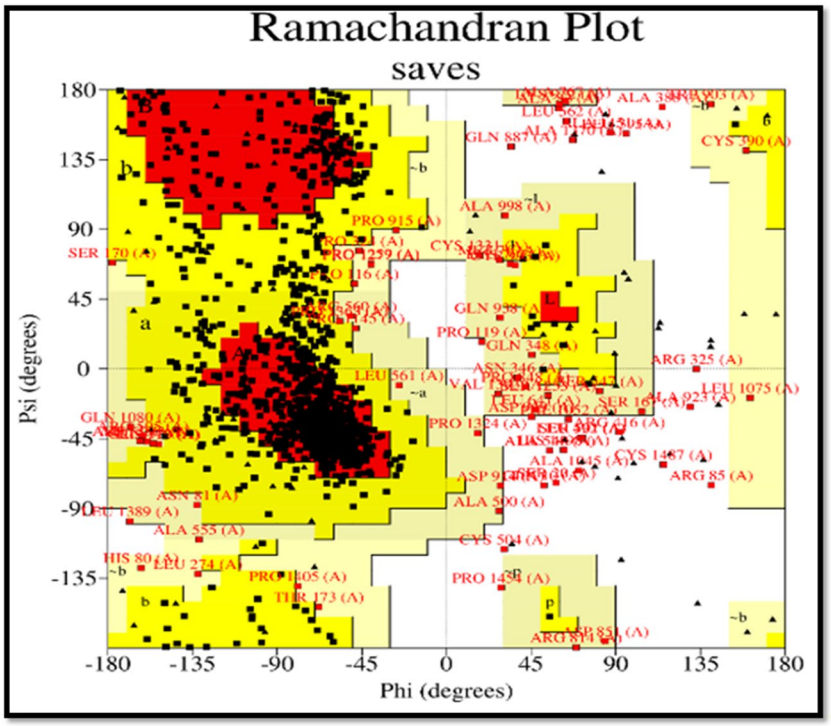

Fig. 13 Structure assessment of hybrid ORF 1 by using Ramachandran plot

the total 1232 non-glycine and non-proline residues, 907 (73.6%) are present in the most favored regions, 266 (21.6%) in additional allowed regions, 37 (3.0%) in the generously allowed region, and 22 (1.8%) in disallowed regions. The plot also includes 122 proline residues and 143 glycine residues, which are represented separately.

The Ramachandran plot serves as a visual representation in protein structure analysis to assess the conformations of protein residues in three dimensions. It provides valuable information about the orientation of amino acid side chains in space through the phi and psi angles. In the case of hybrid ORF 1, the Ramachandran plot was used to evaluate the stereochemistry of the protein and determine if it adopts a proper and stable conformation. Additionally, the plot aids in identifying residues that may be in unfavorable positions, which could affect the protein's folding or stability. This information is crucial for optimizing the protein's expression, stability, and function.

According to Fig. 14's Ramachandran plot for the hybrid ORF 2, 67.4% of the residues are in the highly favorable region, 25.2% are in the additionally allowed region, 4.6% are in the generously permitted region, and 2.8% are in the prohibited region. The structure assessment of hybrid ORF 2 using the Ramachandran plot is illustrated in Fig. 14. It

