

Integrating evolutionary signals and protein structure reveals localized adaptive divergence in *Cereus* lineages

Danilo T Amaral

danilo.trabuco@ufabc.edu.br

Universidade Federal do ABC

João Alfredo Teodoro

Universidade Federal do ABC

Maria Izadora Oliveira Cardoso

Universidade Federal do ABC

Evandro M. Moraes

Federal University of São Carlos

Fernando F Franco

Federal University of São Carlos

Isabel A.S. Bonatelli

Federal University of São Paulo

Research Article

Keywords: Adaptation, Positive Selection, Stability, Variability

Posted Date: April 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9346910/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Understanding how positive selection translates into functional changes in proteins remains challenging, particularly in non-model species. Here, we integrate phylogenetic evidence of lineage-specific positive selection with comparative structural and biophysical analyses to investigate adaptive divergence in the *Cereus fernambucensis*-*C. insularis* clade, which occupies contrasting island and continental environments. Based on a previously published branch-site scan, we selected five orthogroups showing robust statistical support, biological relevance, and full-length sequences for downstream analyses. Orthologous protein structures were predicted and evaluated, followed by quantitative comparisons. Geometrically defined cavities were inferred by pocket detection, and hotspot enrichment near these regions was assessed. Finally, we estimated the stability effects of lineage-specific substitutions using force-field-based predictions restricted to high-confidence structural regions. Across orthogroups, global folds were generally conserved, while structural divergence was predominantly localized to flexible loops and surface-exposed regions. Variability hotspots were non-randomly enriched near predicted cavities across several orthogroups, consistent with preferential diversification of putative interaction-associated surfaces. Stability scans indicated largely mild or near-neutral energetic effects (median $\Delta\Delta G \sim 0-1$ kcal·mol⁻¹), punctuated by a few strongly destabilizing substitutions at specific sites. Within the resolution of FoldX-based estimates, these patterns are consistent with the preservation of overall fold integrity. Our results indicate that adaptive evolution in *Cereus* lineages is primarily associated with localized modifications in surface-accessible regions within thermodynamically permissive regimes, particularly in structurally well-resolved proteins, rather than with large-scale structural innovation. More broadly, these findings highlight both the potential and current limitations of structure-informed evolutionary inference in non-model systems.

1. Introduction

Understanding how organisms adapt to diverse and changing environments is a central goal of evolutionary biology (Jensen et al. 2007). At the molecular level, adaptive genetic changes can shape protein structure and function, influencing organismal fitness by modifying enzymatic activity, regulatory interactions, and responses to environmental stress (López-Maury et al. 2008; Vogt 2022). Advances in genomics and comparative phylogenetic methods have enabled the identification of genes evolving under positive selection (PS) across lineages, providing valuable insights into the genetic basis of adaptation (Liberles et al. 2012; Edwards et al. 2022; Amaral et al. 2024; Jayaraman et al. 2022). However, translating statistical signatures of selection into mechanistic interpretations of how adaptive substitutions affect protein structure, function, and activity remains a major challenge.

Branch-site models and related likelihood-based approaches based on the nonsynonymous/synonymous rate ratio (d_N/d_S ; Yang 2007; Sun and Kozai 2024) have become widely used to detect positive selection acting on specific evolutionary lineages. These methods have been applied to diverse taxa and ecological contexts, revealing genes potentially involved in environmental adaptation, stress tolerance, and niche specialization (Amaral et al. 2024; Ma et al. 2025; Xu et al. 2025).

However, although d_N/d_S is a relevant indicator of selective pressure at the protein level, signatures of PS at the DNA level do not necessarily translate into functional divergence. Many amino acid substitutions may be selectively neutral or buffered by structural constraints, meaning that statistical signals alone provide limited information about their biological relevance (Ng and Henikoff 2006).

Recent advances in protein structure prediction, particularly with the development of AlphaFold (Jumper et al. 2021), have opened new opportunities to investigate the structural consequences of evolutionary change at an unprecedented scale (Varadi and Velankar 2023), helping to bridge the gap between statistical signatures of selection and their functional consequences. High-confidence structural models can now be generated for thousands of proteins, enabling comparative analyses that were previously restricted to a small number of experimentally solved structures. These developments allow researchers to integrate evolutionary, structural, and functional perspectives in a unified analytical framework (Varadi and Velankar 2023; Zhao et al. 2024).

Despite these advances, relatively few studies have combined genome-wide PS scans with comparative structural analyses across multiple lineages (e.g., Orsi et al. 2008; Batarseh et al. 2023; Ren et al. 2025). Most investigations either focus on DNA-level patterns without structural interpretation or analyze individual proteins in isolation. This creates an important methodological gap, as adaptive signals detected at the sequence level often cannot be directly linked to their potential structural or functional consequences. Moreover, many existing studies rely on codon-aware alignments and residue-level mapping of positively selected sites, approaches that require complete coding sequences and therefore cannot be readily applied to transcriptome-based datasets that frequently contain fragmented or partial transcripts. As a result, integrating evolutionary signals with structural inference at a large scale remains technically challenging, particularly for non-model organisms (Schell et al. 2025).

Cereus fernambucensis Lem. and *C. insularis* Hemsl. (Cereeae, Cactaceae) compose a monophyletic group occurring in xeric coastal habitats within the Brazilian Atlantic Forest phytogeographic domain. *Cereus fernambucensis* is distributed in the restinga vegetation along the Brazilian Atlantic coast, whereas *C. insularis* is a narrow endemic restricted to coastal rocky outcrops of the Fernando de Noronha Archipelago. The phylogenetic relationships among populations of both species suggest that *Cereus insularis* originated from northernmost *C. fernambucensis* populations under a progenitor–derivative speciation model (Franco et al. 2024). In a previous comparative transcriptomic and phylogenomic study, Amaral et al. (2024) used branch-site models to identify multiple orthogroups showing signatures of lineage-specific positive selection in the *C. fernambucensis*–*C. insularis* clade. These findings suggested that adaptive evolution has contributed to shaping molecular variation across island and continental populations. However, the structural and functional consequences of these evolutionary patterns remain unknown. This leaves several questions unresolved. For example, even in the absence of codon-level site mapping, do genes exhibiting PS signatures show spatially localized and structurally interpretable divergence among lineages? In particular, it remains unclear whether these adaptive signals are associated with changes in overall protein fold, localized remodeling of surface-accessible regions consistent with altered molecular interactions, or subtle modifications that preserve

overall structural stability. To address this question, we used orthogroups with signatures of positive selection identified for the *C. fernambucensis*-*C. insularis* clade (Amaral et al. 2024) to investigate structural divergence among lineages by integrating comparative structural modeling, sequence variability analysis, pocket detection, and stability inference. Rather than inferring function directly, we tested whether evolutionary heterogeneity was concentrated in structurally permissive regions and near putative cavities, under explicit model-confidence constraints. Using structure predictions and quantitative structural comparison methods, we examine how evolutionary divergence was reflected in three-dimensional protein structures across multiple lineages. Specifically, we aim to: (i) assess the consistency of predicted structures among orthologous proteins; (ii) quantify structural divergence using objective metrics; (iii) identify regions of elevated sequence variability and evaluate their spatial distribution within protein structures; (iv) test whether variable regions form spatial clusters near predicted functional cavities; and (v) estimate the potential impact of lineage-specific substitutions on protein stability in high-confidence structural regions.

2. Methods

2.1. Dataset and selection of positively selected orthogroups

The dataset analyzed in this study was derived from a previously published transcriptomic analysis of adaptive evolution in the *C. fernambucensis*-*C. insularis* clade and related lineages (Amaral et al. 2024). That study identified candidate genes evolving under lineage-specific positive selection using branch-site models. Orthogroups exhibiting statistically significant likelihood ratio tests after false discovery rate correction and supported by Bayes Empirical analyses were considered candidates for adaptive evolution using PAML v.4.9 (Yang 2007). From the original set of positively selected genes, five orthogroups were prioritized for downstream structural and functional analyses (OG0028003, OG0028976, OG0029756, OG0030099, and OG0031271) based on the following criteria: (i) robust statistical support in branch-site tests, (ii) functional annotation suggesting biological relevance, (iii) presence of full-length or near-full-length protein sequences, and (iv) suitability for structural modeling. These orthogroups encode proteins associated with core cellular metabolism, stress response pathways, and regulatory processes (Amaral et al. 2024), which are frequently reported as targets of adaptive evolution in insular and environmentally heterogeneous systems (Tsatsoulis et al. 2013; Wek et al. 2023).

