In Silico Cloning and Bioinformatics Analysis of Shikimate Dehydrogenase Gene from *Medicago sativa*

GuanPing You[1, 2] and JianZhong Huang[1, 2]
[1] College of Life Sciences, Fujian Normal University, Fuzhou, Fujian, China
[2] Engineering Research Center of Industrial Microbiology, Ministry of Education, National and Local United Engineering Research Center of Industrial Microbiology and Fermentation Technology, Fujian Normal University, Fuzhou, Fujian, China

Abstract—The combination of artificial intelligence (AI) and bioinformatics is driving a leap forward in genomics and biological research, especially in the electronic cloning and biological analysis of genes. AI can analyze large-scale genomic data, identify gene variations and predict gene functions through machine learning algorithms, thereby improving the efficiency and accuracy of gene cloning. Electronic cloning technology combines computer modeling and experimental data to simulate the gene expression process, greatly accelerating the progress of gene function research. In the secondary metabolic pathway of plants, shikimate dehydrogenase (SDH) is one of the key enzymes involved in the regulation of the shikimate pathway, which is a key step in the synthesis of important plant secondary metabolites such as phenylpropene compounds, flavonoids and lignin. Shikimate dehydrogenase catalyzes the conversion of shikimate to coumaric acid, which is the basis of plant defense mechanisms, antioxidants and disease resistance. In this study, AI tools were used to deeply analyze the gene expression patterns related to shikimate dehydrogenase, and the shikimate dehydrogenase sequence gene of *Escherichia coli* was used as a probe to clone and analyze the *Medicago sativa* shikimate dehydrogenase gene. The results showed that the cloned shikimate dehydrogenase gene of *M. sativa* was 469 bp in length and had 5 open reading frames (ORFs), of which ORF3 was the longest, with a total length of 258 bp, encoding 85 amino acids. The molecular weight of the protein was 9370.70, and the theoretical isoelectric point pI was 5.67, indicating that it was a functional protein on abiotic membranes. Through further bioinformatics analysis, it was speculated that the gene may play an important role in the secondary metabolism of *M. sativa*, and its expression pattern may be closely related to the growth and environmental adaptability of the plant.

Index Terms—Artificial intelligence, bioinformatics, *Medicago sativa*, shikimate dehydrogenase, in silico cloning

I. INTRODUCTION

Shikimate dehydrogenase (SDH) is a key enzyme in the shikimate pathway, which plays a vital role in the synthesis of aromatic amino acids and their precursors in plants. This pathway is central to plant secondary metabolism, contributing not only to the production of amino acids such as tryptophan, tyrosine, and phenylalanine, but also to certain plant hormones and a vast array of secondary metabolites (Fig. 1). Beyond metabolism, the shikimate pathway influences plant growth, environmental adaptation, stress resistance, and yield. Its industrial significance is also notable, particularly as shikimic acid serves as a key precursor for the synthesis of the antiviral drug oseltamivir (Tamiflu)[1]. Genetic engineering of this pathway offers promising avenues for enhancing the production of valuable compounds and improving agronomic traits[1].

Medicago sativa (alfalfa), an important forage and green manure crop, exhibits shikimate pathway activity that is closely linked to stress tolerance and biomass yield. Although recent genetic studies on *M. sativa* have increasingly focused on stress-related genes, functional characterization of SDH — a pivotal enzyme in secondary metabolism — has remained limited[2]. The integration of artificial intelligence (AI) with bioinformatics has created new opportunities for accelerating gene discovery and functional prediction[3-4]. Yet, current approaches often rely on generalized bioinformatics tools that lack custom analysis tailored to non-model species such as alfalfa, and most conventional methods do not fully leverage AI for predictive functional insight.

To address these limitations, this study employs a targeted in silico cloning strategy combined with multi-level bioinformatics analysis to identify and characterize the SDH gene from *M. sativa*. We present the first report of a putative SDH gene in alfalfa, comprising 469 bp with five open reading frames (ORFs), among which ORF3 encodes an 85-amino acid protein. Our analysis reveals key protein characteristics including a molecular weight of 9370.70, an acidic isoelectric point (pI) of 5.67, hydrophilic nature, and cytoplasmic localization supported by the absence of signal peptides and transmembrane domains. Furthermore, functional motif identification and phylogenetic analysis provide insight into the evolutionary conservation and functional role of SDH in legumes.

