Düzce University
Journal of Science & Technology

*Research Article*

# *De novo* transcriptome assembly reveals three alternative oxidase encoding genes in *Nymphaea alba* L.

Ercan Selçuk ÜNLÜ [a,*], Gülgez Gökçe YILDIZ [b]

[a] *Department of Chemistry, Faculty of Arts and Science, Bolu Abant İzzet Baysal University, Bolu, TURKEY*
[b] *Department of Biology, Faculty of Arts and Science, Bolu Abant İzzet Baysal University, Bolu, TURKEY*
* *Corresponding author's e-mail address: esunlu06@gmail.com*

DOI: 10.29130/dubited.914845

## ABSTRACT

Water lilies are aquatic, ornamental and economically valuable plants classified under *Nymphaea* genus. *Nymphaea alba* L., white water lily, has a special focus since it is a member of basal angiosperms. Alternative oxidase (AOX) proteins are the terminal oxidases in the electron transport chain of plants. Identification of alternative oxidase encoding genes for basal angiosperms is important to increase the quality of phylogenetic studies. However, AOX encoding genes were yet to be discovered for *N. alba*. In this study, we aimed to identify alternative oxidase encoding genes in *N. alba* by performing transcriptome analysis. Annotation of 272,934 unigenes with Trinotate tool revealed 77 transcripts with AOX domains characterized in known alternative oxidases. Blast analysis of these 77 sequences with known AOX proteins revealed three distinct AOX genes (*AOX1*, *AOX2* and *AOX4*) in *N. alba*. After in silico subcellular localization analysis of three identified AOX proteins, AOX1, AOX2 are predicted as mitochondrial while AOX4 is a plastidic alternative oxidase protein. Template-based structural modeling results showed that all identified proteins are statistically similar to known structure models of corresponding AOXs.

***Keywords:*** *AOX, Abant Lake, RNA-Seq, Water lilies.*

# *De Novo* Transkriptom Birleştirme Analizi *Nymphaea alba* L. Türünde Üç Alternatif Oksidaz Kodlayan Gen Olduğunu Göstermektedir

## ÖZ

Nilüfer bitkileri ekonomik değeri yüksek sucul süs bitkileri olup *Nymphaea* ailesi altında sınıflandırılmaktadırlar. *Nymphaea alba* L. (Beyaz nilüfer) bazal angiospermlerin bir üyesi olduklarından özel bir öneme sahiptir. Alternatif oksidaz kodlayan genlerin bazal angiospermlerde belirlenmesi filogenetik çalışmaların kalitesini artırmak için önemlidir. *N. alba* için AOX sentezinden sorumlu genler henüz tanımlanmamıştır. Bu çalışmada transkriptom analizi uygulamaları ile *N. alba* türünde AOX sentezinden sorumlu genlerin tanımlanmasını amaçlanmıştır. Trinotate aracılığı ile 272934 unigenin tanımlaması yapılarak, bilinen alternatif oksidazlarda karakterize edilmiş AOX domainlerini içeren 77 transkriptin dizisi ortaya çıkarılmıştır. Bu 77 transkriptin bilinen AOX proteinlerine karşı Blast analizleri ile *N. alba* türüne ait üç ayrı AOX geni (*AOX1*, *AOX2* and *AOX4*) tespit edilmiştir. In silico hücre içi lokalizasyon analizine göre AOX1 ve AOX2 proteinleri mitokondriyal, AOX4 ise plastidik alternatif oksidaz proteinidir. Şablon bazlı yapısal modelleme sonuçları, tanımlanmış bu proteinlerin model verilerinin, başka türlerde karşılık geldiği AOX proteinlerinin bilinen yapı modellerine istatistiksel olarak benzer olduğunu göstermiştir.

***Anahtar Kelimeler:*** *AOX, Abant Gölü, RNA-Seq, Nilüfergiller.*

Received: 12/04/2021, Revised: 12/08/2021, Accepted: 18/08/2021

# I. INTRODUCTION

Water lilies are ornamental plants found in ponds, rivers and lakes. They are classified under Nymphaeaceae family. About 40 different species have been classified that *Nymphaea* species are considered as economically valuable plants [1]. *Nymphaea alba* L. is preferred in landscape arrangements, vegetable cultivation and medical applications. It was shown that the substances found in plant leaves showed anti-anxiety, antioxidant and anti-carcinogenic effects [2]. In addition, it has been found that *Nymphaea alba* (*N. alba*) leaves and roots have an effect on inflammatory cuts and fecal ulcers [3], [4]. To date, most of the studies on water lilies, including the *N. alba* species, have been focused on general structural features of plants. At the molecular level, the major group of studies is on genetic barcoding and phylogenetic properties of *N. alba*. These studies revealed that *N. alba* has an important place in plant variation analysis by being in the basal angiosperm group [5], [6]. Even though economic value of *N. alba* is high, studies on the cultivation of the plant have not been successful due to the limitations in the propagation properties of *N. alba* [7] and lack of molecular data on the species. Intracellular signal transduction pathways are biochemical communication tools that serve for both maintaining cellular activity during normal conditions and allowing cells to control their defense mechanisms under stress conditions. Mitochondrial retrograde signaling pathway stimulates nuclear gene expression by transmitting biochemical signals to the cell nucleus to protect mitochondrial function. Diverse groups of proteins have been described among different kingdoms as key regulators for the mitochondrial retrograde signaling pathway. In plants, alternative oxidases are proteins synthesized from the *AOX* genes, and they play a key role in the transduction of biochemical signals after activation of mitochondrial retrograde signaling [8]–[10]. In general, alternative oxidases are enzymes of the terminal oxidase the reduction of molecular oxygen to water. Alternative oxidases are members of the di-iron family of non-carboxylases [11], consisting of four conserved helix bundles coordinated for two iron binding sites with four conserved glycated amide and two histidine amino acid motifs.

The high degree conversation of AOX sequences can be found in catalytic core domains of AOX proteins while *N*-terminal region and contacting residues are highly variable among species. Mitochondrial and plastid alternative oxidases are distributed among diverse organisms. It is considered that only a minority of lineages shows complete loss of the AOX encoding genes like Archaebacteria. Phylogenetic origin of alternative oxidases is still not clear since available sequences are limited to certain species. Studies suggest a monophyletic origin in Eukaryotes and Eubacteria. It is also considered that origin of fungal *AOX* genes is independent from Eukaryotes and Eubacteria [12], [13].

Mutation studies on AOX genes in plants have shown that AOX protein activity is necessary for the resistance to various stress conditions such as oxidative stress [14]–[16] cold stress and oxygen deficiency [17]. In addition, plants lacking AOX2 protein activity lost their ability to reproduce and grow. It was also shown that the increase of alternative oxidase expression levels were directly linked to tissues in microsporogenesis and vegetative reproductive cycle [18]. In this context, the genes responsible for alternative oxidase synthesis should be among the genes that should be examined primarily in the examination of reproduction and development characteristics in plants.

In this study, we aimed to annotate *Nymphaea alba* transcriptome data and identify alternative oxidase encoding genes. The data presented in the study can be used as a comprehensive molecular data source.

# II. MATERIALS AND METHODS

## A. PLANT SAMPLES

Fresh *N. alba* samples were collected around noon time from Abant Lake from exact coordinates of N 40º35’58,1352”, E 31 º16’38,1864” in Turkey. After rinsing the samples with pure water, they were dissected into 5 pieces as petal, stamen, leaf, petiole and root, and immediately frozen in liquid nitrogen.

## B. METHODS

### B. 1. Preparation of Pooled cDNA Library

Total *RNA* was isolated from all tissues separately via TRIZol® Plus *RNA* purification kit (Invitrogen). Briefly, samples were homogenized in the presence of guanidine isothiocyanate. Then, homogenates were treated with ethanol. Subsequently, homogenates were transferred to the tubes containing silica-based membrane, and *RNA* molecules were retained in the membrane. *RNA* molecules were separated by centrifuge from the membrane using RNase-free pure water. *mRNA* molecules from isolated total *RNA* samples were separated using magnetic beads containing poly-oligoT (Promega). *cDNA* library was prepared using the TruSeq Stranded mRNA Library Prep kit (Illumina). *mRNA* samples from each tissue were collected in a pool with a concentration of 50 ng/µl. After the combined *mRNA* samples were fragmented, first strand synthesis was performed by means of hybrid primers using the SuperScript II reverse transcriptase enzyme. The second strand synthesis was carried out by incubating the specimen in the specific mixture containing the *DNA* polymerase I and RNase H with the purpose of synthesizing the *DNA* strand by removing the *RNA* strand. Adenine molecules were added to the 3 'ends of the *DNA* fragments to prevent the fusion of the *cDNA* fragments to the ends. Adenine molecules were then added to the ends of the fragments, followed by adapter 1 and adapter 2 ligation reactions. The samples were amplified by *PCR* reaction consisting of 15 cycles.

