Nepal J Biotechnol. 2025 Jul; 13 (1):16-25 Research Article DOI: https://doi.org/10.54796/njb.v13i1.380

Nepal Journal of Biotechnology
Publisher: Biotechnology Society of Nepal
Journal Homepage: https://nepjb.com/index.php/NJB
ISSN (Online): 2467-9313
ISSN (Print): 2091-1130

Computational Design and Bioinformatic Insights into Recombinant Whey Protein Production

Aathish Gurunathan, Harini Subalakshmi M
Department of Industrial Biotechnology, Government College of Technology, Coimbatore, Tamil Nadu, India

Received: 10 Feb 2025; Revised: 14 Jul 2025; Accepted: 21 Jul 2025; Published online: 31 Jul 2025

Abstract

The rising global demand for sustainable protein sources highlights the need for alternatives to traditional whey protein production, which is resource-intensive and environmentally taxing. This study presents an *in-silico* approach for producing recombinant whey proteins—α-lactalbumin and β-lactoglobulin—using *Saccharomyces cerevisiae* as a host. Codon-optimized gene constructs were designed and validated through computational tools, showing enhanced translational efficiency, structural stability, and functional potential. Allergenicity analysis revealed moderate risk, suggesting a need for further refinement. Overall, the findings demonstrate the feasibility of precision fermentation for scalable, eco-friendly whey protein production, with broad applications in functional foods and biotechnology.

Keywords: Whey protein, *In silico* method, Recombinant plasmid construction, *Saccharomyces cerevisiae*, α-lactalbumin, β-lactoglobulin.

Corresponding author, email**:** aathishkg@gmail.com

Introduction

Whey is a soluble milk protein that constitutes approximately 20% of milk content. It comprises essential proteins such as β-lactoglobulin (β-LG), α-lactalbumin (α-LA), bovine serum albumin (BSA), immunoglobulins, and proteose peptones, along with minor components like lactoferrin, lactollin, glycoproteins, lactoperoxidase, and transferrin (1,2). It is particularly rich in branched-chain amino acids (leucine, isoleucine, and valine) and sulfur-containing amino acids (cysteine and methionine), making it a highly reliable source of biologically active proteins and essential nutrients (3). Traditionally, whey protein is produced as a byproduct of cheese and casein manufacturing. During milk coagulation, casein curds separate, leaving behind a liquid known as whey (4). While this process transforms dairy waste into a valuable product, it remains energy-intensive, requiring around 1,250 to 1,660 litres of whey water to yield just 1 kg of whey protein. Additionally, the presence of lactose makes it unsuitable for lactose-intolerant individuals. Globally, an estimated 200 million tons of whey water is produced annually, yet only 50% undergoes further processing, with the remainder often being discarded as waste, contributing to significant environmental concerns (5).

With the global population projected to reach 9.7 billion by 2050, the demand for protein is expected to rise by 23%, increasing total global requirements to approximately 280 million tons annually (6). However, conventional food sources alone may not be sufficient to meet this growing demand, as climate change, pollution, resource depletion, and unsustainable agricultural practices threaten global food security (7). The livestock industry, responsible for nearly half of agriculture-related greenhouse gas emissions, further highlights the urgency of finding alternative protein sources (8). The rise of vegan, vegetarian, and flexitarian diets has intensified interest in plant-based protein alternatives, which also cater to individuals allergic to dairy proteins (9). However, challenges persist regarding the sustainability and nutritional adequacy of plant-based proteins, raising debates on their effectiveness as complete protein substitutes. Given these constraints, precision fermentation has emerged as a highly promising solution for alternative protein production. This technique utilizes genetically modified microorganisms or cell cultures to produce recombinant whey proteins such as α-lactalbumin and β-lactoglobulin. Precision fermentation offers several advantages over conventional protein sources. It reduces the need for farmland and dairy farming, making it more sustainable. It is suitable for vegan diets, provides high purity and consistent quality, and allows for better nutrition by improving amino acid content. It also supports large-scale production with a lower environmental impact.

Likewise, the growing demand for high-protein diets, particularly in sports nutrition, has significantly contributed to the expansion of the whey protein market

(10). In 2023, the global whey protein market was valued at USD 11.79 billion and is expected to grow to USD 22.63 billion by 2032, reflecting a CAGR of 7.5%. The primary factors driving this growth include increased consumer awareness of protein-rich diets and the rising popularity of sports and fitness supplements. Beyond its nutritional benefits, whey protein is also recognized for its therapeutic applications. Studies suggest that whey protein plays a crucial role in antioxidant defence, reducing oxidative stress, and exhibiting potential anti-cancer properties. Additionally, it has been linked to oral health benefits and potential protection against HIV infection, highlighting its importance beyond dietary supplementation (11).

In light of these growing applications, this paper explores an in-silico approach to recombinant whey protein production using *Saccharomyces cerevisiae* as a host organism. Computational techniques were employed to design and validate a recombinant plasmid capable of expressing key whey protein constituents—α-lactalbumin and β-lactoglobulin. The study includes construction, validation, translation efficiency analysis, structural predictions, and allergenicity assessments to optimize the recombinant protein production process. As conventional whey protein production remains resource-intensive and environmentally challenging, advancements in precision fermentation and computational biology offer a promising avenue for sustainable and scalable whey protein synthesis. By leveraging genetic engineering and in-silico modeling, recombinant whey protein production could revolutionize the protein industry, providing a high-quality, sustainable alternative to traditional animal-derived proteins.