2.2. Retrieval and curation of orthologous protein sequences

Orthologous amino acid sequences corresponding to the selected orthogroups were retrieved from the output of OrthoFinder v.2.5.5 (Emms and Kelly 2019) and from curated protein predictions derived from transcriptomes. Protein sequences were obtained from datasets associated with each lineage. For each orthogroup, all available orthologs were extracted and manually inspected to remove truncated, low-

complexity, or poorly aligned sequences. Redundant sequences were filtered using CD-HIT v.4.8.1 (Li and Godzik 2006) with an identity threshold of 100% to ensure non-redundant datasets. Sequence identifiers were standardized to reflect orthogroup membership and lineage origin. The final curated datasets were used as input for subsequent analyses.

2.3. Multiple sequence alignment and variability analysis

Multiple sequence alignments were generated at the amino acid level using MAFFT v7.505 (Katoh et al. 2005) with the L-INS-i strategy, which is optimized for accuracy in datasets with low to moderate sequence divergence. Default gap opening and extension penalties were applied. Alignment quality was assessed visually and by summary statistics, including gap frequency and alignment length. To quantify site-specific variability, Shannon entropy (Shannon 1948) was calculated for each alignment position using custom Python scripts built with BioPython and SciPy. Positions with gap frequencies greater than > 20% were excluded from entropy analyses. Sequence variability hotspots were defined as positions within the upper 5th percentile of entropy values among all non-gapped alignment positions.

2.4. Structural modeling

Three-dimensional protein structures were predicted using AlphaFold (Jumper et al. 2021; available at <https://alphafoldserver.com/>), an optimized high-throughput modeling tool. For each orthogroup and lineage, the corresponding amino acid sequence was submitted to AlphaFold using default parameters. For each protein, the model with the highest average predicted Local Distance Difference Test (pLDDT) score was selected for downstream analyses. Model confidence was evaluated using residue-level pLDDT scores. Regions with mean pLDDT values below 70 were considered low-confidence and were interpreted with caution.

2.5. Quantitative structural comparison and similarity assessment

Structural similarity among orthologous models was assessed using TM-align (Zhang and Skolnick 2005) and Foldseek v.10 (van Kempen et al. 2022) to increase robustness. Pairwise structural alignments were performed for all lineage pairs within each orthogroup. For each alignment, TM-score and root-mean-square deviation (RMSD) values were recorded. TM-score values were interpreted according to standard guidelines, with scores above 0.5 indicating shared global folds and values above 0.7 reflecting high structural similarity. RMSD values were calculated over aligned Ca atoms and used as a complementary measure of structural divergence. Results were compiled into pairwise similarity matrices and visualized as heatmaps using Python-based plotting libraries.

2.6. Identification of structural cavities and pocket-hotspot association

Putative ligand-binding pockets were predicted using FPocket v4.0 (Le Guilloux et al. 2009) with default parameters. For each structural model, FPocket was used to identify cavities based on alpha sphere

detection and Voronoi tessellation. Detected pockets were ranked by fpocket score, volume, and hydrophobicity. These predictions represent geometrically defined cavities and should not be interpreted as direct evidence of functional binding sites. To assess the spatial relationship between predicted pockets and sequence-variability hotspots, hotspot residues were mapped onto corresponding structural coordinates. Hotspot residues were classified as pocket-associated when at least one heavy atom was located within 5 Å of a pocket-defining alpha sphere. Statistical enrichment of hotspots in pocket-associated regions was evaluated using Fisher's exact test by comparing observed and expected frequencies.

2.7. In silico stability analysis

Protein stability was evaluated using FoldX v4 (Schymkowitz et al. 2005), which estimates changes in folding free energy ($\Delta\Delta G$) associated with amino acid substitutions based on empirical force-field approximations and predicted structural models. Before the mutational analyses, all structural models were subjected to the RepairPDB function to optimize side-chain conformations and remove steric clashes. For each orthogroup, lineage-specific amino acid substitutions identified from multiple sequence alignments were introduced using the BuildModel command. For each mutation, five independent FoldX runs were performed, and mean $\Delta\Delta G$ values were calculated to minimize stochastic variation. Analyses were restricted to residues located in high-confidence structural regions (pLDDT > 70). Mutations in low-confidence regions were excluded. $\Delta\Delta G$ values were interpreted according to standard conventions, with positive values indicating destabilizing effects and negative values indicating stabilizing effects, while acknowledging the intrinsic uncertainty associated with force-field-based estimates and predicted backbones.

2.8. Integration of structural, sequence, and functional analyses

Structural similarity metrics, entropy-based variability measures, pocket predictions, and stability estimates were integrated using custom Python scripts. For each orthogroup, the combined datasets were used to identify regions exhibiting convergent evidence of sequence divergence, structural remodeling, and potential functional relevance. The overall analytical workflow, which integrates evolutionary inference, structural modeling, functional site prediction, and stability analysis, is summarized in Figure S1.

3. Results and Discussion

3.1. Selection of positively selected orthogroups and dataset structure

The five selected orthogroups showed consistently high structural quality across lineages, as indicated by high median pLDDT values and stable TM-scores (Fig. 1). These results support strong overall consistency among the predicted structures of orthologous proteins (Objective i). These results indicate

that the predicted models were structurally robust across lineages, providing a reliable basis for downstream comparisons of sequence variation, structural divergence, and stability. This approach is consistent with recent recommendations to incorporate AlphaFold confidence metrics into structure-informed evolutionary studies (Akdel et al. 2022).

The selected orthogroups analyzed here include genes functionally associated with metabolism and stress response, as defined by the annotation and selection framework reported by Amaral et al. (2024). This functional profile is consistent with adaptive scenarios involving environmental heterogeneity, resource limitation, and fluctuating selective pressures. Island systems, in particular, often impose strong constraints on energetic efficiency and physiological resilience, thereby favoring the optimization of metabolic and regulatory networks (Whittaker and Fernández-Palacios 2007; Warren et al. 2015; Losos and Ricklefs 2021).

Moreover, comparative studies in vertebrates, insects, and plants have repeatedly shown that stress-related proteins and metabolic enzymes are recurrent targets of positive selection in insular and fragmented habitats (Rolland et al. 2014; Jeffries et al. 2018; Tigano and Friesen 2016). Our results are consistent with this broader pattern and suggest that similar selective mechanisms may operate in our studied system. The selected orthogroups exhibited contrasting structural and evolutionary profiles (Table 1). OG0028003 and OG0030099 showed the highest model confidence (mean pLDDT = 81.7 and 69.2, respectively) and exhibited strong structural conservation (mean TM-score = 0.901 and 0.933). These features make them the most suitable orthogroups for detailed structure-based functional interpretation. In contrast, OG0029756 and OG0031271 exhibited substantially lower model confidence (mean pLDDT = 48.8 and 61.0, respectively) and high predicted disorder (0.81 and 1.00, respectively), suggesting that these proteins may contain extensive intrinsically disordered regions. This pattern indicates that these later orthogroups likely correspond to conformationally flexible or partially unstructured proteins. As a result, fine-scale structural inference is less reliable in these cases, highlighting an important methodological limit of structure-based analyses, thereby enabling subsequent assessments defined in Objectives (ii–v) (van Der Lee et al. 2014; Wright and Dyson 2015; Chen et al. 2018).

3.2. Structural Model Quality and Lineage-Specific Structural Divergence

Structural models were generated for all selected orthogroups, and residue-level profiles were used to assess local variation in model confidence (Jumper et al. 2021; Tunyasuvunakool et al. 2021). While overall structural robustness is described above (Objective i), here we focus on within-protein confidence patterns. OG0028003 and OG0030099 consistently displayed extensive high-confidence segments, with most residues exceeding pLDDT > 75 and relatively few low-confidence regions. These profiles are consistent with well-defined secondary and tertiary structural elements, enabling more reliable residue-level structural interpretation for these proteins (Akdel et al. 2022).