### B. 2. Transcriptome Sequencing, *De Novo* Assembly and Bioinformatics Analysis

Transcriptome sequencing from cDNA library and preliminary bioinformatics analysis (*de novo* assembly and Trinotate annotation) were performed at Source BioScience core facilities located in Germany. Rest of the bioinformatics analysis was carried out in our laboratory.

Paired end 100bp sequence readings were collected using Illumina HiSeq™ 2000 platform. Raw reads were processed for adapter and quality trimming using Skewer tool (version 0.1) [19] with default parameters. *De novo* assembly protocol followed by using the Trinity package (version 2014-07-17) using default parameters (k-mer=25) [20]. For the assessment of assembly data, assembly data was used as the reference to map the raw sequence reads by using Tophat2 (v2.1.1) [21] with the parameters: max-intron-length: 1000, min-intron-length: 10, microexon-search, b2: very-sensitive, and max-multihits: 1. Assembled contigs were annotated using a collection of well-known tools and databases, such as Swiss-Prot, Pfam, and Eggnog. Results were collected and merged into a comprehensive annotation report using Trinotate software (v3.0.0) (http://trinotate.github.io/) using default parameters. Data was processed through the NCBI non-redundant nucleotide (NT) and protein (NR) databases (https://blast.ncbi.nlm.nih.gov/Blast.cgi), protein families (Pfam) database (http://pfam.xfam.org), in the Clusters of Orthologous Groups of protein database (COG, http://eggnogdb.embl.de), on KEGG Automatic Annotation Server (https://www.genome.jp/kegg/kaas/). Assessment of transcriptome data was carried out using BUSCO tool (v4.0.6) [22] using viridiplantae_odb10 (2019-11-20) and eukaryota_odb10 (2019-11-20) datasets.

### B. 3. Identification of *AOX* Genes

Annotation analysis revealed that 77 contigs from transcriptome data showed *AOX* gene signatures. To confirm and specify AOX encoding genes, we carried out BLAST analysis against known AOX proteins. For this purpose, we executed the standalone version of BLASTP with the Perl code we customized that is compatible with the Bioperl module [23]. To carry out BLAST analysis, the database file was created by the Formatdb tool using 841 different AOX protein sequences downloaded from NCBI protein database. Each of putative AOX sequences was processed through the BLAST code against the AOX protein sequence database.

### B. 4. Bioinformatics Analysis of AOX Sequences

Sequence comparison by multiple sequence alignment analysis was carried out using the Clustal Omega server [24]. Aligned sequences further processed to analyze aligned motif regions using Jalview workbench software [25]. Possible intracellular location predictions of identified AOX proteins were calculated using the TargetP [26] and MitoProt [27] server tools. For phylogenetic analysis we downloaded amino acid sequences of alternative oxidase orthologs for 18 selected species in addition to *N. alba* AOX isoforms. Multiple sequence alignments were carried out using the ClustalW tool (version 2.1) with the default settings [28]. The tree was constructed using FastTree (version 2.1.8) server with the default parameters [29]. Three-dimensional structures of proteins were predicted using RaptorX software [30] with default settings.

# III. RESULTS

## A. *NYMPHAEA ALBA* TRANSCRIPTOME ANALYSIS

### A. 1. Sequencing and *De Novo* Assembly of *N. alba* Transcriptome

Next Generation sequence analysis was carried out from *RNA* samples pooled from different tissues to cover all expressed genes in the whole plant. Pairwise sequence analysis resulted in 346,351,240 raw reads. After the adaptor and quality trimming, the reads were collapsed into 430,618 assembled transcripts(contigs) representing 272,934 unigenes. Mapping the raw reads on assembled transcripts showed that the assembled transcripts are represented by 50.8% overall read mapping rate. The summary for sequence analysis is presented in Table 1.

**Table 1.** Summary for *N. alba* transcriptome assembly and functional annotation statistics.

| ASSEMBY STATISTICS | | | | | |
|---|---|---|---|---|---|
| **Total Unigenes** | **Total Contigs** | **Percent GC** | **Contig N50** | **Median Contig Length** | **Avarage Contig Length** |
| 272,934 | 430,618 | 42.26 | 880 | 354 | 604.33 |

| DATABASE STATISTICS | | | | | |
|---|---|---|---|---|---|
| | **Reference Database** | | | | |
| | **Swiss-Prot (BLASTX)** | **Swiss-Prot (BLASTP)** | **Pfam** | **EggNOG** | **TOTAL** |
| *Contigs* | 20713 | 14212 | 11509 | 9348 | 21205 |
| *Unigenes* | 20341 | 13936 | 11276 | 9297 | 20809 |

*De novo assembly carried out using Trinity package (version 2014-07-17). Annotation report data obtained using Trinotate software. Cluster analysis was carried out by a Perl cluster tool (available from https://github.com/esunlu).*

Figure 1 represents the contig length distribution. The sequence N50, which is larger than the average fragment length, indicates that the transcripts in question are significantly assembled. Raw sequence reads have been deposited at NCBI Sequence Read Archive (*SRA*) under SRP149065 accession (NCBI Bioproject accession: PRJNA472003).

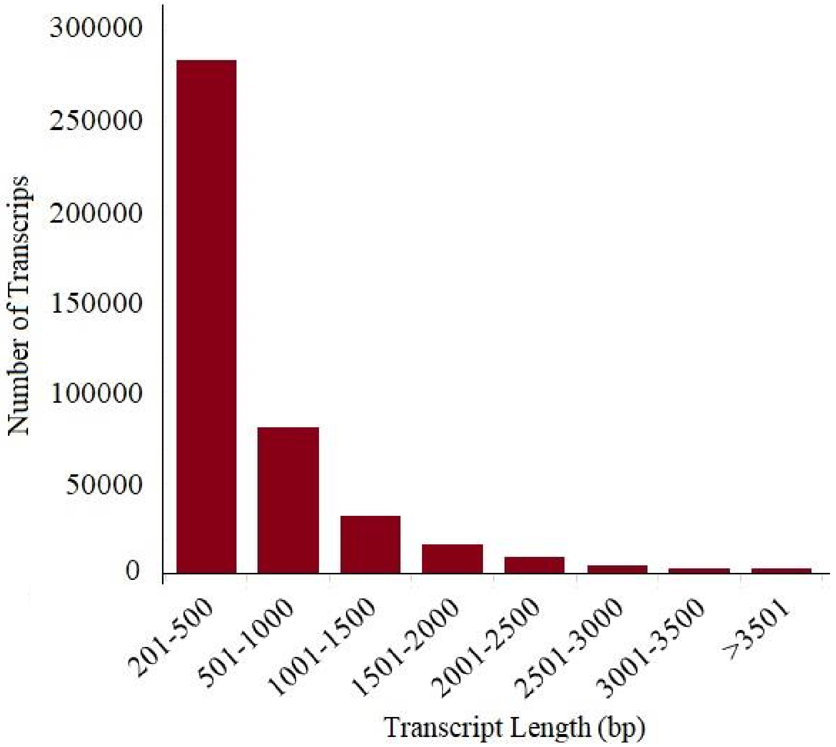


***Figure 1.*** *Length distribution of de novo assembled contigs of Nymphaea alba transcriptome data: Length calculations were carried out using sequences assembled with Trinity tool.*

## A. 2. Functional Annotation of *N. alba* Transcriptome Assembly

Using the Trinotate tool, databases such as Swiss-Prot, PFAM, and EggNOG were scanned. After the removal of contaminants and collapsing the putative unigenes matching to same gene, and a total of 21205 transcript were matched to at least one annotated protein yielding 20809 unigenes. Distribution of annotated transcripts regarding the database type is presented in Table 1. Figure 2 represents the BUSCO analysis summaries. We compared the collapsed and raw assembly data to *N. nucifera* (Chinese lotus) (NCBI RefSeq assembly accession GCF_000365185.1) since it is considered as a close relative to *N. alba* with an available reference genome. We used Viridiplantae and Eukaryota datasets. The results indicated high degree of duplicated genes for *N. alba* and *N. lotus*. Data also shows that collapsing was successful at removing duplicated genes. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GGUL00000000. The version described in this paper is the first version, GGUL01000000.