Methods

Tools and Components Used in Plasmid Construction

This work utilized several computational tools for plasmid construction, including NCBI BLAST, NCBI GenBank, ORF Finder, Proteus2, Protein-Sol, UNAfold web server, AllerTOP and Vector Builder. These tools were employed at various stages of the plasmid development process. For the plasmid backbone selection, an integrating plasmid based on *Saccharomyces cerevisiae* was chosen due to its ability to facilitate stable genomic integration, removing the need for selective pressure for plasmid maintenance. The genes of interest, *LALBA* (α-lactalbumin) and *BLG* (β-lactoglobulin), were selected from *Bos taurus* (domestic cow) due to their significance in whey protein synthesis (12). To enhance expression efficiency in yeast, codon optimization was performed to match the preferred codon usage of *S. cerevisiae*, thus improving translation efficiency and protein yield (13). Additionally, a flexible linker sequence was incorporated between the *LALBA* and *BLG* genes to ensure proper structural stability and functional optimization of the expressed proteins in the yeast system. The plasmid construction process was primarily carried out using the Vector Builder online tool. For clarity and reader convenience, the gene α-lactalbumin is hereafter referred to as LALBA, and β-lactoglobulin as BLG throughout this paper.

Plasmid construction and its components

The recombinant plasmid construction starts by retrieving the required genes of interest from the NCBI GenBank with accession numbers BLG (NM_173929.3) and LALBA (NM_174378.2). Domestic cattle (*Bos taurus*) were chosen as the source for these genes, as their milk contains BLG and LALBA, which are the main constituents of whey protein (12).

A recombinant expression vector was designed using the Vector Builder platform. The vector system was selected for recombinant protein production, specifically optimized for *Saccharomyces cerevisiae*. The gene of interest, *LALBA-BLG* fused with a P2A linker, was inserted into the open reading frame (ORF) region. Codon optimization was performed to enhance expression efficiency in *S. cerevisiae*. A start codon (ATG) was included at the beginning of the gene sequence to ensure proper translation initiation, while a stop codon was placed at the end to terminate translation efficiently(14). The GPD promoter was chosen to drive gene expression, while the CYC1 terminator was used for proper transcription termination(15). Additionally, the *URA3* gene was incorporated as a selection marker(16). A schematic representation of the constructed vector is provided in **Figure 1**.

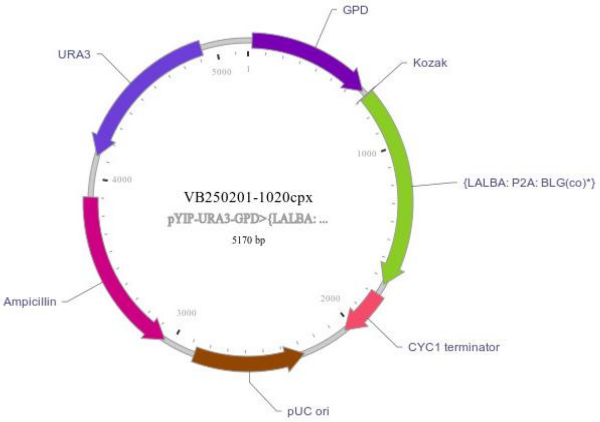


Figure 1. Constructed recombinant vector (plasmid map)

Table 1. Various components of the vector

Component	Position	Size (bp)	Function & Description
GPD Promoter	22-671	650	A strong constitutive promoter from *S. cerevisiae* that drives continuous gene expression.
Kozak Sequence	696-701	6	Enhances translation initiation by facilitating ribosome recognition of the start codon.
Gene of Interest (*LALBA-P2A-BLG*)	702-1730	1029	Codon-optimized fusion of *LALBA* and *BLG* genes linked by P2A for efficient expression in *S. cerevisiae*.
CYC1 Terminator	1789-2042	254	Ensures proper transcription termination and polyadenylation to stabilize mRNA.
pUC Origin of Replication	Complement (2285-2873)	589	Enables high-copy-number replication in *E. coli* for plasmid amplification.
Ampicillin Resistance Gene	Complement (3044-3904)	861	Provides resistance to ampicillin, allowing selection of transformed *E. coli* cells.
URA3 Selection Marker	Complement (4117-4920)	804	Enables yeast cells lacking endogenous URA3 to grow in uracil-deficient media.

The recombinant plasmid was designed with multiple essential components to facilitate gene expression in *Saccharomyces cerevisiae*. The construct includes the *GPD* promoter (22-671 bp), a strong constitutive promoter driving continuous gene expression, followed by a Kozak sequence (696-701 bp) to enhance ribosome recognition and translation initiation. The gene of interest, *LALBA-P2A-BLG* (702-1730 bp), was codon-optimized for *S. cerevisiae* to ensure efficient protein production. The *CYC1* terminator (1789-2042 bp) was incorporated for proper transcription termination and polyadenylation (17). For plasmid replication in *E. coli*, a *pUC* origin (complement 2285-2873 bp) was included to enable high-copy-number amplification. Additionally, an ampicillin resistance gene (complement 3044-3904 bp) was introduced for bacterial selection. In yeast, the *URA3* marker (complement 4117-4920 bp) allows growth in uracil-deficient media, facilitating successful transformation and selection (16). The details described in the above paragraph are briefly summarized in the **Table 1** for clarity and ease of reference.

