Structure-Aware Annotation of Leucine-rich Repeat Domains

Boyan Xu[1, 2], Alois Cerbu[2], Daven Lim[3], Christopher J Tralie[4], and Ksenia Krasileva[1,3]

[1] *Center for Computational Biology, University of California Berkeley, Berkeley, CA 94720, U.S.A.*

[2] *Department of Mathematics, University of California Berkeley, Berkeley, CA 94720, U.S.A.*

[3] *Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA 94720, U.S.A.*

[4] *Department of Mathematics and Computer Science, Ursinus College, Collegeville, PA, USA*

Abstract

Protein domain annotation is typically done by predictive models such as HMMs trained on sequence motifs. However, sequence-based annotation methods are prone to error, particularly in calling domain boundaries and motifs within them. These methods are limited by a lack of structural information accessible to the model. With the advent of deep learning-based protein structure prediction, existing sequenced-based domain annotation methods can be improved by taking into account the geometry of protein structures. We develop dimensionality reduction methods to annotate repeat units of the Leucine Rich Repeat solenoid domain. The methods are able to correct mistakes made by existing machine learning-based annotation tools and enable the automated detection of hairpin loops and structural anomalies in the solenoid. The methods are applied to 127 predicted structures of LRR-containing intracellular innate immune proteins in the model plant *Arabidopsis thaliana* and validated against a benchmark dataset of 172 manually-annotated LRR domains.

Author summary

In immune receptors across various organisms, repeating protein structures play a crucial role in recognizing and responding to pathogen threats. These structures resemble the coils of a slinky toy, allowing these receptors to adapt and change over time. One particularly vital but challenging structure to study is the Leucine Rich Repeat (LRR). Traditional methods that rely just on analyzing the sequence of these proteins can miss subtle changes due to rapid evolution. With the introduction of protein structure prediction tools like AlphaFold 2,

annotation methods can study the coarser geometric properties of the structure. In this study, we visualize LRR proteins in three dimensions and use a mathematical approach to ‘flatten’ them into two dimensions, so that the coils form circles. We then used a mathematical concept called winding number to determine the number of repeats and where they are in a protein sequence. This process helps reveal their repeating patterns with enhanced clarity. When we applied this method to immune receptors from a model plant organism, we found that our approach could accurately identify coiling patterns. Furthermore, we detected errors made by previous methods and highlighted unique structural variations. Our research offers a fresh perspective on understanding immune receptors, potentially influencing studies on their evolution and function.

Introduction

Solenoid domains are a class of protein structures defined by a repeating helical arrangement of their backbone chain. These domains are found in a diverse range of proteins and play important roles in a variety of biological processes, including protein-protein interactions, molecular recognition, and scaffolding [1]. The coil shape of solenoid domains arises from a repeating motif of amino acid residues, known as *tandem repeat units*. The specific amino acid sequence and length of the repeating unit can vary between solenoid domains, resulting in differences in the overall structure and function of the domain. The modular nature of solenoid domains allows for the construction of complex structures by combining different domains in a predictable and controlled manner [2].

Leucine-rich repeat (LRR) domains are a type of curved solenoid domain with repeated units of about 20 - 30 residues long which contain leucine residues in a beta-strand conformation. These domains are found in a wide range of proteins, including cell surface receptors, enzymes, and structural proteins, and are known to play important roles in protein-protein interactions, signal transduction, and immune recognition [3].

Leucine-rich repeats play a critical role in the function of the NOD-like receptor (NLR) family of proteins in the innate immune system of plants and animals [4]. NLRs are intracellular immune receptors that recognize pathogen-derived molecules and activate downstream signaling pathways to initiate an immune response. NLRs are involved in the recognition of a wide range of pathogens, including bacteria, fungi, and viruses. NLRs typically consist of three domains: an N-terminal domain, a central nucleotide-binding domain, and a C-terminal LRR domain. The LRR domain is responsible for recognizing and binding to pathogen-derived molecules, such as effector proteins or pathogen-associated molecular patterns (PAMPs) [5]. In particular, the LRR domains of plant NLRs are highly diverse and can recognize a wide range of pathogen-derived molecules, allowing plants to mount a robust and specific immune response to a broad range of pathogens. Understanding LRR domains in plant NLRs is important for developing strategies to enhance plant immunity and improve crop resistance to pathogens.

The concave surface of the leucine-rich repeat domain is generally responsible for binding to ligands [6]. The amino acid residues on the concave surface of the LRR domain form a specific pattern of hydrophobic, polar, and charged residues that can interact with specific ligands, such as proteins, peptides, carbohydrates, or nucleic acids. The specificity of ligand binding by LRR domains is determined by the overall shape and chemical properties of the concave surface, which can be highly variable between different LRR-containing proteins [7][8]. Additionally, LRR domains can contain variable regions and insertions that can modify the binding specificity and affinity of the domain. More recently, studies such as [9] have shown that "post-LRR" domains which lie at the C-terminal end of the LRR are required for successful plant immune response. Accurate annotation of these domains and their constituent repeat units is thus essential to understanding the components which govern protein shape and binding specificity.

Existing methods for annotating LRR domains give unreliable and inconsistent results due to irregularities in sequence motifs. Profile hidden Markov models (HMMs) are widely used, e.g. by HMMER [10], to annotate protein domains in genomic sequences, but they are sensitive to the size and diversity of the protein family being analyzed and do not perform accurately for rapidly-evolving, highly-divergent families such as LRR [11]. Profile HMMs are also unable to delineate tandem repeat units.