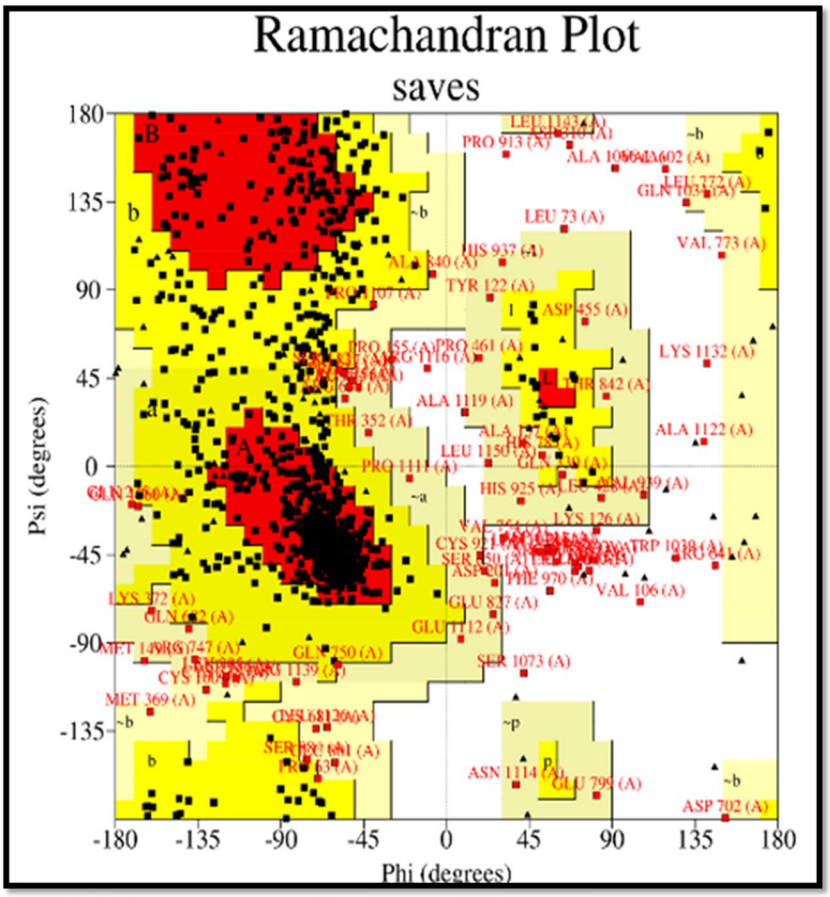

Fig. 14 Structure assessment of hybrid ORF 2 by using Ramachandran plot

showed that 646 (67.4%) residues were present in most favored regions, 242 (25.2%) residues were present in additional allowed regions, 44 (4.6%) residues were present in the generously allowed region, and 27 (2.8%) residues were present in disallowed regions. The total numbers of non-glycine and non-proline residues are 959, where end residue glycine and proline are 2. In detail, 119 glycine residues are shown in triangular forms. We find a total of 1171 residue with different natural of functionalities.

4 Discussion

Microalgae have emerged as a promising feedstock for the production of biofuels, attracting significant attention from the scientific community. With their high oil content and rapid biomass production, microalgae offer a potential solution to our growing energy requirements and the depletion of fossil fuels. Previous researchers have contributed valuable insights into this field, while the exploration of Table 1 and Table 2 demonstrate their potential as a biofuel source, demonstrating the data on the oil content and biomass productivity of different microalgae species. However, the overall process remains costly, from farming to oil extraction. To address this challenge, increasing the oil content of microalgae is crucial to offset production costs and ensure economic viability. Researchers [64] have explored strategies such as strain selection, cultivation system optimization, and oil content enhancement through nutrient manipulation, genetic engineering, and stress induction. Moreover, efforts are underway to develop cost-effective harvesting and extraction methods. Ongoing research and advancements in this field aim to optimize microalgae-based biofuel production, making it more economically viable and sustainable for our energy needs.

In this research, we explored various traditional methods based on the suggestions and insights of researchers from 2019 to 2022. These methods involve the use of different biological organisms for the production of bio-based energy, including bacteria [65], yeast [66], plants [67], and microalgae [68]. Among these organisms, microalgae have been identified as the most promising source of biobased energy. This is due to their fast growth rate, high lipid content, and the ability to be cultivated using inexpensive nutrients such as wastewater, as indicated by earlier researchers [69]. Numerous microalgal strains have been identified for biofuel research, including *Chlorella Crucigenia* [70], *Spirogyra spp.* [71], *Oscillatoria spp.* [72] *Chlamydomonas spp.*

[73], *Euglena spp.* [74], *Cladophora spp.* [75], *Closterium spp.* [76], *Chlorella sorokiniana* [77], *Scenedesmus* sp. [78], *Chlorella protothecides* [79], *Microcystis spp.* [80], *Ankistrodesmus spp.* [81], *and Pediastrum spp.* [82]. Our research focuses on selected microalgae species that exhibit higher lipid content than others, specifically *Chlorella sorokiniana*, *Scenedesmus* sp., and *Chlorella protothecides*. These strains have lipid oil contents of 28.91 ± 2.28%, 38.36 ± 0.80%, and 31.23 ± 1.09%, respectively, as determined through our investigations and results were matched with earlier researcher [57].

Genome-based identification and comparative analysis of enzymes involved in carotenoid biosynthesis in microalgae have implications for biofuel production [83]. By using bioinformatics tools and comparative genomics, researchers can identify and annotate genes encoding these enzymes, which share pathways with biofuel production. Public databases like the NCBI are leveraged to analyze gene sequences, organization, and regulation across microalgal species, providing insights into the evolution and diversity of carotenoid biosynthesis and biofuel production pathways. The researcher [84] is engaged in functional annotation and the delineation of sequence-structure attributes of conjectural proteins associated with carotenoid biosynthesis and the generation of biofuels. Techniques such as sequence homology searches, domain prediction, and structural modeling help assign potential functions to these proteins [84]. This integrated approach enhances our understanding of carotenoid biosynthesis in microalgae and facilitates the identification of genetic engineering targets for increasing carotenoid production and biofuel production efficiency. By leveraging shared pathways and insights gained from carotenoid biosynthesis research, microalgae can be optimized as a sustainable source for biofuel production, contributing to the development of renewable energy solutions [83, 84]. As per investigation of earlier researcher [85], microalgae biodiesel can be recovered using a liquid biphasic flotation system, which separates microalgae biomass from the growth medium based on density differences. This approach provides an efficient method for concentrating microalgae cells for further processing. Biodiesel from microalgae involves lipid extraction and transesterification, resulting in a renewable and sustainable fuel source.