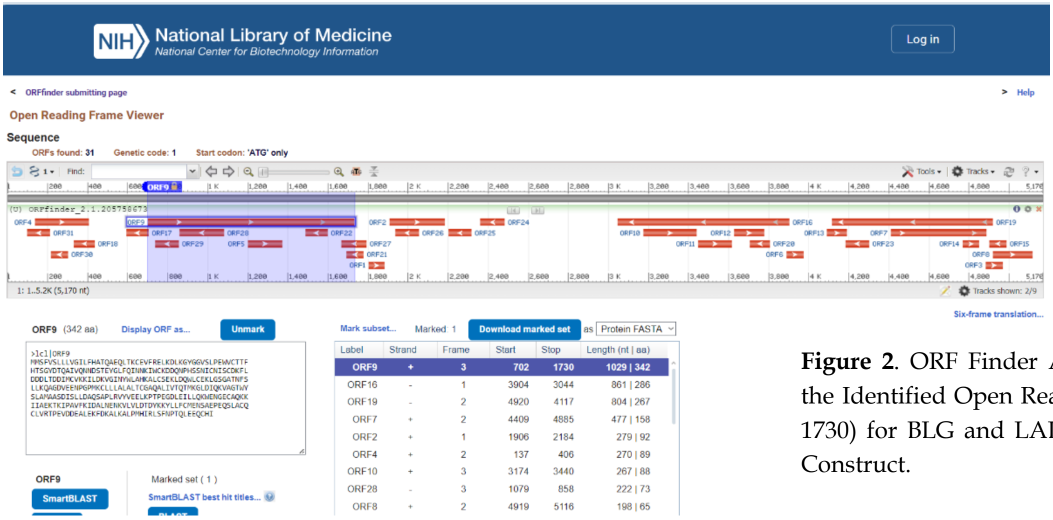
Validation of the construct

To validate the designed recombinant construct, several bioinformatics tools were employed. ORF Finder was used to identify open reading frames within the sequence, followed by a BLAST analysis within ORF Finder to confirm sequence similarity and ensure the correct insertion of the gene of interest. Structural and functional predictions were performed using Proteus2 and ProteinPredict, where Proteus2 was used for secondary structure and disorder prediction, while ProteinPredict provided insights into the amino acid composition of the expressed protein (18). Protein solubility was assessed using Protein-Sol to evaluate the likelihood of proper folding and expression (19). The UNAfold webserver was used to analyse mRNA secondary structures and potential folding patterns, which may influence translation efficiency (20). Additionally, allergenicity predictions were conducted using AllerTOP to assess the potential immunogenicity of the recombinant protein (21). These analyses provided a comprehensive evaluation of the designed expression construct, ensuring its suitability for recombinant protein production.

Results and Discussion

Codon Optimization for Efficient Expression in *Saccharomyces cerevisiae*

The gene sequence was optimized for *Saccharomyces cerevisiae* using the in-built codon optimization tool of Vector Builder. Before optimization, the sequence had a GC content of 44.06% and a Codon Adaptation Index (CAI) of 0.67. After optimization, the GC content was reduced to 34.97%, while the CAI increased to 0.86. This reduction in GC content suggests a decrease in the stability of the mRNA, which may reduce secondary structure formation, facilitating more efficient translation. A study by *Kudla et al.* demonstrated that increased GC content in mRNA hairpin stems near the translation start site dramatically slows translation (22). Similarly, the increase in CAI indicates a stronger alignment with the preferred codon usage of *S. cerevisiae*, likely resulting in improved translation efficiency and higher protein yield. These changes enhance the likelihood of successful protein expression in *S. cerevisiae* by optimizing the sequence for better translational performance.



```
>lcl|ORF9
MMSFVSLLLVGILFHATQAEQLTKCEVFRELKDLKGYGGVSLPEWVCTTF
HTSGYDTQAIVQNNDSTEYGLFQINNKIWCKDDQNPHSSNICNISCDKFL
DDDLTDDIMCVKKILDKVGINYWLAHKALCSEKLDQWLCEKLGSGATNFS
LLKQAGDVEENPGPMKCLLLALALTCGAQALIVTQTMKGLDIQKVAGTWY
SLAMAASDISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWENGECAQKK
IIAEKTKIPAVFKIDALNENKVLVLDTDYKKYLLFCMENSAEPEQSLACQ
CLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI
```

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF9	+	3	702	1730	1029 | 342
ORF16	-	1	3904	3044	861 | 286
ORF19	-	2	4920	4117	804 | 267
ORF7	+	2	4409	4885	477 | 158
ORF2	+	1	1906	2184	279 | 92
ORF4	+	2	137	406	270 | 89
ORF10	+	3	3174	3440	267 | 88
ORF28	-	3	1079	858	222 | 73
ORF8	+	2	4919	5116	198 | 65

Figure 2. ORF Finder Analysis Showing the Identified Open Reading Frame (702-1730) for BLG and LALBA Genes in the Construct.

Gene Presence Confirmation and Sequence Validation Using ORF Finder

After the successful construction of the plasmid, the presence of the gene of interest was confirmed using ORF Finder, which identified the open reading frame (ORF) spanning positions 702 to 1730 in the entire sequence. To further validate the correctness of the sequence, a BLAST search was performed on the identified ORF. The BLAST results revealed that the *BLG* gene showed a 100% sequence identity with *Bos taurus*, with a query coverage of 52%. Similarly, the *LALBA* gene also displayed a 100% sequence identity with *Bos taurus*, with a query coverage of 42%. These results confirm that both genes were correctly inserted into the plasmid, ensuring that the *BLG* and *LALBA* genes from *Bos taurus* are accurately represented in the recombinant plasmid. The high percentage identity and moderate query coverage further suggest the sequence integrity of the cloned genes, supporting their suitability for further expression studies. **Figure 2** shows the results page of the NCBI ORF Finder which displays the identified open reading frames (ORFs).

Structural and Functional Analysis of the Recombinant Protein Secondary Structure Prediction Using Proteus2

The structural prediction of the protein using Proteus2 reveals a balanced composition of secondary structural elements, with a mix of α-helices (26%, 89 residues) and β-sheets (24%, 81 residues) (18). Approximately 50% of the protein is predicted to be flexible or disordered coil regions (172 residues), indicating regions with less defined structure that could contribute to the protein's functional versatility. The absence of signal peptides and membrane-related regions suggests that the protein is likely soluble and non-secretory, rather than being a transmembrane protein or one intended for extracellular secretion. This prediction supports the hypothesis that the protein is stable in aqueous environments and could function in the cytoplasm or other soluble compartments. The protein sequence consists of 342 residues, and no transmembrane helical region or signal peptide was found, suggesting it is not intended for membrane association or secretion. The analysis identified 13 homologs with matching structures in the Protein Data Bank (PDB), with high e-values indicating strong sequence homology, particularly with 1BEBA, 1YUPA, and 1F6RA, which have e-values as low as 5.0E-86, signifying a very strong structural similarity to these known proteins. The sequence alignment for ab-initio predictions involved 49 sequences, providing a solid basis for the structural model, with an overall confidence of 77.3%, indicating a relatively high level of certainty in the results. Figure 3 presents the graphical alignment of soluble PDB homologs. For the reader's convenience, the results and key findings generated by Proteus2 have been summarized in **Table 2**.

