

Article

Predicting the Capsid Architecture of Florendoviruses: Molecular Fossils of Ancient Caulimovirid-like Viruses

Carlos M. Vicent ^{1,2} 

¹ Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Edifici CRAG, Campus UAB, Bellaterra, 08193 Barcelona, Spain; carlos.vicent@cragenomica.es; Tel.: +34-935-636-600

² Consejo Superior de Investigaciones Científicas (CSIC), 08193 Barcelona, Spain

Abstract

Endogenous viral elements derived from members of the family *Caulimoviridae* are widely distributed across plant genomes. Once integrated, they accumulate extensive mutations; most are no longer biologically active and can therefore be regarded as molecular fossils. Their loss of activity precludes the recovery of viral particles, thereby preventing direct experimental characterization of their capsid architecture. Here, we combined sequence-based phylogenetic analyses with artificial-intelligence-assisted protein structure prediction to investigate the coat proteins of florendoviruses, the most abundant endogenous caulimovirid lineage. Phylogenetic analysis placed florendovirus coat proteins in close proximity to those of the genus *Petuvirus*. The structural prediction of the coat protein is limited by the presence of an intrinsically disordered N-terminal region, whose structure cannot be reliably predicted. In contrast, the central capsid domain yields robust predictions, revealing similarity not only to caulimovirus coat proteins but also to retrotransposon Gag proteins. Structural comparisons indicate that the florendovirus coat domain is most closely related to those of *Petuvirus*. Because petuviruses assemble into isometric particles, our data support the hypothesis that the ancestral episomal forms of florendoviruses possessed a similar capsid organization. Our findings extend the study of ancient viruses whose morphology cannot be directly investigated using conventional phylogenetic and AI-assisted structure prediction methods.

Keywords: endogenous viral element; endogenous caulimovirid; florendovirus; AlphaFold 3; AI-assisted structural biology

1. Introduction

Endogenous viral elements (EVEs) are virus-derived DNA sequences that have become incorporated into host chromosomes. When integration occurs in germline cells, these sequences may be transmitted vertically and eventually become fixed within host populations [1]. Retroviruses, including human immunodeficiency virus, undergo obligatory chromosomal integration during replication through the activity of a virus-encoded integrase. By contrast, no plant virus described to date is known to require integration into the host genome as an essential component of its replication cycle. Nevertheless, numerous EVEs have been identified in plant genomes, indicating that viral endogenization has occurred repeatedly during plant evolution [1,2].

Only a small proportion of plant EVEs have retained sufficient structural integrity to remain biologically active. A limited number have been experimentally shown to be transcribed and to generate episomal viral genomes capable of assembling into infectious



Academic Editor: Juraj Gregaň

Received: 10 August 2026

Revised: 9 September 2026

Accepted: 11 September 2026

Published: 14 September 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

particles [3]. Most plant EVEs, however, are represented by incomplete, fragmented, duplicated, or rearranged viral sequences. Following integration, these elements generally undergo progressive molecular degradation through the accumulation of nucleotide substitutions, insertions, deletions, frameshift mutations, premature termination codons, and other disruptions affecting viral open reading frames and regulatory regions. Their transcription and potential reactivation may be further restricted by host-mediated epigenetic mechanisms, including DNA methylation and transcriptional silencing. Consequently, most EVEs are considered replication-defective and are unlikely to produce infectious viral particles. Despite their frequent loss of biological activity, EVEs constitute valuable molecular records of ancient infections [4]. Their identification and phylogenetic analysis can provide information about viral lineages that have become extinct or that no longer circulate in the host species in which their endogenous counterparts are found.

The family *Caulimoviridae* comprises non-enveloped, reverse-transcribing plant viruses with non-covalently closed circular double-stranded DNA genomes [5]. Their replication involves transcription of a pre-genomic RNA intermediate in the nucleus, followed by reverse transcription in the cytoplasm. Integration is not required for this process. Nevertheless, caulimovirid DNA has repeatedly entered plant nuclear genomes during evolution [6]. EVEs derived from members of the family *Caulimoviridae* are commonly designated endogenous caulimovirids, or ECVs, to distinguish them from actively replicating viruses and from members of the genus *Caulimovirus* [7].

The ICTV report for *Caulimoviridae* recognizes 11 genera. Their classification is based on phylogenetic relationships, genome organization, biological characteristics, and virion morphology. Virions are isometric in members of *Caulimovirus*, *Cavemovirus*, *Petuvirus*, *Rosadnavirus*, *Rufiodivirus*, *Solendovirus*, and *Soymovirus*, whereas members of *Badnavirus* and *Tungrovirus* form bacilliform particles. Virion morphology has not yet been established for *Dioscovevirus* or *Vaccinivirus* [5].

