Computational Coiled-Coil Discovery and Modeling

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Mikel Martinez Goikoetxea

aus Iruñea, Spanien

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	20.11.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Andrei Lupas
2. Berichterstatter/-in:	Prof. Dr. Oliver Kohlbacher

Table of Contents

Abstract

Coiled coils are a widespread protein structure motif, consisting of multiple α-helices that wind around a central axis to bury their hydrophobic core. At the sequence level, they are underpinned by short repeats, the most common of which is the 7-residue heptad. By varying the number, composition, and length of their repeats, coiled-coil proteins have access to an outstanding range of structural diversity. In spite of this, their highly regular nature has facilitated the study of the relationship between their sequence and structure, thus becoming a model system for this paradigm of protein science. In this thesis, I address issues with two aspects of coiled-coil research, the discovery and the modeling.

In an effort to discover new coiled-coil families, we bioinformatically investigate the distribution of a hitherto understudied coiled-coil repeat, the 11-residue hendecad, in the proteome of life. To this end, we performed a broad survey for proteins that showed features compatible with hendecad coiled-coil structure, and performed interactive analyses on the resulting dataset. The protein families that we found show that hendecads are more diverse than previously thought, and that this motif expands the topological space accessible to coiled coils.

Further, we address some of the limitations of coiled-coil modeling tools. For this, we evaluated the applicability of AlphaFold, a state-of-the-art protein structure prediction tool, to modeling coiled-coil structures. We benchmarked its performance through two approaches: measuring its accuracy in terms of local geometry, and testing its potential for topological prediction. Our results demonstrate that, even as a general purpose protein structure prediction tool, AlphaFold performs better than coiled-coil specific software. In addition, we also show that it can be leveraged in a coiled-coil framework to improve topological prediction as well as to probe local coiled-coil folding potentials.

Zusammenfassung

Coiled Coils sind ein weit verbreitetes Proteinstrukturmotiv. Sie bestehen aus mehreren Alpha-Helices, die sich um eine mittlere Achse winden, um einen hydrophoben Kern auszubilden. Auf Sequenzebene werden sie durch kurze Sequenzwiederholungen ermöglicht, am häufigsten von Heptaden von 7 Resten. Durch Variation der Anzahl, Zusammensetzung und Länge ihrer Wiederholungen haben Coiled-Coil-Proteine Zugang zu einer außerordentlichen Bandbreite von Strukturen. Dennoch hat ihre äußerst regelmäßige Natur die Untersuchung der Beziehung zwischen ihrer Sequenz und Struktur ermöglicht und sie zu einem Modellsystem der Proteinwissenschaft werden lassen. In dieser Arbeit befassen wir uns mit zwei Aspekten der Coiled-Coil-Forschung: der Entdeckung und der Modellierung.

Wir untersuchen bioinformatisch die Verteilung eines bisher wenig untersuchten Coiled-Coil-Wiederholungsmusters, des 11-Reste-Hendecads, im Proteom des Lebens. Zu diesem Zweck führten wir eine umfassende Untersuchung von Proteinen durch, deren Sequenz mit einer hendekaden Struktur kompatibel sind, und führten interaktive Analysen des resultierenden Datensatzes durch. Die von uns gefundenen Proteinfamilien zeigen, dass Hendecads vielfältiger sind als bisher angenommen und dass dieses Motiv den für Coiled Coils zugänglichen topologischen Raum erweitern könnte.

Darüber hinaus analysieren wir die Grenzen zur Verfügung stehender Modellierungsmethoden für Coiled-Coil-Strukturen. Wir bewerten die Leistungsfähigkeit von AlphaFold, einem hochaktuellen Tool zur generellen Proteinstrukturvorhersage, bei der Berechnung von Coiled-Coil-Strukturen, indem wir dessen Genauigkeit hinsichtlich lokaler Geometrien analysieren und sein Potenzial für topologische Vorhersagen testen. In vergleichenden Analysen stellten wir fest, dass AlphaFold in beiden Punkten besser abschneidet als spezifisch für Coiled Coils entwickelte Modellierungsprogramme. Dies macht AlphaFold zu einem neuen, wichtigen Werkzeug sowohl bei der topologischen Vorhersage von Coiled Coils als auch zur Berechnung lokaler Faltungspotenziale für Coiled-Coil-Strukturen.

Acknowledgements

I am thankful to everyone who made possible this dissertation.

Firstly, I am immensely grateful to my supervisor, Andrei Lupas, for giving me the opportunity to learn from him, and for his continued support and valuable discussions. I cannot recall a single conversation with him from which I emerged without a new insight, and to this day I am still amazed at his ability to generate good ideas (and to point out flaws in the bad ones).

I also extend my thanks to my Thesis Advisory Commitee, who besides Andrei, include Oliver Kohlbacher and John Weir. I am grateful for their comments and their patience. They did their best to try and dissuade me from going down unproductive paths, and I wish I had listened to them sooner.

I have had the privilege of sharing a working ecosystem with many excellent colleagues at the Department of Protein Evolution, and enjoyed both inspiring conversations and supportive interactions with them. I am particularly grateful to Maria, Ioanna, Jens, Joana, Vikram, Stanislaw, Pedro, Matej and Valeria. They made my time as a PhD student pass by quickly, and it is their fault it took me so long to finish. Except Stanislaw, he actually contributed greatly to my PhD, and I am tremendously grateful.

Outside the professional realm, I would like to thank Amaya, Maria, Dorota, Hadeer, Joe, Anasuya, Xixi, and Neha. They distracted me even further, but also contributed great anecdotes which, sadly, cannot be part of this dissertation.

Lastly, and most definitely, I would like to thank my parents and family for their endless love and support. I wish my grandparents would have lived to see the completion of this dissertation, I think they would have enjoyed reading it.

1 Introduction

Proteins are fundamental components of living systems, and have been intensively studied to advance our understanding of biology at the molecular level, as well as for their biomedical and industrial applications. A key idea in protein science is that their primary structure, this is, their sequence, determines their three-dimensional structure, which in turn endows them with function. Proteins exert this function in a variety of ways, including catalyzing chemical reactions, converting between forms of energy (chemical to physical and vice versa), supporting the compartmentalization of cellular structures, providing structural support and sensing changes in the environment. In a way, life can be thought of as the cause and the consequence of the coordinated expression of proteins in time and space.

Due to the difficulty of experimentally determining protein structures, as well as the numerous genome sequencing and automatic gene annotation projects that have been conducted, protein sequence databases have grown exponentially compared to structural databases. Consequently, much of the focus of the field has been devoted to solving the so-called **protein structure prediction problem**, in essence, predicting the three-dimensional fold of a protein, given their sequence. Solving this problem would represent a significant leap in our understanding of the molecular mechanisms underlying protein function, as well as in our ability to design new proteins with novel or improved functions. AlphaFold has made substantial progress in this area, offering highly accurate predictions that have revolutionized the field; however, there is still much to be done to fully understand and predict the vast diversity of protein structures.

Coiled coils are a model system of the relationship between protein sequence and structure, and are considered the best understood protein fold. The short and simple sequence repeats of coiled coils underpin an outstanding structural and functional diversity and, perhaps as a result, are found throughout virtually every proteome. These features make coiled coils an attractive target for the study of protein structure, function and evolution. Notwithstanding the many advances that coiled-coil research has brought in the form of prediction and design capabilities, our understanding of them is far from complete. Now, a solid base of knowledge in combination with the advent of powerful bioinformatics methods enable us to tackle questions and problems that were not possible a mere decade ago. In particular, the two gaps in the field that have motivated the development of the work presented in this thesis are:

- Research on coiled coils has focused primarily on those composed of the commonly occurring **heptad repeats**, neglecting other coiled-coil repeats, such as **hendecads**. These are often found interspersed between long arrays of heptads, and have been rationalized as interruptions in the heptad pattern. However, an emerging body of work has put forward these motifs as functionally relevant features, warranting further investigation into how they differ from canonical heptad repeats, and into their distribution in the proteome of life.

- **Computational modeling** is a fundamental aspect of bioinformatics. Although coiled-coil specific modeling tools exist, they have significant drawbacks, including the inability to be used for prediction purposes, and a requirement for the user to set topological parameters. The breakthrough of **AlphaFold**, an end-to-end protein sequence-to-structure prediction program trained on experimentally-obtained structures, marked a significant leap in solving the protein structure prediction problem, but it has not yet been evaluated in a coiled-coil specific framework for its accuracy and its ability to predict topological parameters.

2 Background

2.1 Preamble

This chapter will provide an overview of coiled-coil research, and lay out the ideas that will be developed in subsequent chapters. It will cover well-established ideas, including a brief historical overview and description of the modern coiled-coil model, as well as emergent knowledge, with a focus on hendecad coiled coils. It also provides an overview of the most significant software programs that have pushed the field forward, as well as some of their limitations yet to be overcome. This will bring us to the frontier of the existing knowledge, and set the stage for the chapters presenting the main results of this study.

2.2 Historical retrospective

Coiled coils are as deeply rooted in the history of structural biology as they are widespread in the proteome of life. For reasons that will be made clear in later subsections, some of the most extended and regularly-repeating proteins are formed by coiled coils, proteins which, understandably, constituted an important stepping-stone for early crystallographic studies. In particular, easy to obtain proteinatious materials such as keratin (hair, horn) or myosin (muscle), were fundamental for the development of the techniques and ideas that eventually resulted in our modern models of protein structure.