An existing tool, LRRPredictor [12], uses an ensemble of 8 machine learning classifiers to determine the residues which comprise the basic LRR motif of the form "LxxLxL" (where "L" refers to Leucine or other hydrophobic amino acid, and "x" can be any amino acid). We found that LRRPredictor often makes mistakes, particularly in identifying divergent motifs near the C- and N-terminal boundaries of the LRR. Because LRRPredictor, like an HMM, is trained on a specific set of LRR sequences taken from Protein Data Bank [13] (PDB), it incorrectly annotates LRR sequences which diverge from its training set.

With AlphaFold 2 [14], a deep-learning-based model, reliable protein structure prediction has become readily available, enabling domain annotation methods with direct access to geometric data from the protein. We leverage this geometric information to annotate essential features of the LRR domain: start/end position, post-LRR detection, repeat unit delineation, and structural irregularities.

From the perspective of differential geometry, a coiling curve in 3D space is characterized by a linearly increasing winding number around a core curve. We therefore detect the coiling LRR region, as the loci where the winding number is sufficiently close to a line of a fixed slope; the post-LRR domain is then decided as C-terminal sequence downstream from the point at which steady winding terminates. The methods section below describes our procedure for computing the winding number across the length of the protein. In contrast to HMM-based or other data driven techniques, our method is completely unsupervised and driven by simple mathematical methods.

Methods

LRR domain annotation via winding number calculation ~~on 2D projection of normal bundle embedding~~

161 NLR protein sequences, i.e. *NLRome*, were obtained from *A. thaliana* Col-0 accession reference proteome. Of these 161 NLRs, 127 had Alphafold-predicted structures available on AlphafoldDB [14][15]. For each structure we extracted amino acid α-carbon positions to obtain a 3D backbone space curve.

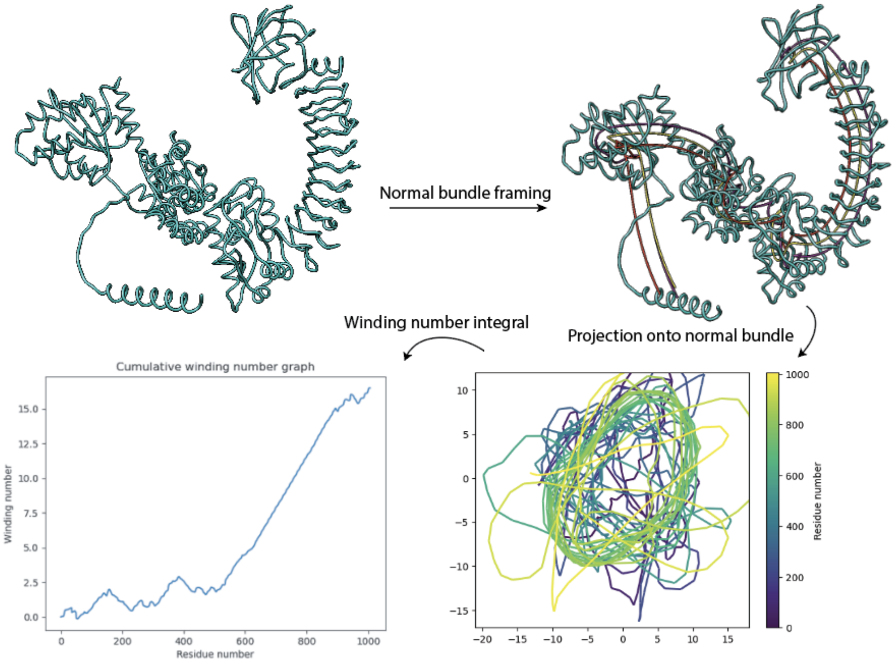

Figure 1: Embedding of protein backbone curve into normal bundle followed by projection onto an orthonormal frame yields a 2D curve containing a flattened slinky shown in lower right. The cumulative winding number, computed using the classic formula from calculus, is computed from the projection. Sloped linear segments of the winding number curve indicate coiling.

To extract the LRR region from the rest of the protein, we leverage the solenoid geometry of the LRR domain in contrast to the post-LRR structure. The non-LRR regions do not wind regularly as the LRR does, so we annotate the LRR by computing the winding number of the coil around the core of the solenoid.

Obtaining the backbone. ~~We obtain the core curve of the coil by applying a Gaussian filter to the backbone 3D space curve.~~ To obtain the backbone of the coil, we first smooth the embedded residue sequence to suppress the loops by using a Gaussian filter. Let t be the residue number of the protein sequence and $\vec{\gamma[t]}$ denote the the residue sequence embedded in 3-dimensional Euclidean space, where each component $\gamma_x[t], \gamma_y[t], \gamma_z[t]$ is a discrete time series. Furthermore, let g_σ be the Gaussian function with standard deviation σ and mean 0, i.e.

$$g_\sigma[t] = \frac{1}{\sqrt{2\pi}\sigma} e^{-t^2/(2\sigma^2)}. \tag{1}$$

We then sample g_σ discretely at indices t between -4σ and 4σ, and we compute the smoothed sequence $\gamma_{i\sigma}$ for the i^{th} component as the discrete convolution $\gamma_i * g_\sigma$, where $*$ denotes the discrete convolution operator; i.e. the convolution $f * g$ of two discrete functions f and g is given by

$$(f * g)[t] = \sum_{j=-\infty}^{\infty} f[t]g[t-j]. \tag{2}$$

For the core curve, we compute $\vec{\gamma_{20}}[t] = (\gamma_{x20}[t], \gamma_{y20}[t], \gamma_{z20}[t])$.