These findings establish a essential genetic resource for future research on metabolic engineering in alfalfa, and demonstrate a bioinformatics workflow that can be augmented with AI tools for improved gene function prediction. This work not only facilitates further functional validation of SDH but also

This work was supported by the National Key Research and Development Program of China under Grant No. 2022YFD1802104. Corresponding author: JianZhong Huang (e-mail: hjz@fjnu.edu.cn). The author GuanPing You(e-mail: 1257887082@qq.com) is with the College of Life Sciences, Fujian Normal University, Fuzhou, Fujian, China.

provides a foundation for enhancing stress resistance and yield in *M. sativa* through molecular breeding.

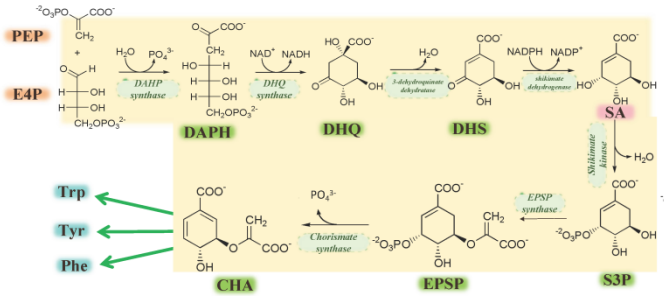

Fig. 1. The shikimate pathway. PEP: Phosphoenolpyruvate; E4P: D-erythrose 4-phosphate; DAHP: 3-deoxy-d-arabino-heptulosonic acid 7-phosphate; DHQ: 3-dehydroquinate; DHS: 3-dehydroshikimate; SA: Shikimate; S3P: Shikimate-3phosphate; EPSP: 5-enolpyruvylshikimate-3-phosphate; CHA: Chorismate; Phe: Phenylalanine; Tyr: Tyrosine; Trp: Tryptophan.

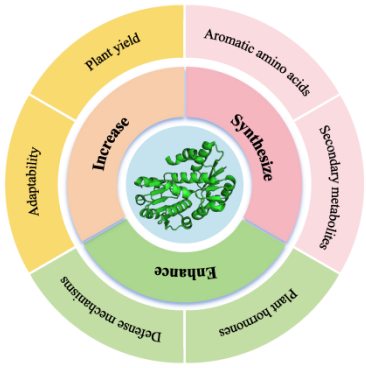

Fig. 2. Application of shikimic acid in plants.

II. Related Work

The shikimate pathway is a fundamental metabolic route in plants, bacteria, and fungi, serving as the crucial bridge between primary carbon metabolism and the biosynthesis of aromatic amino acids and a vast array of secondary metabolites. Within this pathway, shikimate dehydrogenase (SDH, EC 1.1.1.25) catalyzes the reversible reduction of 3-dehydroshikimate to shikimate, utilizing NADPH as a cofactor. This reaction is a critical control point, making SDH a subject of significant interest in both basic research and applied biotechnology.

A.Studies on SDH Genes in Various Species

Extensive research has been conducted on SDH genes across different kingdoms. In bacteria, such as *Escherichia coli,* the *aroE* gene encoding SDH has been well-characterized, and its structure-function relationship has been elucidated, providing a foundational model for understanding enzyme kinetics and mechanism. In plants, SDH has been identified and studied in model species like *Arabidopsis thaliana* and major crops such as rice (Oryza sativa). These studies have confirmed SDH's pivotal role in development and stress responses. For instance, the silencing of SDH in *Arabidopsis* led to severe developmental defects, underscoring its indispensability. Furthermore, recent work has begun to explore SDH in bioenergy crops like Panicum virgatum (switchgrass), highlighting its potential in engineering pathways for improved biomass and stress resilience. While these studies provide a general framework for understanding SDH, functional characteristics can vary significantly between species due to evolutionary divergence and lineage-specific adaptations.