Microalgae possess specific oil-producing genes, and we have identified six such genes from the strains above that act as precursors for biofuel production. We retrieved the relevant information from the NCBI database to obtain their nucleotide sequences, employing the methodology outlined by previous researchers [86]. Subsequently, we utilized the ORF Finder tool to identify the longest ORFs within these genes. The identified ORFs encode their respective functional proteins, following the findings of earlier researchers [87]. To enhance lipid production in different understudied microalgae, we combined these identified ORFs with hybrid ORFs, as suggested by researchers [45]. We followed the same methods employed in predicting our results. We incorporated current and innovative methods based on recent research indications for further analysis. Online resources such as SnapGene [88], Vector NTI® Express Designer [89], Expasy's Translate tool [90], I-TASSER [91], evaluation tools [92], and Ramachandran plot [14] were utilized for various purposes. These included restriction enzyme analysis, hybrid ORF analysis, thermodynamic analysis, clone designing, primary structure prediction, tertiary structure prediction, structure evaluation, and structure assessment.

Our study involved conducting in silico investigations to identify oil-producing genes in multiple microalgal species. To achieve this, we utilized the NCBI genome database and followed the methods established by earlier researchers [74]. Among the selected genes, we identified six specific ones: two ACCD and F751 4275 from *Chlorella protothecides*, C2E21 7193 and C2E21 2849 from *Chlorella sorokiniana*, and two COO60DRAFT 1295191 and COO60DRAFT 1481410 from *Scenedesmus* sp., as presented in Table 3. To further analyze these genes, we employed the BLASTp program to determine their associated superfamilies, comparing them with findings from previous research [44]. To gain a deeper understanding of the genes, we analyzed their ORFs using the ORF Finder tool. This analysis allowed us to identify the length and location of the start and stop codons. Based on our investigations, the resulting hybrid ORF 1 was 5166 bp in length, as depicted in Fig. 1. Subsequently, we combined four ORFs from different genes, namely F751 4275, C2E21 7193, COO60DRAFT 1295191, and COO60DRAFT 1481410, to create the second hybrid ORF, following the methods employed by earlier researchers [51]. Our predicted hybrid ORF 2 had a length of 3516 bp. By utilizing the ORF Finder tool, we analyzed the chosen genes' ORFs and obtained information regarding the length and location of the start and stop codons. To validate our methodology, we compared our results with previous studies.

In our study, we employed SnapGene to conduct a restriction enzyme analysis of hybrid ORF 1 and ORF 2. The analysis revealed that different restriction enzymes could impede hybrid ORF 1 at various points, which was consistent with earlier studies [46]. The thermodynamic evaluation of the hybrid ORFs was carried out using the Vector NTI® Express Designer, following the methods established in previous

research. This technique allowed us to assess the stability of the molecules. Our study presented the outcomes of the thermodynamics analysis in Tables 4 and 5. For instance, hybrid ORF 1 was created with a GC content of 61.2% and a GC% temperature of 84.4, which was higher compared to earlier studies (Table 4), and our results were in line with previous findings.

Similarly, hybrid ORF 2 had a GC content of 73.2% and a GC5 temperature of 89.3, as shown in Table 4. GC concentrations above 60% are generally considered favorable for gene design, protein expression, and primer design in polymerase chain reactions. The stability of the molecules depicted in Tables 4 and 5 was assessed based on the same analysis, and the results were compared with earlier investigations.

The sequence length of hybrid ORF 1 was 5166 bp, which closely matched the findings of earlier researchers. Similarly, hybrid ORF 2 had a sequence length of 3516 bp. For the in-fusion cloning of hybrid ORF 1, the vector pDE-Cas9 with a size of 15,758 bp was employed. In our results, the vector pGL-basic, used for the in-fusion cloning of hybrid ORF 2, had a size of 4818 bp, and the results were compared with earlier research to ensure accuracy. To determine the protein sequences, we translated the hybrid ORF sequences using the Translate feature of Expasy, following the methods established by earlier researchers. The translated sequences produced six frames, three for 5′ to 3′ and three for 3′ to 5′, consistent with previous research. Tertiary structure prediction was performed using bioinformatics approaches, and the results were compared with earlier studies. The Ramachandran plot, which illustrates permissible regions for amino acid backbone dihedral phi angles against psi angles, was employed for further analysis and validation.

Our investigations indicated that both hybrid ORFs could be utilized to produce high lipid contents. However, it is essential to note that no in vitro experiments were conducted in this research. Therefore, further in vitro studies are strongly recommended to validate the functionality of the created hybrid ORF proteins. Additionally, increasing the oil contents in microalgae could involve using genes from diverse microalgae species. Furthermore, the construction of hybrid ORF proteins for other beneficial metabolites, including carbohydrates, can be explored using a similar methodology. Throughout our research, we made effective use of various online resources. The NCBI genome database served as a valuable source of genetic information, while the BLASTp program was instrumental in identifying the superfamilies associated with the specific genes. The ORF Finder tool facilitated the analysis of ORFs, enabling us to determine their characteristics. By incorporating these online resources into our study, we gained valuable insights into the genetic composition and potential functionalities of the identified oil-producing genes. Our study showcases the potential of using in silico investigations, combined with bioinformatics and computational tools, to identify and evaluate oil-producing genes in microalgae. These findings lay the foundation for future experimental studies and advancements in bioenergy research.