Tangent computation. ~~We then compute the tangent vector field of the core curve via a Gaussian derivative, i.e. applying a derivative Gaussian filter by convolving each component $\gamma(t)$ of the core curve~~ We now compute the tangent bundle $\vec{\gamma'_\sigma}[t]$ by computing component-wise derivatives of the original backbone $\vec{\gamma[t]}$. To numerically compute derivatives on such discrete data, we convolve each of them with the *derivative* of a Gaussian, $g'_{\sigma[t]}$, to obtain the derivatives $\gamma'_{i\sigma}$, where $g'_{\sigma[t]}$ is

$$g'_\sigma[t] = -\frac{t}{\sqrt{2\pi}\sigma^3} e^{-t^2/(2\sigma^2)} \tag{3}$$

~~of the Gaussian g_σ via the formula $\gamma * g'_\sigma$ (in our case $\sigma = 1$)~~ where, in practice, we use $\sigma = 1$ to compute $\vec{\gamma'_1}$. By associativity of convolution and the derivative, this way of computing $\gamma'_{i\sigma}$ is equivalent to computing the derivative of $\gamma_{i\sigma}$ smoothed once more by $\gamma * g'_\sigma$. Such a smoothed derivative is a standard trick [16] for computing numerically meaningful derivatives of discrete time series which, from a naive continuous point of view, consist of a bunch of step functions with zero derivative everywhere except at the boundaries between time indices, where the derivative is undefined.

Normal computation and orthonormal framing. Once we have the components of the tangent vector at each residue, we obtain the normal bundle ~~The normal bundle is obtained~~ as the orthogonal complement to the tangent vector field. The backbone curve is embedded into the normal bundle, which is then projected onto the Euclidean plane $\mathbb{R}^2$ by parallel transport of an orthonormal frame along the core curve. Our algorithm for orthonormal framing is as follows:

Parallel transport algorithm for orthonormal framing

1. Complete

2. Randomly initialize a pair of 3D vectors perpendicular to initial tangent vector of core curve. Form the 3×2 matrix A_0 with these two vectors as columns.

3. Given A_t, to obtain A_{t+1}, we project the columns of A_t onto To obtain the next orthonormal frame, project the current orthonormal frame onto the next normal plane. The resulting projection is likely not orthogonal; we replace it with the closest pair of orthonormal vectors (i.e., the 3×2 matrix with orthonormal columns closest in the Frobenius norm). This can be done by computing the singular value decomposition of the 2×3 matrix and replacing its singular values with 1's (the standard solution to the "Orthogonal Procrustes Problem" [17, 18]).

4. Repeat step 2 along the length of the core curve.

The parallel transport algorithm enables the identification of normal bundle fibers and thus projection onto a single 2D plane as shown in Figure 1. The winding number $w(s)$ up to residue s is then computed using the formula:

$$w(s) := \frac{1}{2\pi}\int_0^s d\theta = \frac{1}{2\pi}\int_0^s \frac{1}{x^2+y^2}\left(x\frac{dy}{dt} - y\frac{dx}{dt}\right) dt \tag{4}$$

where $x(t)$ and $y(t)$ are the coordinates of the normal bundle projection as t ranges across the residues of the NLR protein sequence. We compute a discrete version of formula 4 by once again approximating $\frac{d}{dt}$ with a Gaussian derivative ~~and integral with cumulative sum~~, followed by a cumulative sum to approximate the integral in a discrete setting.

Region identification. On the resulting cumulative winding number function we perform a least-squares piecewise linear regression with three pieces: the first being the best horizontal line, the second a sloped line of best fit, and third a horizontal line. The regressed function is similar to a ReLU with threshold or "clipped ReLU," except that it need not be continuous at breakpoints. The horizontal lines flanking the sloped line represent regions of the NLR which are not LRR and do not have gradually increasing winding number. Therefore, the breakpoints in the regression estimate the start and end coordinates of the LRR domain. A plot of the winding number graph and regression, along with a comparison with HMM-based domain annotation from InterPro, is shown in Figure 2.

Four-breakpoint regression detection of hairpin loops or mis-folding in LRR domain

A small number (9 out of 127) of proteins contain hairpin loops or other structural anomalies in the LRR domain which interfere with the clipped ReLU

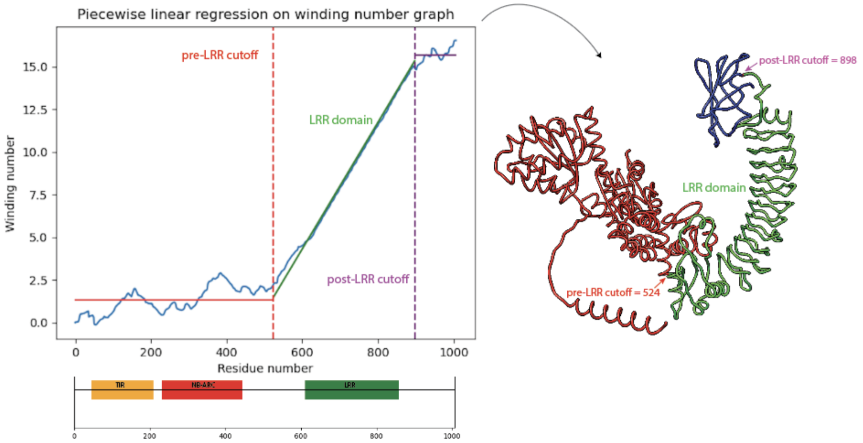

Figure 2: A discontinuous clipped ReLU function is regressed on the graph of the winding number function for *A. thaliana* NLR with TAIR [19] ID AT3G44400.1. The breakpoints of the regression yields the start and end positions of the LRR domain, highlighted in green. InterPro [20] domain annotations are shown below regression plot.

regression, thereby necessitating a piecewise linear regression with more pieces. We detect such anomalous structures by computing the standard deviation of the residuals vector for the sloped section of the 2-breakpoint regression. A large residual standard deviation indicates a break in sloping of the winding number and thus a low-confidence, non-coiling region in the protein structure. For these proteins, we run a similar piecewise linear regression with four breakpoints instead of two, registering a short noncoiling region along the otherwise solenoidal domain, represented by a short horizontal line in the piecewise-linear regression. This short horizontal line represents a large hairpin, insertion, or mis-folding within the LRR domain. See Figure 3.