In the early 1930s, William Astbury was researching the properties of wool at the University of Leeds, funded by the local textile industry. Motivated by the previous successes of crystallography in elucidating the structure of small organic compounds, he sought to relate the material properties of wool, which is mainly formed by keratin, to its X-ray diffraction pattern. He found out that the diffraction pattern of native wool fibers changed when these were stretched. He proposed that the unstretched diffraction pattern corresponded to a helical structure which he coined the *alpha* form, and that subjecting the wool fibers to stretching pulled these helices into an extended state, the *beta* form. Although this model was later refined and expanded, what he described was the essence of our modern concept of α-helices and β-strands, and thus the nomenclature was kept. The diffraction pattern of keratin could be interpreted as a helical structure with a repeating unit of 5.1 Ångstroms,

which would trouble many a researcher in the years to come, and ultimately result in the writing of this thesis.

Thanks to the strong foundation set by the work of Astbury and other eminent crystallographers, several important details of the atomic structure of the *alpha* form, as well as its ubiquity, would be elucidated in the following years, and the interest in the topic would grow to become a race to publish the definitive model. By 1948, the two most prominent contestants were the groups of Lawrence Bragg and Linus Pauling. While bedridden due to a cold during his stay in Oxford, Pauling had the idea of drawing a polypeptide chain on a strip of paper (as a gifted chemist, he had understood the importance of the planarity of the peptide bond), and trying to fold it into a helix. One such attempt resulted in a hydrogen-bonded helix with 3.6 residues per turn and a rise of 1.5 Ångstroms (Å) per residue. He discarded this model because of the incongruent length of the structural repeat, which was 5.4 Å (1.5 x 3.6), and not 5.1 Å as measured by Astbury's X-ray diffraction experiments on the *alpha* form of keratin. Two years later, Bragg, Perutz and Kendrew published a manuscript that described and evaluated various molecular models compatible with this diffraction pattern[1], ending the discussion with "t*he problem is very complex, and one is forced to rely on a number of items of evidence each of which is very slight*". They would come to deeply regret this paper, with Bragg himself regarding it as "*the most ill-planned and abortive in which [he] had ever been involved*". At a conceptual level, their model was fundamentally flawed due to a) their failure in recognizing the planarity of the peptide bond, b) their assumption that the structure had to feature an integer number of residues per turn, and c) their effort in accommodating the 5.1 Angstrom repeating unit.

Later that year, in 1950, Pauling and Corey published a brief communication announcing two upcoming hydrogen-bonded spiral configurations of the polypeptide chain. In 1951, they published, together with Branson, what can essentially be recognized as our modern model of the α-helix[2], barring some details. Notably, the handedness was arbitrarily chosen as left-handed, and it would take until 1963 for Ramachandran to show that the α-helix had to be right-handed to avoid steric clashes[3]. One of the key aspects that lead Pauling to put forward the model that he had essentially envisioned in 1948, was that he decided to ignore the 5.1 Ångstrom meridional arc of keratin, on the grounds of its absence in the diffraction pattern of a synthetic fiber made up of poly-gamma-methyl-L-glutamate, which was clearly of the alpha form.

In the very same 1951 manuscript where he detailed his α-helix model, Pauling also noted that, while it could be structurally described as a helix that distributed 48 residues in 13 turns (48 / 13 = 3.69 residues per turn), in a crystalline arrangement, the helices could be distorted to optimize the packing with neighboring helices, and "*deform them slightly into configurations with a rational number of residues per turn*". Furthermore, he hypothesized the existence of helices with effective structural periodicities of 11/3 (3.67), 15/4 (3.75), and 15/8 (3.6). After the elucidation of a satisfactory atomic model for the α-helix, it was clear that the structure of keratin had to be generally helical, albeit with some distortion, possibly the twisting of the helices.

Towards the end of 1952, Francis Crick published a manuscript suggestively titled "Is alpha-Keratin a Coiled Coil?"[4]. There, he considered how the α-helix model could be reconciled with the 5.1 Å meridional arc of keratin that Astbury measured. He used geometric arguments (he was, indeed, an expert in helices) to show that α-helices could be twisted into a supercoil with a pitch angle of about 18 degrees, thus reducing the characteristic 5.4 Å periodic unit of undistorted α-helices to 5.1 Å (5.4 x cos 18). He supported this notion by pointing out that in such a "coiled coil", the winding of the helices around each other would allow for the continuous interlocking of their side-chains into a packing mode he called "knobs-into-holes", and that such interaction could provide the energy required for the deformation of the helices. Several months later, he published back-to-back "The Fourier Transform of a Coiled-Coil"[5] and "The Packing of alpha-Helices: Simple Coiled-Coils"[6], where he expanded his description of the knobs-into-holes interactions and developed a set of equations to describe the regular backbone of coiled-coil structure. Notably, he also speculated on the physicochemical nature of coiled coils, writing that hydrophobic residues would "t*end to occur between the chains, rather than on the outside of the molecule, since the packing of hydrophobic groups together would allow more hydrogen bonds to be made elsewhere. […]. If this hypothesis were correct, one would expect there to be a tendency towards an [sic] alteration of groups of polar and non-polar residues in tropomyosin, the non-polar occurring at an average interval of 3.5 residues, so that they always pointed 'inside' the molecule*". This model was immediately convincing, and its merits would be confirmed by the determination of the tropomyosin sequence in 1974[7], which showed the expected interval between hydrophobic residues, and the structure of influenza hemagglutinin in 1981[8], which featured knobs-into-holes packing.

This concludes the historical perspective of coiled coils, which is necessarily limited in scope and leaves out many fascinating aspects of the history of coiled-coil research and the people involved

therein[9]. In the next subsection, the modern model of the coiled coil is described, which will lead to the next, more specialized sections.

2.3 Coiled-coil sequences and structures

Coiled coils are a supersecondary structure motif consisting of multiple α-helices that wind around each other to systematically bury their hydrophobic core along the length of the structure. This interaction follows a specific packing geometry, widely considered the hallmark of coiled coils, known as knobs-into-holes. Here, a core residue (knob) places its side-chain into a cavity formed by 4 residues (hole) of a symmetry-related helix. In Crick's description of the canonical coiled-coil model, the optimal knobs-into-holes packing could only be achieved by spacing a core residue every 3.5 residues on average, which indicated that the repeat unit had to be 7 residues long (the least integer multiple of 3.5). Within these constraints, the most uniform core residue distribution is achieved by two blocks of 3- and 4- residues, each starting with a core residue. If coiled coils were to fold as undistorted α-helices, the difference between their core-residue distribution in the sequence (one every 3.5 residues on average) and their structural periodicity (3.63 residues per turn on average) would make it impossible to pack every core residue along the coiled coil, and they would seem to 'drift' slowly out of the interface. Instead, as Crick correctly predicted, coiled-coil helices are slightly distorted, such that their effective periodicity, with respect to the central axis, is reduced to 3.5. Because the required change in periodicity is negative, the α-helices wind in a direction opposite that of their handedness, resulting into a left-handed supercoil. This allows for continuous knobs-into-holes interactions along the core of the coiled coil. The seven residue repeat, which is found in most coiled coils in nature, is known as the *heptad* repeat, and its residues are labeled *a*-*g*, where *a* and *d* are the core residues (Fig. 1).

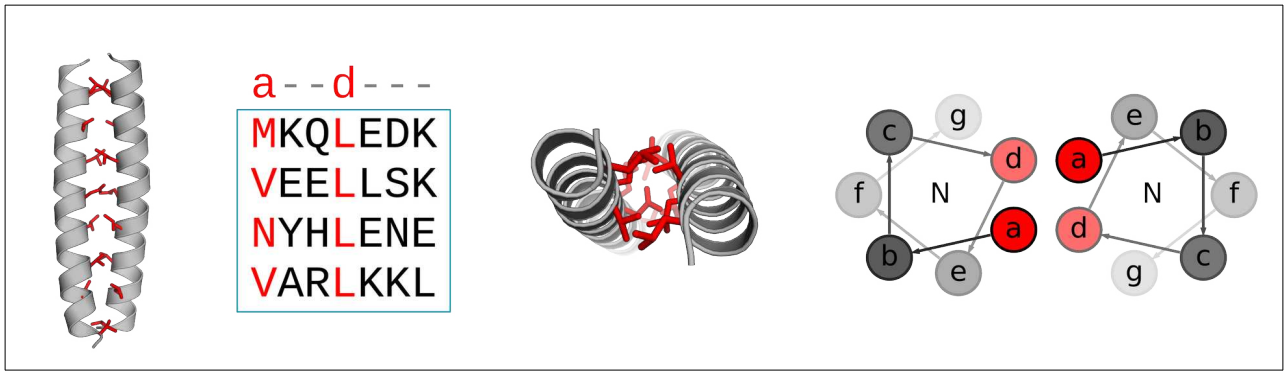


Figure 1: Various representations of heptad coiled coils. From left to right, a cartoon render of the coiled-coil stalk of GCN4 (PDB: 2ZTA) and its corresponding sequence annotation, view from N-terminus, and schematic helical wheel. The core positions a and d are highlighted in red.