5 Conclusion

Our in silico investigations focused on identifying oil-producing genes in multiple microalgal species, utilizing the NCBI genome database and following established research methods. These investigations identified specific genes in *Chlorella protothecides*, *Chlorella sorokiniana*, and *Scenedesmus* sp. and determined their associated superfamilies using the BLASTp program. Subsequent analysis of these genes' ORFs allowed us to create hybrid ORFs, namely hybrid ORF 1 and hybrid ORF 2, by combining different ORFs from the selected genes. We employed various tools and techniques to analyze further and evaluate the hybrid ORFs. SnapGene facilitated a restriction enzyme study, revealing potential obstructive points in hybrid ORF 1. The Vector NTI® Express Designer was used for thermodynamic evaluation, indicating the stability of the hybrid ORFs. Additionally, we utilized online resources such as Expasy to translate ORF sequences into protein sequences and bioinformatics approaches for predicting tertiary structures and assessing the stability of the molecules using the Ramachandran plot.

Our analyses and evaluations consistently matched previous research findings, providing confidence in the accuracy and reliability of our methodology. We determined the sequence lengths of hybrid ORF 1 and hybrid ORF 2, and employed suitable vectors for their in-fusion cloning. Moreover, we confirmed the translation of ORFs into protein sequences and examined their frames and structural characteristics. Based on our investigations, both hybrid ORFs showed promise for producing high lipid contents. However, it is crucial to emphasize that our study remained in the realm of in silico analysis, and further in vitro experiments are strongly recommended to validate the functionality and productivity of the created hybrid ORF proteins. Our research contributes to the understanding of oil-producing genes in microalgae and serves as a crucial starting point for future wet lab experiments involving genetically manipulated microorganisms. Additionally, exploring genes from diverse microalgae species and expanding the application of hybrid ORF proteins for other beneficial metabolites can open new avenues for enhancing microalgae-based biofuel and pharmaceutical production.

Acknowledgements The authors acknowledge the Department of Biotechnology, the University of Okara, for providing the platform for this project.

Author contribution Conceptualization, methodology, validation, formal analysis: Ihtesham Arshad, Muhammad Ahsan, Imran Zafar, Muhammad Sajid, Arslan Sehgal, Waqas Yousaf; writing—original draft preparation, writing—review and editing: Ihtesham Arshad, Muhammad Ahsan, Imran Zafar, Amna Noor, Summya Rashid, Somenath Garai, Meivelu Moovendhan, Rohit Sharma; supervision: Arslan Sehgal, Imran Zafar. All authors have read and agreed to the published version of the manuscript.

