microPublication BIOLOGY

8/19/2025 - *Open Access*

Sequence Analysis and Modeling of the Repetitive Region of the Long Isoform of Clarinet/CLA-1

Benjamin Hunt[1], Timothy Hunt[1], Zhao Xuan[1§]

[1]School of Biology and Ecology, University of Maine, Orono, Maine, United States

[§]To whom correspondence should be addressed: zhao.xuan@maine.edu

Abstract

The *C. elegans* active zone gene *cla-1* encodes three main isoforms. The long isoform, CLA-1L, functions beyond regulating synaptic vesicle exocytosis, including synaptic vesicle clustering and endocytic sorting of a transmembrane autophagy protein. CLA-1L contains a large N-terminal repetitive region. Sequence analysis indicates that this region is enriched with acidic residues and displays disordered structures interspersed with helical sections. While modeling a portion of the repetitive region revealed dynamic transitions between compact and less compact conformations, modeling the entire region suggests a predominantly extended structure. This study sheds light on how CLA-1L carries out its diverse functions.

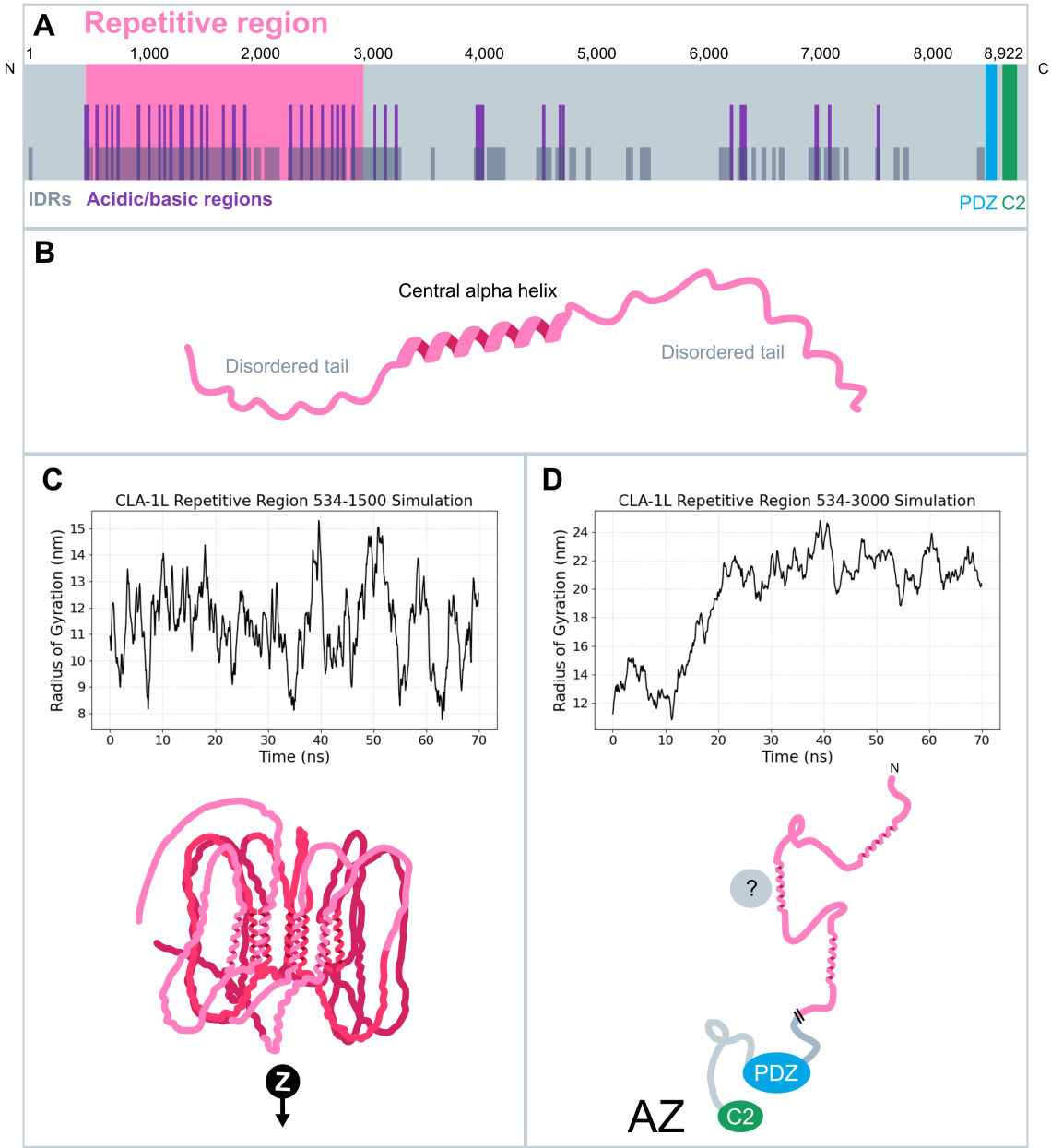

Figure 1. Structural models of the repetitive region of CLA-1L derived from sequence analysis and modeling with AlphaFold 3 and CALVADOS 2:

A. Sequence analysis of CLA-1L indicates that intrinsically disordered regions (IDRs) and acidic/basic regions are enriched in the repetitive region.
B. Each 94-amino acid unit features a central α-helix flanked by intrinsically disordered regions.

C. Simulation of the first ~10 repeats (residues 534–1500) of the CLA-1L repetitive region using CALVADOS 2 (above) and AlphaFold 3 (below). AlphaFold 3 modeling suggests that these ten repeats exhibit favorable interactions between individual α-helices in a parallel arrangement, resulting in a more spatially compact conformation. The CALVADOS 2 simulation suggests alternations between compact and less compact conformations.

D. Simulation of the entire repetitive region using CALVADOS 2 (above) and a hypothetical model of the entire CLA-1L (below). The upward trend in the radius of gyration suggests there is a loose association between some repeats, resulting in a more expanded conformation of the entire repetitive region. We propose a model where the extensive repetitive region enables CLA-1L to adopt a more expanded structure. The intrinsically disordered regions within the repeats may allow for versatile interactions with multiple presynaptic components, promoting multifunctional roles of CLA-1L. This model is supported by spatial localization evidence indicating that the N-terminal region of CLA-1L (which contains the repeats) resides outside the central active zone (Xuan & Colón-Ramos, 2023).