In contrast, the remaining orthogroups exhibited lower average pLDDT values and broader predicted intrinsically disordered regions (IDRs), thereby restricting the reliability of fine-scale structural and energetic analyses. Although lower-confidence regions should be interpreted cautiously, they may reflect biologically meaningful flexibility rather than prediction artifacts. Intrinsically disordered and flexible regions are common in regulatory proteins and stress-response factors and often associated with adaptive diversification (Wright and Dyson 2015; Babu 2016). Thus, localized low-confidence segments do not necessarily compromise the overall structural interpretation, but may instead indicate functional plasticity.

The distribution of pLDDT scores across orthogroups and lineages (Fig. 1) indicates that structural confidence was broadly consistent among populations. This pattern suggests that the structural differences inferred among lineages are unlikely to reflect systematic modeling bias. Integrating structural confidence metrics into evolutionary analyses has been increasingly advocated to reduce false functional inferences based solely on sequence-level signals (Akdal et al. 2022).

To quantitatively assess structural divergence among orthologous proteins, pairwise structural alignments were performed (Objective ii.). Across all orthogroups, most pairwise comparisons exhibited moderate-to-high TM-scores, indicating substantial conservation of global folds. TM-scores above 0.5 typically reflect shared structural topology, whereas values above 0.7 indicate strong fold-level conservation (Zhang and Skolnick 2005). In this context, OG0028003 and OG0030099 showed consistently high TM-scores (> 0.7) and low RMSD values, revealing remarkable structural conservation despite underlying sequence divergence. Quantitatively, OG0028003 and OG0030099 exhibited low structural dispersion (mean RMSD = 2.15 Å and 2.21 Å, respectively), whereas OG0029756 and OG0031271 showed markedly higher divergence (mean RMSD = 5.35 Å and 4.42 Å), reflecting substantial conformational heterogeneity (Table 1). OG0028976 displayed intermediate behavior (mean TM-score = 0.683; RMSD = 2.28 Å), consistent with partial fold conservation combined with enhanced surface flexibility.

This pattern suggests strong functional constraints acting on core structural elements, consistent with the maintenance of essential functional architectures. Similar decoupling between sequence divergence and structural conservation has been widely reported and reflects the intrinsic robustness of protein folds (Chothia and Lesk 1986; Illergård et al. 2009; Liberles et al. 2012). By contrast, orthogroups with lower TM-scores and higher RMSD values showed greater structural variability among lineages. Importantly, this divergence was primarily localized to loop regions and peripheral domains, whereas core secondary structure elements remained largely conserved (Bloom et al. 2006; Echave et al. 2016; Arenas et al. 2013).

Such spatially restricted divergence is particularly consistent with branch-site positive selection models, which often detect adaptive changes affecting limited subsets of residues rather than entire domains (Venkat et al. 2018). The concordance between sequence-level selection signals and localized structural divergence strengthens the biological plausibility of the inferred adaptive events. Heatmaps of pairwise

RMSD and TM-score values (Fig. 2) further illustrate lineage-specific structural differentiation, with clearer divergence patterns observed in orthogroups previously identified as under positive selection. Together, these results show that integrating structural alignment metrics with evolutionary modeling provides a more nuanced view of adaptive divergence, bridging genotype–structure–phenotype relationships (Liberles et al. 2012). Overall, the structural analyses indicate that adaptive evolution in the studied system predominantly operates through localized conformational adjustments rather than through large-scale fold innovation, reinforcing the view that protein evolution is typically constrained by fold stability while permitting functional modulation at peripheral regions (Ferreiro et al. 2025; Huang et al. 2023).

3.3. Sequence variability hotspots and their functional-structural context

Shannon entropy analyses were implemented to identify regions of elevated sequence variability and evaluate their distribution along protein sequences (Objective iii.), without assuming direct evidence of positive selection at individual sites. Entropy values were calculated for each alignment position, and highly variable regions were defined as positions with top-5% entropy values and exhibiting low gap frequencies. Importantly, entropy-based hotspots reflect patterns of sequence variability and are not equivalent to codon-based signatures of positive selection. Across orthogroups, sequence variability hotspots were unevenly distributed along protein sequences and tended to cluster in flexible loops, terminal regions, and surface-exposed segments. In contrast, structurally conserved cores displayed consistently low entropy values, reflecting strong purifying constraints associated with fold stability and functional integrity (Chothia and Lesk 1986; Echave et al. 2016; Liberles et al. 2012).

This spatial segregation between conserved cores and more variable peripheral regions is a well-established hallmark of protein evolution. Core residues are typically constrained by thermodynamic and folding requirements, whereas surface-exposed regions tolerate higher mutational loads and may facilitate lineage-specific functional modulation, particularly when combined with independent evidence of positive selection (Bloom et al. 2006; Goldstein 2011; Nichol et al. 2019). Orthogroups with higher structural conservation exhibited fewer, more localized hotspots, whereas structurally divergent proteins displayed broader regions of elevated variability. This relationship indicates a tight coupling between fold stability and evolutionary flexibility, consistent with theoretical and empirical models of structure–sequence coevolution (Illergård et al. 2009; Arenas et al. 2013).

The entropy profiles and hotspot distributions (Figs. 3 and 4) revealed that sequence diversification is concentrated in structurally permissive regions rather than within the conserved structural core. This pattern is consistent with the general expectation that purifying selection preserves residues essential for structural stability, whereas diversification accumulates in flexible or solvent-exposed regions.

Several substitutions also involve shifts in physicochemical properties, including polarity, charge, and hydrophobicity, between continental and island lineages, potentially altering interaction surfaces or local

binding environments. Together, these results indicate that sequence variability is spatially concentrated in structurally permissive regions rather than uniformly distributed across the protein architecture.

To evaluate the functional relevance of sequence variability, structural cavities, and their spatial relationships with entropy-defined hotspots were systematically assessed (Objective iv). For most orthogroups, major pockets were consistently detected across lineages, indicating that overall binding-site architecture is evolutionarily conserved at the geometric level. Such conservation is expected for functionally essential interaction surfaces and has been observed across diverse protein families (Cukuroglu et al. 2014; Konc and Janežič 2014; Gao et al. 2022). These conserved pockets were frequently located near regions compatible with ligand-binding or regulatory interfaces, suggesting potential functional relevance that remains to be experimentally validated. Statistical enrichment analyses further revealed that sequence variability hotspots were significantly overrepresented near predicted pockets across several orthogroups (Table 2). Several orthogroups exhibited odds ratios > 1.5 , indicating preferential localization of variable residues near interaction cavities. This pattern remained qualitatively consistent after accounting for background residue distributions, supporting a non-random association between sequence variability and functional surfaces. Together, these results indicate that evolutionary diversification preferentially targets surface-exposed regions associated with putative molecular interaction surfaces identified by geometric cavity detection, rather than structurally critical cores. This pattern is consistent with, but does not by itself demonstrate, adaptive functional divergence. Similar hotspot–pocket coupling has been reported in studies of enzyme specificity, protein–protein interactions, and signaling proteins, where adaptive substitutions fine-tune binding properties without compromising fold stability (Aftab et al. 2024).

From an evolutionary perspective, this enrichment supports a model in which positive selection acts primarily on residues mediating intermolecular recognition, substrate affinity, or regulatory modulation. Such regions are expected to experience fluctuating selective pressures in heterogeneous environments, favoring rapid functional optimization (Arenas et al. 2013; Ding et al. 2022; Jimenez-Rosales and Flores-Merino 2018). In contrast, proteins displaying lower structural confidence showed weaker or inconsistent pocket–hotspot associations. This limitation likely reflects uncertainty in structural reconstruction rather than the genuine biological absence of functional coupling. These results emphasize the importance of integrating confidence metrics when interpreting structure–function relationships derived from predicted models (Akdel et al. 2022). The combined analysis of entropy profiles and pocket proximity (Fig. 4; Table 2) thus reveals that adaptive divergence in the studied orthogroups is spatially concentrated around interaction interfaces. This convergence between sequence variability, structural context, and predicted cavity architecture supports the structural plausibility of the inferred adaptive signals.