Data availability All data are available in this manuscript.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Nanda M et al (2021) Integration of microalgal bioremediation and biofuel production: a 'clean up' strategy with potential for sustainable energy resources. Curr Res Green Sustain Chem 4:100128
2. Efroymson RA, Jager HI, Mandal S, Parish ES, Mathews TJ (2021) Better management practices for environmentally sustainable production of microalgae and algal biofuels. J Clean Prod 289:125150
3. Priyadharsini P et al (2022) Genetic improvement of microalgae for enhanced carbon dioxide sequestration and enriched biomass productivity: review on CO2 bio-fixation pathways modifications. Algal Res 66:102810
4. Ambriz-Pérez D, Orozco-Guillen E, Galán-Hernández N, Luna-Avelar K, Valdez-Ortiz A, Santos-Ballardo D (2021) Accurate method for rapid biomass quantification based on specific absorbance of microalgae species with biofuel importance. Lett Appl Microbiol 73(3):343–351
5. Khan S, Naushad M, Iqbal J, Bathula C, Sharma G (2022) Production and harvesting of microalgae and an efficient operational approach to biofuel production for a sustainable environment. Fuel 311:122543
6. Hoang AT, Sirohi R, Pandey A, Nižetić S, Lam SS, Chen WH, ... Pham VV (2022) Biofuel production from microalgae: Challenges and chances. Phytochem Rev 1–38
7. Shahi T, Beheshti B, Zenouzi A, Almasi M (2020) Bio-oil production from residual biomass of microalgae after lipid extraction: the case of Dunaliella sp. Biocatal Agric Biotechnol 23:101494
8. Huang Z et al (2022) Valorisation of microalgae residues after lipid extraction: pyrolysis characteristics for biofuel production. Biochem Eng J 179:108330
9. Arun J et al (2022) Bio-based algal (Chlorella vulgaris) refinery on de-oiled algae biomass cake: a study on biopolymer and biodiesel production. Sci Total Environ 816:151579
10. Naveen S, Gopinath KP, Malolan R, Jayaraman RS, Aakriti K, Arun J (2021) A solar reactor for bio-diesel production from Pongamia oil: studies on transesterfication process parameters and energy efficiency. Chin J Chem Eng 40:218–224
11. Hossain N, Mahlia TMI (2019) Progress in physicochemical parameters of microalgae cultivation for biofuel production. Crit Rev Biotechnol 39(6):835–859
12. Singh R, Arora A, Singh V (2021) Biodiesel from oil produced in vegetative tissues of biomass–a review. Bioresour Technol 326:124772
13. Jagadevan S et al (2018) Recent developments in synthetic biology and metabolic engineering in microalgae towards biofuel production. Biotechnol Biofuels 11(1):1–21
14. Behera B, Selvanayaki S, Jayabalan R, Balasubramanian P (2019) An in-silico approach for enhancing the lipid productivity in microalgae by manipulating the fatty acid biosynthesis. In: Soft computing for problem solving. Springer, pp 877–889
15. Bharadwaj SV, Ram S, Pancha I, Mishra S (2020) Recent trends in strain improvement for production of biofuels from microalgae. In: Microalgae cultivation for biofuels production. Elsevier, pp 211–225
16. Xue J et al (2021) Biotechnological approaches to enhance biofuel producing potential of microalgae. Fuel 302:121169
17. Alishah Aratboni H, Rafiei N, Garcia-Granados R, Alemzadeh A, Morones-Ramírez JR (2019) Biomass and lipid induction strategies in microalgae for biofuel production and other applications. Microb Cell Factories 18(1):1–17
18. Lee SY, Khoiroh I, Vo D-VN, Senthil Kumar P, Show PL (2021) Techniques of lipid extraction from microalgae for biofuel production: a review. Environ Chem Lett 19(1):231–251
19. Satya ADM et al (2023) Progress on microalgae cultivation in wastewater for bioremediation and circular bioeconomy. Environ Res 218:114948
20. Xi Y, Yin L, Luo G (2021) Characterization and RNA-seq transcriptomic analysis of a Scenedesmus obliqnus mutant with enhanced photosynthesis efficiency and lipid productivity. Sci Rep 11(1):1–12
21. Wei C, Wang H, Ma M, Hu Q, Gong Y (2020) Factors affecting the mixotrophic flagellate Poterioochromonas malhamensis grazing on Chlorella cells. J Eukaryot Microbiol 67(2):190–202
22. Ding W et al (2020) Enhanced lipid extraction from the biodiesel-producing microalga Chlorella pyrenoidosa cultivated in municipal wastewater via Daphnia ingestion and digestion. Bioresour Technol 306:123162
23. Yew GY et al (2019) Hybrid liquid biphasic system for cell disruption and simultaneous lipid extraction from microalgae Chlorella sorokiniana CY-1 for biofuel production. Biotechnol Biofuels 12(1):1–12
24. Ren X, Zhao X, Turcotte F, Deschênes J-S, Tremblay R, Jolicoeur M (2017) Current lipid extraction methods are significantly enhanced adding a water treatment step in Chlorella protothecoides. Microb Cell Factories 16(1):1–13
25. Fu L et al (2017) Excessive phosphorus enhances Chlorella regularis lipid production under nitrogen starvation stress during glucose heterotrophic cultivation. Chem Eng J 330:566–572
26. dos Santos RR, Moreira DM, Kunigami CN, Aranda DAG, Teixeira CMLL (2015) Comparison between several methods of total lipid extraction from Chlorella vulgaris biomass. Ultrason Sonochem 22:95–99
27. Han L, Pei H, Hu W, Han F, Song M, Zhang S (2014) Nutrient removal and lipid accumulation properties of newly isolated microalgal strains. Bioresour Technol 165:38–41
28. Shin H-Y, Ryu J-H, Bae S-Y, Crofcheck C, Crocker M (2014) Lipid extraction from Scenedesmus sp. microalgae for biodiesel production using hot compressed hexane. Fuel 130:66–69
29. Karemore A, Pal R, Sen R (2013) Strategic enhancement of algal biomass and lipid in Chlorococcum infusionum as bioenergy feedstock. Algal Res 2(2):113–121
30. Isleten-Hosoglu M, Gultepe I, Elibol M (2012) Optimization of carbon and nitrogen sources for biomass and lipid production by Chlorella saccharophila under heterotrophic conditions and development of Nile red fluorescence based method for quantification of its neutral lipid content. Biochem Eng J 61:11–19