Description

CLA-1 has been identified as an essential active zone protein in *C. elegans*, and contributes to synaptic vesicle clustering and the structural organization of synapses (Xuan et al., 2017). The *cla-1* gene encodes three main isoforms that contain common C-terminal PDZ and C2 domains with homology to the vertebrate active zone proteins Piccolo and RIM (Figure 1A). The large size (8922 amino acids) and an extensive repetitive region at the unique N-terminus make CLA-1L the most enigmatic isoform. While the shorter isoforms are required for active zone assembly and proper synapse development, CLA-1L is essential for synaptic vesicle clustering, supports prolonged neuronal activity, and reduces synaptic depression (Xuan et al., 2017). In addition to a role in facilitating synaptic vesicle exocytosis, CLA-1L has been shown to bridge synaptic domains to regulate the presynaptic sorting of ATG-9, the only transmembrane protein in the autophagy pathway (Xuan & Colón-Ramos, 2023). The large size and extensive repetitive region of CLA-1L hinder traditional methods, such as cloning the genomic region and expressing it as a transgene. Hence, CLA-1L has been challenging to study, resulting in a lack of research on this protein, particularly in the repetitive region.

To provide insight into the structure and function of the repetitive region, we performed sequence analysis and structural modeling. Our analysis revealed that CLA-1L harbors a set of 23 tandem 94-amino acid repeats, which represent almost one-third of the total protein (Figure 1A). We found that intrinsically disordered regions (IDRs) are enriched in the repetitive region—of the 57 total IDRs identified in the CLA-1L sequence, 34 (59.6%) are located entirely or partially within this region. IDRs are typically enriched in polar, acidic, or basic residues, which impede the stable folding of their structure (Uversky et al., 2000). Consistent with IDR characteristics, the repetitive region also showed a strong enrichment in acidic/basic regions—25 out of 38 such regions mapped to the repetitive region. In particular, the acidic residue glutamic acid was overrepresented. IDRs lack a fixed three-dimensional structure, enabling them to interact dynamically with multiple protein partners (Dyson, 2016; van der Lee et al., 2014). These compositional features suggest that the repetitive region may adopt a flexible structure, facilitating diverse protein interactions and potentially contributing to the functional versatility of CLA-1L.