ECVs are widespread in vascular-plant genomes and have been associated with several recognized caulimovirid genera. Some ECVs can be assigned to extant taxa on the basis of sequence similarity and phylogenetic position, whereas many others represent divergent lineages for which no infectious episomal counterpart has yet been identified. The great majority are fragmented and replication-defective. However, replication-competent ECVs have been documented in a small number of plants in which ECVs may generate episomal infections under particular genetic, physiological, or environmental conditions [3,5,7].

Among ECV lineages without known exogenous representatives, florendoviruses (Flora endogenous virus) constitute the most abundant and broadly distributed lineage [6]. Florendovirus-derived sequences have been detected in the genomes of numerous angiosperms and, in certain species, account for a substantial fraction of the nuclear genome [8].

Reconstructed florendovirus genomes display an organization broadly compatible with that of other members of *Caulimoviridae*. Their predicted polyproteins contain conserved functional domains associated with cell-to-cell movement, capsid formation, proteolytic processing, reverse transcription, and RNase H activity. Phylogenetic analyses place florendoviruses close to, but distinct from, members of the genus *Petuvirus* [8].

Coat proteins are central components of caulimovirid virions [5]. They mediate capsid assembly, package and protect the viral DNA genome, and may contribute to genome delivery and interactions with other viral or host proteins. Caulimovirid coat proteins are generally described as modular proteins comprising an N-terminal intrinsically disordered region, a central capsid domain, and a zinc-finger-like motif involved in nucleic-acid binding.

Deep-learning methods for protein-structure prediction have substantially expanded the scope of structural virology. AlphaFold can generate three-dimensional protein models directly from amino acid sequences and has proved particularly useful for annotating poorly characterized viral proteins, identifying conserved structural cores, detecting remote evolutionary relationships, and formulating experimentally testable hypotheses [9,10], although the reliability of each prediction depends on the target, available evolutionary information, and the confidence assigned to individual regions and inter-residue relationships [11]. These approaches are especially valuable for uncultured, extinct, or exclusively endogenous viral lineages whose structures cannot be examined directly. In such cases, predicted structural conservation may retain an evolutionary signal that is less apparent at the primary-sequence level. Nevertheless, predicted models must be interpreted cautiously. Structural prediction should therefore be integrated with sequence phylogenetics, domain annotation or comparative structural searches.

Here, we implemented a structural-bioinformatics framework that combined AlphaFold 3 predictions with sequence-based phylogenetic reconstruction and quantitative structural comparisons to investigate the evolutionary relationships and probable capsid architecture of florendoviruses. This combined approach was designed to determine whether florendovirus coat proteins display a structural organization consistent with their sequence-based placement near petuviruses. The resulting framework provides a proof of concept for the use of predicted protein structures to investigate virion architecture in endogenous extinct viral groups represented only by genomic fossils.

2. Methods

2.1. Sequence Selection

The amino acid sequences analyzed correspond to the central domain of the coat protein (CP; PF22909). Viral CP sequences were retrieved from the NCBI GenBank Virus Database.

Florendovirus CP sequences were obtained from previous analyses [6,8]. Only high-quality genome assemblies were considered, corresponding to complete genome or chromosome-level assemblies according to NCBI criteria (<https://www.ncbi.nlm.nih.gov/datasets/docs/v2/glossary/#assembly-level>; accessed on 12 July 2026). Within these assemblies, only sequences located on chromosomes were retained, whereas sequences located on contigs or scaffolds were excluded. RT-coding DNA sequences were extended by 3000 bp on each side, and regions encoding a complete CP central domain were selected.

The selected sequences from each genome assembly were clustered using CD-HIT [12] with a 90% sequence identity threshold. An identity threshold of 90% was previously used to delineate sequences belonging to the same putative viral species on ECVs [13]. Only clusters containing at least five sequences were retained (Supplementary File S1). In some cases, more than one cluster was selected from a given genome assembly.

Consensus amino acid sequences were generated using EMBOSS Cons v6.6.0.0 with default parameters, including the EBLOSUM62 substitution matrix, a plurality threshold equal to 50% of the total sequence weight, no minimum identity requirement (identity = 0), and a set case threshold equal to 50% of the total sequence weight [14].

2.2. Phylogenetic Analysis

Multiple amino acid sequence alignment was performed using MAFFT v7.511 (<https://mafft.cbrc.jp/alignment/server/index.html> accessed on 12 July 2026), with the L-INS-I strategy [15]. Phylogenetic relationships among CP amino acid sequences were inferred using the maximum-likelihood (ML) method in MEGA12.1 [16]. The analysis was conducted using the a priori selected Jones-Taylor-Thornton (JTT) substitution model,

assuming uniform evolutionary rates among sites and including all alignment positions. The initial tree was generated automatically using the Neighbor-Joining/BioNJ approach and subsequently optimized by nearest-neighbor interchange (NNI). The final tree was left unrooted. Node support was evaluated by 1000 bootstrap replicates.