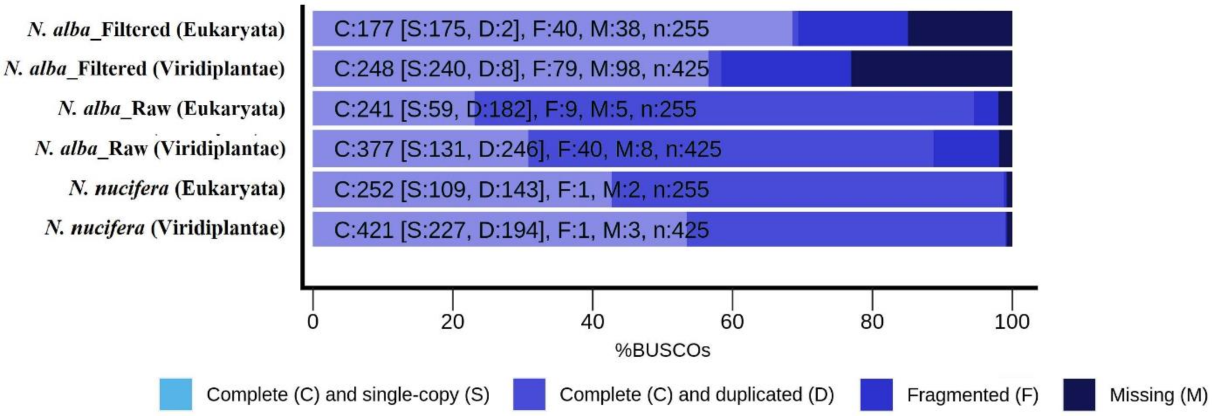


***Figure 2.*** *Summary of Nymphaea alba transcriptome BUSCO assessments: Analysis summarized for filtered and raw assembly data. Nelumbo nucifera reference assemble (RefSeq: GCF_000365185.1) was used as a control reference Datasets used for analysis were indicated between parentheses.*

Clustering *GO* annotations under “biological process”, “molecular function”, “cellular component” terms were carried out by retrieving available GO terms assigned for 20809 unigenes identified from Swiss-Prot BLASTX/BLASTP database. Following cluster analysis, we were able to retrieve 72630 available *GO* term matches for 12676/20809 unigenes. 36.7% of the GO term matches (26639 hits, 9462 unigenes)) with the highest distribution value were obtained for biological process cluster, followed by 34.8% molecular function biological process (25259 hits, 10124 unigenes) and 28.5% cellular component (20732 hits, 977 unigenes) clusters. The general functional distribution graph of the sequences is presented in Figure 3 (a). The summary for the distribution of unigenes affiliated with corresponding *GO* terms is given in Figure 3 (b).

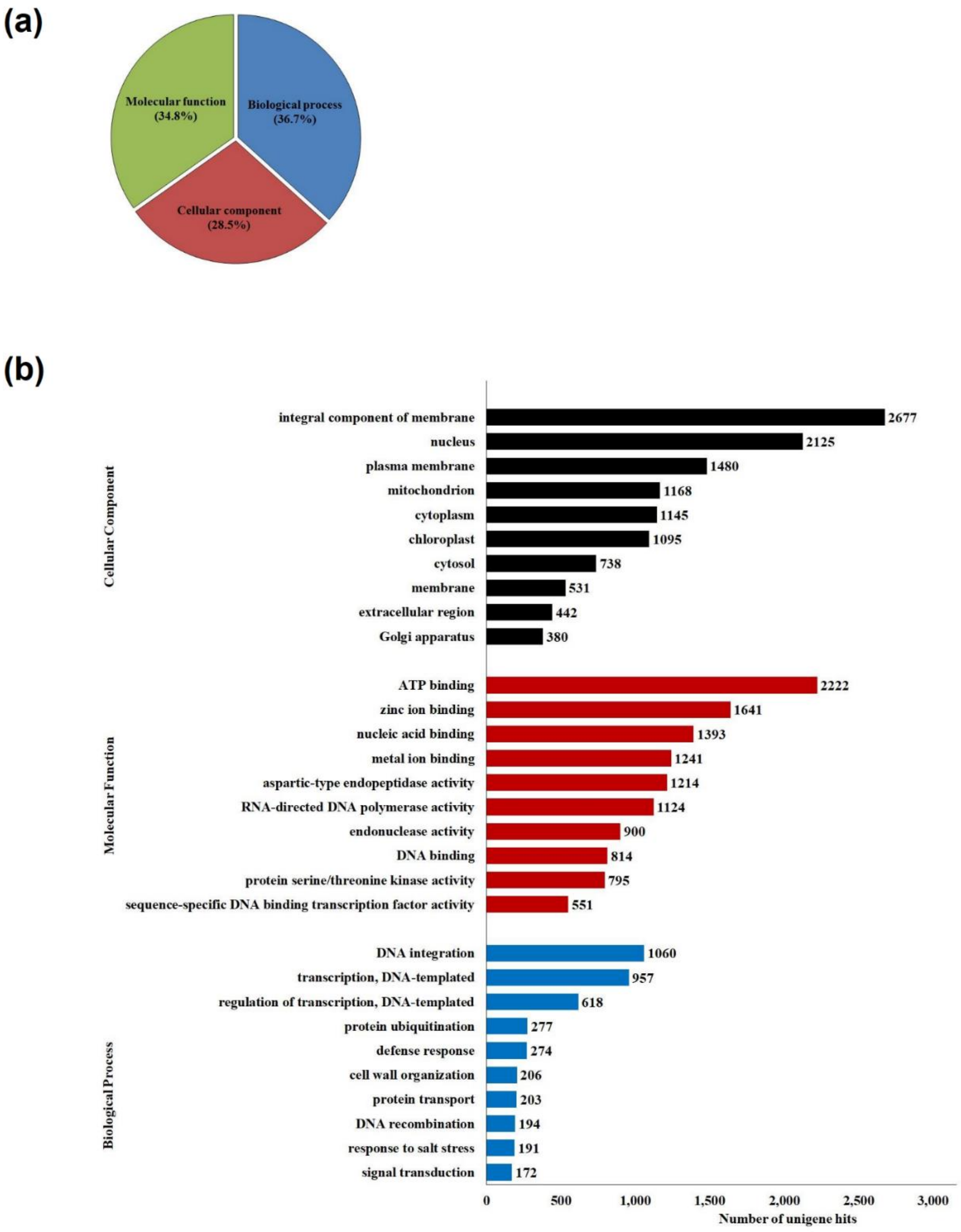


*Figure 3. Functional annotation of Nymphaea alba transcriptome:* ***(a)*** *General GO term distribution of unigenes;* ***(b)*** *Histogram representation of gene ontology classification of Nymphaea alba unigenes. Based on unigene matches in Swiss-Prot database, 10 terms with highest unigene number for three functional categories (Biological Process, Cellular Component and Molecular Function) are represented.*

9297 unigenes annotated through the EggNOG database search were analyzed for their orthologous groups. We predicted a total of 13683 protein matches under *COG*, *KOG*, *NOG*, *euNOG* and *opiNOG* orthologous groups. Detailed analysis of *COG* categories showed that 12469 protein matches were predicted for 8041 unigenes (Figure 4 (a)). "General function prediction only" functional category showed the largest protein matches (18.68%) that followed by protein matches under "Replication, recombination and repair" category (11.52%), "Transcription" category (11.48%), "Signal transduction mechanisms" category (10.54) and "Posttranslational modification, protein turnover, chaperones" category (6.55%), respectively.