B. Research Gap and Objective

While SDH is recognized as a critical enzyme, its specific sequence, structure, and functional attributes in *Medicago sativa*, a legume crop of immense agricultural importance, remain poorly characterized. Previous studies in other species provide a template but cannot directly be extrapolated. The conventional bioinformatics approaches used in many prior SDH studies, while useful, lack the predictive power of modern AI techniques. Therefore, there is a clear need for a comprehensive study that not only identifies and characterizes the *M. sativa* SDH gene using established *in silico* methods but also frames these findings within the modern context of AI-driven biological discovery. This study aims to fill this gap by conducting a foundational bioinformatic characterization of *M. sativa* SDH and explicitly outlining a pathway for its future validation and application using advanced computational intelligence, thereby contributing to the genetic improvement of this vital crop.

III. Analysis Methods and Tools

A. In silico cloning of SDH from M. sativa

The probe amino acid sequence of the *E. coli* SDH used in this study is following:

METYAVFGNPIAHSKSPFIHQQFAQQLNIEHPYGRVL APINDFINTLNAFFSAGGKGANVTVPFKEEAFARADELT ERAALAGAVNTLMRLEDGRLLGDNTDGVGLLSDLERL SFIRPGLRILLIGAGGASRGVLLPLLSLDCAVTITNRTVS RAEELAKLFAHTGSIQALSMDELEGHEFDLIINATSSGIS GDIPAIPSSLIHPGIYCYDMFYQKGKTPFLAWCEQRGSK RNADGLGMLVAQAAHAFLLWHGVLPDVEPVIKQLQEE LSA.

Next, the probe sequence was searched with the M. sativa EST database using the tBlastn tool in NCBI, and gene sequences with a match greater than 50% were selected and downloaded in FASTA format; contig-0 and contig-1 were obtained by gene splicing using BioEdit "CAP conting assembly program". Finally, the *M. sativa* shikimate dehydrogenase cDNA sequence was predicted by ORF Finder in NCBI to determine whether there was a gene with the expected function, and finally the new gene fragment was determined. The analysis tools are shown in Table 1.

TABLE I
TOOLS USED TO PREDICT GENE STRUCTURE

Search content	Tools
Sequence acquisition	NCBI

	https://www.ncbi.nlm.nih.gov/
Sequence splicing	BioEdit
Open reading frame identification	ORFfinder https://www.ncbi.nlm.nih.gov/orffinder/

B. Bioinformatics Analysis of the SDH in M. sativa

The obtained *M. sativa* SDH gene sequence was analyzed using bioinformatics software(*Sequence alignments were auto-assembled via BioEdit CAP, with manual curation of ambiguous regions.* , and an evolutionary tree was constructed by analyzing and predicting the physicochemical properties, hydrophilicity and hydrophobicity, functional sites, transmembrane analysis, signal peptides, subcellular localization, secondary structure, tertiary structure, molecular evolution, etc. of the protein encoded by the gene. The specific analysis content and tool software are shown in Table 2.

TABLE II
TOOLS USED TO PREDICT PROTEIN STRUCTURE AND FUNCTION

Search content	Tools and Parameters
Base composition	Bioedit（Nucleotide Composition, Restriction Map）
Physical and chemical properties	http://web.expasy.org/protparam/
Hydrophilicity/hydrophobicity	http://web.expasy.org/protscale/
Functional sites	https:/web.expasy.org/ prosite/
Transmembrane analysis	TMHMM2.0,Default settings (membrane probability >0.5; N-tail inside) https://services.healthtech.dtu.dk/services/TMHMM-2.0/
Subcellular localization	https://wolfpsort.hgc.jp/
Signal peptide	https://services.healthtech.dtu.dk/service.php?SignalP-5.0
Secondary structure	SOPMAD,Window width=17; Decision constants: Helix (⩾4), Sheet (⩾4) https://npsa-prabi.ibcp.ft/cgi-bin/npsaautomat.pl?page=npsa%20sopma.html
Phosphorylation site	https://services.healthtech.dtu.dk/services/NetPhos-3.1/
Tertiary structure	http://swissmodel.expasy.org/repository/
Homologous evolutionary tree	MEGA11,Neighbor-Joining (NJ) tree; Bootstrap=1000 replicates; Poisson correction