Within this framework, a coiled-coil sequence can be described in terms of what kinds of residues are predominant in any given position. For instance, as Crick anticipated, core positions tend to be enriched in hydrophobic residues, although, as will see later, there are exceptions to this. In turn, coiled-coil structures are typically described in terms of the number and orientation of the helices, as well as the geometry of their core residues. One of the most interesting aspect of coiled coils, is the possibility of studying the relationship between sequence properties, expressed as residues present on a given position, and their effect on the resulting structures. The best example of this is probably the work of Harbury et. al., who in 1993 described the effect of a series of mutations in the core residues of the coiled-coil domain of the yeast transcription factor GCN4[10] (Fig. 1). Their experiments showed that systematically mutating the *a* and *d* positions into isoleucine and leucine respectively (I@a, L@d) yielded the native, dimeric topology, inverting these preferences (L@*a*, I@*d*) yielded tetramers, and placing isoleucine in both core positions (I@a, I@d) yielded trimers. They rationalized these dramatic changes in topology as a consequence of the core packing geometries that are preferred by the different oligomeric states. Thus, parallel dimeric coiled coils tend to pack their core residues at an angle such that the *a* position has more space available than the *d* position, favoring β-branched residues such as valine or isoleucine in *a*. The opposite happens in parallel tetramers, where the *d* position has more space available. Intuitively, trimers feature an intermediate packing angle. Another revealing result of their experiments was that substituting the N16@a residue by an isoleucine yielded a mixture of dimers and trimers, effectively losing the topological specificity in favor of a higher stability. This particular example is representative of a more general feature of coiled coils, that being the presence of hydrophylic residues in the core. Indeed, although the preference for hydrophobics in core positions is the signature feature of coiled coils, hydrophylic residues are not rare in these positions (polar residues represent roughly 25% of all core residues in the CC+ database[11,12], a coiled-coil subset of the PDB). One of the most common

hydrophylic core residues is, precisely, asparagine, which is often found in the *a* position of parallel dimers, and in the *d* position of trimers, sometimes coordinating ions. We find an extreme example of this in the coiled coils of trimeric autotransporter adhesins (TAAs), whose core positions are conspicuously enriched in hydrophylic residues; for example, in some Burkholderia TAAs two thirds of the core positions are occupied by Serine and Threonine residues[13].

After the *a* and *d* core positions, the flanking *e* and *g* positions are the next most critical determinants of coiled-coil topology. They typically consist of polar or charged residues, which depending on the charge complementarity favor or impede a particular helix orientation (parallel or antiparallel). If, instead charged or polar residues, these positions feature hydrophobic residues, the resulting assembly will tend to fold into a higher oligomeric state to more effectively bury the expanded core. The typical case, where one of the flanking positions is unusually enriched in hydrophobic residues, promotes the formation of antiparallel tetramers. This core configuration can be understood in terms of two superimposed core arrangements, sharing one position of the heptad. If the additional hydrophobic position is *e,* then a given helix can be thought of as having two seams of core residues, *a-e* and *a-d*, both of which share the central *a* position. Likewise, if the additional hydrophobic position is *g*, the resulting interface can be thought of as the combination of the *d-a* and *d-g* seams, where the *d* is the central position. These core configurations are known as *a-d-e* or *a-d*-g cores respectively. In order to accommodate such core configurations, the helices rotate axially in order to maximize the burial of the core residues, and in the process, the central position is left pointing towards the center of the bundle (*x* geometry), while the other two flanking positions point sideways (*da* geometry). An important consequence of this axial rotation of the helices is that it comes necessarily with the partial loss of the *knobs-into-holes* interactions and the acquisition of a new geometry of interaction, known as *knobs-to-knobs*, where the side-chains in symmetry-related positions point against one another.

The broad hydrophobic seam of *a-d-e* and *a-d-g* cores can be further expanded into what can be described as two non-overlapping hydrophobic seams, *a-e* and *d-g*, that do not share a central position, and are instead adjacent. Packing this core configuration requires further axial rotation of the helices, and supports the formation of pentamers, hexamers and heptamers[14]. Separating the two seams even further leaves a non-core position between them, which, by convention, is *a* (hydrophobic seams being *b-e* and *d-g*) or *d* (hydrophobic seams at *a-e* and *c-g*). Such core configuration leads to even higher oligomeric states, the largest example being the antiparallel barrel of 12 helices TolC[15]. It is noteworthy that this core arrangement features the maximum

possible angle between heptad positions in a helix, and that further widening the gap between the hydrophobic seams is equivalent to relabeling the core in such a way that a single position separates the seams. More generally, such coiled coils with broad interfaces are called bifaceted, and are labeled type I (hydrophobic seams share one central position, such as *a-d-g* cores), type II (adjacent seams), and type III (seams separated by one position).

Although all these residue preferences have been extensively used for the design, as well as for the prediction of the oligomeric state of coiled coils, they are also illustrative of their unique features. As Harbury's mutants showed, otherwise conservative mutations can dramatically alter the topology of coiled coils. More generally, this reflects the fact that coiled-coil topologies are often isoenergetic, and are thus separated from alternative structural states by a relatively low energy barrier. This can be appreciated in that coiled-coil fragments that are taken outside their native context typically fold in non-physiological topologies (number and orientation of helices), or may fail to fold entirely[16].

This perspective introduces a problem for coiled-coil folding, particularly in the case of long (typically parallel) fibers. If the folding propensity was uniformly distributed along such fiber, the number non-physiological structures that could potentially be formed during the folding process, such as out-of-register assemblies, is considerable. In the best case scenario, these local structures might hinder the folding of the physiological fiber, and in the worst, they could prevent it entirely, trapping the protein into a local energy minimum. These considerations, which do not seem to affect the folding of coiled coils in nature, lead to the development of the *trigger-sequence* hypothesis[17,18]. It introduced the notion that the folding propensity of coiled coils is not uniformly distributed, and that it can, in fact, be localized into short fragments. These would act as nucleation sites that form highly stable helical assemblies with the correct (physiological) topology, which could then be propagated along the fiber. This hypothesis elegantly reconciles the experimental results regarding coiled-coil fragments with the fact that long coiled coils can fold into fibrous structures in an efficient and specific manner. On the other hand, it also implies that coiled-coil sequences are closer to a disordered state than other proteins. Interestingly, it is often the case that disorder-prediction programs detect coiled-coil sequences as disordered regions. This can be attributed to the fact that, apart from trigger sequences, they often are, but also to their sequence properties (low complexity, few hydrophobics), which can yield false disorder predictions. It is important to note that the considerations outlined above do not reflect in any way the thermal stability of the folded assemblies, and that in fact, some of the most thermostable proteins are coiled coils[19].

In this subsection the main determinants of coiled-coil structure and topology have been discussed, with a focus on the *heptad* repeat. By varying the number and composition of their heptad repeats, coiled coils have access to a large structural space, which can be further tuned for properties such as stability or partner specificity, in the case of heteromers. Of course, these properties can also be locallly tuned within a given sequence, in some cases as a neutral change that is *endured* by the protein, and in others as a functional feature. This idea is beautifully illustrated by non-heptad repeats that can nevertheless form coiled coils.

2.4 Non-heptad coiled coils

The origin of the coiled-coil is explained in an entirely natural way as the result of the close-packing of two adjacent alpha-helices, and it is not necessary to postulate a regular sequence of residues repeating every seventh residue. The figure seven comes directly from the nature of an alpha-helix.

--F. H. Crick, "The Packing of alpha-Helices: Simple Coiled-Coils"

After the first coiled-coil sequences and structures validated Crick's insights, the progressively growing number of examples led researchers to realize that the Crick model described an ideal scenario, and that most coiled-coil sequences in nature could not be understood as uninterrupted arrays of heptad repeats. Instead, they often feature insertions of various sizes, which were initially thought of as *interruptions* of the heptad pattern. The most common of these are, ordered by most to least frequent, insertions of 4-residue (*stammer*), 1-residue (*skip*), and 3-residue (*stammer*) fragments. The prevalence of these particular insertion sizes can be explained from the nature of the α-helix, as we will see next.

As it was established in the previous subsection, the supercoiling of coiled-coil helices stems from the distortion required to pack the core residues along the length of the structure. Because α-helices feature a structural periodicity of 3.63 residues per turn on average and heptad repeats have a core residue every 3.5 residues on average, the helices must supercoil in a sense opposite to that of their handedness (one way to think about this is that the periodicity has to be *reduced* from 3.63 to 3.5, thus the *opposite* sense; note that this reduction is relative to the central axis, not the helix itself). If a block of 4 residues, starting with a hydrophobic residue, were to be inserted into an array of heptad repeats (...g *abc* + Hxxx + *defg a*...) without any structural rearrangement, it would take the

position of the original *defg* block without issue, but it would displace the following *a* position out of the core, with *g* in its place. In order to keep the next heptad repeat in register, the *defg* block could be compressed to occupy just one turn, but such localized distortion is energetically disfavourable. Instead, the required distortion is distributed among the 11 (7+4) residues, locally increasing the effective periodicity, with respect to the central axis, to 3.67 residues per turn (7+4 / 2+1 = 11 / 3 = 3.67). Because this effective periodicity is very close to, but higher, than that of undistorted α-helices (3.63), the supercoiling of the helix is locally reduced to an almost negligible degree, and acquires a right-handed direction. Such 11-residue unit, which is the most common non-heptad coiled-coil repeat found in nature, is called a hendecad, and its positions are labeled *a-k* (Fig. 2). In terms of core geometry, the insertion of 4 residues disrupts the knobs-into-holes packing and causes a knobs-to-knobs interaction to be formed. In a manner reminiscent of the *a-d-g* and *a-d-e* cores of the type I bifaceted heptad coiled coils, the constituent helices locally adopt an *x-da* packing geometry, where a residue points towards the central axis (*x*), and the flanking residues point toward its sides (*da*). Within the hendecad labeling scheme, *a* and *h* are located in the center of the core, and *d* and *e* are the flanking positions. Typical hendecad core geometries are those formed by the *a-d-h* or the *a-d-e-h* positions, which are related by axial rotation of the helices.