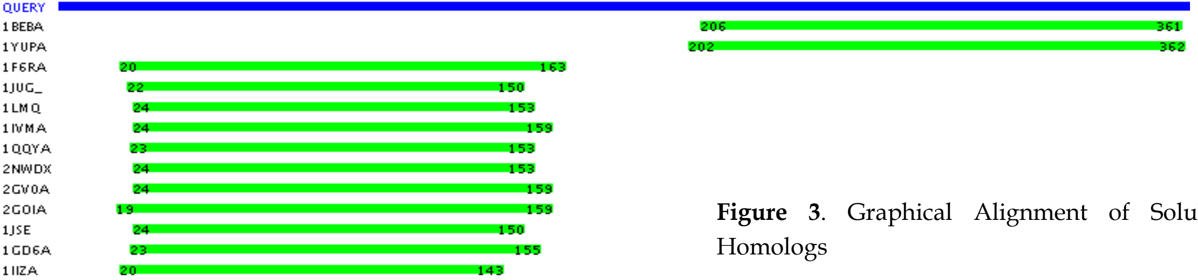


Figure 3. Graphical Alignment of Soluble PDB Homologs

Table 2. Summary of the key findings from the Proteus2 prediction results (18)

Parameter	Result	Interpretation
Number of residues read	342	The protein sequence consists of 342 amino acids.
Transmembrane helices	None	No transmembrane regions, suggesting the protein is not membrane-associated.
Signal peptide	None	Absence of a signal peptide indicates the protein is likely not secretory.
Number of usable PDB homologs	13	Strong structural homology with 13 proteins in the PDB database.
Top 3 PDB homologs (e-values)	1BEBA (e-value = 5.0E-86) 1YUPA (e-value = 1.0E-85) 1F6RA (e-value = 6.0E-73)	High e-values indicate strong homology with these known protein structures.
Number of sequence alignments used for ab-initio predictions	49	49 sequence alignments were used for ab-initio predictions, ensuring a reliable model.
Overall confidence value	77.3%	The overall confidence in the predicted structure is high (77.3%).
Predicted % Helix content	26% (89 residues)	26% of the protein consists of α-helices, important for structural stability and function.
Predicted % Beta sheet content	24% (81 residues)	24% of the protein consists of β-sheets, contributing to the protein's structural framework.
Predicted % Coil content	50% (172 residues)	50% of the protein is in disordered or flexible coil regions, suggesting functional versatility.
Predicted % Signal peptide content	0% (0 residues)	No signal peptide predicted, reinforcing the idea that the protein is not secretory.
Predicted % membrane content	0% (0 residues)	No membrane-associated regions predicted, supporting the hypothesis of a soluble protein.

Physicochemical Property Assessment Using ProtParam

The physicochemical properties of the protein indicate a molecular weight of 38.28 kDa and a theoretical isoelectric point (pI) of 4.85, suggesting that the protein is acidic. The amino acid composition analysis shows that Leucine (13.5%) and Lysine (8.5%) are present in high amounts, potentially influencing the protein's structural integrity and functional interactions. Additionally, the total number of negatively charged residues (Asp + Glu) is 49, while the positively charged residues (Arg + Lys) total 33, implying that the protein carries a net negative charge at physiological pH. The atomic composition reveals that the protein consists of carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and sulfur (S), forming a molecular formula of $C_{1703}H_{2702}N_{434}O_{518}S_{23}$. The extinction coefficient at 280 nm is 45,795 $M^{-1}cm^{-1}$ when assuming all cysteine residues form disulfide bonds, whereas it slightly decreases to 44,920 $M^{-1}cm^{-1}$ when assuming all cysteines remain in the reduced state. This high absorbance value at 280 nm suggests that the protein contains a considerable number of aromatic residues (Trp, Tyr), making it well-suited for quantification via UV spectrophotometry (23).

The instability index (II) is 32.97, classifying the protein as stable. A high aliphatic index (96.99) suggests that the protein is likely to exhibit high thermal stability. The GRAVY (Grand Average of Hydropathicity) score is -0.115, indicating that the protein is slightly hydrophilic, which implies that it is soluble in aqueous environments (24). The protein's estimated half-life—approximately 30 hours in mammalian reticulocytes (in vitro), over 20 hours in yeast (in vivo), and more than 10 hours in *E. coli* (in vivo)—indicates excellent inherent stability across diverse biological systems (25). The amino acid composition present in the protein is tabulated as **Table 3**.

Table 3. Various composition of amino acid present in the protein

Amino Acid	Percentage (%)
Alanine (A)	7.6
Arginine (R)	1.2
Asparagine (N)	4.4
Aspartic acid (D)	7.0
Cysteine (C)	4.4
Glutamine (Q)	5.3
Glutamic acid (E)	7.3
Glycine (G)	4.7
Histidine (H)	1.8
Isoleucine (I)	5.6
Leucine (L)	13.5
Lysine (K)	8.5
Methionine (M)	2.3
Phenylalanine (F)	3.2
Proline (P)	3.5
Serine (S)	5.3
Threonine (T)	5.3
Tryptophan (W)	1.8
Tyrosine (Y)	2.3
Valine (V)	5.3

Protein Solubility Prediction Using Protein-Sol

The predicted scaled solubility of the protein is 0.522, which suggests moderate solubility in aqueous environments. A value above 0.45 is generally considered soluble, indicating that this protein is likely to remain in solution under physiological conditions rather than forming aggregates. The theoretical isoelectric point (pI) is 4.84, which is consistent with previous ProtParam results (26). This means that the protein will carry a net negative charge at physiological pH (7.4), making it more likely to interact with positively charged molecules or environments. Overall, the moderate solubility prediction suggests that this protein can be expressed and purified efficiently in an aqueous system without significant aggregation issues, making it a good candidate for biochemical and structural studies (19).

