Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 03 | Mar-2026
ISSN (O) 3107-6696

Role of Bioinformatics Tools and Techniques in the Identification of Agro-Economically Important Genes in Clitoria ternatea Using Glycine max as a Model Crop

Dr. V. K. Singh

Information Officer, Centre for Bioinformatics, School of Biotechnology, Institute of Science, Banaras Hindu University, Varanasi, Uttar Pradesh 221005

---***---

Abstract- The integration of bioinformatics in plant genomics has accelerated the identification of genes responsible for agronomically important traits. *Clitoria ternatea* (butterfly pea) is an economically valuable leguminous plant known for medicinal, ornamental, and nutraceutical properties, while *Glycine max* (soybean) serves as a well-studied model crop with extensive genomic resources. Comparative genomics and bioinformatics tools enable the identification of homologous genes between these species, facilitating the discovery of genes associated with stress tolerance, yield, and secondary metabolite production. Tools such as sequence alignment programs, genome annotation platforms, and functional genomics databases play a crucial role in this process. This paper discusses key bioinformatics tools, databases, and computational approaches used for identifying agro-economically important genes in *C. ternatea* using *G. max* as a reference model. The rapid development of genomic technologies and computational biology has significantly advanced the identification of agriculturally important genes in plants. Bioinformatics plays a crucial role in analyzing large-scale genomic datasets and predicting gene functions through comparative analysis and sequence annotation. *Clitoria ternatea*, commonly known as butterfly pea, is an economically important leguminous plant with medicinal, ornamental, and nutritional value. However, genomic information for this species has historically been limited. Recently, the availability of transcriptome shotgun assembly (TSA) and whole genome shotgun (WGS) sequence datasets for *Clitoria ternatea* in public databases has created new opportunities for gene discovery and functional analysis. Comparative genomics using well-characterized model crops such as *Glycine max* (soybean) provides an effective strategy for identifying genes associated with important agronomic traits. Soybean possesses a well-annotated genome and extensive genetic databases, making it suitable for identifying orthologous genes in related legumes. Bioinformatics tools such as sequence alignment algorithms, gene prediction software, genome annotation platforms, and functional analysis databases are widely used to identify candidate genes responsible for traits such as stress tolerance, nitrogen fixation, growth regulation, and secondary metabolite biosynthesis. This study highlights the role of computational tools and publicly available genomic resources in identifying agro-economically important genes in *Clitoria ternatea* through comparative analysis with soybean. The integration of bioinformatics techniques with plant genomics can accelerate crop improvement programs, enhance understanding of plant metabolic pathways, and support sustainable agricultural development.

Keywords: Bioinformatics, comparative genomics, gene identification, *Clitoria ternatea*, *Glycine max*, plant genomics, functional annotation, legume genomics, stress tolerance genes, agricultural biotechnology.

1. Introduction

Agricultural productivity and crop improvement depend on the identification of genes responsible for traits such as yield, stress tolerance, disease resistance, and metabolic pathways. Bioinformatics combines computational tools, databases, and algorithms to analyze large biological datasets generated by genomic and transcriptomic studies. *Clitoria ternatea* is a leguminous plant widely used in herbal medicine, natural food coloring, and ornamental horticulture. Despite its economic importance, genomic information for this species remains limited. However, comparative genomics with well-studied legumes such as *Glycine max* provides a valuable strategy to identify candidate genes involved in agriculturally significant traits. Soybean is among the most extensively studied legumes with large genomic datasets and expressed sequence tag (EST) collections available for functional genomics studies.
Using soybean as a model organism allows researchers to identify homologous genes, regulatory pathways, and gene families that may also exist in *C. ternatea*.