2.3. AI-Assisted Structural Inference

Structural predictions of the CP central domain were generated using the AlphaFold Server web service, powered by AlphaFold 3 v3.0.3 (<https://alphafoldserver.com/>). AlphaFold 3 has been described as a deep-learning framework for predicting biomolecular structures and interactions, including proteins and other biomolecular components [11]. Models were generated using the default configuration settings. For each single-chain query sequence, five models were produced, and the model with the highest pLDDT confidence score was selected for subsequent analyses. Only high-confidence models (pTM $\geq$ 0.80) were retained.

Pairwise structural distances among the retained models were estimated as root-mean-square deviation (RMSD) values after structural superposition using the MatchMaker algorithm implemented in ChimeraX 1.11.1 [17,18]. The resulting RMSD distance matrix was converted into a structural dendrogram using the UPGMA algorithm implemented in DendroUPGMA web server (<http://genomes.urv.cat/UPGMA/> accessed on 12 July 2026) [19]. Structural similarity searches were performed using the heuristic PDB search against the Protein Data Bank with default parameters in the DALI server (<http://ekhidna2.biocenter.helsinki.fi/dali/> accessed on 12 July 2026) [20].

3. Results and Discussion

Caulimovirus coat protein (CP) sequences are not particularly well conserved at the amino acid level, especially when compared with other viral proteins such as reverse transcriptase (RT). Consequently, they have not been widely used in phylogenetic analyses. Nevertheless, two conserved domains have been identified within the CP. The most highly conserved region corresponds to a C-terminal zinc-finger motif; however, this region comprises fewer than 20 amino acids and is therefore of limited value for phylogenetic inference. The second conserved domain (PF22909, Caulimovirus coat protein, conserved domain) is located in the central region of the protein. Although this domain is neither highly conserved nor particularly long, comprising approximately 104 amino acids, it enables the detection of sequence conservation among both episomal and endogenous members of the *Caulimoviridae*. Moreover, it exhibits a degree of similarity to the GAG proteins of retrotransposons, including a conserved α -helical core [21].

Because protein structure is generally more conserved than the primary amino acid sequence, combining PF22909-based sequence analysis with structural prediction may improve the evolutionary characterization of highly divergent caulimovirid coat proteins. This approach is particularly useful for endogenous lineages whose coding sequences have accumulated mutations following their integration into plant genomes. Nevertheless, conservation of the central α -helical fold should not be interpreted as evidence of identical oligomeric organization, capsid geometry, or overall virion morphology.

Because of the presence of an intrinsically disordered region at the N-terminus (NID, N-terminal Intrinsically Disordered Domain), AlphaFold 3 predictions generated from complete coat protein sequences did not yield consistently high-confidence models, with predicted template modelling scores below 0.80. In contrast, analyses restricted to the central PF22909 domain produced high-confidence structural predictions in most cases, indicating that this region constitutes a more structurally constrained and reliably modelled core. Therefore, considering its phylogenetic informativeness, detectable conservation

across the sequences examined, and improved prediction confidence, the PF22909 domain was selected for all subsequent sequence-based and structure-based analyses. Restricting the analysis to this homologous domain also minimized the potential influence of disordered, highly variable, or poorly modelled regions. The contrast between limited amino acid sequence conservation and stronger preservation of the predicted three-dimensional fold provides a clear rationale for applying structure-prediction methods to investigate florendovirus coat proteins and formulate hypotheses concerning their evolutionary relationships and probable capsid architecture.

Coat protein (CP) sequences representing all species currently recognized within the family *Caulimoviridae* were retrieved from the NCBI GenBank Virus database, with one representative sequence selected for each species (Supplementary File S1). Florendovirus sequences were obtained from previously published datasets generated through searches based on conserved reverse transcriptase (RT) domains [6,8]. For each genomic sequence encoding an identifiable RT domain, the corresponding genomic interval was extended by 3000 nucleotides in both the upstream and downstream directions to recover the surrounding viral coding region. Only intervals encoding a complete CP sequence containing the conserved PF22909 domain were retained for subsequent analyses.

To minimize errors associated with incomplete or poorly assembled genomic regions, only high-quality plant genome assemblies were considered. Furthermore, the analysis was restricted to viral sequences located on assembled chromosomes, whereas sequences identified on unplaced contigs or scaffolds were excluded. CP sequences recovered from each plant genome were subsequently grouped into clusters using a pairwise amino acid sequence similarity threshold of 90%, an identity threshold previously used to delineate sequences belonging to the same putative viral species on ECVs [13]. To ensure that each cluster represented a sufficiently abundant and reproducible endogenous lineage, only clusters containing at least five CP sequences were retained. After applying these quality-control, completeness, and abundance criteria, the final dataset comprised 41 sequence clusters derived from 34 plant species. These clusters were selected for all subsequent phylogenetic and structural analyses, and the complete sequence dataset and associated metadata are provided in Supplementary File S1.