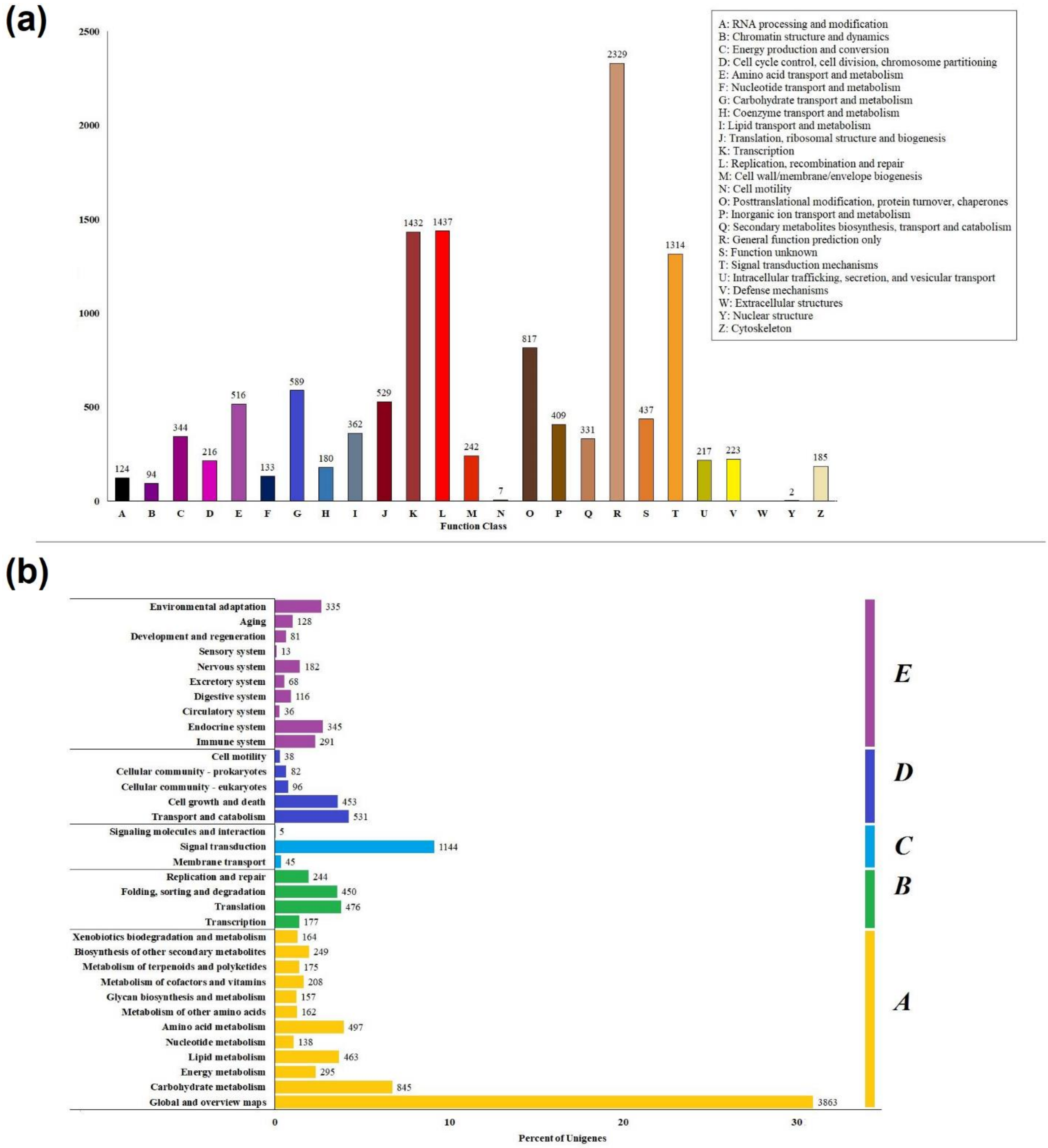


***Figure 4.*** *Gene ontology and KEGG distributions of Nymphaea alba transcriptome:* ***(a)*** *COG distributions of Nymphaea alba transcriptome* ***(b)*** *KEGG pathway distribution of Nymphaea alba transcriptome. Custom Perl codes were used to calculate the number of unigenes classified under five KO categories: Metabolism (A), Genetic Information Processing (B), Environmental Information Processing (C), Cellular Processes (D) and Organismal Systems (E).*

2566 *KO* entries were detected for 6382 of 20809 identified unigenes. The unigenes along with corresponding *KO* entries were processed under five main *KEGG* categories. The analysis resulted in 12552 matches representing 309 *KEGG* pathways. The highest number of hits was represented under metabolism category (57.49%) that was followed by organismal systems (12.71%), genetic information processing (10.73%), cellular processes (9.56%) and environmental information processing (9.51%) categories, respectively (Figure 4 (b)).

## B. IDENTIFICATION AND IN SILICO CHARACTERIZATION OF AOX GENES

### B. 1. Identification of AOX Encoding Genes

BLAST analysis against known alternative oxidase proteins revealed three distinct AOX encoding genes. Three protein sequences are defined by their high similarity to the proteins belonging to different AOX family members. The transcript sequences of these proteins were extracted from the sequence assembly data. After comparing similarity scores of the sequences to known AOX protein sequences, we classified them as *AOX1*, *AOX2* and *AOX4*, respectively. Detailed sequence information is also available from NCBI GenBank database with MK575586 (*AOX1)*, MK575587 (*AOX2*) and MK575588 (*AOX4*) accession numbers. BLAST results revealed that named *AOX* transcripts show close similarity to known functional domain regions of 10 species. Multiple sequence alignment comparison shows that functional AOX domains (LET, NERMHL LEEEA and RADE__H) are well conserved for *N. alba* AOX proteins. The amino acids required for iron-binding (E$^{202}$, E$^{241}$, H$^{244}$, E$^{292}$, E$^{343}$ and H$^{346}$ for AOX1; E$^{214}$, E$^{253}$, H$^{256}$, E$^{304}$, E$^{355}$ and H$^{358}$ for AOX2; E$^{256}$, E$^{295}$, H$^{298}$, E$^{347}$, E$^{416}$ and H$^{419}$ for AOX4) were also conserved for each peptide sequence of *AOX* transcripts (Figure 5).