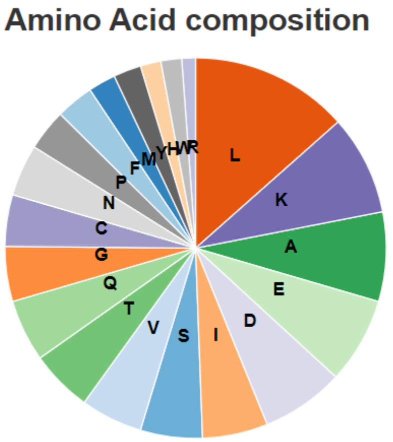


Figure 4. Composition of the amino acids present in the protein.

The windowed charge score analysis of the protein sequence shown in the **Figure 5** reveals a pattern of alternating charged regions, with distinct clusters of positively and negatively charged amino acids. The majority of the sequence exhibits negative charge scores, particularly in the middle and towards the end, indicating regions rich in negatively charged residues. In contrast, positively charged clusters appear at specific positions (~100, ~160, ~240, and ~320), suggesting localized regions of electrostatic potential. This charge distribution may play a crucial role in the protein's structural stability, interaction with other biomolecules, or functional properties such as binding affinity.

The windowed fold propensity analysis shown in the *Figure 6* highlights variations in the structural tendencies of the protein along its sequence. Several regions, particularly between residues 150-200, exhibit high fold propensity, indicating the likelihood of forming stable secondary structures such as α-helices or β-sheets. Conversely, certain segments, notably around residues 40-80 and 230-250, display negative scores, suggesting flexible or disordered regions. These variations in structural propensity could influence the protein's overall stability, functional flexibility, and interaction dynamics, potentially affecting its role in biological processes.

The provided bar graph from *Figure 7* illustrates deviations from the population average for various parameters, including amino acid composition and biochemical properties, with yellow bars representing positive deviations and green bars indicating negative deviations. Notably, high deviations for Alanine (A), Isoleucine (I), Lysine (K), and Glutamine (Q) suggest an enrichment of these residues, while biochemical properties such as Kanamycin Resistance (KmR), Polymyxin Resistance (PmN), and membrane-related features also show positive deviations. Conversely, Glycine (G), Histidine (H), Arginine (R), and Serine (S) are underrepresented, alongside biochemical properties

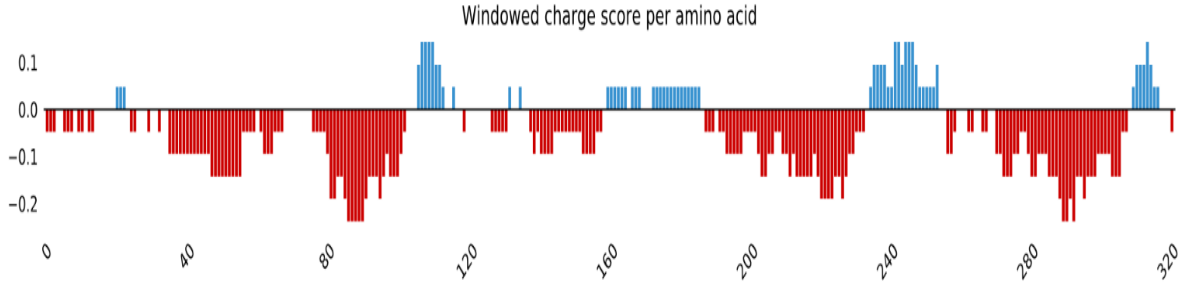


Figure 5. Windowed charge score analysis

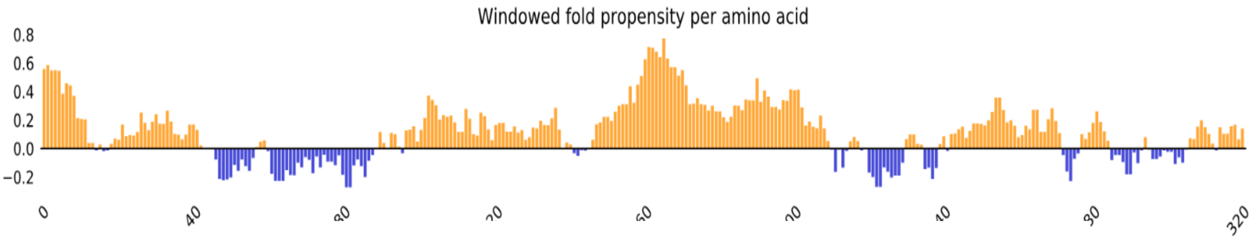


Figure 6. Variations in the structural tendencies of the protein

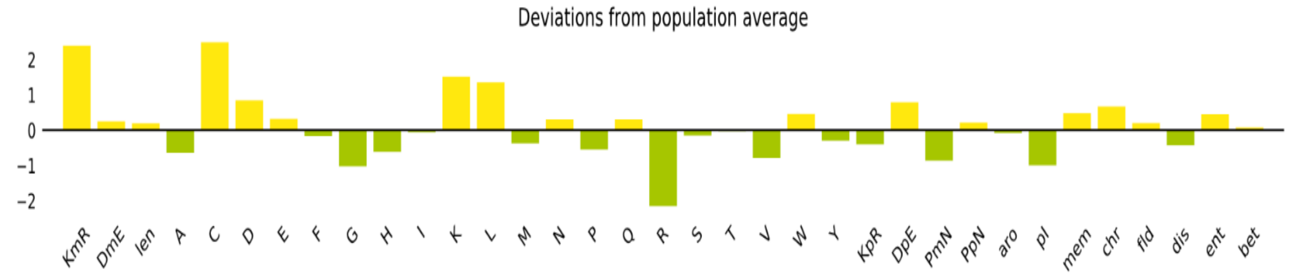


Figure 7. Deviations from the population average for various parameters

like Isoelectric Point (pI), aromatic content (aro), and fold-related properties (fld), which exhibit negative deviations. This analysis indicates that the protein has a distinct amino acid composition, favouring hydrophobic residues such as Isoleucine and Alanine while being deficient in polar residues like Glycine and Arginine. The observed deviations in biochemical properties, particularly resistance markers and membrane association, suggest a potential functional role related to structural stability or antimicrobial resistance mechanisms.