31. Zhao G, Yu J, Jiang F, Zhang X, Tan T (2012) The effect of different trophic modes on lipid accumulation of Scenedesmus quadricauda. Bioresour Technol 114:466–471
32. Goswami RD, Kalita M (2011) Scenedesmus dimorphus and Scenedesmus quadricauda: two potent indigenous microalgae strains for biomass production and CO2 mitigation—a study on their growth behavior and lipid productivity under different concentration of urea as nitrogen source. J Algal Biomass Util 2(4):2–4
33. Yang J, Li X, Hu H, Zhang X, Yu Y, Chen Y (2011) Growth and lipid accumulation properties of a freshwater microalga, Chlorella ellipsoidea YJ1, in domestic secondary effluents. Appl Energy 88(10):3295–3299
34. Zafar I et al (2022) Genome-wide identification and analysis of GRF (growth-regulating factor) gene family in Camila sativa through in silico approaches. J King Saud Univ Sci 34(4):102038
35. Agarwal D, Zafar I, Ahmad SU, Kumar S, Sundaray JK, Rather MA (2022) Structural, genomic information and computational analysis of emerging coronavirus (SARS-CoV-2). Bull Natl Res Cent 46(1):1–16
36. Rafique Q et al (2023) Reviewing methods of deep learning for diagnosing COVID-19, its variants and synergistic medicine combinations. Comput Biol Med 163:107191
37. Barten RJ, Wijffels RH, Barbosa MJ (2020) Bioprospecting and characterization of temperature tolerant microalgae from Bonaire. Algal Res 50:102008
38. Rayan RA, Zafar I, Tsagkaris C (2021) Artificial intelligence and big data solutions for COVID-19. In: Intelligent data analysis for COVID-19 pandemic. Springer, pp 115–127
39. Ali S et al (2023) Amomum subulatum: a treasure trove of anticancer compounds targeting TP53 protein using in vitro and in silico techniques. Front Chem 11:1174363
40. Ahmad HM et al (2022) Characterization of fenugreek and its natural compounds targeting AKT-1 protein in cancer: pharmacophore, virtual screening, and MD simulation techniques. J King Saud Univ Sci 34(6):102186
41. Ali S et al (2023) Predicting the effects of rare genetic variants on oncogenic signaling pathways: a computational analysis of HRAS protein function. Front Chem 11:1173624
42. Rather MA, Dutta S, Guttula PK, Dhandare BC, Yusufzai S, Zafar MI (2020) Structural analysis, molecular docking and molecular dynamics simulations of G-protein-coupled receptor (kisspeptin) in fish. J Biomol Struct Dyn 38(8):2422–2439
43. Thiyagarajan S, Arumugam M, Kathiresan S (2020) Identification and functional characterization of two novel fatty acid genes from marine microalgae for eicosapentaenoic acid production. Appl Biochem Biotechnol 190(4):1371–1384
44. Charon J, Kahlke T, Larsson ME, Abbriano R, Commault A, Burke J, ... Holmes EC (2022) Diverse RNA viruses associated with diatom, eustigmatophyte, dinoflagellate, and rhodophyte microalgae cultures. J Virology 96(20):e00783-22
45. Nigam M, Yadav R, Awasthi G (2021) In-silico construction of hybrid ORF protein to enhance algal oil content for biofuel. In: Advances in Biomedical Engineering and Technology. Springer, pp 67–89
46. Dehghani J, Adibkia K, Movafeghi A, Pourseif MM, Omidi Y (2020) Designing a new generation of expression toolkits for engineering of green microalgae; robust production of human interleukin-2. BioImpacts: BI 10(4):259
47. Icgen B, Yilmaz F (2016) Design a cadA-targeted DNA probe for screening of potential bacterial cadmium biosorbents. Environ Sci Pollut Res 23(6):5743–5752
48. Lu G, Moriyama EN (2004) Vector NTI, a balanced all-in-one sequence analysis suite. Brief Bioinform 5(4):378–388
49. Rossolillo P, Kolesnikova O, Essabri K, Zamorano GR, Poterszman A (2021) Production of multiprotein complexes using the baculovirus expression system: Homology-based and restriction-free cloning strategies for construct design. In: Multiprotein Complexes. Springer, pp 17–38
50. Bianco M, Ventura G, Calvano CD, Losito I, Cataldi TR (2022) Discovery of marker peptides of spirulina microalga proteins for allergen detection in processed foodstuffs. Food Chem 393:133319
51. Raza S et al (2019) In silico analysis of four structural proteins of aphthovirus serotypes revealed significant B and T cell epitopes. Microb Pathog 128:254–262
52. Dave M, Daga A, Rawal R (2015) Structural and functional analysis of AF9-MLL oncogenic fusion protein using homology modeling and simulation based approach. Int J Pharm Pharm Sci 7(12):155–161
53. Mahapatra SR, Dey J, Jaiswal A, Roy R, Misra N, Suar M (2022) Immunoinformatics-guided designing of epitope-based subunit vaccine from Pilus assembly protein of Acinetobacter baumannii bacteria. J Immunol Methods 508:113325
54. Samad A, Ahammad F, Nain Z, Alam R, Imon RR, Hasan M, Rahman MS (2022) Designing a multi-epitope vaccine against SARS-CoV-2: An immunoinformatics approach. J Biomol Struct Dyn 40(1):14–30
55. Ahmad SU et al (2022) A comprehensive genomic study, mutation screening, phylogenetic and statistical analysis of SARS-CoV-2 and its variant omicron among different countries. J Infect Public Health 15(8):878–891
56. Dey J, Mahapatra SR, Lata S, Patro S, Misra N, Suar M (2022) Exploring Klebsiella pneumoniae capsule polysaccharide proteins to design multiepitope subunit vaccine to fight against pneumonia. Expert Rev Vaccines 21(4):569–587
57. A.-y. Liu, C. Wei, L.-L. Zheng, L.-R. Song, "Identification of high-lipid producers for biodiesel production from forty-three green algal isolates in China," Prog Nat Sci Mater, vol. 21, no. 4, pp. 269-276, 2011.
58. Hsieh H-J, Su C-H, Chien L-J (2012) Accumulation of lipid production in Chlorella minutissima by triacylglycerol biosynthesis-related genes cloned from Saccharomyces cerevisiae and Yarrowia lipolytica. J Microbiol 50(3):526–534
59. Sharma T, Gour RS, Kant A, Chauhan RS (2015) Lipid content in Scenedesmus species correlates with multiple genes of fatty acid and triacylglycerol biosynthetic pathways. Algal Res 12:341–349
60. Li YX, Zhao FJ, Yu DD (2015) Effect of nitrogen limitation on cell growth, lipid accumulation and gene expression in Chlorella sorokiniana. Braz Arch Biol Technol 58:462–467
61. Sun Z, Chen YF, Du J (2016) Elevated CO 2 improves lipid accumulation by increasing carbon metabolism in Chlorella sorokiniana. *Plant Biotechnol J* 14(2):557–566
62. Zhang H et al (2021) Trophic transition enhanced biomass and lipid production of the unicellular green alga Scenedesmus acuminatus. Front Bioeng Biotechnol 9:638726
63. Gao C et al (2014) Oil accumulation mechanisms of the oleaginous microalga Chlorella protothecoides revealed through its genome, transcriptomes, and proteomes. BMC Genomics 15(1):1–14
64. Moshood TD, Nawanir G, Mahmud F (2021) Microalgae biofuels production: a systematic review on socioeconomic prospects of microalgae biofuels and policy implications. Environ Chall 5:100207
65. Hwangbo M, Chu K-H (2020) Recent advances in production and extraction of bacterial lipids for biofuel production. Sci Total Environ 734:139420
66. Ko JK, Lee JH, Jung JH, Lee S-M (2020) Recent advances and future directions in plant and yeast engineering to improve lignocellulosic biofuel production. *Renew Sust Energ Rev* 134:110390