III. AMINO ACID COMPOSITION ANALYSIS OF SDH IN *M. SATIVA*

Amino acid	Number	Proportion	Amino acid	Number	Proportion
Ala(A)	9	10.6%	Ile(I)	7	8.2%
Arg(R)	3	3.5%	Leu(L)	8	9.4%
Asn(N)	2	2.4%	Lys(K)	3	3.5%
Asp(D)	5	5.9%	Met(M)	3	3.5%
Cys(C)	3	3.5%	Phe(F)	4	4.7%
Gln(Q)	3	3.5%	Pro(P)	4	4.7%
Glu(E)	4	4.7%	Ser(S)	7	8.2%
Gly(G)	8	9.4%	Thr(T)	2	2.4%
His(H)	4	4.7%	Trp(M)	2	2.4%
Val(V)	1	1.2%	Tyr(Y)	3	3.5%

A. In silico cloning results of shikimate dehydrogenase from M. sativa

The EST sequences obtained by TBLASTN were saved in FASTA format. The sequences were spliced using Bioedit software to finally obtain a contig with a length of 469bp. Then, through the prediction tool ORF Finder in NCBI, the *M. sativa* SDH sequence has 5 ORFs, of which the longest ORF is 258bp long. The protein it encodes contains 86 amino acids, and the sequence is:

MDELEGHEFDLIINATSSGISGDIPAIPSSLIHPGIYCYD MFYQKGKTPFLAWCEQRGSKRNADGLGMLVAQAAHA FLLWHRCSA. Sequences with high homology were found through BLAST. Using the Bioedit tool, the base composition can be obtained by analyzing the electronic cloning splicing sequence, among which the proportion of adenine A is 28.14%; the proportion of cytosine C is 27.08%; the proportion of guanine G is 22.17%; and the proportion of thymine T is 22.60%. Analysis of restriction enzyme positions of the electronically cloned spliced sequence revealed that the restriction enzyme sites of the nucleotide sequence of the *M. sativa* SDH include AflIII, AlwI, AseI, and the like.

B. Analysis of the physicochemical properties of M. sativa shikimate dehydrogenase protein

The ExPASy-Protparam tool was used to analyze the physicochemical properties of *M. sativa* shikimate dehydrogenase. The analysis showed that the number of amino acids was 85, the molecular weight of the protein was 9370.70, the theoretical isoelectric point pl was 5.67, the molecular formula was $C_{419}H_{636}N_{112}O_{121}S_6$, the instability coefficient was 35.85, it was a stable protein, the fat coefficient was 82.82, and the average hydrophobicity was -0.031. Its amino acid composition is shown in Table 3. There are 9 negatively charged amino acids (Asp + Glu), 6 positively charged amino acids (Arg + Lys), Ala accounts for the largest proportion, 10.6%, there is no pyrrolysine (Pyl) and selenocysteine (Sec), and the protein is an acidic protein.

C. Prediction and analysis of hydrophilicity/hydrophobicity of M. sativa shikimate dehydrogenase

The ExPASy-ProtScale online software was used to predict

the hydrophilicity/hydrophobicity of the amino acid protein encoded by the *M. sativa* SDH gene. The results are shown in Figure 3. Analysis of the figure shows that the larger the negative value, the weaker the hydrophilicity of the protein; conversely, the larger the positive value, the stronger the hydrophobicity of the protein. The value between +2 and -3 indicates that the amino acid is amphoteric. The highest score of the polypeptide chain is at the 69th position of the protein, which is +1.689, and the lowest score is at the 58th and 59th positions of the protein, which is -2.533, so it is a hydrophilic protein. It can also be observed that the negative peak is significantly higher than the positive peak, and it is inferred that the *M. sativa* shikimate dehydrogenase is hydrophilic. This is consistent with the analysis results of the ExPASy-Protparam software.