mRNA Structure Prediction Using UNAfold Web Server

The dot plot (*Figure 8*) is generated using UNAfold web server(20). *Figure 8* illustrates the predicted secondary structure of the WHEY protein at 37°C, providing insights into its folding stability and interaction potential. The plot displays base-pair interactions, with the X and Y axes (ranging from 1 to 1029) representing sequence positions. The lower triangular region depicts the most stable folding conformation, with an optimal free energy of -257.1 kcal/mol, indicating a highly stable structure. The upper triangular region showcases suboptimal structures with slightly higher energy values, represented by a colour gradient ranging from -257.1 kcal/mol (most stable) to -246.8 kcal/mol (least stable). Additionally, the free energy (ΔG) of 12.0 kcal/mol suggests the energy required to unfold the structure. These findings align with previous studies on protein stability, emphasizing the thermodynamic robustness of WHEY protein. The presence of alternative folding patterns further supports its structural flexibility, reinforcing its functional significance in biological systems (27).

Functional and Structural Insights from PredictProtein

The results from the PredictProtein online tool provide insights into the structural characteristics of the analysed protein (28). The solvent accessibility prediction illustrated in the *Figure 9* indicates that 46% of residues are exposed, 43% are buried, and 11% fall into the intermediate category, suggesting a well-balanced distribution between surface-exposed and core residues. Additionally, the secondary structure prediction depicted in the *Figure 10* reveals that the protein is primarily composed of 62% random or undefined structures (others), 21% alpha-helices, and 17% beta-strands. This structural composition suggests a predominantly disordered or flexible conformation, with a moderate presence of stable helices and strands contributing to the protein's overall fold.

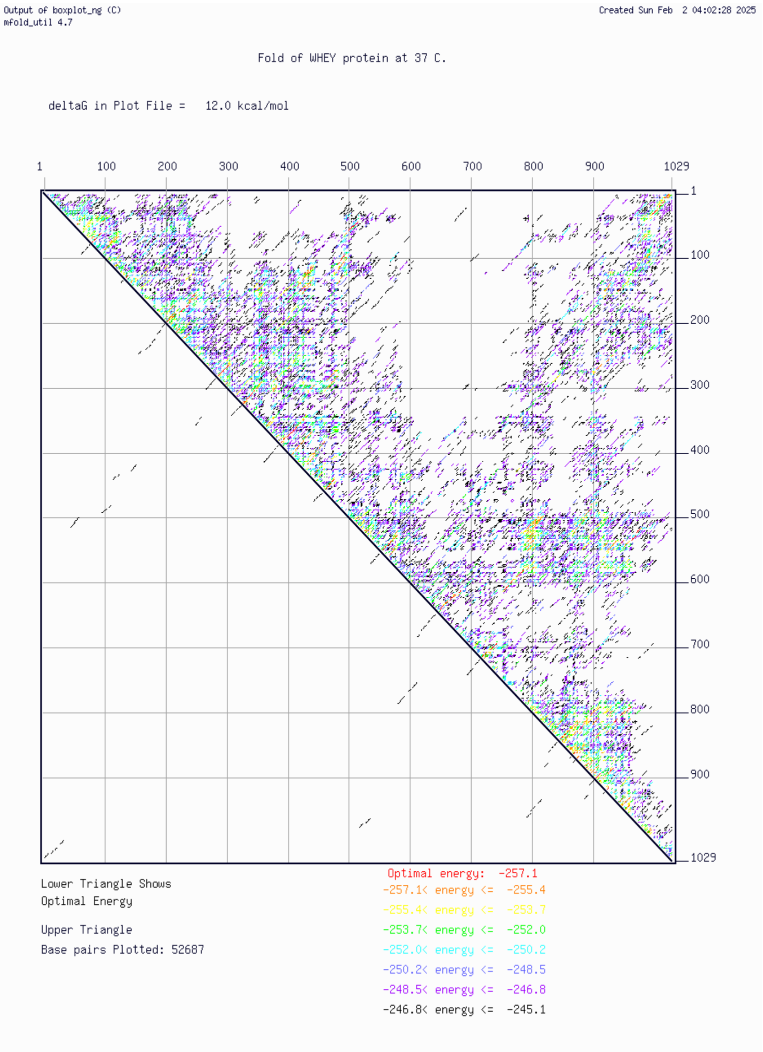


Figure 8. Dot plot generated using UNAfold web server(20)

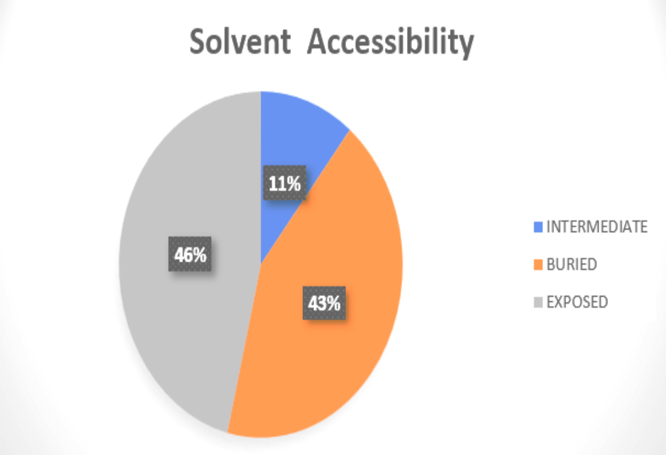


Figure 9. solvent accessibility (predict protein tool)