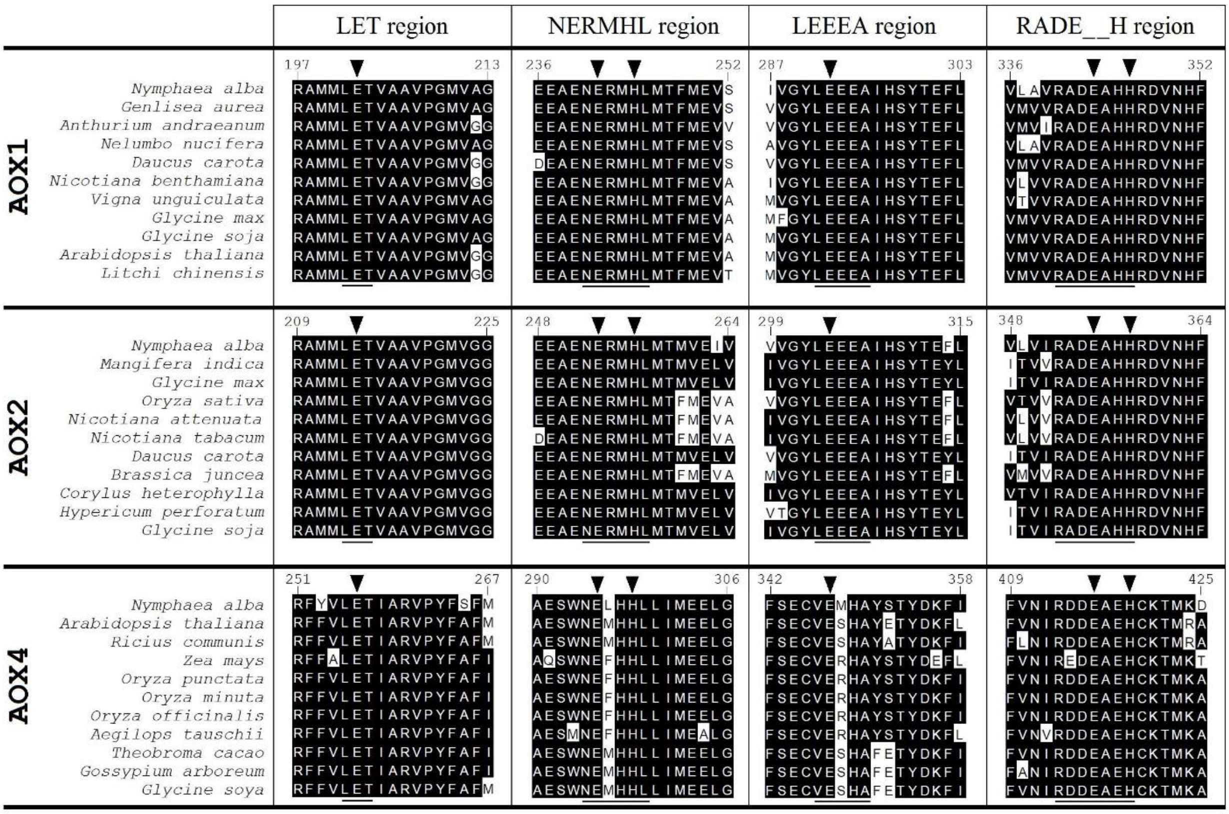
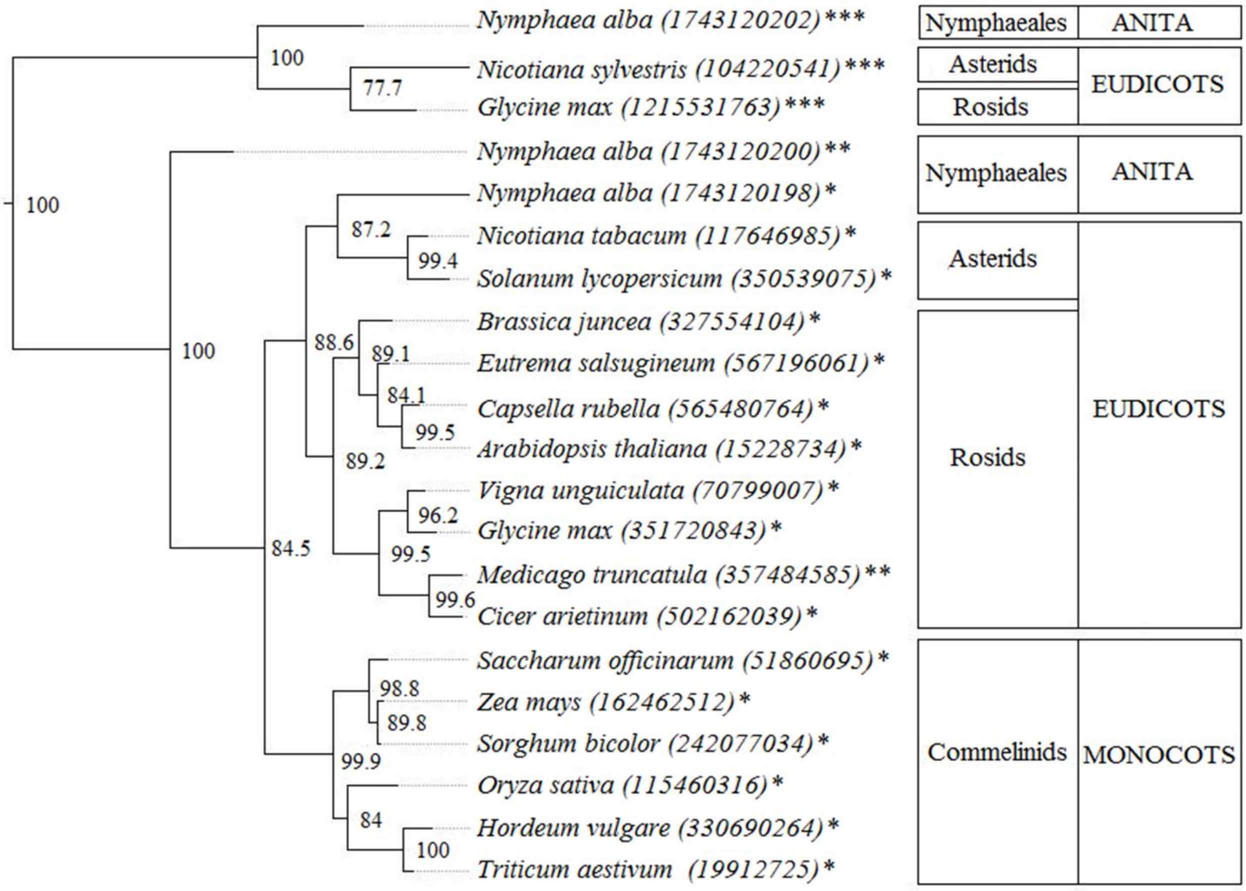

LET region
NERMHL region
LEEEA region
RADE__H region
AOX1
197 213 | 236 252 | 287 303 | 336 352
Nymphaea alba RAMMLETVAAVPGMVAG EEAENERMHLMTFMEVS IVGYLEEEAIHSYTEFL VLAVRADEAHHRDVNHF
Genlisea aurea RAMMLETVAAVPGMVAG EEAENERMHLMTFMEVS VVGYLEEEAIHSYTEFL VMVVRADEAHHRDVNHF
Anthurium andraeanum RAMMLETVAAVPGMVGG EEAENERMHLMTFMEVV VVGYLEEEAIHSYTEFL VMVIRADEAHHRDVNHF
Nelumbo nucifera RAMMLETVAAVPGMVAG EEAENERMHLMTFMEVS AVGYLEEEAIHSYTEFL VLAVRADEAHHRDVNHF
Daucus carota RAMMLETVAAVPGMVGG DEAENERMHLMTFMEVS VVGYLEEEAIHSYTEFL VMVVRADEAHHRDVNHF
Nicotiana benthamiana RAMMLETVAAVPGMVGG EEAENERMHLMTFMEVA IVGYLEEEAIHSYTEFL VLVVRADEAHHRDVNHF
Vigna unguiculata RAMMLETVAAVPGMVAG EEAENERMHLMTFMEVA MVGYLEEEAIHSYTEFL VTVVRADEAHHRDVNHF
Glycine max RAMMLETVAAVPGMVAG EEAENERMHLMTFMEVA MFGYLEEEAIHSYTEFL VMVVRADEAHHRDVNHF
Glycine soja RAMMLETVAAVPGMVAG EEAENERMHLMTFMEVA MVGYLEEEAIHSYTEFL VMVVRADEAHHRDVNHF
Arabidopsis thaliana RAMMLETVAAVPGMVGG EEAENERMHLMTFMEVA MVGYLEEEAIHSYTEFL VMVVRADEAHHRDVNHF
Litchi chinensis RAMMLETVAAVPGMVGG EEAENERMHLMTFMEVT MVGYLEEEAIHSYTEFL VMVVRADEAHHRDVNHF
AOX2
209 225 | 248 264 | 299 315 | 348 364
Nymphaea alba RAMMLETVAAVPGMVGG EEAENERMHLMTMVEIV VVGYLEEEAIHSYTEFL VLVIRADEAHHRDVNHF
Mangifera indica RAMMLETVAAVPGMVGG EEAENERMHLMTMVELV IVGYLEEEAIHSYTEYL ITVVRADEAHHRDVNHF
Glycine max RAMMLETVAAVPGMVGG EEAENERMHLMTMVELV IVGYLEEEAIHSYTEYL ITVIRADEAHHRDVNHF
Oryza sativa RAMMLETVAAVPGMVGG EEAENERMHLMTFMEVA VVGYLEEEAIHSYTEFL VTVVRADEAHHRDVNHF
Nicotiana attenuata RAMMLETVAAVPGMVGG EEAENERMHLMTFMEVA IVGYLEEEAIHSYTEFL VLVVRADEAHHRDVNHF
Nicotiana tabacum RAMMLETVAAVPGMVGG DEAENERMHLMTFMEVA IVGYLEEEAIHSYTEFL VLVVRADEAHHRDVNHF
Daucus carota RAMMLETVAAVPGMVGG EEAENERMHLMTMVELV VVGYLEEEAIHSYTEYL ITVIRADEAHHRDVNHF
Brassica juncea RAMMLETVAAVPGMVGG EEAENERMHLMTFMEVA MVGYLEEEAIHSYTEFL VMVVRADEAHHRDVNHF
Corylus heterophylla RAMMLETVAAVPGMVGG EEAENERMHLMTMVELV IVGYLEEEAIHSYTEYL VTVIRADEAHHRDVNHF
Hypericum perforatum RAMMLETVAAVPGMVGG EEAENERMHLMTMVELV VTGYLEEEAIHSYTEYL ITVIRADEAHHRDVNHF
Glycine soja RAMMLETVAAVPGMVGG EEAENERMHLMTMVELV IVGYLEEEAIHSYTEYL ITVIRADEAHHRDVNHF
AOX4
251 267 | 290 306 | 342 358 | 409 425
Nymphaea alba RFYVLETIARVPYFSFM AESWNELHHLLIMEELG FSECVEMHAYSTYDKFI FVNIRDDEAEHCKTMKD
Arabidopsis thaliana RFFVLETIARVPYFAFM AESWNEMHHLLIMEELG FSECVESHAYETYDKFL FVNIRDDEAEHCKTMRA
Ricius communis RFFVLETIARVPYFAFM AESWNEMHHLLIMEELG FSECVESHAYATYDKFI FLNIRDDEAEHCKTMRA
Zea mays RFFALETIARVPYFAFI AQSWNEFHHLLIMEELG FSECVERHAYSTYDEFL FVNIREDEAEHCKTMKT
Oryza punctata RFFVLETIARVPYFAFI AESWNEFHHLLIMEELG FSECVERHAYSTYDKFI FVNIRDDEAEHCKTMKA
Oryza minuta RFFVLETIARVPYFAFI AESWNEFHHLLIMEELG FSECVERHAYSTYDKFI FVNIRDDEAEHCKTMKA
Oryza officinalis RFFVLETIARVPYFAFI AESWNEFHHLLIMEELG FSECVERHAYSTYDKFI FVNIRDDEAEHCKTMKA
Aegilops tauschii RFFVLETIARVPYFAFI AESMNEFHHLLIMEALG FSECVERHAYSTYDKFL FVNVRDDEAEHCKTMKA
Theobroma cacao RFFVLETIARVPYFAFI AESWNEMHHLLIMEELG FSECVESHAFETYDKFI FVNIRDDEAEHCKTMKA
Gossypium arboreum RFFVLETIARVPYFAFI AESWNEMHHLLIMEELG FSECVESHAFETYDKFI FANIRDDEAEHCKTMKA
Glycine soya RFFVLETIARVPYFAFM AESWNEMHHLLIMEELG FSECVESHAFETYDKFI FVNIRDDEAEHCKTMKA