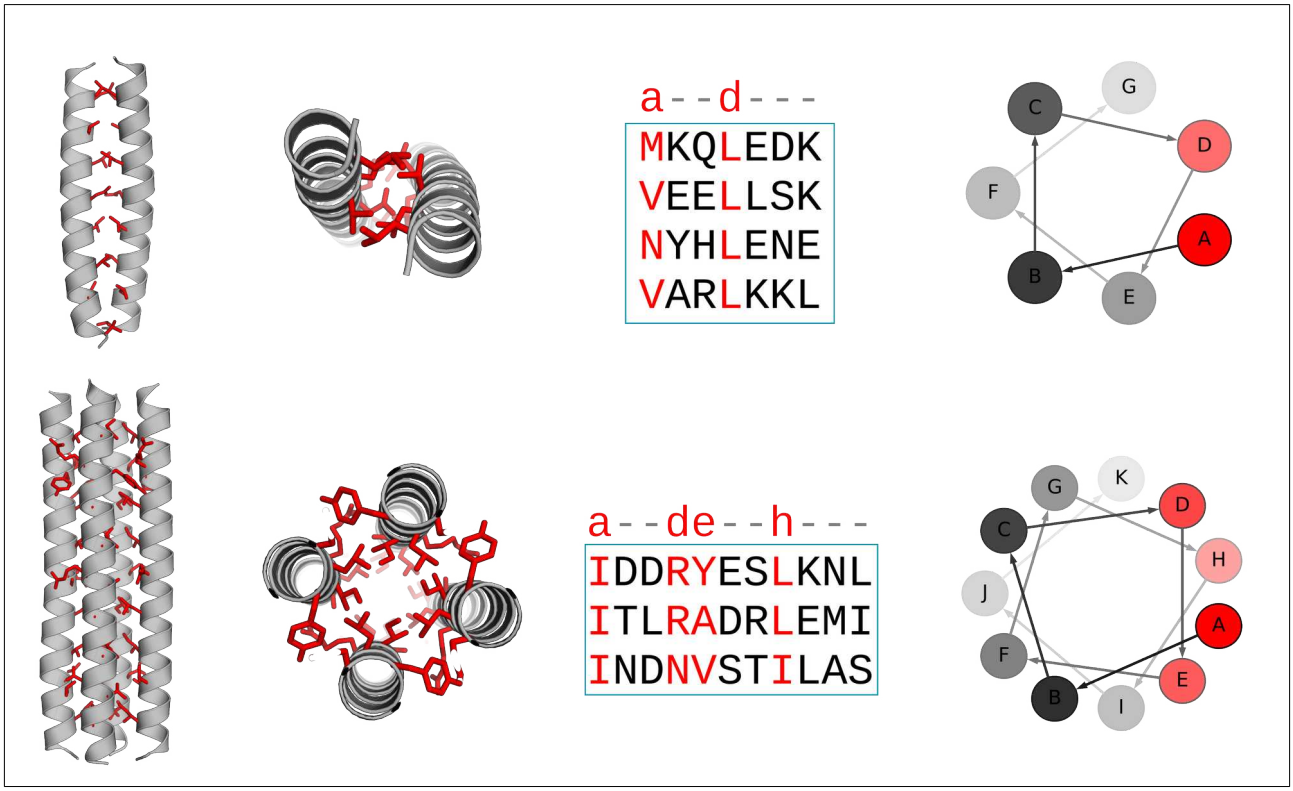


Figure 2: Comparison between various representations of heptads (as seen in Fig. 1) and hendecads. From left to right, a cartoon render of the coiled-coil stalks of (top) GCN4 (PDB: 2ZTA) and (bottom) tetrabrachion (PDB: 1FE6), their corresponding sequence annotation, view from N-terminus, and schematic helical wheels. The core positions are highlighted in red.

More generally, all coiled-coil structures are underpinned by sequence repeats composed of blocks of 3- and 4- residues, each starting with a core residue. If these elements are alternated, as in the case of the heptad repeat (3+4 = 7), the helices are able to form knobs-into-holes interactions along the length of the structure, but any repetition of the same block type (3+3 or 4+4) will cause a knobs-to-knobs interaction to be formed, as well as a change in the degree and direction of the supercoiling in order to ensure the burial of the core. The possible sequence repeats compatible with coiled-coil structure are only limited by the difference between the average spacing of their core residues and the structural periodicity of an undistorted α-helix (3.63). Thus, sequence periodicities of 3.6 (18/5) or 3.67 (11/3) are essentially straight, periodicities of 3.5 (7/2) or 3.57 (25/7) are left-handed, and periodicities of 3.75 (15/4) or 3.8 (19/5) are right-handed (Fig. 3). If this difference is higher than about 0.23, the coiled-coil helix must bend beyond its breaking point to ensure the burial of the core.

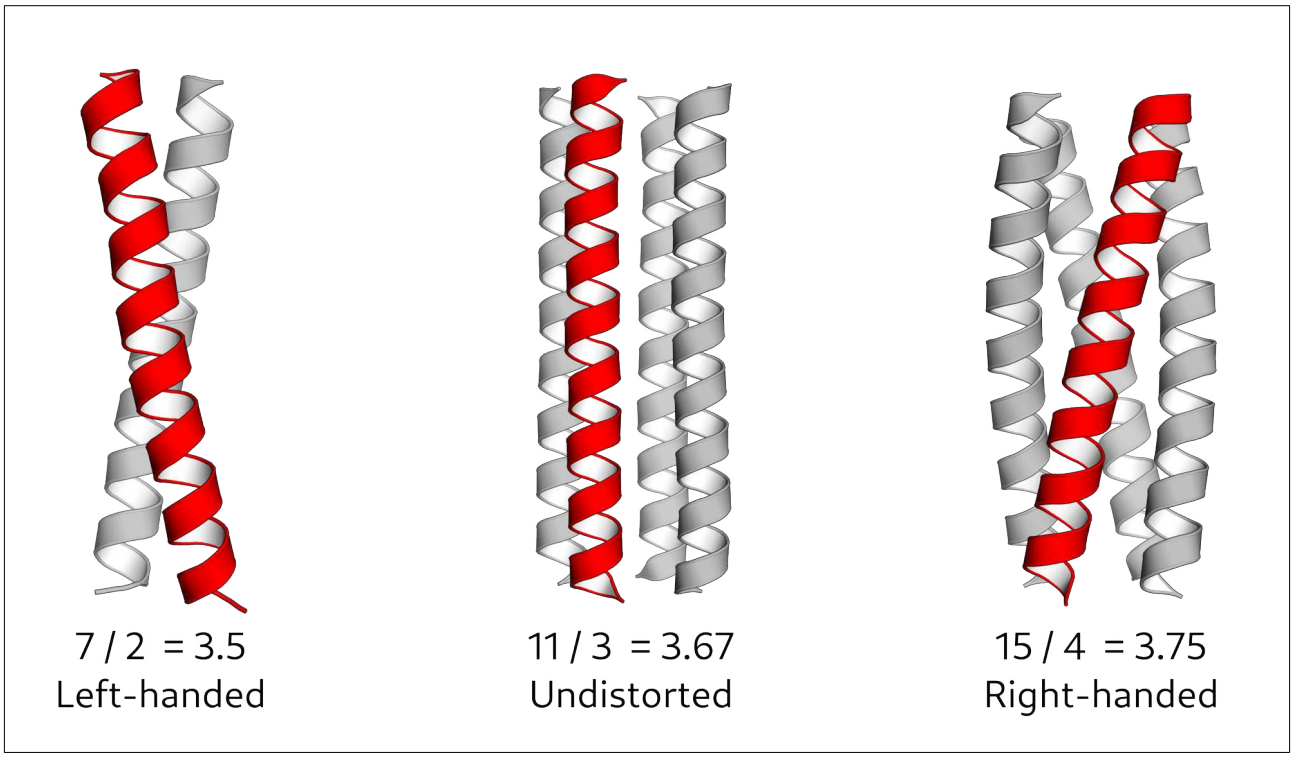


Figure 3: Highlight of the bundle supercoiling that results from different coiled-coil periodicities . Left) Heptad stalk of GCN4 (PDB: 2ZTA). Center) Hendecad stalk of tetrabrachion (PDB: 1FE6). Right) Pentadecad stalk of VASP (PDB: 1USD).

The reason why the *stutter* is so common in coiled coils (compared to other insertions), is probably that it is easier to acommodate than other insertions. For example, an insertion of 3 residues (stammer) would result in an average spacing between core residues of 3.33 (7+3 / 2+1), which

would cause the coiled coil to locally increase the supercoiling degree beyond that of the heptad. An insertion of one residue (skip), would also be energetically disfavourable for a single heptad, resulting in a spacing of 2.67 (7+1 / 2+1 = 2.67), which lays well beyond what an α-helix can be effectively deformed into. However, a skip residue may be comfortably delocalized along several heptads (7+7+1 / 2+2+0 = 15/4 = 3.75), thus becoming equivalent two stutters (7+4+4 / 2+1+1 = 15/4 = 3.75).

A consequence of the introduction of knobs-to-knobs interactions is that, due to the more constrained nature of the *x* geometry, a steric limitation is imposed. This is clearly seen in the case of parallel coiled-coil dimers with knobs-to-knobs layers, where residues in *x* point directly towards each other, and thus must feature small side-chains (such as glycine or alanine) in order to enable enough distance between them to avoid steric clashes. Alternatively, the coiled-coil helices must assemble into an oligomeric state higher than a dimer in order to accommodate residues with larger side-chains. This explains why hendecad coiled coils usually form trimers or tetramers, rather than dimers, which are the most common form of heptad coiled coils. Coiled-coil structures may also cope with the steric constraint in other ways, such as by breaking symmetry, locally shifting their side-chains out of register, or switching to an antiparallel topology.