3.4. Structural stability and integrated patterns of adaptive divergence

Representative structural models were subjected to systematic mutational scanning across orthogroups, and changes in folding free energy were calculated for lineage-specific variants, providing comparative rather than absolute estimates of energetic effects (Objective v.). Analyses were restricted to regions with high structural confidence (pLDDT > 70) to minimize prediction artifacts associated with uncertain backbone conformations (Jumper et al. 2021; Akdel et al. 2022). Across orthogroups, most substitutions exhibited mildly destabilizing or near-neutral effects ($\Delta\Delta G \approx 0-1.5 \text{ kcal}\cdot\text{mol}^{-1}$), although a small subset showed pronounced destabilization ($> 4 \text{ kcal}\cdot\text{mol}^{-1}$). Median $\Delta\Delta G$ values ranged from approximately 0.0 to 1.0 $\text{kcal}\cdot\text{mol}^{-1}$ (Table 3), with OG0028003 and OG0028976 displaying slightly higher central tendencies (median $\Delta\Delta G \approx 1.00$ and 0.76, respectively) than OG0029756 and OG0030099 (median $\Delta\Delta G < 0.25$). However, several individual sites exhibited substantially higher energetic impacts, including OG0028003 I223A ($\Delta\Delta G \approx 3.28 \text{ kcal}\cdot\text{mol}^{-1}$) and OG0030099 D55V ($\Delta\Delta G \approx 4.47 \text{ kcal}\cdot\text{mol}^{-1}$), revealing localized hotspots of structural sensitivity. These values fall within ranges commonly interpreted as mildly destabilizing or near-neutral, indicating limited impact on global folding stability within the resolution of FoldX-based predictions. This pattern is consistent with theoretical and empirical evidence indicating that most naturally occurring substitutions in functional proteins are constrained to remain close to the stability threshold required for proper folding and activity (Bloom et al. 2006; Goldstein 2011; Tokuriki and Tawfik 2009).

A subset of lineage-specific variants displayed moderate to strong destabilizing effects, exemplified by OG0030099 D55V ($\Delta\Delta G \approx 4.5 \text{ kcal}\cdot\text{mol}^{-1}$) and OG0028003 I223A ($\Delta\Delta G \approx 3.28 \text{ kcal}\cdot\text{mol}^{-1}$). These regions are known to tolerate greater energetic perturbations due to their lower contribution to global fold stability and their frequent involvement in regulatory and interaction processes (Echave et al. 2016; Nichol et al. 2019; Pillai et al. 2022). Destabilizing substitutions were rarely observed within core secondary structure elements or densely packed hydrophobic cores. This spatial bias supports the view that adaptive divergence preferentially targets peripheral regions while preserving essential structural scaffolds. Similar patterns have been reported across diverse protein families and are considered a hallmark of structurally constrained adaptive evolution (Liberles et al. 2012; Arenas et al. 2013). Overall, the FoldX results (Table 3) suggest that adaptive substitutions in this system generally operate within a thermodynamically permissive regime, as inferred from relative stability trends rather than absolute energetic values, with only a few high-impact substitutions. Although OG0031271 exhibited elevated predicted disorder and comparatively lower structural confidence, representative sites were nevertheless subjected to exploratory stability analyses, which should be interpreted cautiously given the limited reliability of energy calculations in highly flexible systems. In this orthogroup, predicted $\Delta\Delta G$ values were consistently near-neutral ($|\Delta\Delta G| < 0.5 \text{ kcal}\cdot\text{mol}^{-1}$), reinforcing its role as a methodological reference and highlighting the limits of energy-based inference in highly flexible systems.

When examined individually, the analyzed orthogroups revealed complementary adaptive trajectories. OG0028003 represented a highly conserved structural scaffold with localized surface diversification and moderate energetic tolerance, suggesting functional fine-tuning without architectural disruption. OG0030099 exhibited similar fold stability but greater interface-associated variability, consistent with adaptive modulation of interaction surfaces. OG0028976 displayed intermediate structural conservation

combined with elevated disorder, indicating enhanced regulatory or signaling flexibility. In contrast, OG0029756 showed low structural confidence and extensive disorder, restricting detailed functional inference, whereas OG0031271 primarily served as a methodological reference, illustrating the limits of structure-based evolutionary interpretation in highly flexible systems, as reflected by its consistently near-neutral energetic responses despite elevated disorder.

Beyond the specific biological patterns observed in the studied orthogroups, our analyses also illustrate a general framework for interpreting signals of positive selection using structural information. In this approach, statistical evidence of lineage-specific positive selection identifies candidate genes, while subsequent analyses place amino acid substitutions within their structural and biophysical context. Specifically, sequence variability is mapped onto predicted protein structures to evaluate whether divergence concentrates in structurally permissive regions or near interaction-associated surfaces, and stability analyses assess whether substitutions occur within thermodynamically tolerable regimes. This framework links positive selection signals, amino acid substitutions, structural context, and potential functional consequences, thereby bridging sequence-level evolutionary inference with protein structure and biophysical constraints (Liberles et al. 2012; Echave et al. 2016; Jayaraman et al. 2022).

By integrating structural similarity metrics (Objective ii), sequence variability profiles (Objective iii), pocket enrichment analyses (Objective iv), and stability estimates (Objective v), within the framework of structurally consistent models (Objective i), we identified consistent patterns linking evolutionary divergence to protein architecture. Orthogroups exhibiting strong evidence of positive selection in previous analyses tended to display localized structural divergence near surface-accessible and functionally relevant regions. These regions were characterized by elevated sequence entropy, enrichment near predicted interaction pockets, and moderate effects on stability, with occasional pronounced destabilization at specific hotspots, collectively suggesting adaptive remodeling of molecular interaction surfaces. Such coordinated patterns are consistent with contemporary models of protein evolution in which adaptive change is concentrated at functional interfaces rather than distributed uniformly across structures (Nichol et al. 2019; Pillai et al. 2022). This mode of evolution enables fine-tuning of binding specificity, regulatory responsiveness, or interaction networks while maintaining global fold stability. In contrast, highly conserved orthogroups exhibited low sequence variability, strong structural conservation, and minimal stability perturbations, indicating stronger functional and biophysical constraints. These proteins likely occupy central positions in essential cellular pathways, where even minor perturbations may incur substantial fitness costs (Liberles et al. 2012; Echave et al. 2016).

The observed coupling between evolutionary signals, structural context, and energetic effects supports the biological relevance of the inferred adaptive events. Rather than reflecting stochastic sequence divergence, the detected patterns are consistent with selective optimization of protein surfaces under lineage-specific ecological and physiological pressures. Thus, the integrative analyses indicate that adaptive evolution in the analyzed lineages is predominantly driven by subtle yet functionally meaningful remodeling of protein surfaces, mediated by coordinated changes in sequence variability, structural

configuration, and thermodynamic stability. This evolutionary strategy minimizes deleterious pleiotropic effects while enabling continuous functional innovation, reinforcing the view that most adaptive protein evolution proceeds through incremental modifications of existing molecular frameworks rather than through radical architectural shifts (Arenas et al. 2013; Liberles et al. 2012; Jayaraman et al. 2022).

3.5. Surface-centered adaptive divergence across lineages

When considered jointly, the structural, evolutionary, and biophysical analyses reveal a consistent pattern linking lineage-specific divergence to protein spatial organization. Across orthogroups, variability hotspots and lineage-specific substitutions are predominantly concentrated in surface-exposed regions and flexible loops rather than within structurally constrained cores (Figs. 3 and 4). The majority of high-entropy sites occur in loop regions or solvent-exposed surfaces, reinforcing the observation that diversification is largely restricted to structurally permissive regions. This spatial segregation between conserved cores and variable peripheral regions is consistent with established models of protein evolution in which purifying selection maintains residues essential for fold stability whereas diversification accumulates in solvent-accessible segments that can tolerate higher mutational loads (Chothia and Lesk 1986; Bloom et al. 2006; Echave et al. 2016; Liberles et al. 2012).