***Figure 5.*** *Multiple sequence alignment of conserved active domain in predicted Nymphaea alba AOX proteins: Clustal Omega multiple sequence alignment analysis results are shown for known AOX motifs (LET, NERMHL, LEEEAH and RADE_ _AH) sequences. Exact motif regions are indicated with underline. Black rectangle shows suggested iron binding sites. Numbers above the alignment shows amino acid coordinates for AOX sequences of N. alba.*

We also analyzed the phylogenetic distribution of AOX orthologs among distant plant species (Figure 6). AOX4 orthologs merged under separate monophyletic group (BS=100) that was diverged from mitochondrial AOX orthologs not only for *N. alba* but also for other species included in analysis. Data also shows mitochondrial AOX isoforms are structured in two monophyletic (Rosids (Eudicots) and Commelinids (Monocots)) and one paraphyletic (Asterids) clusters. The evolutionary positions of *N. alba* AOX isoforms are structured as a separate polyphyletic group (ANITA grade) from the others.



Nymphaea alba (1743120202)***
Nicotiana sylvestris (104220541)***
Glycine max (1215531763)***
Nymphaea alba (1743120200)**
Nymphaea alba (1743120198)*
Nicotiana tabacum (117646985)*
Solanum lycopersicum (350539075)*
Brassica juncea (327554104)*
Eutrema salsugineum (567196061)*
Capsella rubella (565480764)*
Arabidopsis thaliana (15228734)*
Vigna unguiculata (70799007)*
Glycine max (351720843)*
Medicago truncatula (357484585)**
Cicer arietinum (502162039)*
Saccharum officinarum (51860695)*
Zea mays (162462512)*
Sorghum bicolor (242077034)*
Oryza sativa (115460316)*
Hordeum vulgare (330690264)*
Triticum aestivum (19912725)*
100
77.7
100
100
87.2
99.4
88.6
89.1
84.1
99.5
89.2
96.2
99.5
99.6
84.5
98.8
89.8
99.9
84
100
Nymphaeales
ANITA
Asterids
Rosids
EUDICOTS
Nymphaeales
ANITA
Asterids
Rosids
EUDICOTS
Commelinids
MONOCOTS

***Figure 6.*** *Phylogenetic tree analysis of AOX peptide sequences: Nymphaea alba sequences were analyzed against different AOX orthologs (*AOX1; **AOX2; ***AOX4). Numbers in parenthesis represents the NCBI GI number of AOX records for the corresponding species.*

## B. 2. In Silico Prediction of Subcellular Locations of AOX Proteins

Intracellular location profiles of the putative AOX proteins were estimated using the TargetP server tool [26]. Table 2 summarizes the calculated data to predict subcellular locations of predicted AOX proteins. The analytical relevance we found giving consistent results to the AOX protein groups identified in other species. To test mitochondrial localization predictions AOX1 and AOX2 sequences were reanalyzed with the MitoPROT program [27], which was designed to perform mitochondrial location analysis only in order to confirm the data. In our analysis, AOX1 0.9804 and AOX2 0.9672, mitochondrial location estimates were obtained.

***Table 2.*** *In silico prediction of AOX protein subcellular localization.*

| Protein | Length | cTP | mTP | SP | Other | RC | Predicted Location |
|---|---|---|---|---|---|---|---|
| AOX1 | 345 | 0.511 | 0.766 | 0.011 | 0.014 | 4 | M |
| AOX2 | 342 | 0.179 | 0.792 | 0.015 | 0. 021 | 2 | M |
| AOX4 | 353 | 0.930 | 0.085 | 0.015 | 0.019 | 1 | C |

*Data obtained using TargetP server tool. C, chloroplast; M, mitochondria; cTP, chloroplast transit peptide; mTP, mitochondrial targeting peptide, SP, signal peptide, RC, Reliability class.*

### B. 3. Predictions of 3-D structures of *N. alba* AOX proteins.

Three-dimensional folding structures of *N. alba* AOX proteins have been constructed to examine whether the identified AOX proteins show similar folding structure to known AOX protein constructs by using The RaptorX program. The results obtained from the analysis are shown in Table 3 and the three-dimensional structure models are shown in Figure 7. *uGDT* (*GDT*), *uSeqId* (*SeqId*) and p-values conform with statistical significance of constructed models. Three-dimensional structure models show similar folding structures that were previously proposed for alternative oxidases [31].

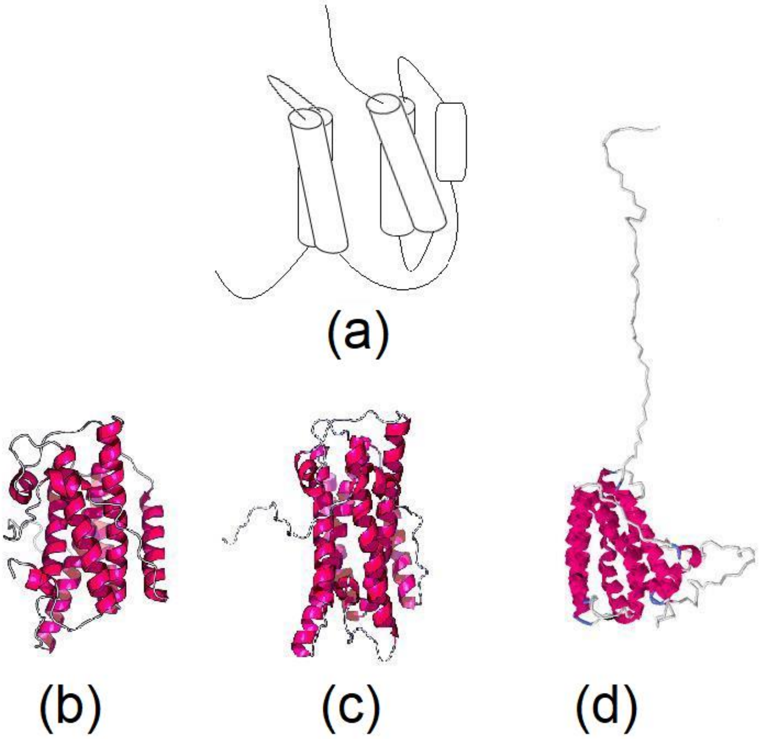


***Figure 7.*** *Three-dimensional structure predictions for Nymphaea alba AOX proteins: Structure models for each identified AOX proteins were predicted using RaptorX server.* ***(a)*** *Known model of AOX proteins (Redrawn from [31] ;* ***(b)*** *structure model for AOX1;* ***(c)*** *structure model for AOX2;* ***(d)*** *structure model for AOX4.*

***Table 3.*** *Three-dimensional structure calculation results of the identified AOX protein sequences.*

| Protein | P-value | uGDT(GDT) | uSeqId(SeqId) | Score |
|---|---|---|---|---|
| AOX1 | 1.21e-06 | 196(85) | 91(39) | 194 |
| AOX2 | 1.12e-06 | 192(83) | 93(40) | 204 |
| AOX4 | 1.09e-05 | 155(55) | 47(17) | 150 |

*Data obtained predicted AOX structure models represented in Figure 5. The p-value is the likelihood of a predicted model; GDT, global distance test value; uGDT, unnormalized GDT score; SeqId, number of identical residues in the alignment; uSeqId, normalized SeqId value; Score, alignment score.*