We then used AlphaFold 3 to predict the three-dimensional structures of the first 10 repeats and the full repetitive region of CLA-1L. We further assessed the compactness of both smaller and larger repeat segments using CALVADOS 2, a coarse-grained model designed to simulate the behavior of intrinsically disordered proteins (Tesei & Lindorff-Larsen, 2022). AlphaFold 3 predicted each 94-amino acid repeat unit to adopt a consistent secondary structure, consisting of a central α-helix flanked by two intrinsically disordered tails (Figure 1B). When modeling the first 10 repeats of the repetitive region (residues 534 to 1500 in CLA-1L) using AlphaFold 3, the α-helical cores of adjacent repeats exhibited attractive interactions, leading to a stacked or clustered arrangement in a parallel configuration (Figure 1C and Movie 1). This model was supported by CALVADOS 2 simulations, which showed that the radius of gyration—an indicator of protein compactness—fluctuated, suggesting dynamic transitions between more compact and less compact conformations (Bagewadi et al., 2023; Lobanov et al., 2008). The oscillatory behavior of the radius of gyration suggests that the repeats do not adopt a well-defined three-dimensional structure, consistent with the region being largely composed of IDRs. The CALVADOS 2 simulation of the entire repetitive region (534-3000aa; ~23 continuous repeats with two gaps and two truncated repeats) suggests an extended structure (Figure 1D), as indicated by the upward trend in the radius of gyration. AlphaFold 3 modeling of the entire repetitive region predicted low-confidence structures of α-helical cores distributed in clusters. Since AlphaFold 3 remains fundamentally limited to stable, well-structured proteins that adopt a relatively fixed form (*AlphaFold 3*, n.d.; Riley et al., 2023), we favor the model that the repetitive region adopts a more extended conformation as the repeat number increases (Figure 1D). Furthermore, an extended conformation is more plausible, as we

previously observed that the C- and N-termini of CLA-1L localize to different spatial regions of the presynapse, with the C terminus being concentrated in the active zone and the N-terminus, containing the repetitive region, reaching into the lateral active zone or periactive zone (Xuan & Colón-Ramos, 2023).

Repetitive regions occur in 14% of all proteins and high-incidence repeats are associated with unique eukaryotic functions (Marcotte et al., 1999). One example of a protein with a functionally significant repetitive region is ankyrin-1 (ANK1). It has 24 tandem 33-amino acid ankyrin repeats, which play an important role in linking spectrin-actin base membrane skeletons to the plasma membrane (Gallagher et al., 1997; Lux et al., 1990). Since CLA-1L shares its C-terminus with shorter isoforms, and deletion of the long isoform alone disrupts synaptic vesicle clustering and presynaptic sorting of ATG-9, this suggests that the unique N-terminus of CLA-1L—composed largely of a repetitive region—plays a key role in these processes. Our sequence analysis revealed that the repetitive region of CLA-1L is enriched in IDRs. Thus, IDRs may provide structural features that enable CLA-1L to carry out its diverse functions in synaptic vesicle clustering and the endocytic sorting of ATG-9. First, IDRs facilitate liquid–liquid phase separation, an emerging mechanism underlying various forms of presynaptic compartmentalization, including synaptic vesicle clustering (Choi et al., 2024). A recent study showed that the giant active zone scaffold protein Piccolo (~5000 amino acids) undergoes phase separation to extract synaptic vesicles from synapsin condensates and deliver them to the active zone surface upon Ca^{2+} entry—a mechanism that supports synaptic vesicle transport from the reserve pool to the readily releasable pool (RRP) (Qiu et al., 2024). CLA-1L may maintain synaptic vesicle clustering by regulating phase separation through its IDRs, and its large size may allow it to span distinct synaptic vesicle pools and facilitate synaptic vesicle shuttling between them—critical for sustaining neurotransmission. Second, the structural flexibility of IDRs allows them to modulate their function in response to specific cellular contexts, effectively serving as adaptable hubs that facilitate diverse protein interactions and signaling pathways (Dyson, 2016; van der Lee et al., 2014). In the context of ATG-9 endocytosis, we found that CLA-1L genetically interacts with periactive zone endocytic proteins such as Eps15/EHS-1 and intersectin/ITSN-1. Double mutants exhibit a synergistic effect on the presynaptic mislocalization of ATG-9 (Xuan et al., 2023). The repetitive region of CLA-1L, may regulate ATG-9 endocytosis by interacting with endocytic factors at the periactive zone. The α-helices at the center of each repeat unit may serve as semi-rigid cores that provide modular stability. Depending on the cellular context, these α-helices could also mediate interactions with one another or with other presynaptic proteins, thereby coordinating vesicle retrieval from the reserve pool, sorting of endocytic intermediates, and maintenance of the recycling pool. Overall, the large size of CLA-1L, along with its unique repetitive region, may enable it to serve both a structural role in maintaining synaptic vesicle clustering and a regulatory role in the endocytic sorting of ATG-9 at presynaptic sites. Further studies, such as TurboID-based proximity labeling, will help identify proteins interacting with CLA-1L, shedding light on the mechanisms by which CLA-1L performs its multifunctional roles.