Although notable exceptions have been documented, most ECVs are integrated within repetitive, gene-poor regions of plant genomes and are therefore generally considered unlikely to exert a substantial effect on host viability or fitness [7]. Following endogenization, these viral sequences consequently evolve in a manner broadly comparable to pseudogenes. They progressively accumulate nucleotide substitutions, insertions, deletions, frameshifts, and premature termination codons at rates largely determined by the evolutionary dynamics of the host genome. The extent of sequence degeneration therefore depends, among other factors, on the age of the integration event, with older insertions generally exhibiting more extensive molecular decay.

The accumulation of post-integration mutations can obscure the original viral sequence and complicate the reconstruction of phylogenetic relationships, particularly when ancient or highly degraded integrations are examined. This problem becomes even more consequential when the objective is to reconstruct the proteins encoded by these elements, because individual mutations may disrupt conserved domains and alter predicted secondary structures. To minimize the influence of mutations acquired after integration and to approximate the ancestral viral sequence more reliably, a consensus sequence was generated for each cluster. A similar approach was used previously, for example, to study the human endogenous retrovirus HERV-K [22]. These cluster-level consensus sequences were used in all subsequent phylogenetic and structural analyses.

Maximum-likelihood phylogenetic analysis of the coat protein sequences resolved three major, well-defined clades (Figure 1). The overall tree topology was broadly congruent with previous phylogenetic reconstructions based on other conserved viral domains, indicating that the coat protein retains a sufficient evolutionary signal to recover the principal relationships among the analyzed lineages [6]. Florendovirus sequences consistently clustered with those of petuviruses, and this relationship received strong bootstrap support. Given that petuviruses assemble into isometric virions, the robust phylogenetic proximity between the two lineages is consistent with the hypothesis that florendoviruses may have possessed a comparable capsid organization. Nevertheless, phylogenetic relatedness alone does not conclusively establish virion morphology because related coat proteins may differ in their oligomeric interfaces, subunit arrangement, or higher-order assembly. This hypothesis therefore requires further evaluation through comparative structural analyses of the corresponding coat protein domains.