The advancement of modern biotechnology has revolutionized the study of plant genomes and the identification of genes responsible for economically important traits. In recent years, bioinformatics has become an essential tool in plant genomics research because it enables scientists to analyze large volumes of biological data efficiently. Bioinformatics integrates biological knowledge with computational algorithms, statistical analysis, and database management to interpret genomic and molecular data. The rapid growth of next-generation sequencing technologies has produced massive datasets of genomic sequences, transcriptomes, and proteomes from various organisms. These datasets require advanced computational tools to identify genes, predict protein functions, and understand biological processes. Agricultural productivity largely depends on the identification and manipulation of genes responsible for important traits such as yield, stress tolerance, nutrient efficiency, and disease resistance. Traditional plant breeding methods rely primarily on phenotypic selection and hybridization, which can be time-consuming and often limited by environmental variability. The integration of genomic information into breeding programs has accelerated crop improvement by enabling researchers to directly target genes associated with desirable traits.

Leguminous plants play a vital role in global agriculture because they contribute to food security, soil fertility, and sustainable farming systems. Members of the Fabaceae family are known for their ability to fix atmospheric nitrogen through symbiotic relationships with soil bacteria. This biological nitrogen fixation reduces the need for synthetic fertilizers and enhances soil productivity. Among these legumes, *Clitoria ternatea*, commonly known as butterfly pea, is gaining increasing attention due to its multiple agricultural and medicinal benefits. The plant is widely distributed in tropical and subtropical regions and is particularly known for its vibrant blue flowers containing high concentrations of anthocyanin pigments. These pigments are widely used as natural food colorants and possess antioxidant properties. Despite its significant economic and medicinal value, genomic information available for *Clitoria ternatea* has been limited compared with major crop species. However, recent advances in sequencing technologies have resulted in the deposition of transcriptome and genome assemblies in public databases. The availability of a transcriptome shotgun assembly (TSA) dataset for *Clitoria ternatea* (NCBI accession GLLN00000000.1) and a whole genome shotgun (WGS) assembly (JAXHEB000000000.1) provides valuable genomic resources for computational analysis and gene identification and *in-silico* characterization. These datasets provide an opportunity to explore the genetic basis of important agronomic and metabolic traits in this species. To overcome limitations in genome annotation, researchers often rely on comparative genomics using model plants with well-characterized genomes. Comparative genomics involves analyzing genetic similarities and differences between species to identify conserved genes and biological pathways.

Glycine max, commonly known as soybean, is one of the most important legume crops worldwide and serves as a valuable model for studying legume genomics. The soybean genome has been fully sequenced and extensively annotated, providing a rich resource for plant genomic studies. Because *Clitoria ternatea* and soybean belong to the same plant family, they share evolutionary and genetic similarities. These similarities allow researchers to identify orthologous genes between the two species. Orthologous genes are genes in different species that evolved from a common ancestral gene and generally retain similar biological functions. Bioinformatics tools such as BLAST, gene prediction algorithms, and functional annotation platforms allow researchers to analyze sequence data obtained from TSA and WGS datasets and compare them with known genes from soybean. This approach enables the identification of candidate genes involved in metabolic pathways, stress tolerance, nitrogen fixation, and secondary metabolite biosynthesis in *Clitoria ternatea*.

2. Importance of *Clitoria ternatea* and *Glycine max*

2.1 Economic significance of *Clitoria ternatea*

Clitoria ternatea is widely cultivated for several important uses. It is a rich source of natural blue pigment, primarily due to the presence of anthocyanins, which are commonly used as natural food colorants, primarily ternatins, which provide a stable, intense blue color. The plant is also valued for its medicinal compounds and has long been used in traditional medicine for its potential antioxidant, anti-inflammatory and cognitive health benefits. In agriculture, it plays an important role in soil improvement because it is a nitrogen-fixing legume that enhances soil fertility and supports sustainable farming systems. Additionally, ***Clitoria ternatea*** is grown for ornamental purposes due to its attractive blue flowers and is increasingly used in nutraceutical products for its health-promoting properties.