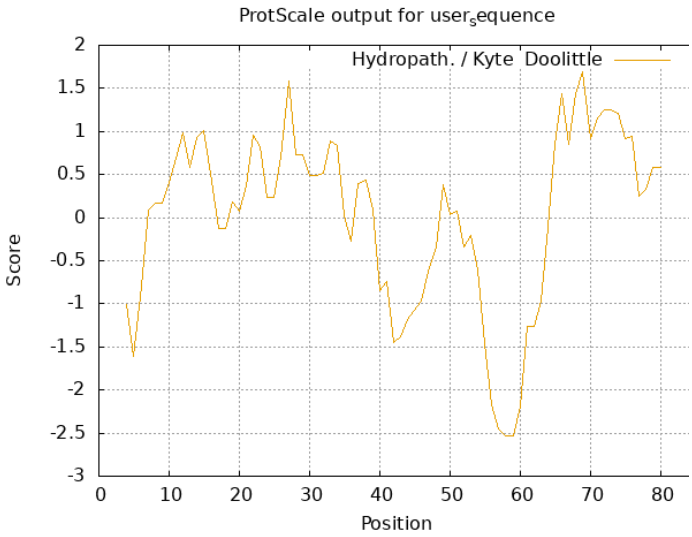

Fig. 3. The hydrophilicity/hydrophobicity prediction results of *M. sativa* SDH.

D. Prediction and analysis of signal peptide and transmembrane domain of M. sativa SDH

The signal peptide is a peptide segment consisting of 20 to 30 amino acid residues at the N-terminus of the nascent peptide chain of a secretory protein. It determines the modification of certain amino acid residues and is often used to guide the transmembrane transfer of proteins. The signal peptide structure of *M. sativa* SDH was predicted using the Signalp5.0 online software. The prediction results are shown in Figure 4. The results show that the probability of the protein being a signal peptide is 0.0015, and it can be inferred that the protein is a non-secretory protein，and the absence of signal peptides (SignalP5.0 score=0.0015) and transmembrane domains (TMHMM) confirms cytosolic localization. This aligns with SDH's role in cytoplasmic shikimate metabolism [1,3] and suggests direct interaction with cytoplasmic substrates like shikimate.

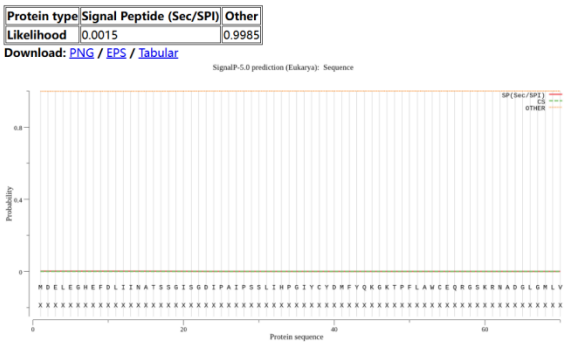

Fig. 4. Protein signal peptide prediction results of *M. sativa* SDH.

The transmembrane domain is the main part where the membrane-intrinsic protein binds to the membrane lipids. It is generally composed of about 20 hydrophobic amino acids to form an alpha helix, which is fixed to the cell membrane and acts as an anchor. The TMHMM-2.0 online tool was used to predict the transmembrane domain of the protein. The results are shown in Figure 5. At 1.0, it is the outer boundary of the cell membrane, and 0 is inside the cell membrane. This study predicted a protein sequence with a length of 85 amino acids, but no predicted transmembrane helices (TMHs) were found, indicating that the protein is likely to contain no transmembrane regions, or these regions may be too short or do not meet the prediction criteria of the TMHMM model. Overall, this protein may be a non-transmembrane protein or located inside the cell. Further, the protein was annotated and the functional site prediction was performed using ExPASy-Prosite. The results showed that the 14-17, 58-63, and 67-72 amino acids were the predicted functional sites of the protein.

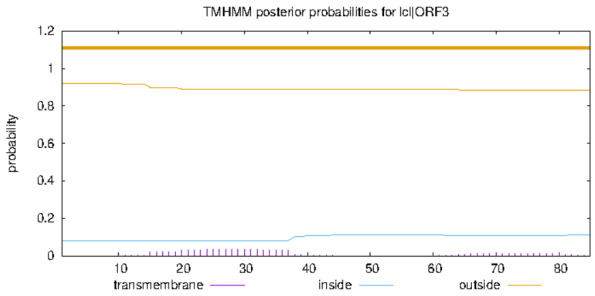

Fig. 5. Protein transmembrane domain prediction results of *M. sativa* SDH.

E. Structural prediction and analysis of M. sativa SDH

The local spatial structure of the polypeptide main chain of *M. sativa* SDH was analyzed using the online software SOPMA. The results showed that in the secondary structure of the protein, the largest proportion was random coils, accounting for 48.24%, with 41; followed by α helices, accounting for 31.76%, with 27; there were 17 extended chains, accounting for 20%; there was no β fold.