Due to the manner in which non-canonical coiled-coil repeats are typically found in coiled-coil domains --that is, in low copy number, and flanked by heptad repeats-- they have been generally assumed to be *tolerated* deviations rather than functional features. Although that may be true for many of them, a growing body of work provides examples that indicate the opposite situation. For instance, it has been shown that the function of the human EEA1 (Early Endosome Antigen 1), whose long coiled-coil stalk is thought to transition between the flexible and extended states, is impeded by the removal of non-heptad discontinuities[20], possibly via the disruption of the propagation of conformational changes along the fiber. More generally, the conserved presence of a non-canonical coiled-coil segment within a protein family is suggestive of its functional role. Such is the case with KfrA, where its non-canonical coiled-coil segment was proposed to promote an unstable region required for filament formation[21]. Non-canonical coiled coils may also constitute functionally relevant features in a periodicity-specific manner. For instance, hendecad repeats have been shown to be play an important role in septin-septin heteromeric pair recognition, due to the unfavorable interaction between heptad and hendecad coiled coils[22]. Interestingly, it has also been shown that the hendecad periodicity is the idoneous for the design of heterochiral coiled-coil assemblies, due to its characteristic small degree of supercoiling[23].

While more often than not arrays of heptads are interrupted by sparse non-heptad repeats, there are very few reliable examples of structurally characterized extended non-heptad coiled coils. These include the surface layer protein tetrabrachion[19], the EibD trimeric autotransporter adhesin[24], and the phi-X174 phage H protein[25]. In contrast, there are abundant examples of repetitive sequences with features compatible with non-heptad coiled-coil structure, but that cannot be confidently detected by coiled-coil prediction software.

2.5 Coiled-coil computation

The widespread nature of coiled coils and the regularity of their sequence repeats have fostered the development of a number of software programs to aid in the prediction, analysis, and modeling of these motifs, the most relevant of which are summarized below. The aim of this subsection is to highlight how these tools have advanced the field, as well as the limitations that are yet to be overcome.

The first computational coiled-coil prediction program was COILS[26], which implemented a scoring scheme that used a heptad substitution matrix and a sliding window approach to scan sequences for coiled-coil forming propensities. The matrix was derived from the residue preferences of the heptads of representative groups of coiled-coil families, such as tropomyosins, myosins, and keratins (this is how the initial MTK matrix was produced, although additional matrices were subsequently introduced), weighted by the relative occurrence of the respective residues in the sequence database. The final coiled-coil probability prediction was computed via statistical analysis of the different distributions of the COILS scores in coiled-coil and globular sequences. Although it was the first of many such methods, COILS has remained relevant as a quick and accurate coiled-coil detection tool, as seen in the fact that it is still used by current bioinformatic databases for sequence annotation, such as UniProt (https://www.uniprot.org/help/coiled).

Soon after, other programs were developed, making use of the ever-expanding set of available databases and methods. Some of the most prominent ones were, in chronological order, PairCoil[27], based on pairwise residue correlations, Multicoil[28], which could distinguish between dimeric and trimeric bundles, MARCOIL[29], based on windowless Hidden Markov Models, PCOILS[30], which extended COILS with profiles, and LOGICOIL[31], which extended the range of the predicted topologies to tetramers and antiparallel dimers. More recently, the widespread adoption of deep-

learning lead to the development of DeepCoil[32], a convolutional neural network trained on one-hot encoded sequences, its update DeepCoil2, which uses an updated architecture in combination with SeqVec embeddings, and CoCoNat[33], which uses ProtT5 and ESM2 embeddings for the prediction of coiled-coil topology. All these methods are able to generate automatic annotations at the residue level, expressed in terms of heptad register positions, although they offer varying balances between accuracy and speed. It is difficult to make strong claims regarding their relative performances, since the testing datasets of older programs tend to intersect with the training sets of newer ones. Furthermore, they are capable of performing different predictions, including the presence, annotation, and topology of coiled coils. A common theme among these programs is that they are mainly trained on typical heptads, and that they do not explicitly implement biophysics in their inferences. The bias towards heptad repeats is a direct consequence of the biased distribution of coiled-coil repeats in nature, which overwhelmingly consist of heptads. Consequently, coiled-coil prediction programs tend to underperform when applied to non-canonical repeats or coiled coils with uncommon compositions, such as trimeric autotransporter adhesins or coiled coils with bifaceted helices.

The regularity of coiled-coil structures has also allowed the development of tools for their structural analysis and modeling. For example, SOCKET[34] is able to detect knobs-into-holes interactions based on structural coordinates, which makes it exceptionally useful in the automatic detection of coiled-coil domains in proteins of known (or predicted) structure. It has also been used to automatically create databases of reference coiled-coil structures[35,36]. The fact that coiled-coil backbones can be described by the Crick parametric equations has been exploited in tools that can analyze coiled-coils structures in terms of their bundle periodicity (supercoiling) and the degree of axial rotation of the helices, which is a descriptive feature of the geometry of the core. These include Twister[37], SamCC[38] and CCCP[39], which differ in their methodological approach and limitations, but can generally be used to describe coiled-coil structures at the resolution of individual residues.

Besides the analysis of coiled-coil structures, the parametric description of coiled coils has also been applied in the generation of structures for the purpose of modeling. For example, BeammotifCC[40] implemented a generalized set of the Crick equations that made it possible to model coiled coils of varying periodicity along the length of the structure. An important limitation of that program was that the core assignment had to be specified, severely limiting its use cases. ISAMBARD[41] reimplemented these equations with an emphasis on protein design, providing a

modular system with which to generate backbone coordinates, graft the side-chains of a given sequence, and optimize the structural parameters such as radius and pitch according to an energy function. This program is particularly applicable to the design of highly-symmetric and regular bundles, such as coiled-coil barrels[14] but cannot be used to broadly model coiled-coil domains, which typically feature discontinuities in their periodicity. A different approach to modeling coiled coils was employed by CCfold[42], by implementing a fragment-threading approach. Despite overcoming some of the limitations of parametric modeling, CCfold is constrained to modeling dimers and trimers, and tends to impose coiled-coil structure to non coiled-coil sequences.

3 Objectives

The objective of this thesis is to address two issues in coiled-coil research, I) the scarcity of examples of globally non-heptad coiled coils, and II) the limitations of current coiled-coil modeling programs. These issues are intimately related by the fact that new coiled-coil examples are the best benchmark of our coiled-coil predicting and modeling tools, and that, in turn, the performance of our tools is limited by the set of known coiled coil examples. I aim to advance the field by providing new examples of hitherto unknown coiled-coil families, and by incorporating a state-of-the-art modeling tool, AlphaFold, into a coiled-coil framework. The work presented in this cumulative dissertation is detailed in four research papers, grouped by topic as outlined below.

The chapter titled **On coiled-coil discovery** presents our efforts in expanding our coiled-coil bestiary by performing a broad search for sequences likely to form hendecad coiled coils in the proteome of life. Here, our aim was to collect a set of reliable examples from which to learn about the sequence properties and the functional features that hendecad coiled coils show in nature. The published papers that detail the work presented this chapter are:

- New protein families with hendecad coiled coils in the proteome of life
- A conserved motif suggests a common origin for a group of proteins involved in the cell-division of Gram+ bacteria

Our work regarding coiled-coil discovery benefited greatly from the advent of AlphaFold, although it also made us aware of its limitations. The chapter titled **On the modeling of coiled coils** presents our systematic benchmark of the performance of AlphaFold in the task of modeling coiled-coil domains, as well as our efforts at extending its usability. We measured its geometric accuracy and its topological prediction capabilities, and explored whether it could be used to scan protein sequences for local coiled-coil forming potentials. This work is detailed in following papers:

- Applicability of AlphaFold2 in the modelling of coiled-coil domains (preprint)
- CCfrag: Scanning folding potential of coiled-coil fragments with AlphaFold (preprint)

These four papers can be found in the **Appendix** section.

4 On coiled-coil discovery

4.1 Introduction

In nature, coiled coils are predominantly formed by heptad repeats. Although it is common to find non-heptad repeats interspersed between them, there are just a handful of examples of globally non-heptad coiled coils, some of which have been pointed out in the **Background** section. Currently available coiled-coil prediction tools tend to perform poorly on such sequences, which is to be expected, since they have been trained on datasets that reflect the relative scarcity of non-heptad coiled coils in nature. On the other hand, the fact that these tools perform poorly on such coiled coils means that they cannot be employed to scan protein sequence databases for new examples, with which to, potentially, understand their properties, expand our coiled-coil repertoire for protein design, and build better prediction programs.

To deal with this cyclic dependency problem, we set out to generate a reliable set of examples of proteins families that featured long segments of hendecads, the most common non-heptad coiled-coil repeats in nature. For this, we filtered the protein sequence database for features compatible with hendecad structure, such as 11-residue periodicity and lack of β-strand prediction, and then performed interactive cluster analyses on the resulting set.

4.2 Results

We scanned the NR database, filtered to 50% maximum identity, for sequences that had at least three consecutive 11-residue repeats, as detected by REPwin[43]; this program computes a self-vs-self alignment and reports repetitive regions that show significant internal sequence similarity, allowing for a fast detection of moderately divergent repeats. We filtered this set further by requiring that these repeating segments would not be predicted as β-strands by PSIPRED[44]. This resulted in a set of about 40k sequences, which we clustered in CLANS[45] for visualization and interactive analysis (Fig. 4). The clustering was performed using all-vs-all pairwise comparison with BLAST[46], although we masked the repetitive segments in order to reduce the number of spurious matches.