2.2 Importance of *Glycine max* as a model crop

Glycine max (soybean) is a major legume crop cultivated worldwide and is widely used as a genomic reference species. Its genome has been extensively sequenced and well annotated, making it a valuable resource for genetic and molecular studies. In addition, a large amount of transcriptomic and proteomic data is available for soybean, which helps researchers study gene expression and protein functions in detail. Many gene families related to stress tolerance, seed development, and metabolism have also been thoroughly characterized in this species. Because of these well-developed genomic resources, comparative studies using soybean allow scientists to identify orthologous genes in other related legume species and improve the understanding of gene functions and evolutionary relationships.

3. Bioinformatics Approaches for Gene Identification

Bioinformatics provides multiple computational approaches to discover important genes:

3.1 Comparative Genomics

Comparative genomics identifies conserved genes between species through sequence similarity and synteny analysis. Tools such as ortholog clustering and collinearity analysis help detect shared gene families and evolutionary relationships between legumes.

3.2 Sequence Alignment and Homology Search

Sequence alignment methods compare DNA or protein sequences from *C. ternatea* with known sequences in soybean databases. The widely used algorithm BLAST (Basic Local Alignment Search Tool) compares nucleotide or protein sequences against genomic databases to detect homologous genes and infer functional similarities between species.

3.3 Gene Prediction

Gene prediction algorithms identify coding regions within genomic DNA sequences. Programs such as FGENESH/GENSCAN/AUGUSTUS detect gene structures including exons, introns, and splice sites in genomic sequences using computational models.

3.4 Functional Annotation

After identifying gene sequences, functional annotation assigns biological roles to genes. Tools like Blast2GO combine BLAST similarity searches with Gene Ontology (GO) classification to provide automated functional annotation of genes and proteins.

4. Databases Used in Legume Bioinformatics

Several genomic and transcriptomic databases support gene discovery:

Database	**Purpose**
NCBI GenBank	Sequence repository
Phytozome	Plant comparative genomics
Gramene	A resource for plant comparative genomics
Ensembl Plants	Genome-scale information for plants
Legume Information System	Legume genomic resources
UniProt	Protein functional information
Gene Ontology (GO)	Functional classification

These databases store genome sequences, transcriptome datasets, and annotation data for crops including soybean.

5. Identification of Agro-Economically Important Genes

Using bioinformatics approaches, researchers can identify genes involved in key agricultural traits.

5.1 Stress Resistance Genes

Genes associated with tolerance to drought, salinity, and temperature stresses can be identified using a combination of bioinformatics and experimental approaches. One strategy involves homology searches with well-characterized stress-responsive genes from ***Glycine max*** (soybean), allowing the prediction of candidate genes in related species. Additionally, transcriptome analysis under specific stress conditions can reveal genes that are differentially expressed in response to environmental challenges. Together, these approaches help uncover key molecular players involved in stress adaptation, providing valuable targets for breeding or engineering stress-resilient plants.

5.2 Nitrogen Fixation Genes

As legumes, both species possess symbiotic nitrogen fixation pathways. Comparative analysis helps identify genes involved in nodulation and nitrogen metabolism.

5.3 Secondary Metabolite Biosynthesis

Clitoria ternatea produces valuable bioactive compounds, including anthocyanins and flavonoids, which have nutritional and medicinal significance. Through bioinformatics analyses, researchers can identify the key enzymes involved in flavonoid biosynthetic pathways, shedding light on the molecular steps that lead to the production of these compounds. Additionally, such analyses can reveal regulatory transcription factors that control the expression of these biosynthetic genes, providing insights into how metabolite production is regulated and offering potential targets for metabolic engineering to enhance the yield of these valuable phytochemicals.

5.4 Yield and Growth-Related Genes

Candidate genes involved in important biological processes such as seed development, plant growth, and nutrient transport can be predicted through comparative genomics approaches. By comparing genomic information with well-characterized gene families from *Glycine max* (soybean), researchers can identify homologous or orthologous genes in related species. This comparison helps in predicting the potential functions of genes and provides valuable insights into the molecular mechanisms that regulate key developmental and physiological processes in plants.