~~Eigenvectors of graph Laplacian on mutual nearest neighbors yield solenoid phase estimation~~Solenoid phase estimation

In the previous sections, we used piecewise linear regression on the cumulative winding number to isolate the LRR domain. In this section, we seek to obtain an angular coordinate, or *phase estimation*, on the LRR domain sequence. We first compute smoothed tangent vectors on the truncated LRR solenoid curve by convolving the position of the curve with the derivative of a Gaussian to obtain a tangent vector field, just as we did with the region annotation. To accentuate periodic features in the coil [21], we perform a sliding window embedding of window size 24 (roughly the length of the LRR period) with delay time 1 on each component of the tangent vector field~~, using the formula~~The formula for

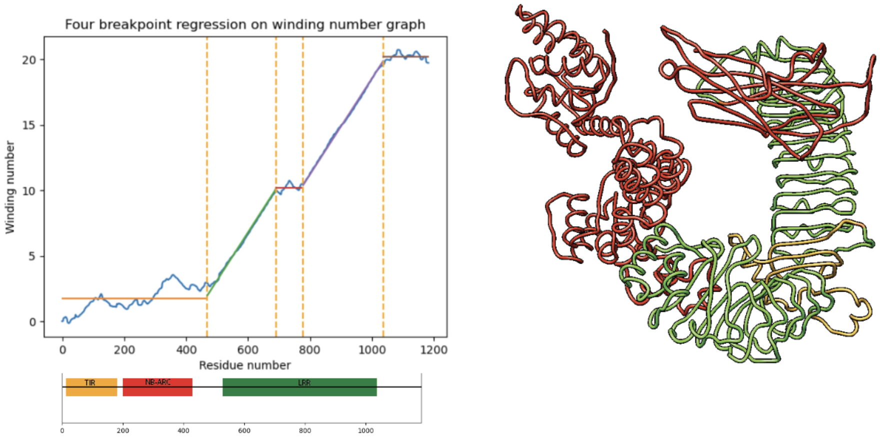

Figure 3: Four breakpoint piecewise linear regression enables detection of a non-coiling structure (highlighted in yellow at right) which deviates from the usual coiling in the LRR domain. Below regression plot, HMM-based InterPro domain annotations fail to detect non-coiling region within LRR domain. TAIR ID is AT1G72840.2.

such a sliding window embedding of some sequence $f[t]$ is

$$\mathrm{SW}_1^{24} f[t] := \begin{bmatrix} \gamma[t] \\ \gamma[t+1] \\ \gamma[t+2] \\ \vdots \\ \gamma[t+24] \end{bmatrix} \in \mathbb{R}^{24+1} \tag{5}$$

~~applied to each of the 3 components of the tangent vector field~~We concatenate together $\mathrm{SW}_1^{24}\,\gamma'_{i\sigma}$ for each of the three components of the tangent vector $\gamma'_{i\sigma}$, resulting in a ~~curve~~ sequence in 75-dimensional Euclidean space. We then construct a 50-mutual-nearest-neighbors graph on the sliding window embedding.

From the mutual-NN graph we compute leading eigenvectors of the unweighted graph Laplacian[22], shown in Figure 4. Intuitively, the graph Laplacian is a generalization of a discrete second derivative operator to graphs. For the same reason that sines and cosines are eigenfunctions of the second derivative operator with associated eigenvalues proportional to the frequency, eigenvectors of the graph Laplacian on a graph of a circle are sine/cosine pairs, up to a phase, that go through an integer number of cycles over one revolution of the circle, and lower frequency pairs have smaller eigenvalues [23]. We expect a near circular graph in the mutual-NN graph in the periodic LRR region, and the Laplacian eigenvectors are known to degrade gracefully in the presence of imperfections. Therefore, we expect the two eigenvectors with the smallest eigenvalue to be approximately periodic and $\pi/4$-phase shifted. If we use the two entries of these eigenvectors as x and y coordinates, respectively, we obtain a projection of the LRR coil onto a circle winding in the plane. Our phase estimation θ along the LRR coil is simply obtained as $\theta = \tan^{-1}(\frac{y}{x})$, as shown in Figure 4 below.

We note that a similar phase-estimation scheme with the graph Laplacian of mutual nearest neighbors has been used to order photographs along a loop [24] and to parameterize periodic videos [21]. Furthermore, a spiritually similar but more computationally intensive topological phase estimation based on cohomology [25, 26] has been applied to motion capture data [27] and to recovering phase based on head orientation from neural data[28].

~~Generically, the leading pair of 0th and 1st eigenvectors of the graph Laplacian are out-of-phase periodic functions with frequency matching the expected frequency of the LRR coil (Figure 4), and thus yield projection onto principal axes perpendicular to the core of the coil (Figure 4). Given these x and y coordinates, we compute phase estimation θ to obtain a phase estimation along the LRR coil as shown in Figure 4 below.~~

Results

Cumulative winding number reveals errors made by ML-based LRR repeat unit delineator

We ran the LRR annotation tool LRRPredictor [12] on the 127 NLRs from *A. thaliana* to obtain predicted locations of the LRR motif “LxxLxL”. Let $R_1, \ldots, R_k$ denote the starting residues for the LRR motifs predicted by LRRPredictor. The analogous measurement in our model is to record the residues at which our cumulative winding number w crosses integers.