Variability hotspots were frequently located near predicted structural cavities (Fig. 5), and enrichment analyses indicate that these sites were overrepresented near pocket-associated residues in several orthogroups (Table 2). Such coupling between sequence variability and predicted interaction surfaces suggests that adaptive divergence preferentially targets molecular interfaces rather than structurally critical cores. Stability analyses further support this interpretation, as most lineage-specific substitutions produced mild or near-neutral energetic effects in high-confidence structural regions (Table 3), consistent with adaptive change occurring within thermodynamically permissive regimes (Tokuriki and Tawfik 2009; Bloom et al. 2006; Goldstein 2011).

Together, these results indicate that adaptive evolution in the *C. fernambucensis* – *C. insularis* clade primarily operates through localized biochemical remodeling of protein surfaces while preserving global fold architecture. Such surface-centered diversification may allow lineage-specific tuning of molecular interactions under heterogeneous ecological conditions while minimizing destabilizing effects on protein structure (Liberles et al. 2012; Echave et al. 2016; Jayaraman et al. 2022).

4. Conclusions

By integrating evidence of positive selection with structural modeling, sequence-variability analyses, pocket detection, and stability inference, this study provides a mechanistic framework for interpreting adaptive protein evolution within the *C. fernambucensis*-*C. insularis* clade. Our results show that lineage-specific adaptation predominantly operates through localized remodeling of surface-exposed and interaction-associated regions, particularly in orthogroups exhibiting high structural confidence, while highly flexible proteins highlight the current limits of structure-informed evolutionary inference. This pattern indicates that functional diversification is primarily achieved through subtle modifications to

molecular interfaces rather than through large-scale structural rearrangements. Methodologically, our integrative approach demonstrates that combining AlphaFold confidence metrics with structural, energetic, and evolutionary analyses improves the biological interpretability of positive selection signals. By integrating sequence variation, three-dimensional structure, and stability constraints, this framework offers a robust strategy for investigating adaptive molecular evolution in heterogeneous and insular systems and can be extended to other taxa and ecological contexts.

Declarations

Funding

This work was supported by São Paulo Research Foundation (FAPESP 2023/05589-4 to DTA, 2025/17270-8 to JAT, 2024/19266-5 to MIOC, and 2020/15161-3 to FFF) and CAPES 001 (MIOC).

Competing Interests

The authors declare no competing interests.

Authors contributions

DTA conceived the idea for this study. DTA, MIOC, and JAT performed data collection and analyses. All authors contributed with numerous conceptions, writing, and the intellectual development of the paper, made multiple revisions, and approved the final draft.

Funding information

This work was supported by São Paulo Research Foundation (FAPESP 2023/05589-4 to DTA, 2025/17270-8 to JAT, 2024/19266-5 to MIOC, and 2020/15161-3 to FFF) and CAPES 001 (MIOC).

Author Contribution

DTA conceived the idea for this study. DTA, MIOC, and JAT performed data collection and analyses. All authors contributed with numerous conceptions, writing, and the intellectual development of the paper, made multiple revisions, and approved the final draft.

Acknowledgments

We thank the funding agencies and the authors' institutions for their support. We also thank Ms. Liamar Benevides for inspiring this work.

Data Availability

Custom scripts and detailed file descriptions are available in GitHub (<https://github.com/BBMDO/CeEVoS>).

References

1. Aftab A, Sil S, Nath S, Basu A, Basu S (2024) Intrinsic disorder and other malleable arsenals of evolved protein multifunctionality. *J Mol Evol* 92:669–684
2. Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B et al (2022) A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol* 29:1056–1067
3. Amaral DT, Bonatelli IAS, Romeiro-Brito M, Telhe MC, Moraes EM, Zappi DC et al (2024) Comparative transcriptome analysis reveals lineage- and environment-specific adaptations in cacti from the Brazilian Atlantic Forest. *Planta* 260:4
4. Anisimova M, Yang Z (2007) Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol Biol Evol* 24:1219–1228
5. Arenas M, Dos Santos HG, Posada D, Bastolla U (2013) Protein evolution along phylogenetic histories under structurally constrained substitution models. *Bioinformatics* 29:3020–3028
6. Babu MM (2016) The contribution of intrinsically disordered regions to protein function. *Cell Mol Life Sci* 73:3095–3108
7. Batarseh TN, Batarseh SN, Morales-Cruz A, Gaut BS (2023) Comparative genomics of the *Liberibacter* genus reveals widespread diversity in genomic content and positive selection history. *Front Microbiol* 14:1206094
8. Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci USA* 103:5869–5874
9. Chen J, Kriwacki RW (2018) Intrinsically disordered proteins: structure, function and therapeutics. *J Mol Biol* 430:2275
10. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* 5:823–826
11. Cukuroglu E, Engin HB, Gursoy A, Keskin O (2014) Hot spots in protein–protein interfaces: Towards drug discovery. *Prog Biophys Mol Biol* 116:165–173
12. Delpont W, Poon AF, Frost SD, Kosakovsky Pond SL (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26:2455–2457
13. Ding Y, Perez-Ortiz G, Peate J, Barry SM (2022) Redesigning enzymes for biocatalysis: exploiting structural understanding for improved selectivity. *Front Mol Biosci* 9:908285
14. Echave J, Spielman SJ, Wilke CO (2016) Causes of evolutionary rate variation among protein sites. *Nat Rev Genet* 17:109–121

15. Edwards SV, Robin VV, Ferrand N, Moritz C (2022) The evolution of comparative phylogeography: putting the geography (and more) into comparative population genomics. *Genome Biol Evol* 14:evab176
16. Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20:238
17. Ferreiro D, Pazos E, Arenas M (2025) Trends in substitution models of protein evolution for phylogenetic inference. *Mol Phylogenet Evol* 108473
18. Gao M, Nakajima An D, Parks JM, Skolnick J (2022) AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat Commun* 13:1744
19. Goldstein RA (2011) The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins* 79:1396–1407
20. Huang Y, Ma Q, Sun J, Zhou LN, Lai CJ, Li P et al (2023) Comparative analysis of Diospyros plastomes: insights into genomic features, mutational hotspots, and adaptive evolution. *Ecol Evol* 13:e10301
21. Illergård K, Ardell DH, Elofsson A (2009) Structure is three to ten times more conserved than sequence. *Proteins* 77:499–508
22. James JE, Lascoux M (2025) Amino acid properties, substitution rates, and the nearly neutral theory. *Genome Biol Evol* 17:evaf025
23. Jayaraman V, Toledo-Patiño S, Noda-García L, Laurino P (2022) Mechanisms of protein evolution. *Protein Sci* 31:e4362
24. Jeffries KM, Cannon RE, Verhille CE, Dabruzzi TF, Britton MT, Durbin-Johnson BP et al (2019) Divergent transcriptomic signatures in response to salinity exposure. *Evol Appl* 12:1212–1226
25. Jensen JD, Wong A, Aquadro CF (2007) Approaches for identifying targets of positive selection. *Trends Genet* 23:568–577
26. Jimenez-Rosales A, Flores-Merino MV (2018) Tailoring proteins to re-evolve Nature: A short review. *Mol Biotechnol* 60:946–974
27. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O et al (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589
28. Katoh K, Kuma KI, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511–518
29. Konc J, Janežič D (2014) Binding site comparison for function prediction and pharmaceutical discovery. *Curr Opin Struct Biol* 25:34–39
30. Le Guilloux V, Schmidtke P, Tuffery P (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* 10:168
31. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of sequences. *Bioinformatics* 22:1658–1659

32. Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E et al (2012) The interface of protein structure and molecular evolution. *Protein Sci* 21:769–785
33. López-Maury L, Marguerat S, Bähler J (2008) Tuning gene expression to changing environments. *Nat Rev Genet* 9:583–593
34. Ma G, Shi M, Li Y, Wang S, Zeng X, Jia Y (2025) Diverse adaptation strategies to stress in coastal sediments. *Environ Res* 271:121073
35. Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions. *Annu Rev Genomics Hum Genet* 7:61–80
36. Nichol D, Robertson-Tessi M, Anderson AR, Jeavons P (2019) Model genotype–phenotype mappings. *J R Soc Interface* 16
37. Orsi RH, Sun Q, Wiedmann M (2008) Genome-wide analyses of *Listeria monocytogenes*. *BMC Evol Biol* 8:233
38. Pillai AS, Hochberg GK, Thornton JW (2022) Simple mechanisms for the evolution of protein complexity. *Protein Sci* 31:e4449
39. Ren C, Comes HP, Zhu S, Zhang X, Jiang W, Fu C et al (2025) Genome-wide patterns of local adaptation. *New Phytol* 247:1503–1519
40. Rolland J, Condamine FL, Jiguet F, Morlon H (2014) Faster speciation in the tropics. *Science* 343:746–749
41. Schell T, Greve C, Podsiadlowski L (2025) Genome sequencing for non-model organisms. *Front Zool* 22:7
42. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server. *Nucleic Acids Res* 33:W382–W388
43. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423
44. Sun X, Kozai T (2024) Different selection levels of mitogenomes. *Diversity* 16:715
45. Tigano A, Friesen VL (2016) Genomics of local adaptation with gene flow. *Mol Ecol* 25:2144–2164
46. Tokuriki N, Tawfik DS (2009) Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol* 19:596–604
47. Tsatsoulis A, Mantzaris MD, Bellou S, Andrikoula M (2013) Insulin resistance: an evolutionary perspective. *Metabolism* 62:622–633
48. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A et al (2021) Protein structure prediction for the human proteome. *Nature* 596:590–596
49. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK et al (2014) Classification of intrinsically disordered proteins. *Chem Rev* 114:6589–6631
50. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Gilchrist CL, Söding J et al (2022) Foldseek: fast and accurate protein structure search. *bioRxiv*
51. Varadi M, Velankar S (2023) Impact of AlphaFold Protein Structure Database. *Proteomics* 23:2200128

52. Venkat A, Hahn MW, Thornton JW (2018) Multinucleotide mutations cause false inferences of positive selection. *Science* 361:64–69
53. Vogt G (2022) Environmental adaptation and epigenetics. *Epigenomes* 7:1
54. Warren BH, Simberloff D, Ricklefs RE, Aguilée R, Condamine FL, Gravel D et al (2015) Islands as model systems in ecology and evolution. *Ecol Lett* 18:200–217
55. Wek RC, Anthony TG, Staschke KA (2023) Translational control and stress response. *Antioxid Redox Signal* 39:351–373
56. Whittaker RJ, Fernández-Palacios JM (2007) Island biogeography: ecology, evolution, and conservation. Oxford University Press
57. Wright PE, Dyson HJ (2015) Intrinsically disordered proteins in cellular signalling. *Nat Rev Mol Cell Biol* 16:18–29
58. Xu W, Luo D, Peterson K, Zhao Y, Yu Y, Ye Z et al (2025) Ecological niche models for forest adaptation. *Biol Rev* 100:1754–1781
59. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591
60. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm. *Nucleic Acids Res* 33:2302–2309
61. Zhao N, Wu T, Wang W, Zhang L, Gong X (2024) Methods for predicting protein complex structure. *Interdiscip Sci Comput Life Sci* 16:261–288

Tables

Tables 1 to 3 are available in the Supplementary Files section.

Figures

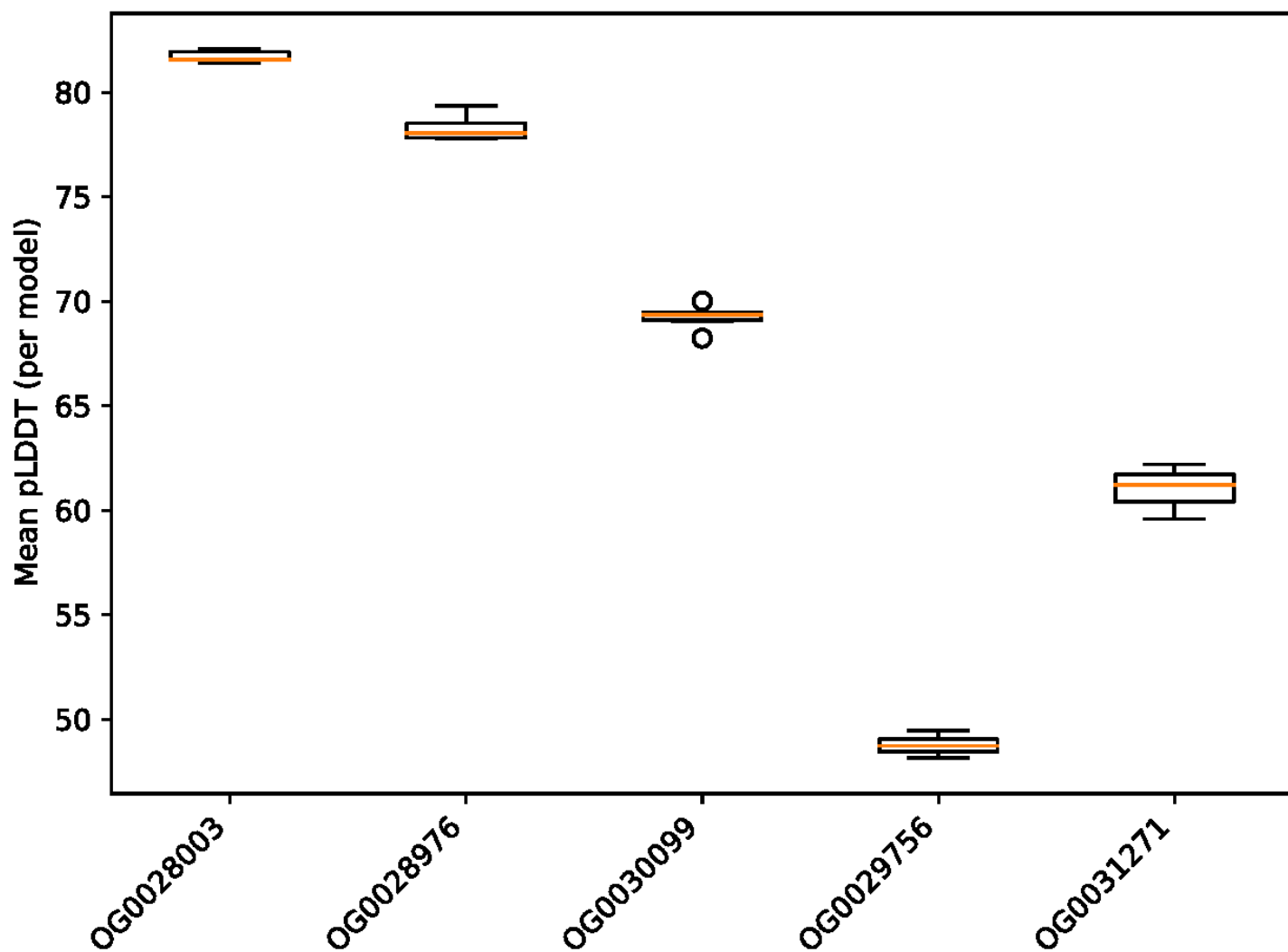


Figure 1

Boxplots show the distribution of mean per-model pLDDT scores for each analyzed orthogroups.