6. Bioinformatics Workflow for Gene Identification

A typical workflow includes:

1. Sequence Data Collection: Obtain genomic or transcriptomic sequences of *C. ternatea*
2. Quality Control and Assembly: Sequence trimming and assembly of contigs
3. Gene Prediction: Identify coding regions using gene prediction software
4. Homology Search: Compare sequences with soybean genes using BLAST
5. Functional Annotation: Assign Gene Ontology terms using Blast2GO
6. Phylogenetic Analysis: Construct evolutionary relationships among gene families
7. Marker Development: Identify SSRs and SNPs for breeding programs
8. Structural Analysis: Molecular Modeling and molecular docking for functional analysis

The identification of agro-economically important genes in *Clitoria ternatea* can be effectively performed using a bioinformatics-based comparative genomics approach. Public genomic databases and computational tools allow retrieval of sequences, similarity searches, gene structure prediction, and functional annotation. Sequence data for *Clitoria ternatea* are available in publicly accessible genomic resources, including the transcriptome shotgun assembly (TSA) dataset (GLLN00000000.1) and the whole genome shotgun (WGS) assembly (JAXHEB000000000.1) at NCBI. These datasets provide comprehensive transcriptomic and genomic sequence information suitable for gene identification and comparative analysis.

7. Applications in Crop Improvement

The identification of agro-economically important genes through bioinformatics has several important practical applications in crop improvement. It supports marker-assisted breeding by enabling the selection of desirable traits at the genetic level, thereby accelerating breeding programs. This approach also contributes to the development of stress-resistant varieties that can better tolerate environmental challenges such as drought, salinity, and pests. In addition, bioinformatics aids in genetic engineering strategies aimed at enhancing the production of valuable metabolites. It also plays a significant role in conservation and genetic diversity analysis by helping researchers understand variation within and between plant populations. Molecular markers such as SSRs have already been used to study genetic diversity in ***Clitoria ternatea***, highlighting the effectiveness of genomic tools in supporting breeding and improvement programs.

8. Challenges and Future Prospects

Despite these advantages, several challenges still remain. One major limitation is the scarcity of comprehensive genomic resources available for ***Clitoria ternatea***, which restricts detailed genetic and molecular studies. In addition, many genes in this species still lack complete functional annotation, making it difficult to fully understand their biological roles. Furthermore, there is a growing need for integrated multi-omics approaches, combining genomics, transcriptomics, proteomics, and metabolomics, to obtain a more comprehensive understanding of gene functions and regulatory networks.

9. Conclusion

Bioinformatics tools and computational techniques have become indispensable for identifying agriculturally important genes in plants. By using *Glycine max* as a model crop, researchers can employ comparative genomics to identify homologous genes in *Clitoria ternatea*. Tools such as BLAST, FGENESH, GENSCAN, AUGUSTUS and Blast2GO enable sequence alignment, gene prediction, and functional annotation, facilitating the discovery of genes responsible for stress tolerance, metabolic pathways, and yield improvement. Continued development of genomic resources and bioinformatics methodologies will further enhance crop improvement and sustainable agriculture.