Methods

The protein sequence for the long isoform of Clarinet (**accessionID W6RTA4-1**) was retrieved from the UniProt database, which contains existing annotations (e.g., acid/basic regions, disordered regions), derived from UniProt's annotation pipeline that identifies features by utilizing external tools, including: TMHMM, SignalP, Phobius, Coils, and MobiDB-lite (*UniProt*, 2022). The acid and basic regions were analyzed in Excel to identify the relative composition between different amino acid residues, the percentage of acidic and basic residues in the regions, and the number of aspartic acid and glutamic acid residues in each region. Repetitive region analysis was conducted on the entire sequence using a local installation of T-REKs for the discovery and alignment of novel repeats in sequences using a K-means algorithm (Jorda & Kajava, 2009). The repetitive region annotations for T-REKs were then added to CLA-1L using Geneious Prime (https://www.geneious.com/). These annotations were leveraged to guide targeted structural predictions for sections of CLA-1L with AlphaFold 3. Since the accuracy of AlphaFold 3 decreases significantly over 2000 amino acids, the annotations were used to make predictions for a portion of the repetitive region of CLA-1L. UCSF ChimeraX (Meng et al., 2023) was used to facilitate inspection of individual regions and to create visualization movies. The CALVADOS system was utilized to simulate the repetitive region of CLA-1L using the methods provided by Tesei and Lindorff-Larsen with three replicates for each section (Tesei & Lindorff-Larsen, 2022). Scripts available at: https://github.com/XuanLab123/CLA-1L

Acknowledgements: We thank Samantha Barrick (the University of Maine) and Christine Li (The City College of New York) for comments on the manuscript. We thank the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco for providing free access to UCSF ChimeraX.

Extended Data

Description: A movie created using UCSF ChimeraX to visualize the structural model of the first ten repeats (aa 534-1501) of CLA-1L.. Resource Type: Audiovisual. File: Movie_1_CLA-1L_Alphafold_Structure_534-1501.mp4. DOI: 10.22002/78d7k-v8417

References

AlphaFold 3: Exciting Advance yet Unresolved Major Issues Remain | Deep Origin. (n.d.). Retrieved June 9, 2025, from https://deeporigin.com/blog/alphafold-3-exciting-advance-yet-unresolved-major-issues-remain

Bagewadi ZK, Yunus Khan TM, Gangadharappa B, Kamalapurkar A, Mohamed Shamsudeen S, Yaraguppi DA. 2023. Molecular dynamics and simulation analysis against superoxide dismutase (SOD) target of Micrococcus luteus with secondary metabolites from Bacillus licheniformis recognized by genome mining approach. Saudi Journal of Biological Sciences 30: 103753. DOI: 10.1016/j.sjbs.2023.103753

Dyson HJ. 2016. Making Sense of Intrinsically Disordered Proteins. Biophysical Journal 110: 1013-1016. DOI: 10.1016/j.bpj.2016.01.030

Gallagher PG, Tse WT, Scarpa AL, Lux SE, Forget BG. 1997. Structure and Organization of the Human Ankyrin-1 Gene. Journal of Biological Chemistry 272: 19220-19228. DOI: 10.1074/jbc.272.31.19220