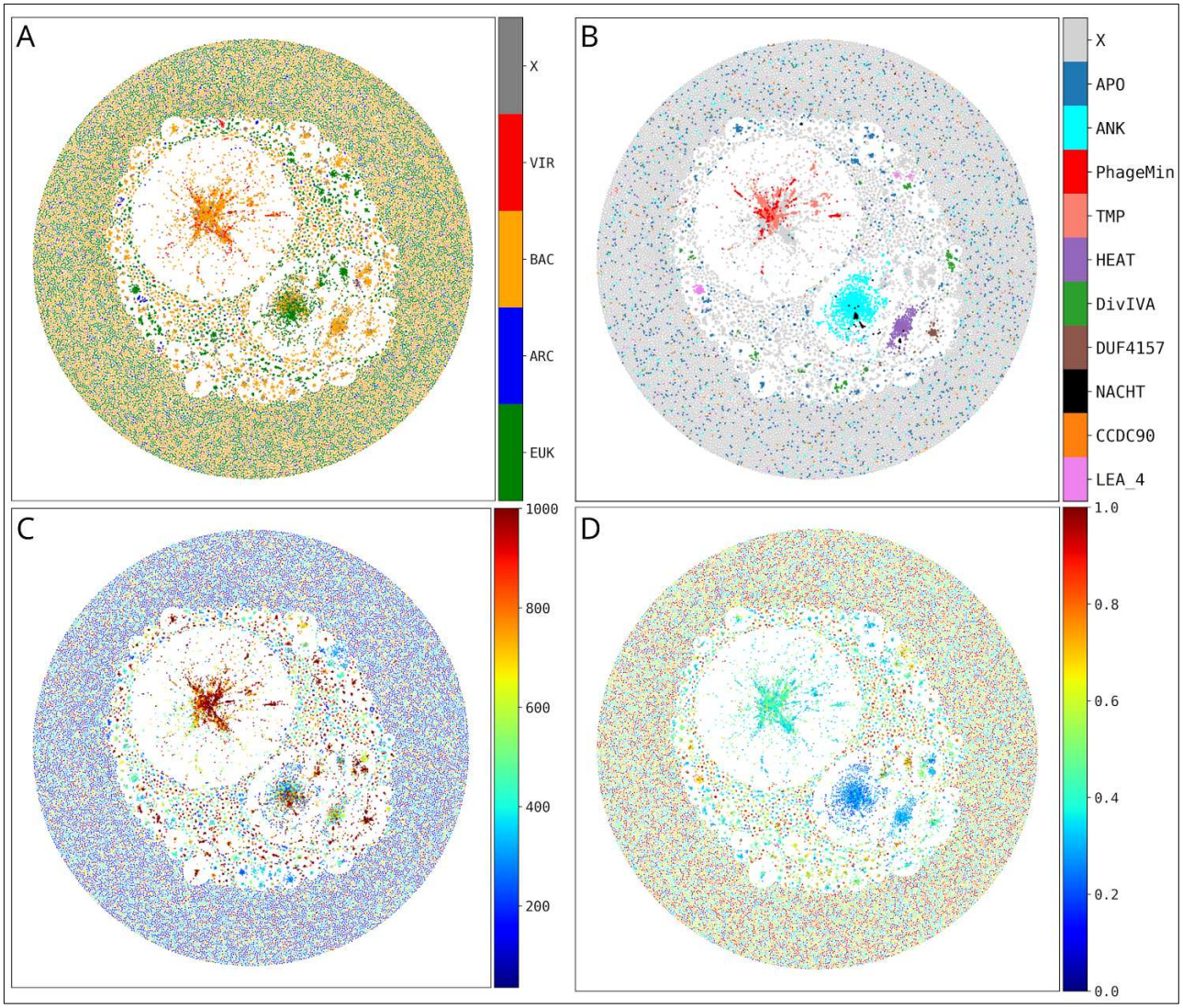


Figure 4: The CLANS cluster map of our dataset, at a p-value threshold of 1E-14. The four panels show the map colored by (A) taxonomy, (B) top 10 detected PFAM domains, (C) sequence length, and (D) identity of the three consecutive most identical 11-residue repeats.

The largest, central cluster in our map consisted of phage tail tape measure proteins (TMP), which are thought to oligomerize as extended tubes and play a fundamental role in determining the length of the phage tail[47,48]. The sequences in this group featured extended segments of 11-residue repeats with wide hydrophobic cores compatible with *adeh* geometry (Fig. 5A), although enriched in proline and glycine relative to the average coiled coil. These features are atypical in coiled-coil proteins due to their destabilizing effects on α-helices, but could be easily accommodated in large oligomers, such as the tubes that the TMP proteins are thought to form. Although we were able to obtain robust AlphaFold models for the α-helical solenoid domain that these proteins feature N-terminally (Fig. 5B), modeling the coiled-coil segments proved unsuccessful (Fig. 5C).

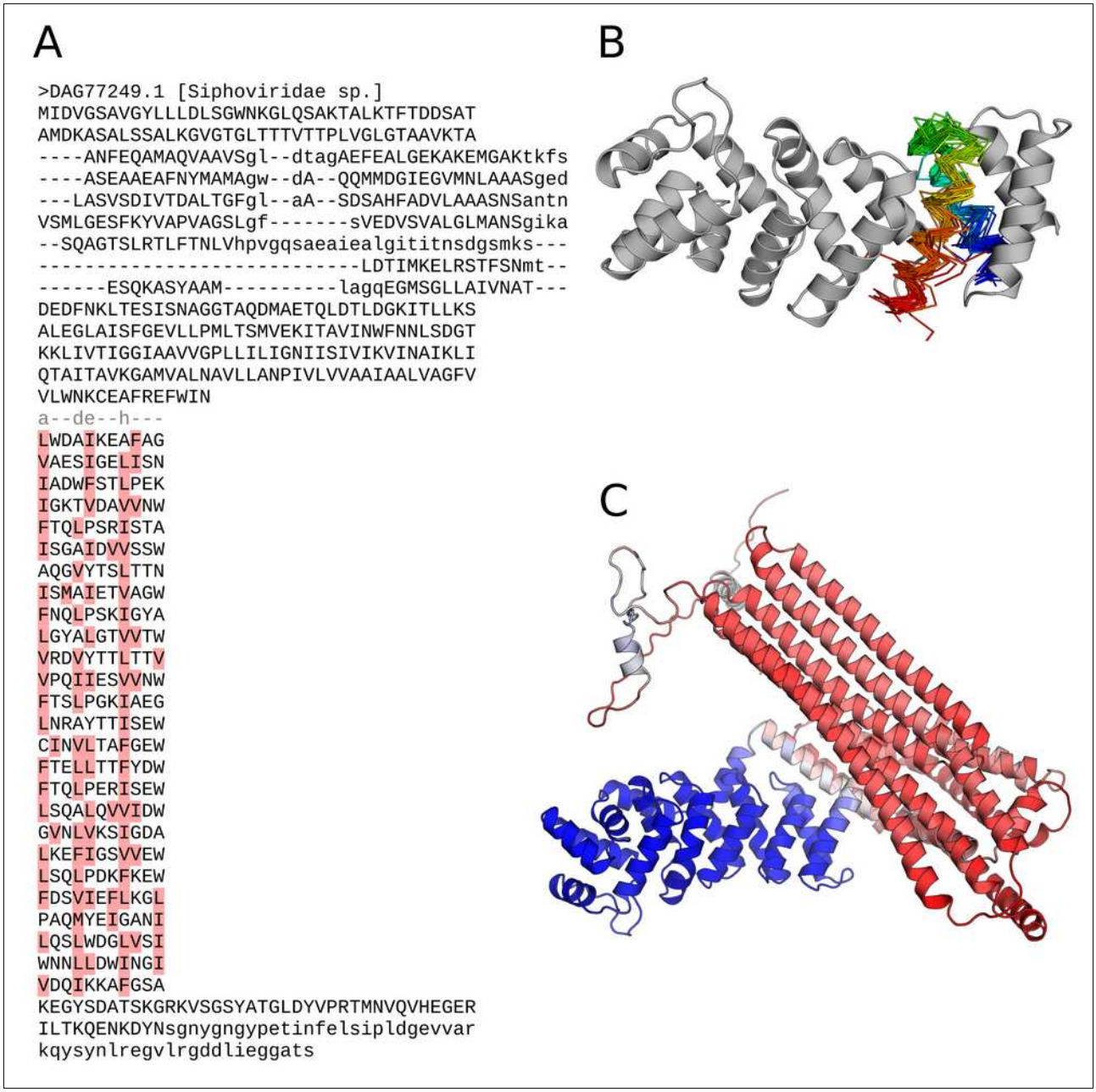


Figure 5: Graphical summary of the viral TMP proteins in our cluster map. (A) Sequence annotation of a representative of the largest subcluster of TMP proteins (DAG77249.1) with, N-terminally, an alignment of the helical hairpins that form the solenoid and, C-terminally, the hendecad stalk annotation. (B) Detail of the N-terminal α-helical solenoid domain; in gray, the AlphaFold model of DAG77249.1; in superimposed rainbow-colored ribbons, model hairpins of representatives from all 42 viral subclusters. (C) AlphaFold monomer prediction of DAG77249.1 (full sequence), colored by pLDDT, from lowest (red) to highest (blue).

The longest hendecad coiled-coil sequences that we found were located in a small cluster containing proteins belonging to *Hyphomicrobiales,* a suborder of alphaproteobacteria. This group was characterized by the presence of two N-terminal transmembrane helices, which form an antiparallel hairpin, followed by a long hendecad stalk, and a disordered region connecting this to the C-terminal short helical domain (Fig. 6A,B). Based on sequence features and various modelling attempts, we determined the trimer to be the most likely oligomeric state. Echoing the case of the

viral TMP proteins, we were able to obtain high-confidence AlphaFold models for the N- and the C-terminal domains, but not for the hendecad stalk, which appeared as extended helices without coiled-coil interactions (Fig. 6C).