10. References

- Abreu, M. L., Vieira, R. A., Rocha, N. S., Araujo, R. P., Glória, L. S., Fernandes, A. M., de Lacerda, P. D., & Júnior, A. G. (2014). *Clitoria ternatea* L. as a Potential High Quality Forage Legume. *Asian-Australasian journal of animal sciences*, *27*(2), 169–178. https://doi.org/10.5713/ajas.2013.13343
- Aljanabi S. (2001). Genomics and plant breeding. *Biotechnology annual review*, *7*, 195–238. https://doi.org/10.1016/s1387-2656(01)07038-7
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
- Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., & Yeh, L. S. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic acids research*, *32*(Database issue), D115–D119. https://doi.org/10.1093/nar/gkh131
- Bernal-Gallardo, J. J., & de Folter, S. (2024). Plant genome information facilitates plant functional genomics. Planta, 259(5), 117. https://doi.org/10.1007/s00425-024-04397-z
- Bolser, D. M., Staines, D. M., Perry, E., & Kersey, P. J. (2017). Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomic Data. *Methods in molecular biology (Clifton, N.J.)*, *1533*, 1–31. https://doi.org/10.1007/978-1-4939-6658-5_1
- Bolser, D., Staines, D. M., Pritchard, E., & Kersey, P. (2016). Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data. *Methods in molecular biology (Clifton, N.J.)*, *1374*, 115–140. https://doi.org/10.1007/978-1-4939-3167-5_6
- Campbell, J. D., Cameron, C. T., Cleary, A. M., Dash, S., Lavelle, E., Farmer, A., Huang, W., Novak, S., Prom, C., Weeks, N. T., Cannon, S. B., & Nelson, R. T. (2026). The SoyBase, LegumeInfo, and PeanutBase databases in support of legume research and crop improvement. *Genetics*, iyag020. Advance online publication. https://doi.org/10.1093/genetics/iyag020
- Conesa, A., & Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International journal of plant genomics*, *2008*, 619832. https://doi.org/10.1155/2008/619832
- Gonzales, M. D., Archuleta, E., Farmer, A., Gajendran, K., Grant, D., Shoemaker, R., Beavis, W. D., & Waugh, M. E. (2005). The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic acids research*, *33*(Database issue), D660–D665. https://doi.org/10.1093/nar/gki128

- Gonzales, M. D., Archuleta, E., Farmer, A., Gajendran, K., Grant, D., Shoemaker, R., Beavis, W. D., & Waugh, M. E. (2005). The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic acids research*, *33*(Database issue), D660–D665. https://doi.org/10.1093/nar/gki128

- Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, *40*(Database issue), D1178–D1186. https://doi.org/10.1093/nar/gkr944
- Hamilton, J. P., & Buell, C. R. (2012). Advances in plant genome sequencing. The Plant journal : for cell and molecular biology, 70(1), 177–190. https://doi.org/10.1111/j.1365-313X.2012.04894.x
- Hoff, K. J., & Stanke, M. (2019). Predicting Genes in Single Genomes with AUGUSTUS. *Current protocols in bioinformatics*, *65*(1), e57. https://doi.org/10.1002/cpbi.57
- Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R. J., Franklin, L. D., He, J., Xu, D., May, G., & Stacey, G. (2010). An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. The Plant journal : for cell and molecular biology, 63(1), 86–99. https://doi.org/10.1111/j.1365-313X.2010.04222.x