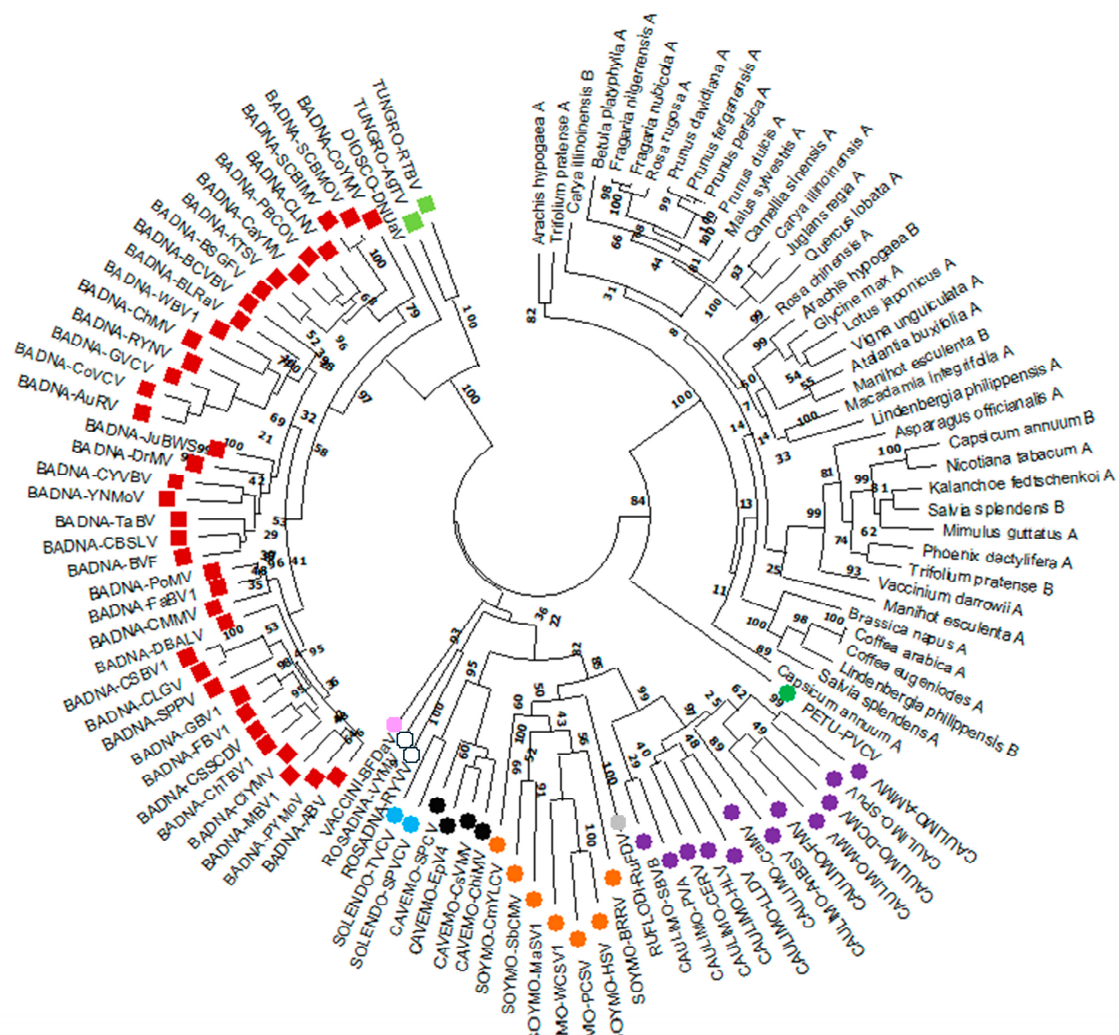


Figure 1. Phylogenetic relationships among caulimovirid coat proteins. Maximum-likelihood phylogenetic tree inferred from the amino acid sequences of the conserved central region of the coat protein, corresponding to Pfam domain PF22909. Branch support was assessed using 1000 bootstrap replicates. Labels in uppercase indicate episomal viruses, whereas labels in lowercase correspond to florendoviruses. Square symbols denote viruses with bacilliform virions (red, *Badnavirus*; green, *Tungrovirus*). Rounded symbols denote viruses with icosahedral virions (white, *Rosadnavirus*; blue, *Solendovirus*; black, *Cavemovirus*; orange, *Soymovirus*; grey, *Ruflodiviridae*; violet, *Caulimovirus*; green, *Petuvirus*). Florendovirus sequences are shown without symbols.

To investigate this relationship further, AlphaFold 3 was used to predict the structure of the central PF22909 coat protein domain. Only models with high global confidence, defined by a predicted template modelling score of $pTM \geq 0.80$, were retained, resulting in a final dataset of 58 predicted structures. All models were then subjected to pairwise structural comparison using MatchMaker in ChimeraX. Hierarchical clustering of the resulting structural similarity scores generated a dendrogram whose overall topology was largely congruent with the sequence-based phylogeny (Figure 2). In particular, the predicted central coat protein domains of florendoviruses consistently clustered with those of petuvirus species. This agreement between sequence-based and structure-based relationships provides support for a close evolutionary association between the two lineages and is consistent with the hypothesis that florendoviruses possessed a capsid architecture broadly similar to that of petuviruses.

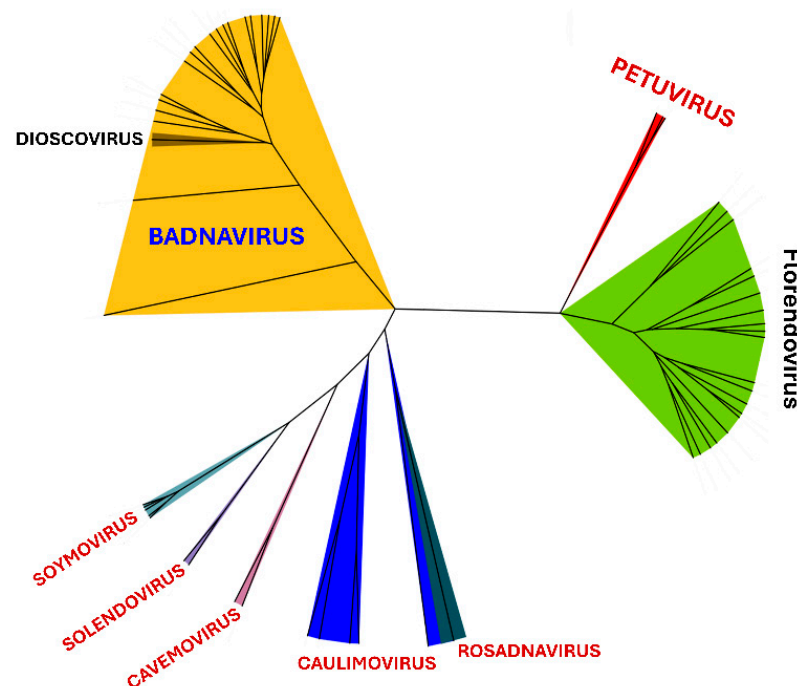


Figure 2. Structure-based relationships among caulimovirid coat proteins. Dendrogram generated from pairwise structural comparisons of AlphaFold 3-predicted models of the conserved central coat protein domain. Structural dissimilarity was expressed as root-mean-square deviation (RMSD) values calculated using MatchMaker in ChimeraX. Taxonomic groups are color-coded as follows: Badnavirus, orange; Petuvirus, red; florendoviruses, green; Soymovirus, light blue; Solendovirus, violet; Cavemovirus, pink; Caulimovirus, blue; and Rosadnavirus, grey.