To compare the two prediction schemes, we evaluate our cumulative winding number at the residues returned by LRRPredictor. That is, we form the list of numbers $(w(R_1), \ldots, w(R_k))$. If the models are in agreement, the running difference $(w(R_2) - w(R_1), \ldots, w(R_k) - w(R_{k-1})$ should equal the all-ones vector $(1, \ldots, 1)$ (that is, the structure should wind exactly once around the core between residues R_j and R_{j+1}). The “discrepancy”

$$D(R_1, \ldots, R_k) := \sqrt{\sum_{j=1}^{k} (w(R_j) - w(R_{j-1}) - 1)^2} \tag{6}$$

quantifies the extent to which this is not the case. A number of LRRPredictor outputs contained false predictions in which consecutive motif start sites R_j and R_{j-1} appear close together—often only a couple residues apart. Such duplicate predictions result in a high discrepancy $D(R_1, ..., R_k)$ because the difference $w(R_j) - w(R_{j-1})$ as computed in formula (6) above is close to 0.

To test the validity of our winding number computation, we ran the discrepancy computation on the LRRPredictor outputs on the 127 *A. thaliana* reference proteome NLRs as well as the training dataset for LRRpredictor, a manually-annotated “ground truth” dataset of LRR motifs on 172 experimentally-derived

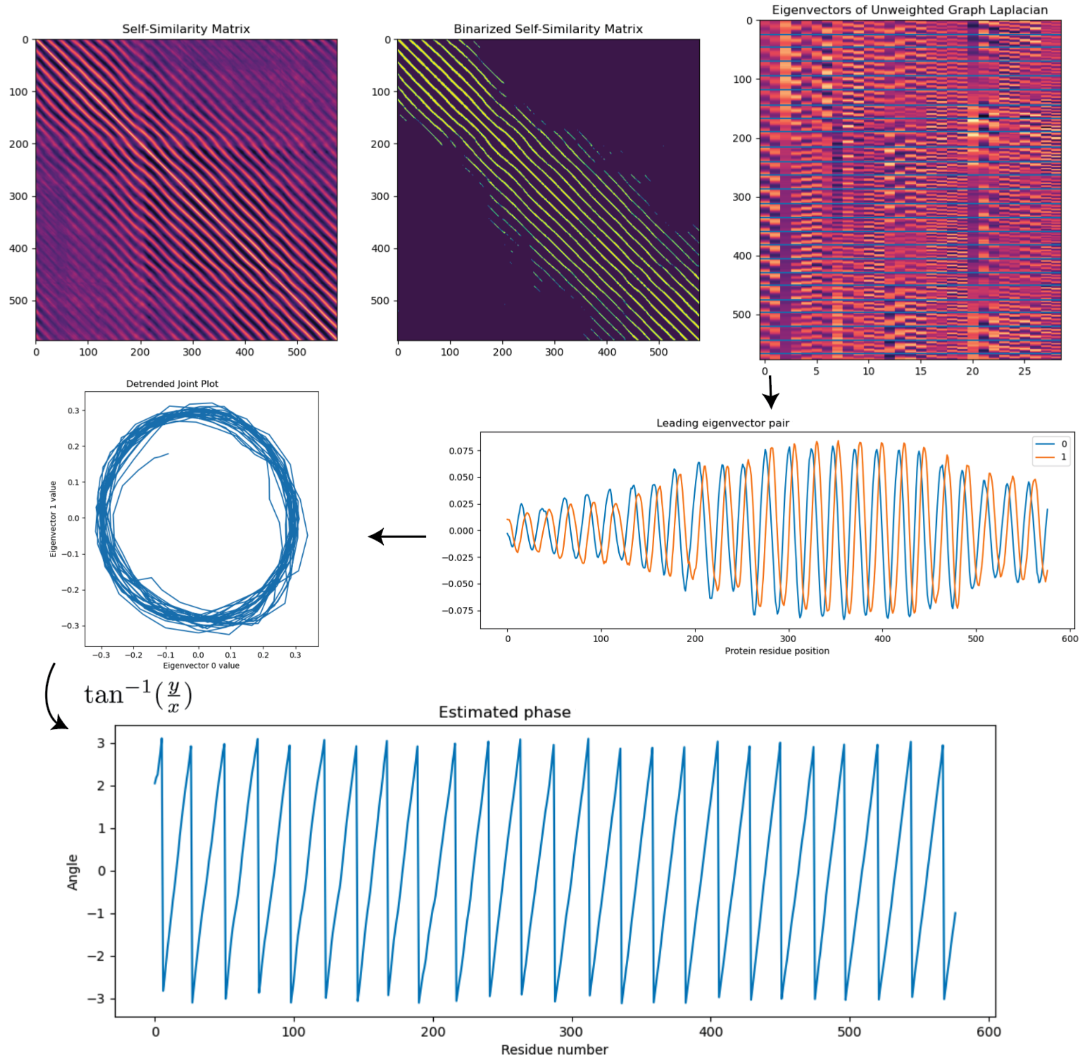

Figure 4: Graph Laplacian eigenvectors of mutual nearest neighbor graph on LRR solenoid curve tangent vectors. LRRPredictor residues are shown as blue horizontal lines on eigenmatrix plot. The 0^{th} and 1^{st} eigenvectors have period matching the expected period of the solenoid as determined by LRRPredictor. Leading eigenvectors of graph Laplacian are periodic and are $\pi/4$-phase shifted, thereby yielding projections of LRR coil onto a winding around a circle in a 2D-plane. Phase estimation using the formula $\theta = \tan^{-1}(\frac{y}{x})$ of LRR coil at bottom taking values between $-\pi$ and π.

LRR structures taken from Protein Data Bank. These PDB protein structures were derived from a diverse set of organisms comprising bacteria, fungi, plants, and animals.