- Maia, N. M. A., Andressa, I., Cunha, J. S., Costa, N. A., de Oliveira, E. B., Leite Júnior, B. R. C., & Vieira, É. N. R. (2025). *Clitoria ternatea*: Perspectives on Its Application in Foods and Potential Health Benefits. *Plants (Basel, Switzerland)*, *14*(21), 3322. https://doi.org/10.3390/plants14213322
- Morrell, P. L., Buckler, E. S., & Ross-Ibarra, J. (2011). Crop genomics: advances and applications. Nature reviews. Genetics, 13(2), 85–96. https://doi.org/10.1038/nrg3097
- Oguis, G. K., Gilding, E. K., Jackson, M. A., & Craik, D. J. (2019). Butterfly Pea (*Clitoria ternatea*), a Cyclotide-Bearing Plant With Applications in Agriculture and Medicine. *Frontiers in plant science*, *10*, 645. https://doi.org/10.3389/fpls.2019.00645
- Olson, A., Kumari, S., Wei, X., Chougule, K., Lu, Z., Tello-Ruiz, M. K., Kumar, V., Van Buren, P., Olson, A., Kim, C., Braynen, J., Zhang, L., Dyer, S., Alvarez-Jarreta, J., Saraf, S., Contreras-Moreira, B., Naamati, G., Ernst, C., Papatheodorou, I., George, N., … Ware, D. (2026). Gramene 2025: expanded comparative genomics and pathway resources, integrated search, and pan-genome portals for crop research. *Nucleic acids research*, *54*(D1), D1720–D1732. https://doi.org/10.1093/nar/gkaf1260
- Paterson, A. H., Bowers, J. E., Burow, M. D., Draye, X., Elsik, C. G., Jiang, C. X., Katsar, C. S., Lan, T. H., Lin, Y. R., Ming, R., & Wright, R. J. (2000). Comparative genomics of plant chromosomes. *The Plant cell*, *12*(9), 1523–1540. https://doi.org/10.1105/tpc.12.9.1523
- Redsun, S., Hokin, S., Cameron, C. T., Cleary, A. M., Berendzen, J., Dash, S., Brown, A. V., Wilkey, A., Campbell, J. D., Huang, W., Kalberer, S. R., Weeks, N. T., Cannon, S. B., & Farmer, A. D. (2022). Doing Genetic and Genomic Biology Using the Legume Information System and Associated Resources. *Methods in molecular biology (Clifton, N.J.)*, *2443*, 81–100. https://doi.org/10.1007/978-1-0716-2067-0_4
- Sayers, E. W., Beck, J., Bolton, E. E., Brister, J. R., Chan, J., Comeau, D. C., Connor, R., DiCuccio, M., Farrell, C. M., Feldgarden, M., Fine, A. M., Funk, K., Hatcher, E., Hoeppner, M., Kane, M., Kannan, S., Katz, K. S., Kelly, C., Klimke, W., Kim, S., … Sherry, S. T. (2024). Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, *52*(D1), D33–D43. https://doi.org/10.1093/nar/gkad1044
- Sayers, E. W., Beck, J., Bolton, E. E., Brister, J. R., Chan, J., Connor, R., Feldgarden, M., Fine, A. M., Funk, K., Hoffman, J., Kannan, S., Kelly, C., Klimke, W., Kim, S., Lathrop, S., Marchler-Bauer, A., Murphy, T. D., O'Sullivan, C., Schmieder, E., Skripchenko, Y., … Pruitt, K. D. (2025). Database resources of the National Center for Biotechnology Information in 2025. *Nucleic acids research*, *53*(D1), D20–D29. https://doi.org/10.1093/nar/gkae979
- Siadjeu, C., & Pucker, B. (2023). Medicinal plant genomics. *BMC genomics*, *24*(1), 429. https://doi.org/10.1186/s12864-023-09542-8
- Sreekala, S., & Muraleedharan, U. D. (2021). Cationic Clitoria ternatea Seed Peptide as a Potential Novel Bioactive Molecule. *Protein and peptide letters*, *28*(11), 1259–1271. https://doi.org/10.2174/0929866528666210922124735
- UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic acids research*, *43*(Database issue), D204–D212. https://doi.org/10.1093/nar/gku989
- UniProt Consortium (2025). UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic acids research*, *53*(D1), D609–D617. https://doi.org/10.1093/nar/gkae1010
- Wang, H., Chen, M., Wei, X., Xia, R., Pei, D., Huang, X., & Han, B. (2024). Computational tools for plant genomics and breeding. *Science China. Life sciences*, *67*(8), 1579–1590. https://doi.org/10.1007/s11427-024-2578-6
- Wang, J., Chen, Y., & Zou, Q. (2023). Comparative Genomics and Functional Genomics Analysis in Plants. *International journal of molecular sciences*, *24*(7), 6539. https://doi.org/10.3390/ijms24076539
- Yao, H., Guo, L., Fu, Y., Borsuk, L. A., Wen, T. J., Skibbe, D. S., Cui, X., Scheffler, B. E., Cao, J., Emrich, S. J., Ashlock, D. A., & Schnable, P. S. (2005). Evaluation of five ab initio gene prediction programs for the discovery of maize genes. *Plant molecular biology*, *57*(3), 445–460. https://doi.org/10.1007/s11103-005-0271-1
- Yi, X., Du, Z., & Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic acids research, 41(Web Server issue), W98–W103. https://doi.org/10.1093/nar/gkt281