The SWISS-MODEL online homology modeling method was used to analyze the *M. sativa* SDH protein, predict its tertiary structure, and use the ball-and-stick model to display all aromatic amino acids. The results are shown in Figure 6, A is the predicted image of the tertiary structure of the protein, and the blue part shown in B is the position of the aromatic amino acids in the protein stick-and-ball model.

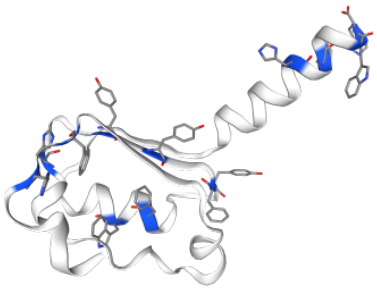

Fig. 6. Protein tertiary structure and predicted position of aromatic amino acids.

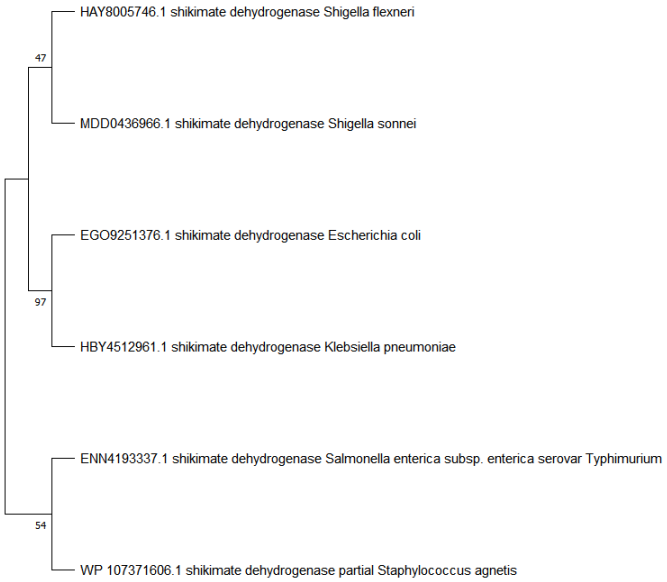

Fig. 7. Phylogenetic analysis of *M. sativa* SDH.

F. Construction of the phylogenetic tree of M. sativa SDH

Using the results from ORF Finder in NCBI, we obtained homologous sequences of various species through BLAST analysis, and then used the software MEGA11 to analyze the evolutionary tree structure (Figure 7). We searched for protein sequences similar to the protein sequence and performed multiple sequence alignment on them to construct the molecular evolutionary phylogenetic tree of the protein in plant species (using the neighbor-joining method, NJ).

Phylogenetic analysis indicates a clear evolutionary relationship among the SDH genes from the bacterial species examined. Sequences from members of the Enterobacteriaceae family—including *Escherichia coli* , *Shigella flexneri* , *Shigella sonnei* , *Klebsiella pneumonia*e , and *Salmonella enterica* — form a tightly clustered monophyletic clade with strong bootstrap support (values of 97 and 47), demonstrating high sequence similarity and suggesting a relatively recent common ancestor. In contrast, *Staphylococcus agnetis* (phylum Firmicutes) represents a more distantly related lineage. Its position as an outer group, with a bootstrap value of 54 at the divergent node, is consistent with its taxonomic distinction from the Enterobacteriaceae, underscoring the divergence between these bacterial groups.

IV. CONCLUSION

This study focused on the SDH of *M. sativa* and successfully obtained the full-length sequence of the gene using electronic cloning technology. Through a series of bioinformatics analysis tools and methods, the biological characteristics of the gene and the protein it encodes were systematically analyzed[5]. The results of gene sequence analysis revealed the characteristics of its base composition, potential restriction sites and sequence variation information, laying the foundation for subsequent gene function research and application. In terms of protein sequence analysis, we deeply explored key factors including functional sites, domains, physicochemical properties, hydrophilic and hydrophobic properties. In addition, signal peptide prediction, transmembrane structure analysis ~~and subcellular localization analysis~~ also provide an intuitive understanding of the distribution and action location of the protein in the cell. Further secondary and tertiary structure predictions provide a spatial structural basis for revealing the functional mechanism of the protein, while the phylogenetic tree constructed based on multiple sequence alignment and molecular evolution analysis reveals the evolutionary relationship of the gene, providing important clues for inferring its evolutionary process and genetic differences between species.