Jorda J, Kajava AV. 2009. T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. Bioinformatics 25: 2632-2638. DOI: 10.1093/bioinformatics/btp482

Lobanov MIu, Bogatyreva NS, Galzitskaia OV. 2008. [Radius of gyration is indicator of compactness of protein structure]. Mol Biol (Mosk) 42(4): 701-6. PubMed ID: 18856071

Lux SE, John KM, Bennett V. 1990. Analysis of cDNA for human erythrocyte ankyrin indicates a repeated structure with homology to tissue-differentiation and cell-cycle control proteins. Nature 344: 36-42. DOI: 10.1038/344036a0

Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. 1999. A census of protein repeats. Journal of Molecular Biology 293: 151-160. DOI: 10.1006/jmbi.1999.3136

Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, Ferrin TE. 2023. UCSF ChimeraX: Tools for structure building and analysis. Protein Science 32: 10.1002/pro.4792. DOI: 10.1002/pro.4792

Riley AC, Ashlock DA, Graether SP. 2023. The difficulty of aligning intrinsically disordered protein sequences as assessed by conservation and phylogeny. PLOS ONE 18: e0288388. DOI: 10.1371/journal.pone.0288388

Tesei G, Lindorff-Larsen K. 2023. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. Open Research Europe 2: 94. DOI: 10.12688/openreseurope.14967.2

UniProt. (2022, December 23). UniProt. https://www.uniprot.org/help/sam

Uversky VN, Gillespie JR, Fink AL. 2000. Why are ?natively unfolded? proteins unstructured under physiologic conditions?. Proteins: Structure, Function, and Genetics 41: 415-427. DOI: 10.1002/1097-0134(20001115)41:3<415::aid-prot130>3.0.co;2-7

van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al., Babu. 2014. Classification of Intrinsically Disordered Regions and Proteins. Chemical Reviews 114: 6589-6631. DOI: 10.1021/cr400525m

Xuan Z, Colón-Ramos DA. 2023. The active zone protein CLA-1 (Clarinet) bridges two subsynaptic domains to regulate presynaptic sorting of ATG-9. Autophagy 19: 2807-2808. DOI: 10.1080/15548627.2023.2229227

Xuan Z, Manning L, Nelson J, Richmond JE, Colón-Ramos DA, Shen K, Kurshan PT. 2017. Clarinet (CLA-1), a novel active zone protein required for synaptic vesicle clustering and release. eLife 6: 10.7554/elife.29276. DOI: 10.7554/eLife.29276

Choi J, Rafiq NM, Park D. 2024. Liquid–liquid phase separation in presynaptic nerve terminals. Trends in Biochemical Sciences 49: 888-900. DOI: 10.1016/j.tibs.2024.07.005

Qiu H, Wu X, Ma X, Li S, Cai Q, Ganzella M, et al., Zhang. 2024. Short-distance vesicle transport via phase separation. Cell 187: 2175-2193.e21. DOI: 10.1016/j.cell.2024.03.003

Xuan Z, Yang S, Clark B, Hill SE, Manning L, Colón-Ramos DA. 2023. The active zone protein Clarinet regulates synaptic sorting of ATG-9 and presynaptic autophagy. PLOS Biology 21: e3002030. DOI: 10.1371/journal.pbio.3002030

Funding: Benjamin Hunt was supported by UMaine CUGR fellowship AY 2024-2025.

Author Contributions: Benjamin Hunt: data curation, methodology, writing - original draft. Timothy Hunt: data curation, methodology, writing - original draft. Zhao Xuan: conceptualization, supervision, writing - review editing.

Reviewed By: Anonymous

Nomenclature Validated By: Anonymous

WormBase Paper ID: WBPaper00068469

History: Received June 20, 2025 **Revision Received** August 1, 2025 **Accepted** August 14, 2025 **Published Online** August 19, 2025 **Indexed** September 2, 2025

microPublication BIOLOGY

8/19/2025 - *Open Access*

Citation: Hunt B, Hunt T, Xuan Z. 2025. Sequence Analysis and Modeling of the Repetitive Region of the Long Isoform of Clarinet/CLA-1. microPublication Biology. 10.17912/micropub.biology.001712