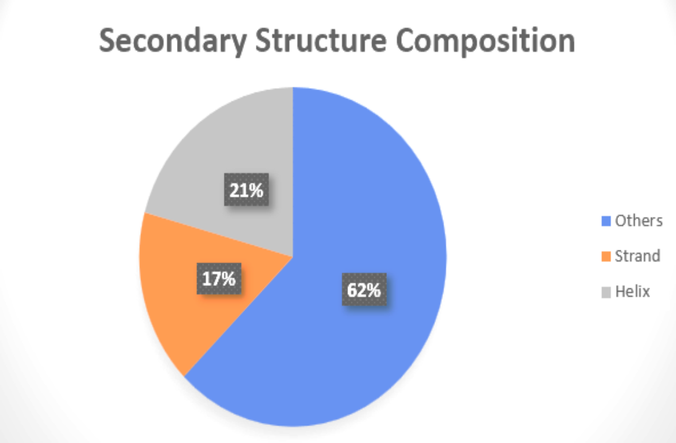


Figure 10. Secondary structure composition

Homology Modeling Using SWISS-MODEL

The homology modeling results from SWISS-MODEL provide structural insights based on sequence similarity with known templates, generating three models with different confidence levels. Model 1, built using the **A0A836D465.1.A** template, belongs to the Lipocalin/cytosolic fatty-acid binding domain-containing protein and has a GMQE score of 0.48 with a sequence identity of 95.51%, indicating reasonable structural reliability *Figure 11* shows the model 1 built using Swiss-model. Model 2, using the **6nkq.1.A** template for BLG, is a monomer with a GMQE of 0.44 and a QMEANDisCo Global score of 0.86 ± 0.07, showing high confidence is depicted in the *Figure 12*. With a 100.00% sequence identity, this model closely mirrors its template. Model 3, based on the **2g4n.4.A** template of LALBA, also exists as a monomer, with a GMQE of 0.27 and a QMEANDisCo Global score of 0.78 ± 0.08 is depicted in the *Figure 13*. Although its GMQE is lower than the other models, it maintains a 100.00% sequence identity, supporting its structural accuracy. The structural comparisons suggest that the protein shares functional similarities with lipocalins, BLG, and LALBA families, which could influence its binding properties and stability. The structural overlays reveal distinct differences in folding patterns and secondary structure alignments among the models, highlighting potential functional implications.(24,28–32)

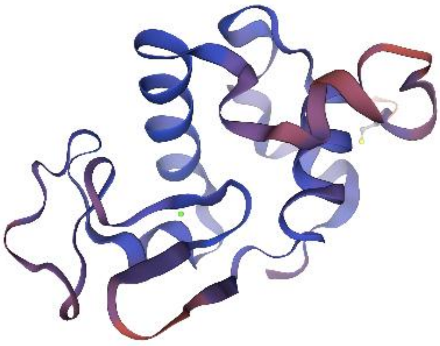


Figure 11. Model 1 from Swiss model

Figure 12. Model 2 from Swiss model

Figure 2. Model 3 from Swiss model

Allergenicity Assessment Using AllerTOP

The AllerTOP tool analysis identified the LALBA precursor protein and BLG as the most similar proteins, suggesting a structural and functional resemblance to known allergens. Based on this similarity, the tool classified the analysed protein as a probable allergen, indicating a potential allergenic response in sensitive individuals. This classification suggests that the protein may share conserved epitopes with common milk allergens, which could have implications for food safety and immunogenicity studies (21).

While the studied protein exhibits potential allergenicity, its numerous functional advantages, including high purity and safety, animal-free production, consistent quality across batches, and high yields can be achieved in a cost-effective and scalable manner, highlight its significance in various biotechnological applications. Despite its allergenic potential, further research could explore strategies to mitigate this limitation, such as molecular modifications or controlled expression systems. These findings not only underscore the protein's broad utility but also open new avenues for future investigations aimed at optimizing its safe and effective use in food safety and immunogenicity studies.

Conclusions and future directions

This study successfully demonstrates the application of computational design and bioinformatic tools in constructing a recombinant plasmid for expressing whey proteins LALBA and BLG in *Saccharomyces cerevisiae*. One of the most notable outcomes is the significant improvement in translational efficiency post-codon optimization, evidenced by an increased Codon Adaptation Index (CAI) and reduced GC content. These enhancements are expected to promote efficient protein expression in yeast by minimizing unfavorable mRNA secondary structures.

Structural and functional predictions further confirmed the designed protein's suitability for recombinant production. The secondary structure analysis revealed a balanced presence of α-helices and β-sheets, along with a substantial proportion of flexible coil regions, indicating the potential for structural adaptability. Physicochemical assessments, including thermal and solubility properties, support the protein's robust behavior in diverse biological environments. Furthermore, the protein demonstrated stability across multiple host systems, with estimated half-lives exceeding 10 hours in E. coli, 20 hours in yeast, and 30 hours in mammalian cells.

A critical finding was the predicted allergenicity of the recombinant protein, which aligns with known milk allergen characteristics. Although this presents a limitation, it also offers a foundation for future protein engineering strategies aimed at epitope modification or suppression. Addressing this issue will be essential for ensuring consumer safety, especially if the protein is intended for direct food applications.

The implications of this study extend to the broader context of sustainable food technology. By leveraging precision fermentation and in-silico modeling, it is possible to bypass resource-intensive dairy production while maintaining functional and nutritional qualities of whey proteins. This approach aligns with global efforts toward carbon-neutral protein sources and holds promise for scaling up recombinant protein production in food, nutraceutical, and therapeutic sectors.

Future research should focus on experimental validation of the designed construct in *S. cerevisiae*, including expression trials, purification, and functional assays. Additionally, immunogenicity testing and epitope engineering should be pursued to mitigate allergenic responses. Incorporating post-translational modification prediction and glycosylation pattern modeling could further refine the construct for real-world application.