# IV. DISCUSSION

## A. *DE NOVO* TRANSCRIPTOME ANNOTATION OF *N. ALBA*

In this study, pooled *RNA* samples were sequenced to annotate expressed *mRNA* sequences of *N. alba*. *De novo* assembly protocol resulted in 272,934 unigenes. Unigenes are defined as the sequences that cannot be extended on either end during the *de novo* assembly process, thus can be considered as a transcript encoded from a reference gene [32]. According to quality assessments a little bit more than the half of raw sequence reads were covered in the assembled transcripts. The BUSCO analysis showed high degree of duplicated genes. The data can be explained by whole genome duplication event during the evolutionary process and high polyploidy levels. Indeed, it is known that polyploidy is common among Nymphaeales [33] and *N. nucifera* [34]. However, available genome data is limited for *Nymphaea* species (especially *Nymphaea alba*), it would be unwise to drive a concrete conclusion. In addition, data suggest the annotated transcripts covers %77 of the genes listed in Viridiplantae database and % 85.1 of the genes in Eukaryota database. Since the study covers the expressed transcripts, more comprehensive data can be obtained by sequencing the transcripts from samples collected from diverse environmental conditions or by whole genome sequencing analysis.

Functional annotation of assembled sequences was carried out by searching the sequences in several databases (Swiss-Prot (BLASTX/BLASTP); EggNOG; PFAM). We were able to annotate 20809 of assembled unigenes with a functional protein which is close to estimated number of genes for plant species in average. For GO term clustering, we were able to assign at least one GO term for 60.9% of 20809 unigenes. The highest distribution was obtained for biological process clustered unigenes. Report on transcriptome annotation among diverse plants can vary due to the state of plant samples upon sampling [35]. The results suggest that the unigenes are functionally important for structural regulation, cellular signaling and development. 2.56% of clustered unigenes were closer to bacteria, archaea and fungi. That could be due to evolutionary consequences, but there is always a chance of contamination issues for transcriptome studies carried out with plant samples collected from the field [36].

We also compared the phylogenetic topology of AOX orthologs. Tree topologies of AOX1 and AOX2 shows that AOX4 evolutionary diverged from the descendants of mitochondrial AOX isoforms. On the other hand, data suggest that evolution of mitochondrial AOX gene families were originated from a common ancestor. Structure topology is consistent with the previous study where land plants are clearly grouped under separate monophyletic groups [13]. The polyphyletic position of mitochondrial and chloroplast AOX isoforms of *N. alba* is expected due to the evolutionary position of the species. As a member of ANITA grade, *N. alba* is classified under the Nymphaeales order which is evolved from the lineage leading to most of known plants.

## B. IDENTIFICATION OF N. ALBA ALTERNATIVE OXIDASE ENCODING GENES

From assembled transcripts, 77 contigs with AOX signatures were analyzed with BLAST against known AOX proteins. Comparative approach revealed three transcripts (*AOX1*, *AOX2*, *AOX4*) with full-length ORF sequence information encoding AOX proteins. Our analysis failed to identify AOX3 as a conserved AOX protein. AOX3 was first identified in *Glycine max* and 5 other plant species according to NCBI database by June 2020 [37] . It was also considered as a variant of *AOX1* gene family [38]. Studies suggest that expression of AOX genes can be different among groups, depending on the specific growth conditions and the tissues [39]. In this study, the plant samples were collected from nature and we had no control on growth conditions. This may have limited the number of isolated AOX variants.

Sequence alignment analysis showed that amino acids responsible for iron-binding sites were well conserved. Glutamate and histidine residues required for iron-binding, and catalytic activity were conserved within four motif regions (LET, NERMH, LEEEA and RADE_ _H regions) [14], [40], [41]. Although they were conserved among compared species, there is an obvious difference for sets of amino acids within motif regions when AOX1 and AOX2 proteins were compared to AOX4 sequence. Identified AOX4 protein is a plastid terminal oxidase that is similar to previously identified immutant AOX protein in *Arabidopsis thaliana* and some other plant species (Berthold and Stenmark 2003). The differences in AOX4 motif regions are also conserved among other immutant AOX proteins [42]. Evolutionary divergence of *AOX4* gene shows a similar pattern obtained for chloroplast genome comparisons in ANITA clade [43]. On the other hand, phylogenetic tree topologies are different for AOX1 and AOX2 when compared to each other and to AOX4 protein. Data suggest that evolution of three alternative oxidases were independent from each other. Indeed, more detailed data for diverse plant species is required before giving a precise conclusion.

Three-dimensional structure analysis of identified AOX proteins showed folding structure that the estimated models conformed to the characteristic structure models of known AOX proteins. When these structures are examined, the characteristic sequential alpha-helix strand form is preserved in both AOX proteins. These constructs show high similarity to known 3-dimensional AOX models, and they support that the proteins are AOX proteins [14], [31], [41]. Data suggest that the model presented for all three proteins are statistically correct. Especially, AOX1 and AOX2 show that the model is constructed correctly in terms of all parameters. However, AOX4 *uSeqId* (*SeqId*) value is low even though the p-value and *uGDT* (*GDT*) values are above the required threshold values. The *uSeqID* value is essentially a measure of the level of affinity of the sequence used as a source of the protein of interest. Since AOX4 proteins are less studied compared to other AOX proteins, it is expected to have a lower uSeqID value. The most important value in these parameters is the *uGDT* (*GDT*) parameter, and the results above 50 are an important criterion for indicating the accuracy of the model [44]. Indeed, more accurate structural validation models can only be achieved by analyzing functional changes on purified AOX proteins supported with the data obtained from mutant AOX proteins.

# V. CONCLUSIONS

In conclusion, this is the first study presenting a comprehensive gene annotation for *Nymphaea alba*. The data presented will be useful for detailed functional genomic approaches. Especially, *qPCR*-based investigation of AOX gene expression profiles under diverse seasonal and growth-dependent conditions should increase our understanding for *N. alba* behavior at molecular level. Three identified AOX genes can also be targeted for functional studies to improve cultivation techniques and for evolutionary studies considering *N. alba* as a member of the basal angiosperm group.

ACKNOWLEDGEMENTS: This study was financially supported by Bolu Abant İzzet Baysal University [Grant no: 2014.03.03.714]. We also thank Ph.D. students Özge KAYA, Ömer Can ÜNÜVAR, and Yunus ŞAHİN for helping through field sampling.

# VI. REFERENCES

[1] H. Luo *et al.*, “The expression of floral organ identity genes in contrasting water lily cultivars,” *Plant Cell Rep.*, vol. 30, no. 10, pp. 1909–1918, 2011.

[2] B. S. Thippeswamy, B. Mishra, V. P. Veerapur, and G. Gupta, “Anxiolytic activity of Nymphaea alba Linn. in mice as experimental models of anxiety,” *Indian J. Pharmacol.*, vol. 43, no. 1, pp. 50–55, 2011.

[3] N. Khan and S. Sultana, “Inhibition of potassium bromate-induced renal oxidative stress and hyperproliferative response by *Nymphaea alba* in Wistar rats,” *J. Enzyme Inhib. Med. Chem.*, vol. 20, no. 3, pp. 275–283, 2005.

[4] N. Khan and S. Sultana, “Anticarcinogenic effect of Nymphaea alba against oxidative damage, hyperproliferative response and renal carcinogenesis in Wistar rats,” *Mol. Cell. Biochem.*, vol. 271, no. 1–2, pp. 1–11, 2005.

[5] A. Chaveerach, T. Tanee, and R. Sudmoon, “Molecular identification and barcodes for the genus *Nymphaea*,” *Acta Biol. Hung.*, vol. 62, no. 3, pp. 328–340, 2011.

[6] D. E. . Soltis *et al.*, “Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences,” *Bot. J. Linn. Soc.*, vol. 133, no. 4, pp. 381–461, 2000.

[7] P. Lakshmanan, “In vitro establishment and multiplication of Nymphaea hybrid 'James Brydon',” *Plant Cell. Tissue Organ Cult.*, vol. 36, no. 1, pp. 145–148, 1994.

[8] R. Clifton, A. H. Millar, and J. Whelan, “Alternative oxidases in Arabidopsis: a comparative analysis of differential expression in the gene family provides new insights into function of non-phosphorylating bypasses.,” *Biochim. Biophys. Acta*, vol. 1757, no. 7, pp. 730–41, 2006.