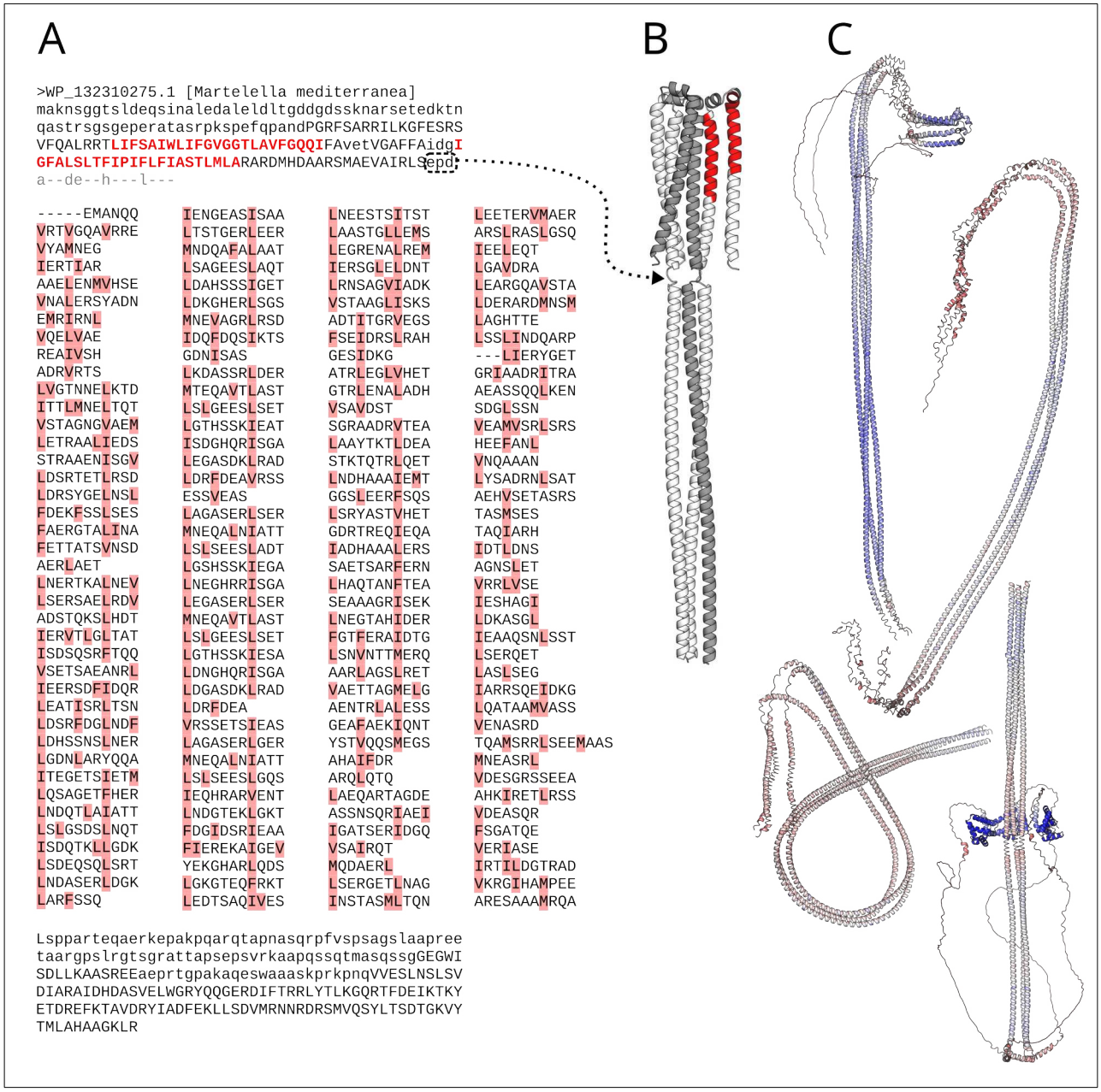


Figure 6: Graphical summary of the MACH proteins. (A) Sequence annotation of the member of the cluster with the longest uninterrupted hendecad stalk (WP_132310275.1), showing N-terminally the two predicted transmembrane helices and C-terminally the hendecad stalk annotation. (B) Trimeric AlphaFold model of the N-terminal part of the protein; the predicted transmembrane helices are colored red in one subunit. (C) Trimeric AlphaFold models of 500-residue overlapping fragments covering the full sequence, colored by pLDDT, from lowest (red) to highest (blue).

Another cluster of sequences with predicted transmembrane helices included ZorA proteins from diverse proteobacteria, which had been previously discovered as part of a bacterial antiphage defense system[49]. These proteins were described as being conspicuously similar to the MotA proteins, which form a proton channel in combination with MotB and play a role in bacterial cell motility[50]. ZorA proteins feature three transmembrane segments N-terminally, followed by a coiled-coil domain which predominantly consists of heptads N-terminally, but shifts to hendecads towards the C-terminus (Fig. 7A,B). AlphaFold models of the N-terminal domain showed that its preferred oligomeric state is the pentamer, in agreement with the oligomerization state adopted by MotA. We were able to obtain high-confidence models for the N-terminal transmembrane domain and for the heptad segment, but not for the hendecad stalk, which, once again, was modeled as a bundle of helices with low confidence and poor side-chain packing (Fig. 7C).

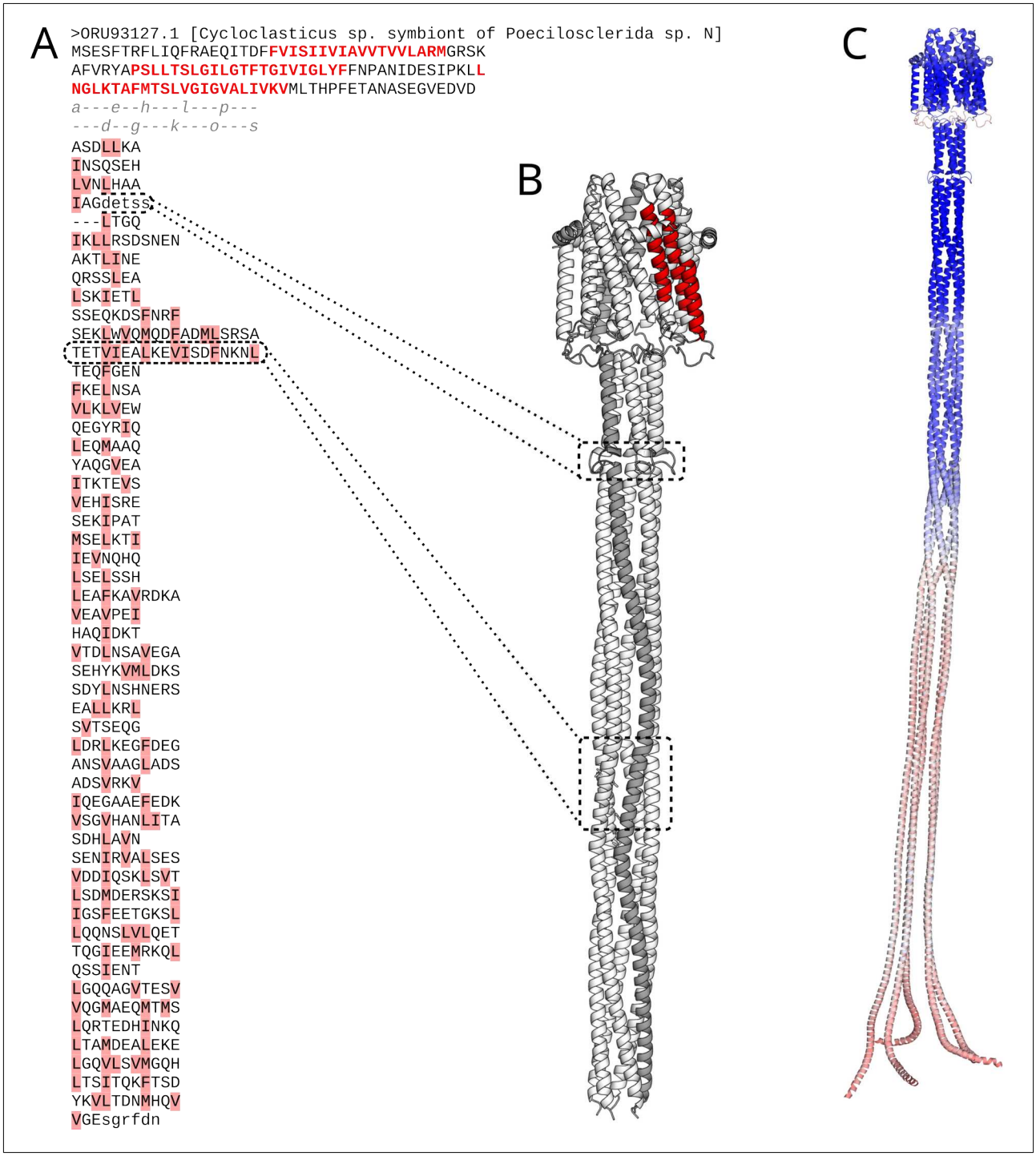


Figure 7: Graphical summary of the ZorA proteins in our cluster map. (A) Sequence annotation of a representative of the ZorA cluster, ORU93127.1, showing the three predicted transmembrane helices (red) and the periodicity of the hendecad stalk, with the two seams of interacting residues marked above the stalk sequence. (B) Pentameric AlphaFold model of the N-terminal part of the protein, the predicted transmembrane helices are coloured red in one subunit. The presence of two 19-residue repeats in a segment of the stalk that has primarily heptad periodicity leads to a strong local perturbation of the supercoil. (C) Pentameric AlphaFold model of ORU93127.1 assembled from two overlapping predictions and colored by pLDDT, from lowest (red) to highest (blue).