These research results not only provide key data support for our in-depth understanding of the SDH mechanism of *M. sativa* growth and development, but also provide a theoretical basis for the gene function and metabolic regulation of *M. sativa*. By revealing the specific function of this gene in the metabolic pathway of *M. sativa*, these data provide an important reference for *M. sativa* variety improvement, metabolic regulation and stress resistance research. At the same time, the gene sequence and protein characteristic analysis obtained in this study

provide a solid theoretical basis for subsequent functional verification experiments, and provide new ideas and directions for further exploring the application potential of *M. sativa* in agricultural production.

V. Discussion

While this study utilized conventional bioinformatics tools to successfully clone and characterize the *M. sativa SDH* gene, the integration of AI in future work could profoundly deepen the interpretation of our findings and accelerate functional validation. The specific features of the *M. sativa SDH* sequence uncovered here—such as its acidic pI (5.67), hydrophilic nature, absence of transmembrane domains, and key functional motifs—provide an ideal foundation for AI-driven predictive modeling.

For instance, the amino acid sequence we identified could be used as direct input for deep learning models like AlphaFold [9] to generate a high-accuracy tertiary structure model, moving beyond the preliminary model presented in this study. This could reveal the spatial arrangement of catalytic residues and suggest potential binding mechanisms for substrates or inhibitors. Furthermore, AI algorithms could analyze our phylogenetic results in a broader context, identifying conserved regulatory elements across legumes that control SDH expression under stress conditions [11, 16].

The non-secretory, cytosolic localization predicted for the SDH protein indicates its role in intracellular metabolism. AI models could integrate this subcellular localization data with public transcriptomic datasets to build predictive models of how *M. sativa* SDH expression correlates with drought or pathogen challenge, thereby guiding targeted experimental validation [12]. Finally, the unique sequence motifs we reported could help train convolutional neural networks to identify SDH genes with similar regulatory features in other crops, supporting comparative genomics and precision breeding efforts [8, 13].

In conclusion, the bioinformatic profile of *M. sativa* SDH established in this work provides the essential data layer upon which AI and machine learning can be deployed to transition from *in silico* characterization to *in planta* functional analysis and metabolic engineering.

While this study utilized conventional bioinformatics tools to successfully clone and characterize the *M. sativa SDH* gene, the integration of AI in future work could profoundly deepen the interpretation of our findings and accelerate functional validation. The specific features of the *M. sativa* SDH sequence uncovered here—such as its acidic pI (5.67), hydrophilic nature, absence of transmembrane domains, and key functional motifs—provide an ideal foundation for AI-driven predictive modeling.

For instance, the amino acid sequence we identified could be used as direct input for deep learning models like AlphaFold [9] to generate a high-accuracy tertiary structure model, moving beyond the preliminary model presented in this study. This approach has demonstrated remarkable success in predicting protein structures with atomic-level accuracy, as exemplified by AlphaFold’s performance in the CASP14 competition. Such a model could reveal the spatial arrangement of catalytic residues and suggest potential binding mechanisms for substrates or inhibitors, thereby facilitating targeted mutagenesis or inhibitor design. Furthermore, AI algorithms could analyze our phylogenetic results in a broader context, identifying conserved regulatory elements across legumes that control SDH expression under stress conditions [11, 16]. For example, random forest models could be employed to integrate transcriptomic data from public repositories like PhytoMine, enabling the prediction of SDH expression patterns under drought or pathogen challenge and guiding subsequent experimental validation [12].

Additionally, the unique sequence motifs we reported could help train convolutional neural networks to identify SDH genes with similar regulatory features in other crops, supporting comparative genomics and precision breeding efforts [8, 13]. Beyond expression prediction, AI-powered tools such as DeepCRISPR [15] could leverage the gene sequence information to design high-efficiency sgRNAs for CRISPR-based knockout or editing of *M. sativa* SDH, thereby functionally validating its role in stress adaptation or metabolic flux. DeepCRISPR has already been successfully applied in animal and plant systems to improve the efficiency and specificity of gene editing, suggesting its strong potential for use in alfalfa.