67. do Vale Borges A, Fuess LT, Alves I, Takeda PY, Damianovic MHRZ (2021) Co-digesting sugarcane vinasse and distilled glycerol to enhance bioenergy generation in biofuel-producing plants. *Energy Convers Manag* 250:114897
68. Das PK, Rani J, Rawat S, Kumar S (2022) Microalgal co-cultivation for biofuel production and bioremediation: current status and benefits. *BioEnergy Research* 15(1):1–26
69. Saratale RG et al (2022) Microalgae cultivation strategies using cost–effective nutrient sources: recent updates and progress towards biofuel production. Bioresour Technol 361:127691
70. Azeez NA et al (2021) Biodiesel potentials of microalgal strains isolated from fresh water environment. *Environ Chall* 5:100367
71. Ge S, Madill M, Champagne P (2018) Use of freshwater macroalgae Spirogyra sp. for the treatment of municipal wastewaters and biomass production for biofuel applications. *Biomass Bioenergy* 111:213–223
72. Nordin N, Yusof N, Samsudin S (2017) Biomass production of Chlorella sp., Scenedesmus sp., and Oscillatoria sp. in nitrified landfill leachate. *Waste Biomass Valorization* 8(7):2301–2311
73. Aucoin HR, Gardner J, Boyle NR (2016) Omics in chlamydomonas for biofuel production. In: Lipids in plant and algae development. Springer, pp 447–469
74. Mahapatra DM, Chanakya H, Ramachandra T (2013) Euglena sp. as a suitable source of lipids for potential use as biofuel and sustainable wastewater treatment. *J Appl Phycol* 25(3):855–865
75. Trung VT, Ly BM, Hang NT (2013) Research to produce ethanol from seaweed biomass Cladophora sp. J Mater Sci Eng, B 3(10B)
76. Buakhiaw B, Sanguanchaipaiwong V (2017) Effect of media on acetone-butanol-ethanol fermentation by isolated Clostridium spp. *Energy Procedia* 138:864–869
77. Eladel H, Abomohra AE-F, Battah M, Mohmmed S, Radwan A, Abdelrahim H (2019) Evaluation of Chlorella sorokiniana isolated from local municipal wastewater for dual application in nutrient removal and biodiesel production. *Bioprocess Biosyst Eng* 42(3):425–433
78. Pancha I et al (2015) Salinity induced oxidative stress enhanced biofuel production potential of microalgae Scenedesmus sp. CCNM 1077. *Bioresour Technol* 189:341–348
79. Darpito C et al (2015) Cultivation of Chlorella protothecoides in anaerobically treated brewery wastewater for cost-effective biodiesel production. *Bioprocess Biosyst Eng* 38(3):523–530
80. Madusanka DAT, Manage PM (2018) Optimising a solvent system for lipid extraction from cyanobacterium Microcystis spp.: future perspective for biodiesel production
81. Zhao Y, Qiao T, Gu D, Zhu L, Yu X (2022) Stimulating biolipid production from the novel alga Ankistrodesmus sp. by coupling salt stress and chemical induction. *Renew Energy* 183:480–490
82. Pham TL (2019) Biosorption combined with lipid production and growth inhibition of copper on the microalgal Pediastrum sp. J Viet Env 11(1):15–20
83. Narang PK et al (2022) Genome-based identification and comparative analysis of enzymes for carotenoid biosynthesis in microalgae. World J Microbiol Biotechnol 38:1–22
84. Narang PK et al (2021) Functional annotation and sequence-structure characterization of a hypothetical protein putatively involved in carotenoid biosynthesis in microalgae. S Afr J Bot 141:219–226
85. Aron NSM, Chew KW, Ang WL, Ratchahat S, Rinklebe J, Show PLJF (2022) Recovery of microalgae biodiesel using liquid biphasic flotation system. Fuel 317:123368
86. Misra N, Panda PK, Parida BK, Mishra BK (2016) dEMBF: a comprehensive database of enzymes of microalgal biofuel feedstock. *PLoS One* 11(1):e0146158
87. Liu Y, Bao H, Zhu M-L, Hu C-X, Zhou Z-G (2022) Subcellular localization and identification of acyl-CoA: lysophosphatidylethanolamine acyltransferase (LPEAT) in the arachidonic acid-rich green microalga, Myrmecia incisa Reisigl. *J Appl Phycol* 34(2):837–855
88. Fathy W et al (2021) Recombinant overexpression of the Escherichia coli acetyl-CoA carboxylase gene in Synechocystis sp. boosts lipid production. *J Basic Microbiol* 61(4):330–338
89. Pan X et al (2020) Enhancing astaxanthin accumulation in Xanthophyllomyces dendrorhous by a phytohormone: metabolomic and gene expression profiles. *Microb Biotechnol* 13(5):1446–1460
90. Popova L et al (2018) In silico analyses of transcriptomes of the marine green microalga Dunaliella tertiolecta: identification of sequences encoding P-type ATPases. *Mol Biol* 52(4):520–531
91. Durante L, Hübner W, Lauersen KJ, Remacle C (2019) Characterization of the GPR1/FUN34/YaaH protein family in the green microalga Chlamydomonas suggests their role as intracellular membrane acetate channels. *Plant Direct* 3(6):e00148
92. Gouda M, Huang Z, Liu Y, He Y, Li X (2021) Physicochemical impact of bioactive terpenes on the microalgae biomass structural characteristics. *Bioresour Technol* 334:125232