Funding

No funding was received to support this research work or the preparation of this manuscript.

Conflicts of interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Data availability

Not applicable for this research.

References

1. Gangurde H, Patil P, Chordiya M, Baste N. Whey protein. Scholars' Research Journal. 2011;1(2):69.
2. Kilara A, Panyam D. Peptides From Milk Proteins and Their Properties. Crit Rev Food Sci Nutr. 2003 Oct;43(6):607–33.
3. Mir Khan U, Selamoglu Z. Nutritional and Medical Perspectives of Whey Protein: A Historical Overview. Journal of Pharmaceutical Care. 2020 Feb 18;
4. Smithers GW. Whey and whey proteins—From 'gutter-to-gold.' Int Dairy J. 2008 Jul;18(7):695–704.
5. Rotz CA, Montes F, Chianese DS. The carbon footprint of dairy production systems through partial life cycle assessment. J Dairy Sci. 2010 Mar;93(3):1266–82.
6. Lagrange V, Whitsett D, Burris C. Global Market for Dairy Proteins. J Food Sci. 2015 Mar 10;80(S1).
7. Crippa M, Solazzo E, Guizzardi D, Monforti-Ferrario F, Tubiello FN, Leip A. Food systems are responsible for a third of global anthropogenic GHG emissions. Nat Food. 2021 Mar 8;2(3):198–209.
8. Wood P, Tavan M. A review of the alternative protein industry. Curr Opin Food Sci. 2022 Oct;47:100869.
9. Ismail BP, Senaratne-Lenagala L, Stube A, Brackenridge A. Protein demand: review of plant and animal proteins used in alternative protein product development and production. Animal Frontiers. 2020 Oct 30;10(4):53–63.
10. Fortune business insights. Whey protein market size, share and industry analysis, by types, application and regional forecast 2024-2032. 2023 [cited 2025 Feb 4]; Available from: https://www.fortunebusinessinsights.com/methodology/whey-protein-market-106555

11. Yalcin A. Emerging Therapeutic Potential of Whey Proteins and Peptides. Curr Pharm Des. 2006 May 1;12(13):1637–43.
12. Patel S. Functional food relevance of whey protein: A review of recent findings and scopes ahead. J Funct Foods. 2015 Dec;19:308–19.
13. Yang S, Song L, Wang J, Zhao J, Tang H, Bao X. Engineering Saccharomyces cerevisiae for efficient production of recombinant proteins. Engineering Microbiology. 2024 Mar;4(1):100122.
14. Mauro VP. Codon Optimization in the Production of Recombinant Biotherapeutics: Potential Risks and Considerations. BioDrugs. 2018 Feb 1;32(1):69–81.
15. Feng X, Marchisio M. Saccharomyces cerevisiae Promoter Engineering before and during the Synthetic Biology Era. Biology (Basel). 2021 Jun 6;10(6):504.
16. Sameshima-Yamashita Y, Yarimizu T, Uenishi H, Tanaka T, Kitamoto H. Uracil-auxotrophic marker recycling system for multiple gene disruption in Pseudozyma antarctica. Biosci Biotechnol Biochem. 2022 May 24;
17. Song W, Li J, Liang Q, Marchisio MA. Can terminators be used as insulators into yeast synthetic gene circuits? J Biol Eng. 2016 Dec 16;10(1):19.
18. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389–402.
19. Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J. Protein–Sol: a web tool for predicting protein solubility from sequence. Bioinformatics. 2017 Oct 1;33(19):3098–100.
20. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003 Jul 1;31(13):3406–15.
21. Dimitrov I, Flower DR, Doytchinova I. AllerTOP - a server for in silico prediction of allergens. BMC Bioinformatics. 2013 Apr 17;14(S6):S4.
22. Babendure JR, Babendure JL, Ding JH, Tsien RY. Control of mammalian translation by mRNA structure near caps. RNA. 2006 May;12(5):851–61.
23. Reinmuth-Selzle K, Tchipilov T, Backes AT, Tscheuschner G, Tang K, Ziegler K, et al. Determination of the protein content of complex samples by aromatic amino acid analysis, liquid chromatography-UV absorbance, and colorimetry. Anal Bioanal Chem. 2022 Jun 23;414(15):4457–70.
24. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein Identification and Analysis Tools on the ExPASy Server. In: The Proteomics Protocols Handbook. Totowa, NJ: Humana Press; 2005. p. 571–607.
25. Saikat ASM, Uddin MdE, Ahmad T, Mahmud S, Imran MdAS, Ahmed S, et al. Structural and Functional Elucidation of IF-3 Protein of Chloroflexus aurantiacus Involved in Protein Biosynthesis: An In Silico Approach. Biomed Res Int. 2021 Jan 2;2021(1).
26. Liu X, Qin X, Wang Y, Zhong J. Physicochemical properties and formation mechanism of whey protein isolate-sodium alginate complexes: Experimental and computational study. Food Hydrocoll. 2022 Oct;131:107786.
27. Seelig J, Seelig A. Protein Unfolding—Thermodynamic Perspectives and Unfolding Models. Int J Mol Sci. 2023 Mar 13;24(6):5457.
28. Bernhofer M, Dallago C, Karl T, Satagopam V, Heinzinger M, Littmann M, et al. PredictProtein - Predicting Protein Structure and Function for 29 Years. Nucleic Acids Res. 2021 Jul 2;49(W1):W535–40.
29. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022 Jan 7;50(D1):D439–44.
30. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018 Jul 2;46(W1):W296–303.
31. Bienert S, Waterhouse A, de Beer TAP, Tauriello G, Studer G, Bordoli L, et al. The SWISS-MODEL Repository—new features and functionality. Nucleic Acids Res. 2017 Jan 4;45(D1):D313–9.
32. Studer G, Rempfer C, Waterhouse AM, Gumienny R, Haas J, Schwede T. QMEANDisCo—distance constraints applied on model quality estimation. Bioinformatics. 2020 Mar 1;36(6):1765–71.