Representative examples of the predicted structures are shown in Figure 3. The models were broadly consistent with the structural organization proposed by Krupovic and Koonin (2017) [21] for this conserved region, which comprises a predominantly α -helical core formed by five α -helices. The predicted florendovirus domain exhibited particularly strong structural similarity to the corresponding domain of the petunia vein-clearing virus (PVCV), a member of the genus *Petuvirus* (Figure 3A). A closely related fold was also observed in mirabilis mosaic virus (MMV), a member of the genus *Caulimovirus*, although the relative position of helix α_{10} differed slightly between the two models. By comparison, the predicted structure of the coconut vein-clearing virus (CoVVCV) domain, belonging to the genus *Badnavirus*, displayed more pronounced differences, primarily in the length and spatial orientation of helix α_8 , as well as in the positions of helix α_{10} and the 3/10 helix. Despite these local variations, all models retained a recognizable α -helical core, supporting the evolutionary conservation of this region across diverse caulimovirids. The particularly

close structural correspondence between florendovirus and petuvirus coat protein domains is consistent with the hypothesis that florendoviruses potentially assembled icosahedral capsids. Similar trends were observed across all other structural comparisons, as can be observed in the complete pairwise RMSD matrix provided in Supplementary File S2.

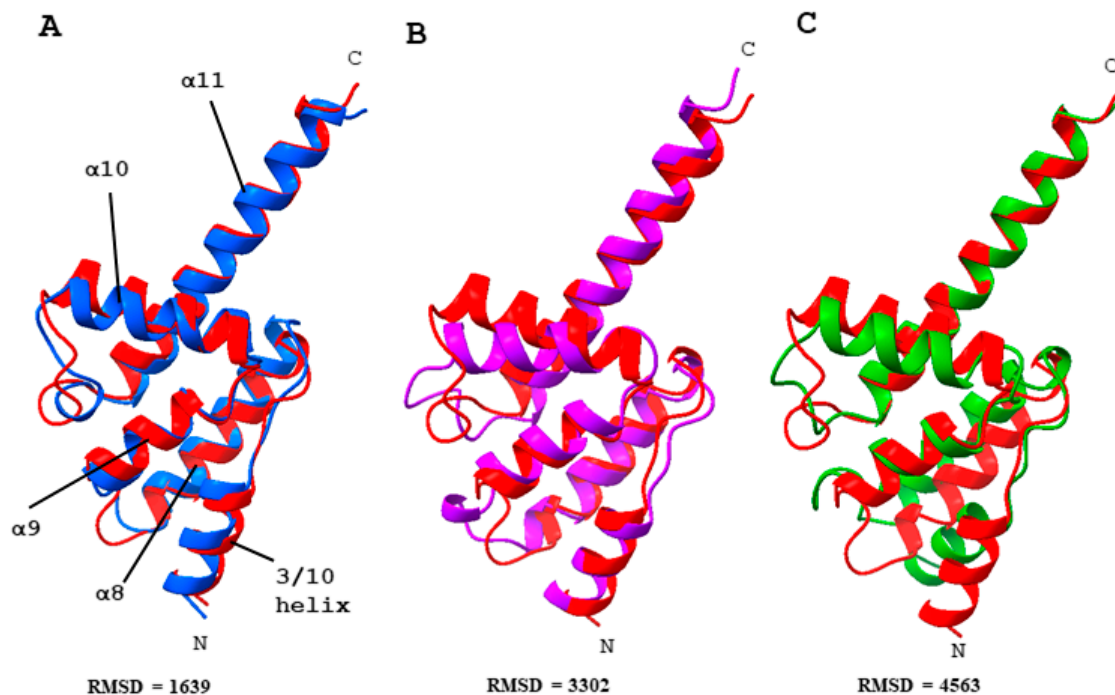


Figure 3. Structural modelling and alignment of the central coat protein domain of florendovirus *Manihot_esculenta_A*. The predicted structure of the florendovirus domain, shown in red, was superimposed on representative models from three caulimovirid genera: (A) petunia vein-clearing virus (PVCV; genus *Petuvirus*), shown in blue; (B) mirabilis mosaic virus (MMV; genus *Caulimovirus*), shown in violet; and (C) coconut vein-clearing virus (CoVCV; genus *Badnavirus*), shown in green. N and C indicate the N- and C-terminal residues of the modelled domains, respectively. Labels in panel A identify the conserved secondary-structure elements described for caulimovirid coat proteins [21]. RMSD denotes the root-mean-square deviation calculated for each pairwise structural comparison.

Additional searches for structural homologues were performed using the DALI server [20], with the predicted PF22909 domain structure from florendovirus *Manihot_esculenta_A* as the query. These searches identified several Gag and Gag-like proteins among the closest structural matches, including PEG10, MOAP1, and the Ty1 and Ty3 retrotransposon proteins from yeast. This structural correspondence is consistent with the previously proposed homology between caulimovirid coat proteins and the Gag proteins of retroviruses and long terminal repeat retrotransposons [21].