[9] B. H. Simons, F. F. Millenaar, L. Mulder, L. C. Van Loon, and H. Lambers, “Enhanced Expression and Activation of the Alternative Oxidase during Infection of Arabidopsis withPseudomonas syringae pv tomato,” *Plant Physiol.*, vol. 120, no. 2, 1999.

[10] C.-R. Li *et al.*, “Unravelling mitochondrial retrograde regulation in the abiotic stress induction of rice ALTERNATIVE OXIDASE 1 genes,” *Plant. Cell Environ.*, vol. 36, no. 4, pp. 775–788, 2013.

[11] D. A. Berthold and P. Stenmark, “Membrane-bound diiron carboxylate proteins,” *Annu. Rev. Plant Biol.*, vol. 54, no. 1, pp. 497–517, 2003.

[12] A. E. McDonald and G. C. Vanlerberghe, “Origins, evolutionary history, and taxonomic distribution of alternative oxidase and plastoquinol terminal oxidase,” *Comp. Biochem. Physiol. - Part D Genomics Proteomics*, vol. 1, no. 3, pp. 357-364, 2006.

[13] R. Pennisi, D. Salvi, V. Brandi, R. Angelini, P. Ascenzi, and F. Polticelli, “Molecular Evolution of Alternative Oxidase Proteins: A Phylogenetic and Structure Modeling Approach,” *J. Mol. Evol.*, vol. 82, no. 4-5, pp. 207-218, 2016.

[14] J. N. Siedow and A. L. Umbach, “The mitochondrial cyanide-resistant oxidase: structural conservation amid regulatory diversity,” *Biochim. Biophys. Acta - Bioenerg.*, vol. 1459, no. 2, pp. 432–439, 2000.

[15] A. L. Umbach, F. Fiorani, and J. N. Siedow, “Characterization of Transformed Arabidopsis with Altered Alternative Oxidase Levels and Analysis of Effects on Reactive Oxygen Species in Tissue,” *Plant Physiol.*, vol. 139, no. 4, pp. 1806-1820, 2005.

[16] D. P. Maxwell, Y. Wang, and L. McIntosh, “The alternative oxidase lowers mitochondrial reactive oxygen production in plant cells.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 14, pp. 8271–6, 1999.

[17] M. Ribas-Carbo, R. Aroca, M. A. Gonzàlez-Meler, J. J. Irigoyen, and M. Sánchez-Díaz, “The Electron Partitioning between the Cytochrome and Alternative Respiratory Pathways during Chilling Recovery in Two Cultivars of Maize Differing in Chilling Sensitivity,” *Plant Physiol.*, vol. 122, no. 1, pp. 199–204, 2000.

[18] T.-T. Chai, D. Simmonds, D. A. Day, T. D. Colmer, and P. M. Finnegan, “A GmAOX2b antisense gene compromises vegetative growth and seed production in soybean,” *Planta*, vol. 236, no. 1, pp. 199–207, 2012.

[19] H. Jiang, R. Lei, S.-W. Ding, and S. Zhu, “Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads,” *BMC Bioinformatics*, vol. 15, no. 1, p. 182, 2014.

[20] B. J. Haas *et al.*, “*De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.,” *Nat. Protoc.*, vol. 8, no. 8, pp. 1494–512, 2013.

[21] C. Trapnell and S. L. Salzberg, "How to map billions of short reads onto genomes.," *Nat. Biotechnol.*, vol. 27, no. 5, pp. 455–457, 2009.

[22] R. M. Waterhouse *et al.*, "BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics.," *Mol. Biol. Evol.*, vol. 35, no. 3, pp. 543–548, 2018.

[23] J. E. Stajich *et al.*, "The Bioperl toolkit: Perl modules for the life sciences.," *Genome Res.*, vol. 12, no. 10, pp. 1611–8, 2002.

[24] F. Sievers *et al.*, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.," *Mol. Syst. Biol.*, vol. 7, p. 539, 2011.

[25] A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton, "Jalview Version 2--a multiple sequence alignment editor and analysis workbench.," *Bioinformatics*, vol. 25, no. 9, pp. 1189–91, 2009.

[26] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence," *J. Mol. Biol.*, vol. 300, no. 4, pp. 1005–1016, 2000.

[27] M. G. Claros and P. Vincens, "Computational Method to Predict Mitochondrially Imported Proteins and their Targeting Sequences," *Eur. J. Biochem.*, vol. 241, no. 3, pp. 779–786, 1996.

[28] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res.*, vol. 22, no. 22, pp. 4673-4680, 1994.

[29] M. N. Price, P. S. Dehal, and A. P. Arkin, "Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix," *Mol. Biol. Evol.*, vol. 26, no. 7, pp. 1641- 1650, 2009.

[30] M. Källberg *et al.*, "Template-based protein structure modeling using the RaptorX web server.," *Nat. Protoc.*, vol. 7, no. 8, pp. 1511–22, 2012.

[31] A. E. McDonald, "Alternative oxidase: an inter-kingdom perspective on the function and regulation of this broadly distributed 'cyanide-resistant' terminal oxidase," *Funct. Plant Biol.*, vol. 35, no. 7, p. 535, 2008.

[32] B. B. Patnaik *et al.*, "Sequencing, *De Novo* Assembly, and Annotation of the Transcriptome of the Endangered Freshwater Pearl Bivalve, Cristaria plicata, Provides Novel Insights into Functional Genes and Marker Discovery," *PLoS One*, vol. 11, no. 2, p. e0148622, 2016.

[33] J. Pellicer, L. J. Kelly, C. Magdalena, and I. J. Leitch, "Insights into the dynamics of genome size and chromosome evolution in the early diverging angiosperm lineage Nymphaeales (water lilies).," *Genome*, vol. 56, no. 8, pp. 437–449, 2013.

[34] R. Ming *et al.*, "Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.)," *Genome Biol.*, vol. 14, no. R41, pp. 1-11, 2013.

[35] M. E. Bolger, B. Arsova, and B. Usadel, "Plant genome and transcriptome annotations: from misconceptions to simple solutions," *Brief. Bioinform.*, vol. 12, p. bbw135, 2017.

[36] N. J. B. Brereton *et al.*, "Comparative Transcriptomic Approaches Exploring Contamination Stress Tolerance in Salix sp. Reveal the Importance for a Metaorganismal *de Novo* Assembly Approach for Nonmodel Plants.," *Plant Physiol.*, vol. 171, no. 1, pp. 3–24, 2016.

[37] J. Whelan, A. H. Millar, and D. A. Day, “The alternative oxidase is encoded in a multigene family in soybean,” *Planta*, vol. 198, no. 2, pp. 197–201, 1996.

[38] Y. Ito, D. Saisho, M. Nakazono, N. Tsutsumi, and A. Hirai, “Transcript levels of tandem-arranged alternative oxidase genes in rice are increased by low temperature,” *Gene*, vol. 203, no. 2, pp. 121–129, 1997.

[39] D. Saisho, E. Nambara, S. Naito, N. Tsutsumi, A. Hirai, and M. Nakazono, “Characterization of the gene family for alternative oxidase from Arabidopsis thaliana,” *Plant Mol. Biol.*, vol. 35, no. 5, pp. 585–596, 1997.

[40] T. Magnani *et al.*, “Cloning and functional expression of the mitochondrial alternative oxidase of *Aspergillus fumigatus* and its induction by oxidative stress,” *FEMS Microbiol. Lett.*, vol. 271, no. 2, pp. 230–238, 2007.

[41] J. N. Siedow, A. L. Umbach, and A. L. Moore, “The active site of the cyanide-resistant oxidase from plant mitochondria contains a binuclear iron center,” *FEBS Lett.*, vol. 362, no. 1, pp. 10–14, 1995.

[42] D. A. Berthold, M. E. Andersson, and P. Nordlund, “New insight into the structure and function of the alternative oxidase,” *Biochim. Biophys. Acta - Bioenerg.*, vol. 1460, no. 2–3, pp. 241–254, 2000.

[43] V. V Goremykin, K. I. Hirsch-Ernst, S. Wölfl, and F. H. Hellwig, “The Chloroplast Genome of Nymphaea alba: Whole-Genome Analyses and the Problem of Identifying the Most Basal Angiosperm,” *Mol. Biol. Evol.*, vol. 21, no. 7, pp. 1445–1454, 2004.

[44] J. Peng and J. Xu, “Raptorx: Exploiting structure information for protein alignment by statistical inference,” *Proteins Struct. Funct. Bioinforma.*, vol. 79, no. S10, pp. 161–171, 2011.