We found consistently low discrepancy values for the ground truth set with mean 0.127. By comparison, *A. thaliana* NLRome discrepancy values were generally low with mean 0.373, but exhibited higher values in cases where LRRpredictor made mistakes. Figure 5 below shows a pair of overlaid histograms comparing discrepancy values for both the validation dataset and NLRome dataset (S1 and S2 Table). The discrepancy values are much lower on the LRRPredictor ground truth dataset compared to the NLRome dataset, implying that our technique makes fewer mistakes than LRRPredictor does on new data. Figure 6 demonstrates how the discrepancy is able to catch duplicate motif predictions made by LRRPredictor. These results demonstrate not only the winding number's ability to accurately model the LRR coil, but also its generalizability to non-NLR LRR's derived from species other than *A. thaliana*.

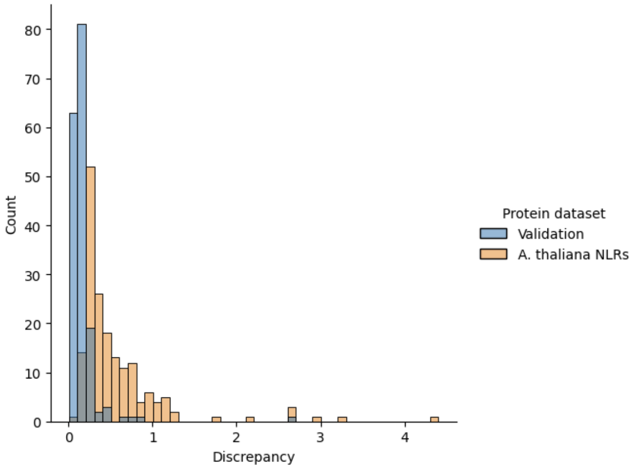

Figure 5: In yellow, histogram of discrepancies for LRRpredictor outputs on 127 *A. thaliana* NLRs showing a large peak around the mode. In blue, a histogram of discrepancies for manually-annotated LRR repeat units used as the training set for the LRRpredictor model. This ground truth dataset produces low overall discrepancy compared to LRRpredictor model outputs, thereby demonstrating the ability of the cumulative winding number computation to faithfully recapitulate the periodicity of the LRR coil.

Structural anomaly detection by sliding window L2 distance from Laplacian eigenvector winding number to line

Many LRR coils have hairpin loops and other structural anomalies which deviate from coiling. In these anomalous regions, the leading eigenvectors deviate from

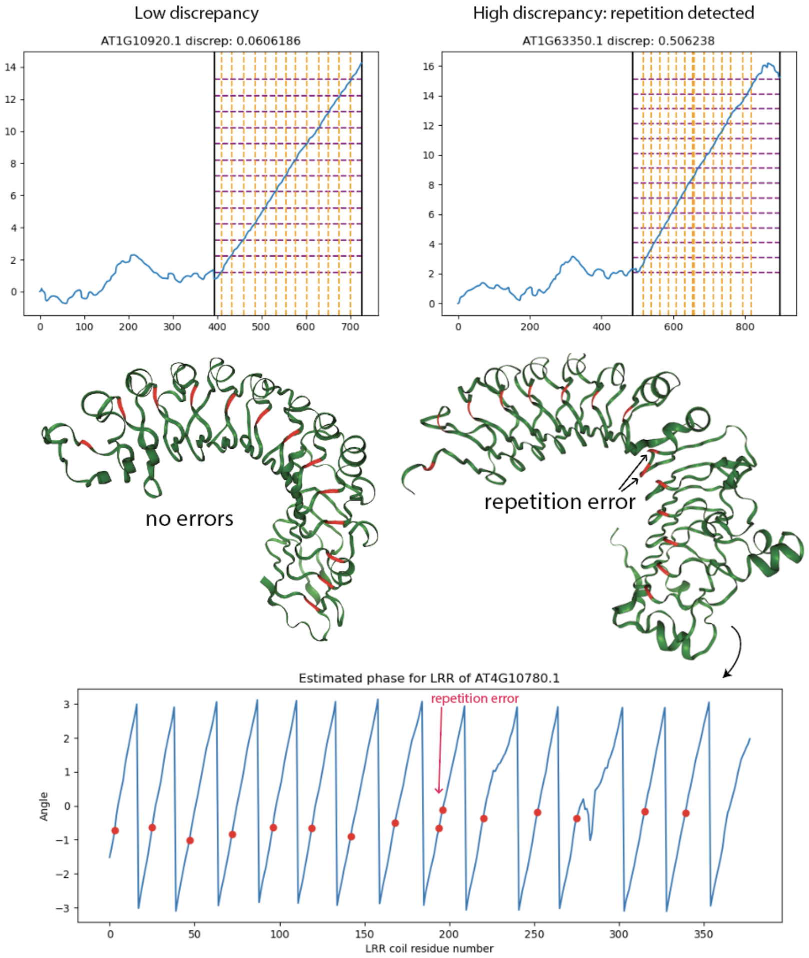

Figure 6: **LRRPredictor discrepancy computation reveals proteins with erroneously repeated predictions**. NLRs with high-discrepancy LRRPredictor outputs tend to carry repetition errors or missing motif annotations. Orange vertical lines overlaid on winding number plot depict LRRPredictor residues, while purple horizontal lines depict the integer-spaced grid which best approximates the winding number graph evaluated at LRRPredictor residues. A repetition error can be seen in the grid representation as a doubled orange line around residue 685. At bottom, LRRPredictor residues are mapped onto graph Laplacian eigenvector phase estimation, revealing an pair of duplicates with adjacent phase.

their usual periodic behavior. Applying the winding number formula (4) above to the pair of leading graph Laplacian eigenvectors leads to a cumulative winding number within the LRR domain which is better able to discern small hairpins compared to the previous winding number computation based on normal bundle projection. As shown in Figure 7 below, we detect a small hairpin as a spike in L2 distance between the winding number and its median slope.