The structural alignment of the central coat protein domain of florendovirus *Manihot_esculenta_A* with the retrotransposon-derived protein PEG10 (Figure 4) illustrates the pronounced similarity between them. PEG10 is a human protein encoded by a domesticated GAG retrotransposon-derived gene [23]. These findings provide additional evidence that the central PF22909 domain retains a deeply conserved structural signature linking caulimovirid coat proteins to the broader evolutionary diversity of reverse-transcribing viruses and retroelements.

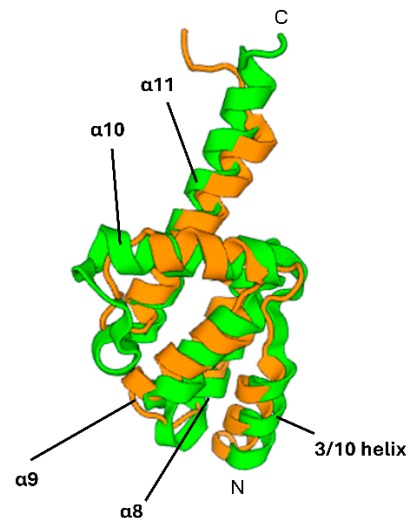


Figure 4. Structural alignment of the central coat protein domain of florendovirus *Mani-hot_esculenta_A* with the retrotransposon-derived protein PEG10. The predicted florendovirus structure, shown in green, was superimposed on the structure of PEG10, shown in orange, using the 3D Superimposition facility in the DALI server. N and C indicate the N- and C-terminal residues of the aligned domains, respectively. Labels identify the conserved secondary-structure elements described for caulimovirid coat proteins by Krupovic and Koonin (2017) [21].

4. Conclusions

Although the present study is based on computationally predicted protein structures rather than experimentally determined virion or capsid structures, the overall agreement between the sequence-based phylogenetic reconstruction and the structure-based comparative analyses provides convergent support for the hypothesis that the ancestral episomal forms of florendoviruses assembled into isometric virions. In particular, the consistent phylogenetic proximity of florendovirus and petuvirus coat proteins, together with the pronounced structural similarity observed between their conserved central domains, strengthens this interpretation. Nevertheless, conservation of the central coat protein fold cannot, by itself, establish the precise geometry, symmetry, or oligomeric arrangement of the complete ancestral capsid. These properties also depend on intersubunit interfaces, terminal protein regions, proteolytic maturation, and interactions with additional virion-associated components. More broadly, our findings demonstrate that the integration of conventional phylogenetic reconstruction with computational protein structure prediction and quantitative structural comparison can provide biologically informative insights into highly divergent endogenous viral lineages. This approach is particularly valuable when intact viral particles are unavailable and direct structural characterization is therefore not experimentally feasible. Rather than constituting definitive evidence of ancestral virion morphology, the predicted structures provide a framework for generating an explicit, biologically plausible, testable hypothesis concerning capsid evolution and organization. Similar integrative strategies may consequently facilitate the structural and evolutionary characterization of other extinct, uncultivated, or exclusively endogenous viruses preserved only as genomic fossils in their host species.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/sci8090259/s1>. CP sequences used in this study are provided in Supplementary File S1. AlphaFold 3 input sequences, confidence metrics and pairwise RMSD matrix are available as Supplementary File S2. Predicted structures are available as Supplementary File S3.

Funding: The author declared that financial support was received for this work and its publication. This research received support from Grant PID2022-143167NB-I00 funded by MICIU/AEI/10.13039/501100011033 and by “ERDF/EU”. This work was also supported by grant CEX2019-000902-S funded by MCIN/AEI/10.13039/501100011033, and by the CERCA Programme/Generalitat de Catalunya.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: CP sequences used in this study are provided in Supplementary File S1. AlphaFold 3 input sequences, confidence metrics and pairwise RMSD matrix are available as Supplementary File S2.