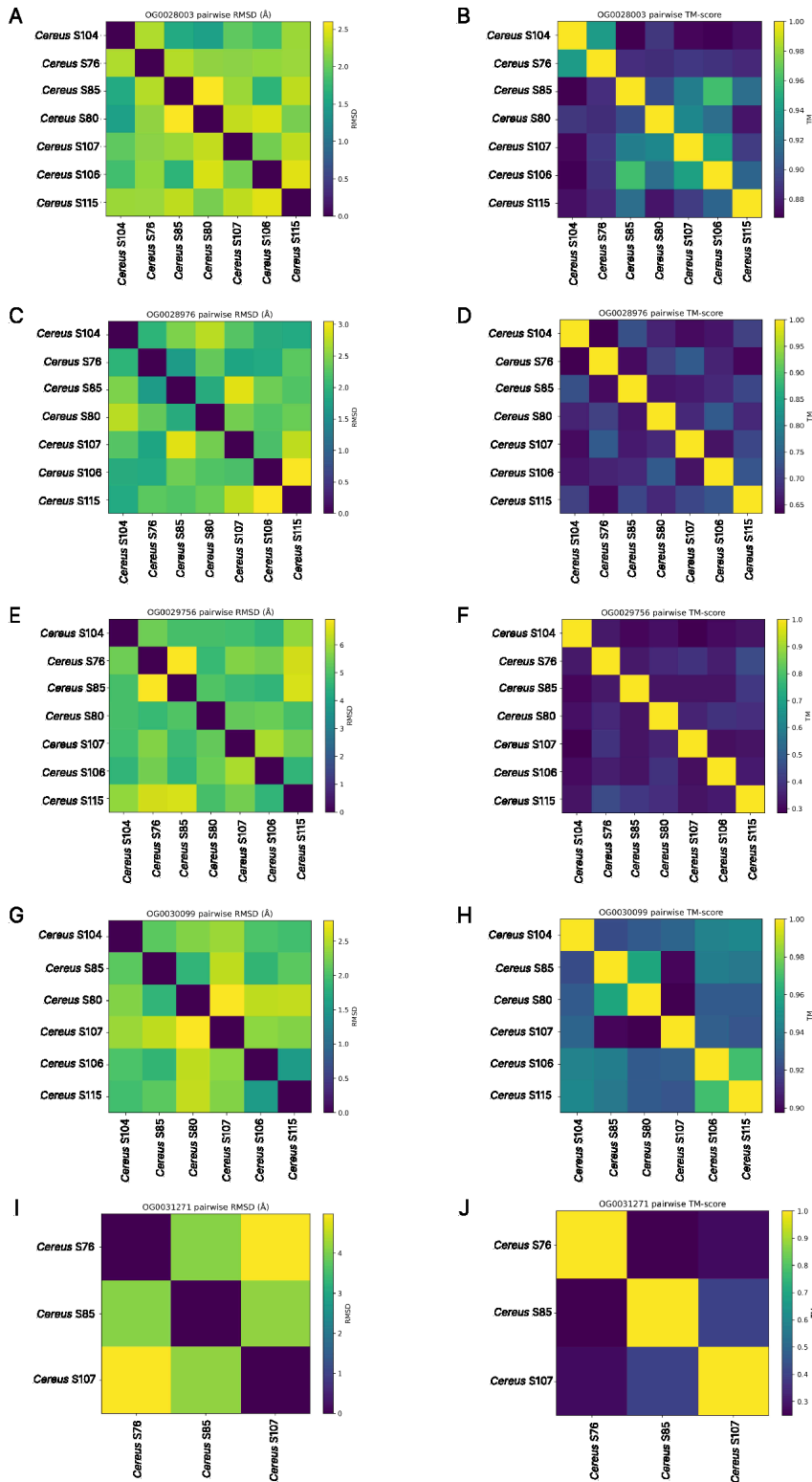


Figure 2

Heatmaps showing pairwise RMSD (Å) and TM-score values among predicted protein structures from selected *Cereus* orthogroups. Panels A–B: OG0028003; C–D: OG0028976; E–F: OG0029756; G–H: OG0030099; I–J: OG0031271. RMSD indicates structural divergence, while TM-score reflects global fold similarity. Comparisons were performed among samples S104, S76, S85, S80, S107, S106, and S115.

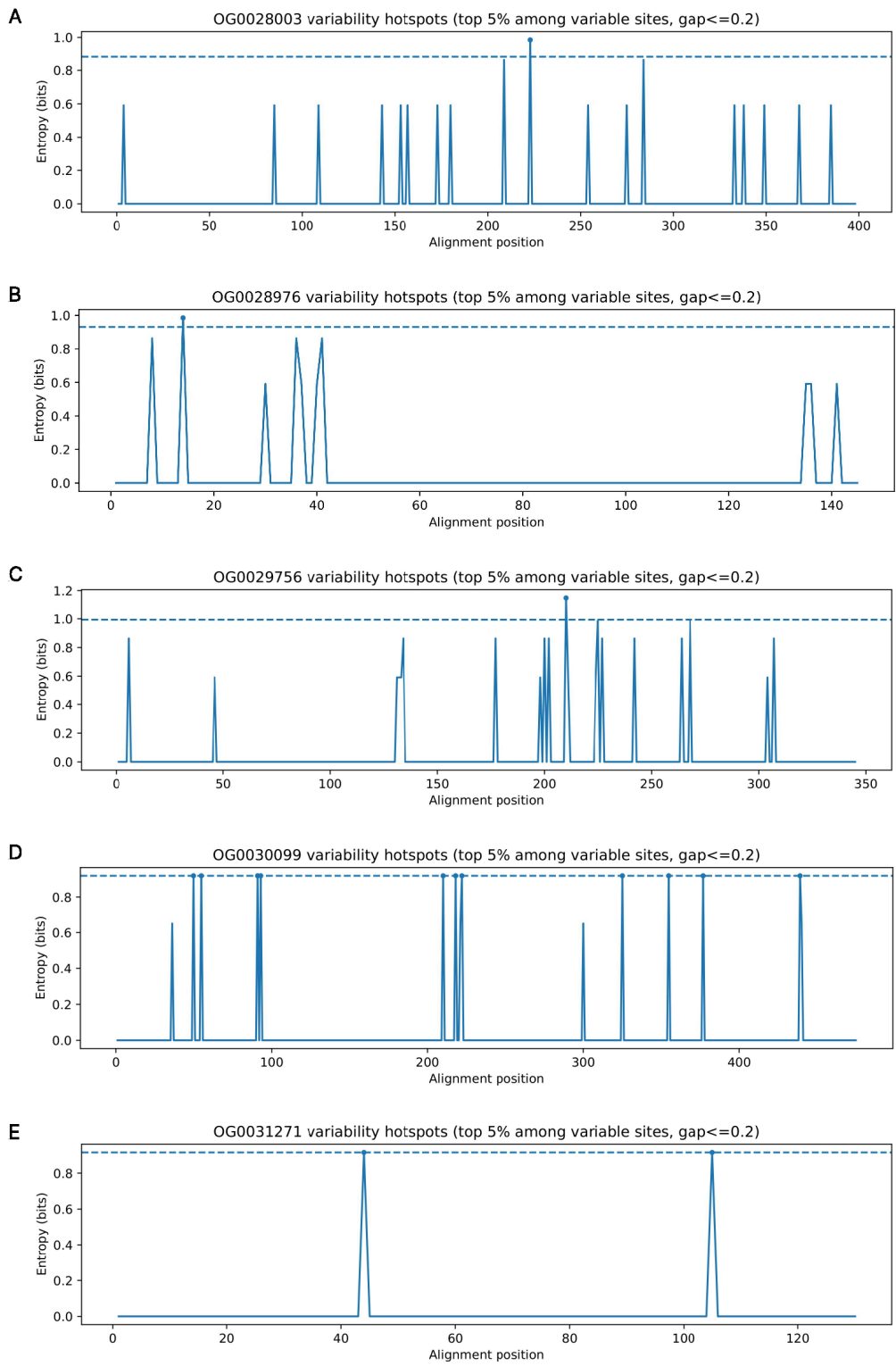


Figure 3

Shannon entropy profiles across multiple sequence alignments for selected orthogroups: (A) OG0028003, (B) OG0028976, (C) OG0029756, (D) OG0030099, and (E) OG0031271. Dashed lines indicate the top 5% most variable sites after excluding gap-rich positions. Peaks represent sequence variability hotspots.