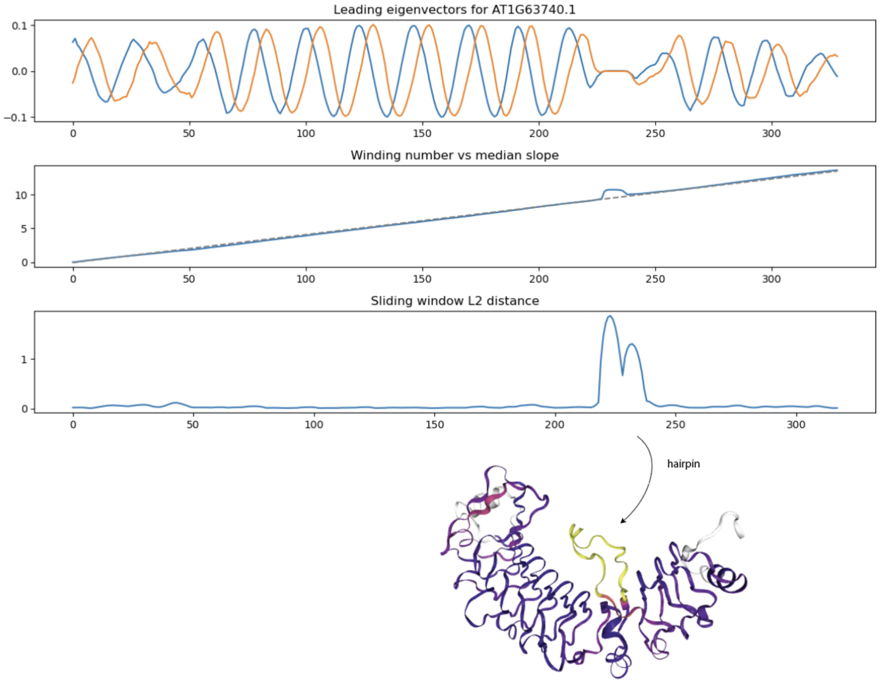

Figure 7: Sliding window L2 distance (SWL2D) from winding number to median secant line detects small hairpins/insertions in LRR coil domain. Structure at bottom is colored according to SWL2D where yellow values are higher.

Discussion

The emergence of AlphaFold 2 has catalyzed a paradigm shift in protein structure prediction, facilitating access to genome-wide high-quality structural predictions. Traditional sequence homology-based domain annotation techniques, like LRRPredictor, often face challenges with LRRs, especially in proteins with high divergence and significant mutations. While evolutionary divergence might veil the sequence homology of LRR units, their core structural topology, characterized by 20-30 amino acid stretches typically involved in protein-protein interactions, often remains conserved, acting as a distinct structural signature.

This study uses AlphaFold 2 to generate a 3D space curve from a protein sequence, which subsequently is projected into the 2D plane by identifying a series of “slinky” cross-sections. Through computing the cumulative winding number on the resultant 2D curve and employing piecewise linear regression, the linearly sloped region, identified as the LRR domain, is discerned. Our method pivots on the nuanced application of geometric data analysis to illuminate structural motifs that remain elusive to sequence analysis alone.

The use of geometric and topological concepts in our method aligns with previous studies that have explored Topological Data Analysis (TDA) in protein structure and dynamics [29, 30]. For instance, SINATRA Pro has been used to identify biophysical signatures in protein dynamics by detecting topological differences between protein structures [29]. Similarly, TopologyNet integrates TDA with deep learning for biomolecular property predictions [30]. Our approach builds on these foundational ideas by leveraging large-scale AI/ML-derived databases like AlphaFoldDB, showcasing the potential of combining AI-based structural predictions with geometric and topological analyses for advanced domain annotation.

Our method yields several kinds of precise results: a) it identifies the start and end sites of the LRR domain with greater accuracy than HMM-based methods, b) it annotates repeat units more reliably than the existing LRRPredictor, c) it identifies misannotations by other annotation/prediction tools, and d) it reveals structural anomalies within the LRR domain that deviate from conventional coiling behaviors. These findings not only underscore the utility of our approach but also present a robust framework for delving into the intricate structural patterns intrinsic to LRR domains.

Our methods are general enough to adapt well to the detection of other solenoid domains. The outcomes from our approach serve as a foundation for structure-guided annotation of proteins containing LRR or other solenoid domains, which are often elusive to HMMs. In the broader context, our findings offer a comparative lens through which the evolution and function of NLRs, including repeat shuffling, can be scrutinized across various lineages [31]. Moreover, the precision and reliability of our annotation methodology can potentially serve as a catalyst for propelling research in disease resistance across a spectrum of plant species by furnishing detailed insights into the structure and function of LRR domains in NLR proteins.

While we benchmarked our work on LRR domains in NLR proteins, the intrinsic methodology has the capacity for broader applications, extending to other solenoid protein domains like armadillo (ARM), tetratricopeptide (TPR), and ankyrin (ANK) repeats, all of which feature distinctive repeat sequences and structural configurations. The amalgamation of advanced protein structure prediction technologies and nuanced mathematical models, as demonstrated in our approach, underscores the potential for widening our understanding of protein function across varied biological systems.

Our method does come with limitations. For instance, while it can detect non-coiling structural anomalies within the LRR domain, the origin, authenticity, and potential functionality of these regions remain ambiguous. Moreover, our structure-based annotation method, albeit effective for domains with a straightforward geometric description like LRRs, might not be universally applicable to other protein domains without developing a new geometric model tailored to them. This underscores a potential limitation when juxtaposing sequence-based versus structure-based domain annotation, highlighting a future avenue warranting exploration: developing geometric models for other protein domains.

Declarations

Acknowledgements

We thank Daniil Prigozhin for running LRRPredictor on *A. thaliana* NLRome, Chandler Sutherland and the Krasileva Lab for providing feedback and suggestions on this project and resulting manuscript.