Acknowledgments: During the preparation of this manuscript/study, the author used Copilot GPT-5.5 for the purposes of language-editing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Feschotte, C.; Gilbert, C. Endogenous viruses: Insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* **2012**, *13*, 283–296. [[CrossRef](#)] [[PubMed](#)]
2. Chiba, S.; Kondo, H.; Tani, A.; Saisho, D.; Sakamoto, W.; Kanematsu, S.; Suzuki, N. Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. *PLoS Pathog.* **2011**, *7*, e1002146. [[CrossRef](#)] [[PubMed](#)]
3. Staginnus, C.; Richert-Pöggeler, K.R. Endogenous pararetroviruses: Two-faced travelers in the plant genome. *Trends Plant Sci.* **2006**, *11*, 485–491. [[CrossRef](#)] [[PubMed](#)]
4. Aiewsakun, P.; Katzourakis, A. Endogenous viruses: Connecting recent and ancient viral evolution. *Virology* **2015**, *479–480*, 26–37. [[CrossRef](#)] [[PubMed](#)]
5. Teycheney, P.Y.; Geering, A.D.W.; Dasgupta, I.; Hull, R.; Kreuze, J.F.; Lockhart, B.; Muller, E.; Olszewski, N.; Pappu, H.; Pooggin, M.M.; et al. ICTV Virus Taxonomy Profile: Caulimoviridae. *J. Gen. Virol.* **2020**, *101*, 1025–1026. [[CrossRef](#)] [[PubMed](#)]
6. de Tomás, C.; Vicient, C.M. Genome-wide identification of Reverse Transcriptase domains of recently inserted endogenous plant pararetrovirus (Caulimoviridae). *Front. Plant Sci.* **2022**, *13*, 1011565. [[CrossRef](#)] [[PubMed](#)]
7. Vassilief, H.; Geering, A.D.W.; Choisne, N.; Teycheney, R.Y.; Maumus, F. Endogenous Caulimovirids: Fossils, Zombies, and Living in Plant Genomes. *Biomolecules* **2023**, *13*, 1069. [[CrossRef](#)]
8. Geering, A.D.W.; Maumus, F.; Copetti, D.; Choisne, N.; Zwickl, D.J.; Zytynicki, M.; McTaggart, A.R.; Scalabrin, S.; Vezzulli, S.; Wing, R.A.; et al. Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nat. Commun.* **2014**, *5*, 5269. [[CrossRef](#)] [[PubMed](#)]
9. Gutnik, D.; Evseev, P.; Miroshnikov, K.; Shneider, M. Using AlphaFold predictions in viral research. *Curr. Issues Mol. Biol.* **2023**, *45*, 3705–3732. [[CrossRef](#)]
10. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [[CrossRef](#)] [[PubMed](#)]
11. Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A.J.; Bambrick, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500. [[CrossRef](#)] [[PubMed](#)]
12. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [[CrossRef](#)] [[PubMed](#)]
13. Vassilief, H.; Serfraz, S.; Choisne, N.; Geering, A.D.W.; Lefevre, P.; Teycheney, P.Y.; Maumus, F. Endogenous viral elements trace the ancient origins and early evolution of the Caulimoviridae. *PLoS Pathog.* **2026**, *22*, e1014340. [[CrossRef](#)] [[PubMed](#)]
14. Madeira, F.; Madhusoodanan, N.; Lee, J.; Eusebi, A.; Niewielska, A.; Tivey, A.R.N.; Lopez, R.; Butcher, S. The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Res.* **2024**, *52*, W521–W525. [[CrossRef](#)] [[PubMed](#)]
15. Katoh, K.; Rozewicki, J.; Yamada, K.D. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **2019**, *20*, 1160–1166. [[CrossRef](#)] [[PubMed](#)]
16. Kumar, S.; Stecher, G.; Suleski, M.; Sanderford, M.; Sharma, S.; Tamura, K. MEGA12: Molecular Evolutionary Genetic Analysis Version 12 for Adaptive and Green Computing. *Mol. Biol. Evol.* **2024**, *41*, msae263. [[CrossRef](#)] [[PubMed](#)]
17. Meng, E.C.; Goddard, T.D.; Pettersen, E.F.; Couch, G.S.; Pearson, Z.J.; Morris, J.H.; Ferrin, T.E. UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci.* **2023**, *32*, e4792. [[CrossRef](#)] [[PubMed](#)]

18. Krissinel, E.; Henrick, K. Multiple Alignment of Protein Structures in Three Dimensions. In *Computational Life Sciences*; Berthold, M.R., Glen, R., Diederichs, K., Kohlbacher, O., Fischer, I.B., Eds.; CompLife, LNBI 3695; Springer: Berlin/Heidelberg, Germany, 2005; pp. 67–78.
19. Garcia-Vallve, S.; Palau, J.; Romeu, A. Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in *Escherichia coli* and *Bacillus subtilis*. *Mol. Biol. Evol.* **1999**, *16*, 1125–1134. [[CrossRef](#)] [[PubMed](#)]
20. Holm, L. Using Dali for Protein Structure Comparison. In *Structural Bioinformatics*; Gáspári, Z., Ed.; Methods in Molecular Biology; Humana: New York, NY, USA, 2020; Volume 2112, pp. 29–42.
21. Krupovic, M.; Koonin, E.V. Homologous Capsid Proteins Testify to the Common Ancestry of Retroviruses, Caulimoviruses, Pseudoviruses, and Metaviruses. *J. Virol.* **2017**, *91*, e00210–17. [[CrossRef](#)] [[PubMed](#)]
22. Lee, Y.N.; Bieniasz, P.D. Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathog.* **2007**, *3*, e10. [[CrossRef](#)] [[PubMed](#)]
23. Ono, R.; Kobayashi, S.; Wagatsuma, H.; Aisaka, K.; Kohda, T.; Kaneko-Ishino, T.; Ishino, F. A retrotransposon-derived gene, PEG10, is a novel imprinted gene located on human chromosome 7q21. *Genomics* **2001**, *73*, 232–237. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.