In conclusion, the bioinformatic profile of *M. sativa* SDH established in this work provides the essential data layer upon which AI and machine learning can be deployed to transition from *in silico* characterization to *in planta* functional analysis and metabolic engineering.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (Grant No. 2022YFD1802104). The authors would like to thank all members of the College of Life Sciences, Fujian Normal University, for their valuable assistance and collaboration during the preparation of this manuscript.

References

[1] F. X. Niu, Y. P. Du, Y. B. Huang, *et al*. “Recent advances in the production of phenylpropanoic acids and theirderivatives by genetically engineered microorganisms,” *Synthetic Biology Journal*, vol. 1, no. 3, pp. 337-357, 2020.

[2] X. Wang, N. Y. Zhu, W. Jiang, S. Y. Si, “Identification of novel anti-tuberculosis lead compound targeting shikimate kinase, ” *Acta Pharmaceutica Sinica*, vol. 53, no. 6, pp. 878-886, 2018.

[3] Y. Y. Gong, S. Q. Guo, H. M. Shu, *et al*, “Analysis of Molecular Evolution and Gene Structure of EPSPS Protein in Plant Shikimate Pathway,”*Chinese Bulletin of Botany*, vol. 50, no. 3, pp. 295-309, 2015 .

[4] H. H. Xia and J. Z. Huang, "In Silico Cloning and Analysis of Shikimate Dehydrogenase Gene from Panicum virgatum," *J. Fujian Normal Univ. (Nat. Sci. Ed.)*, vol. 41, no. 3, pp. 97-102, 2025.

[5] F. S. G. Hashemi, M. R. Ismail, M. R. Yusop, M. S. G. Hashemi, M. H. N. Shahraki, H. Rastegari, G. Miah, & F. Aslani. "Intelligent mining of large-scale bio-data: Bioinformatics applications," *Biotechnology & Biotechnological Equipment*, vol. 32, pp. 10-29.2018.

[6] T. Huang, L. Chen, M. Zheng, & J. Song. "Integrated analysis of multiscale large-scale biological data for investigating human disease," *BioMed Research International*, 2015.

[7] H. Poon, C. Quirk, K. Toutanova, & W. Yih. "NLP for precision medicine," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1-2, 2017.

[8] Y. Cai, J. Wang, & L. Deng. "SDN2GO: An integrated deep learning model for protein function prediction," *Frontiers in Bioengineering and Biotechnology*, vol.8, pp. 391. 2020.

[9] A. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, & D. Hassabis. "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 8, pp. 706-710, 2020.

[10] H. Tong, A. Küken, & Z. Nikoloski. "Integrating molecular markers into metabolic models improves genomic selection for Arabidopsis growth," *Nature Communications*, vol 11, no 2410, 2020.

[11] C. Hill, T. Czauderna, M. Klapperstück, U. Roessner, & F. Schreiber. "Metabolomics, standards, and metabolic modeling for synthetic biology in plants," *Frontiers in Bioengineering and Biotechnology*, vol. 3, no. 167, 2015.

[12] P. Wang, B. M. Moore, S. Uygun, M. D. Lehti-Shiu, C. S. Barry, & S. Shiu. "Optimizing the use of gene expression data to predict plant metabolic pathway memberships," *bioRxiv*. vol.15,no. 204222, 2020.

[13] S. Sun, C. Wang, H. Ding, Q. Zou. "Machine learning and its applications in plant molecular studies," *Briefings in Functional Genomics*, vol. 19, no. 1, pp. 40-48,2019.

[14] H. Deng, Y. Jia, & Y. Zhang. "Protein structure prediction," *International Journal of Modern Physics B*, vol. 31, no. 1741007, pp. 16-19, 2017.

[15] K. Plaimas, J. Mallm, M. Oswald, F. Svara, V. Sourjik, R. Eils, & R. König. "Machine learning based analyses on metabolic networks supports high-throughput knockout screens," *BMC Systems Biology*, vol. 2, no. 67, 2008.

[16] Q. Song, J. Lee, S. Akter, M. Rogers, R. Grene, & S. Li. "Prediction of condition-specific regulatory genes using machine learning," *Nucleic Acids Research*, vol. 48, no. 11, 2020.