Funding

BX has been supported by Department of Energy Computational Sciences Graduate Fellowship grant number DE-SC0020347. KVK has been supported by funding from the Innovative Genomics Institute (https://innovativegenomics.org/), the Gordon and Betty Moore Foundation (https://www.moore.org/), grant number 8802, and the National Institute of Health New Innovator Director's Award, grant number DP2AT011967. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Code availability

A Jupyter notebook for running the winding number LRR annotator is available at `https://github.com/amcerbu/LRR-Annotation/tree/main`

Author contributions

BX contributed to conceptualization, data curation, formal analysis, methodology, investigation, software, validation, visualization, and writing. AC contributed to formal analysis, methodology, software, visualization, and writing. DL contributed to data curation, formal analysis, investigation, methodology, and validation. CJT contributed to formal analysis, methodology, software, visualization, validation, and writing. KK contributed to conceptualization, funding acquisition, supervision, and writing. BX wrote original draft of manuscript and other authors contributed to revisions.

References

[1] Jang H, Stevens P, Gao T, Galperin E. The leucine-rich repeat signaling scaffolds Shoc2 and Erbin: cellular mechanism and role in disease. The FEBS journal. 2021;288(3):721-39.

[2] Park K, Shen BW, Parmeggiani F, Huang PS, Stoddard BL, Baker D. Control of repeat-protein curvature by computational protein design. Nature structural & molecular biology. 2015;22(2):167-74.

[3] Ng A, Xavier RJ. Leucine-rich repeat (LRR) proteins: integrators of pattern recognition and signaling in immunity. Autophagy. 2011;7(9):1082-4.

[4] Jones JD, Vance RE, Dangl JL. Intracellular innate immune surveillance devices in plants and animals. Science. 2016;354(6316):aaf6395.

[5] Tamborski J, Krasileva KV. Evolution of plant NLRs: from natural history to precise modifications. Annual review of plant biology. 2020;71:355-78.

[6] Padmanabhan M, Cournoyer P, Dinesh-Kumar S. The leucine-rich repeat domain in plant innate immunity: a wealth of possibilities. Cellular microbiology. 2009;11(2):191-8.

[7] Prigozhin DM, Krasileva KV. Analysis of intraspecies diversity reveals a subset of highly variable plant immune receptors and predicts their binding sites. The Plant Cell. 2021;33(4):998-1015.

[8] Barragan AC, Weigel D. Plant NLR diversity: the known unknowns of pan-NLRomes. The Plant Cell. 2021;33(4):814-31.

[9] Saucet SB, Esmenjaud D, Van Ghelder C. Integrity of the post-LRR domain is required for TIR-NB-LRR function. Molecular Plant-Microbe Interactions. 2021;34(3):286-96.

[10] Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research. 2011;39(suppl_2):W29-37.

[11] Bateman A, Coggill P, Finn RD. DUFs: families in search of function. Acta Crystallographica Section F: Structural Biology and Crystallization Communications. 2010;66(10):1148-52.

[12] Martin EC, Sukarta OC, Spiridon L, Grigore LG, Constantinescu V, Tacutu R, et al. LRRpredictor—a new LRR motif detection method for irregular motifs of plant NLR proteins using an ensemble of classifiers. Genes. 2020;11(3):286.

[13] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic acids research. 2000;28(1):235-42.

[14] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9.

[15] Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Research. 2024;52(D1):D368-75.

[16] Mokhtarian F, Mackworth AK. A theory of multiscale, curvature-based shape representation for planar curves. IEEE transactions on pattern analysis and machine intelligence. 1992;14(8):789-805.

[17] Wahba G. A least squares estimate of satellite attitude. SIAM review. 1965;7(3):409-9.

[18] Kabsch W. A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography. 1976;32(5):922-3.

[19] Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, et al. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. genesis. 2015;53(8):474-85.

[20] Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. Nucleic acids research. 2021;49(D1):D344-54.

[21] Tralie CJ, Berger M. Topological eulerian synthesis of slow motion periodic videos. In: 2018 25th IEEE international conference on image processing (ICIP). IEEE; 2018. p. 3573-7.

[22] Chung FR. Spectral graph theory. vol. 92. American Mathematical Soc.; 1997.

[23] Godsil C, Royle GF. Algebraic graph theory. vol. 207. Springer Science & Business Media; 2001.

[24] Averbuch-Elor H, Cohen-Or D. Ringit: Ring-ordering casual photos of a temporal event. ACM Transactions on Graphics (TOG). 2015;34(3):1-11.

[25] De Silva V, Vejdemo-Johansson M. Persistent cohomology and circular coordinates. In: Proceedings of the twenty-fifth annual symposium on Computational geometry; 2009. p. 227-36.

[26] Perea JA. Sparse circular coordinates via principal $\mathbb{Z}$-bundles. In: Topological Data Analysis: The Abel Symposium 2018. Springer; 2020. p. 435-58.

[27] Vejdemo-Johansson M, Pokorny FT, Skraba P, Kragic D. Cohomological learning of periodic motion. Applicable algebra in engineering, communication and computing. 2015;26(1):5-26.

[28] Rybakken E, Baas N, Dunn B. Decoding of Neural Data Using Cohomological Feature Extraction. Neural Computation. 2019;31:68-93.

[29] Tang WS, da Silva GM, Kirveslahti H, Skeens E, Feng B, Sudijono T, et al. A topological data analytic approach for discovering biophysical signatures in protein dynamics. PLoS computational biology. 2022;18(5):e1010045.

[30] Cang Z, Wei GW. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. PLoS computational biology. 2017;13(7):e1005690.

[31] Steuernagel B, Witek K, Krattinger SG, Ramirez-Gonzalez RH, Schoonbeek Hj, Yu G, et al. The NLR-annotator tool enables annotation of the intracellular immune receptor repertoire. Plant Physiology. 2020;183(2):468-82.