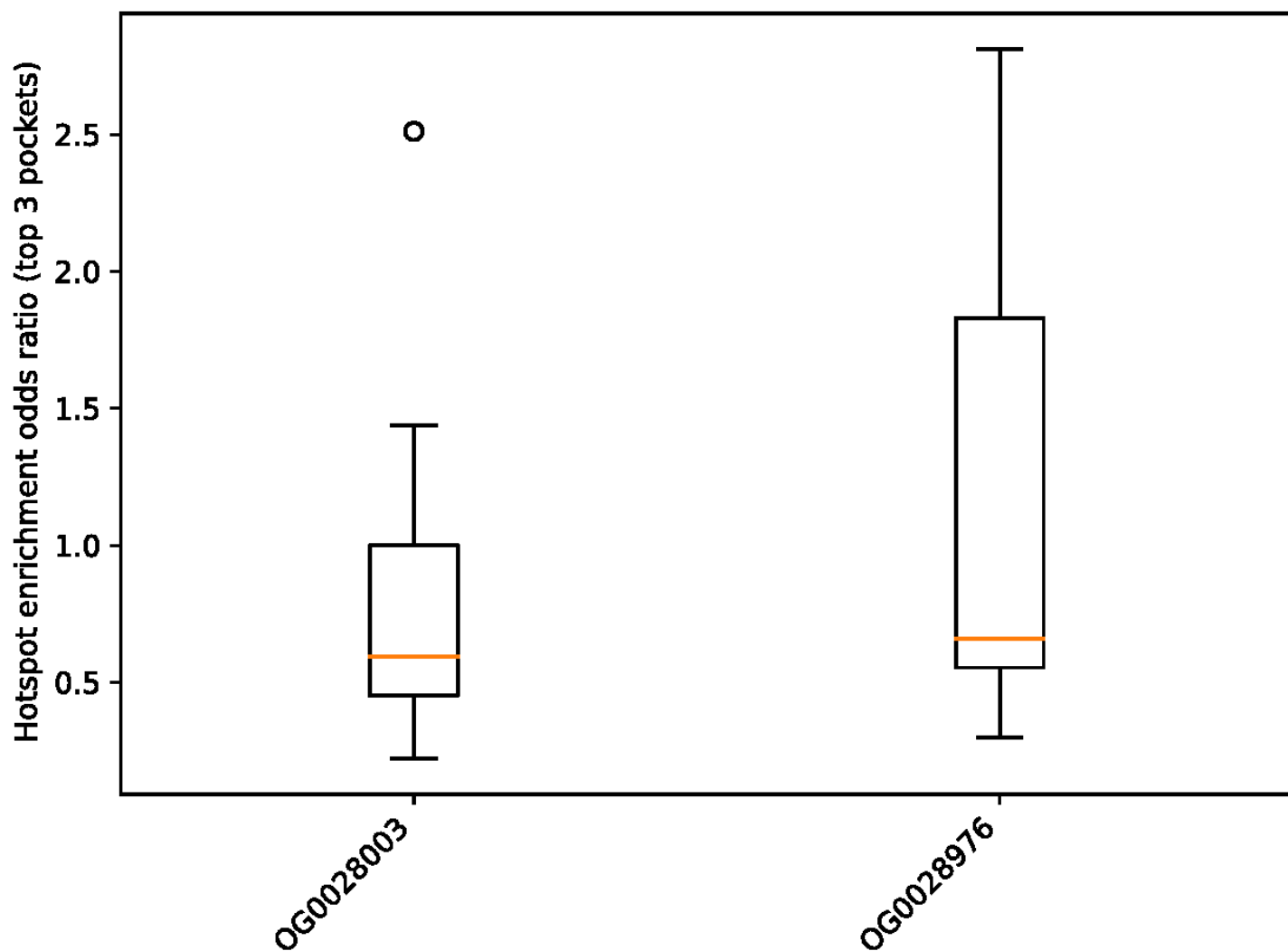


Figure 4

Bar plot showing hotspot enrichment odds ratios for residues under positive selection located within the top three predicted pockets identified by fpocket.

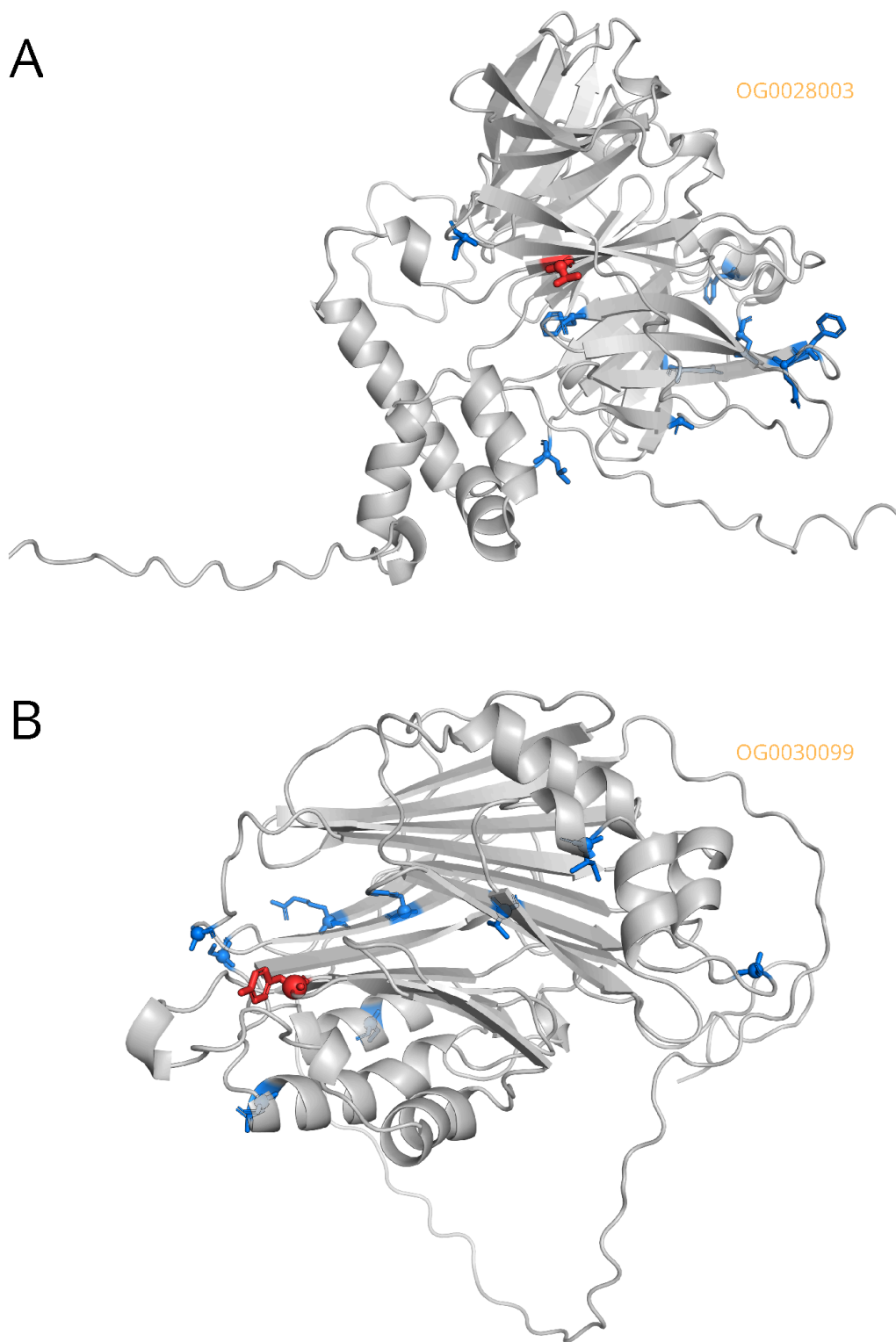


Figure 5

Predicted structures of two representative orthogroups: (A) OG0028003 and (B) OG0030099. Protein backbones are shown in grey, lineage-specific substitutions in blue, and hotspot residues in red. Substitutions are primarily located in surface-exposed regions rather than in structurally constrained cores.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Tables.xlsx](#)
- [SMprotCereus.docx](#)