Authors and Affiliations

Ihtesham Arshad[1] · Muhammad Ahsan[2] · Imran Zafar[3] · Muhammad Sajid[1] · Sheikh Arslan Sehgal[4] · Waqas Yousaf[5] · Amna Noor[1] · Summya Rashid[6] · Somenath Garai[7] · Meivelu Moovendhan[8] · Rohit Sharma[9]

✉ Sheikh Arslan Sehgal
arslansehgal@yahoo.com

✉ Rohit Sharma
rohitsharma@bhu.ac.in

Ihtesham Arshad
ihteshamarshad86@gmail.com

Muhammad Ahsan
ahsan_m@outlook.com

Imran Zafar
Bioinfo.pk@gmail.com

Muhammad Sajid
sajid@uo.edu.pk

Waqas Yousaf
waqasyousaf2012@gmail.com

Amna Noor
Amnachaudhary127@gmail.com

Summya Rashid
frenlysara@gmail.com

Somenath Garai
sgarai@bhu.ac.in

Meivelu Moovendhan
moovendhan85@gmail.com

[1] Department of Biotechnology, Faculty of Life Sciences, University of Okara, Okara 56300, Pakistan

[2] Institute of Environmental and Agricultural Sciences, University of Okara, Okara 56300, Pakistan

[3] Department of Bioinformatics and Computational Biology, Virtual University, Lahore, Punjab 54700, Pakistan

[4] Department of Bioinformatics, Institute of Biochemistry, Biotechnology and Bioinformatics, The Islamia University of Bahawalpur, Bahawalpur 63051, Pakistan

[5] Institute of Molecular Biology and Biotechnology (IMBB), Department of Botany, The University of Lahore, Lahore, Punjab, Pakistan

[6] Department of Pharmacology & Toxicology, College of Pharmacy, Prince Sattam Bin Abdulaziz University, P.O. Box 173, Al-Kharj 11942, Saudi Arabia

[7] Department of Chemistry, Institute of Science, Banaras Hindu University, Varanasi 221005, India

[8] Col Dr. Jeppiaar Research Park Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu 600 119, India

[9] Department of Rasashastra and Bhaishajya Kalpana, Faculty of Ayurveda, Institute of Medical Sciences, Banaras Hindu University, Varanasi 221005, India