Serendipituously, we discovered similarities between various groups of bacterial sequences annotated as Scy[51] and FilP[52]. These sequences featured long stretches of alanine-rich 11-residue

repeats, yet did not cluster together at the p-value threshold that we employed to construct the map, possibly due to the repeat-masking procedure. We noticed that these proteins shared a common N-terminal motif, which by sequence searches also was present in DivIVA[53] and GpsB[54]; these had been crystallized as fundamentally identical short heptad coiled-coil dimers (Fig. 8A,B). Intriguingly, these four proteins had been studied for their implication in the cell division process of Gram-positive bacteria, resulting in ample experimental evidence supporting their pairwise interaction[55,56]. In order to investigate the distribution and features of their conserved N-terminal motif, we performed a broad survey for proteins containing it. We used MEME[57], a tool for motif enrichment analysis, to ascertain that the motif was conserved within each group. Then, we scanned the protein sequence database for the conserved motif, which we determined to be of the form RGYDxxEVD (Fig. 8C,D).

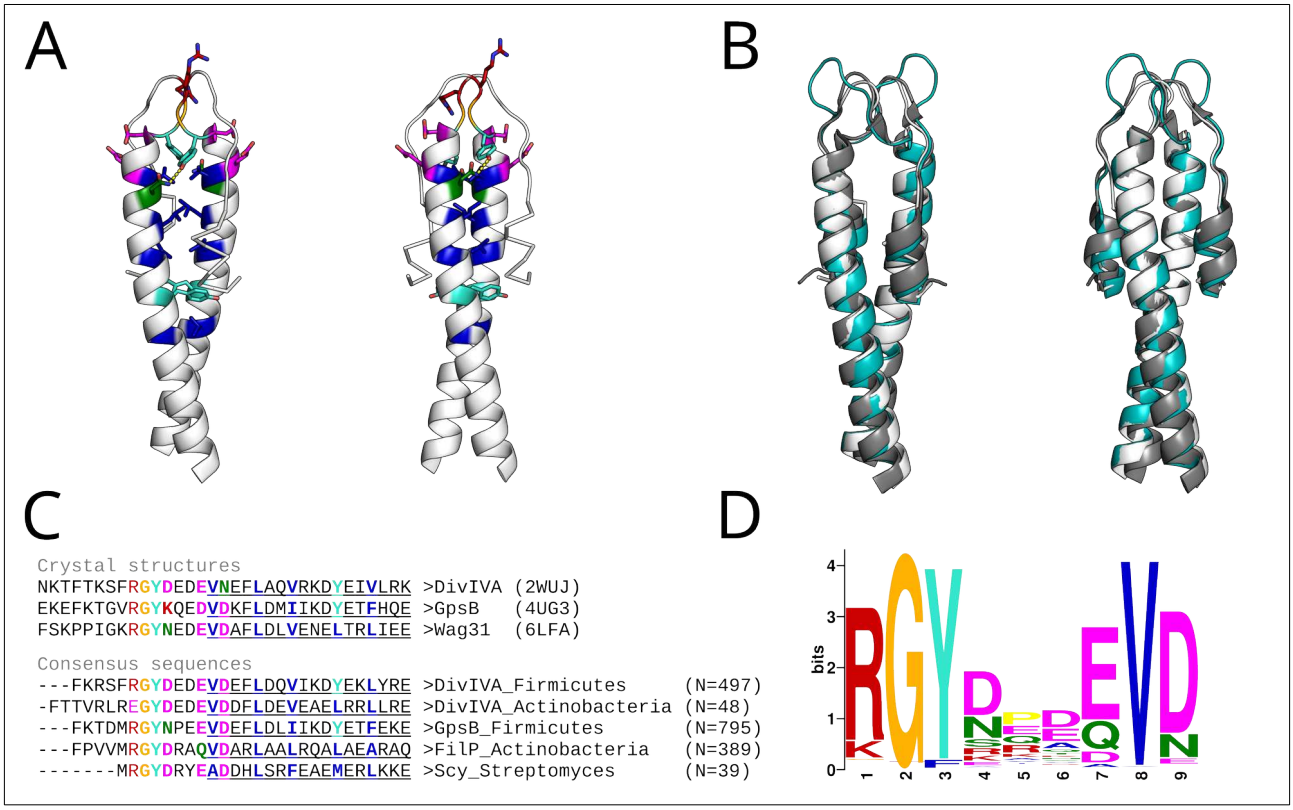


Figure 8: Summary of the similarities between DivIVA-like proteins. (A) Cartoon renders of the DivIVA N-terminal domain (2WUJ); color-coded to match panel C, are the residues of the conserved motif as well as the core residues of the two following heptad coiled coils. (B) Superimposition of DivIVA (2WUJ, white), GpsB (4UG3, gray) and Wag31 (6LFA, teal); RMSD of the superimposition to 2WUJ is1.0 and 2.2 Angstrom respectively. (C) Alignment of the sequences for the structures shown in panel B to the consensus sequences (N =number of sequences) of representatives for the major groups of DivIVA homologs; the conserved motif and the core residues of the two following heptad coiled coils are highlighted in colors, and the Quick2D α-helical prediction is shown as underlined characters. (D) Logo representation of the top scoring motif found by MEME in the set of DivIVA homologs.

In the majority of the resulting sequences, the conserved motif was followed by a short heptad coiled coil, a combination we named a DivIVA-like domain. Performing interactive cluster analysis on these sequences showed the domain to be present in three main groups, Scy/FilP, DivIVA/GpsB, and a third, hitherto unknown family, containing multiple copies of the DivIVA-like domain (we found examples with 2, 4 and 8)(Fig. 9). Based on sequence features and taxonomic distribution, we propose that an ancestral form of DivIVA must have been present in a common ancestor at the time before the division between the main lineages of Gram-positive bacteria, Firmicutes and Actinobacteria. In the firmicute branch, DivIVA was duplicated to form GpsB, which is thought to coordinate the synthesis of peptidoglycan. Both proteins share a fundamentally identical architecture, and have been reported to interact with various proteins involved in bacterial cell division[56]. Within Actinobacteria, another duplication event of DivIVA gave rise to FilP, which additionally, incorporated a longer rod segment, primarily composed of hendecad repeats; within Streptomycetales, an additional duplication of FilP gave rise to Scy, which incorporated a significantly longer hendecad segment. Multiple intragenic duplications of a DivIVA-like protein have also resulted into polyDIVs, which, unlike all the other members of the DivIVA-like superfamily, are not obligate dimers, and can likely assemble into intra-chain pseudodimers[58].

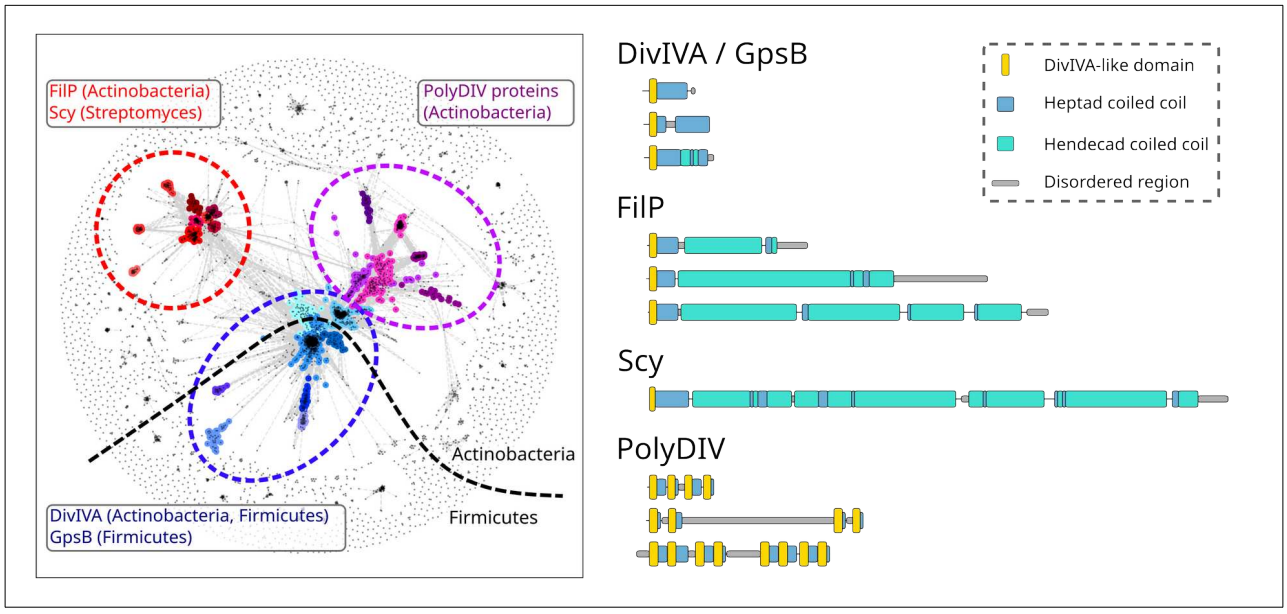


Figure 9: Overview of the DivIVA-like superfamily. (Left) CLANS cluster map of sequences containing the DivIVA-like motif; the different groups are color-coded, and the dashed line indicates taxonomic representation either from Actinobacteria or Firmicutes. (Right) schematic view of the architecture of representative members of each group.

It is noteworthy that the hendecad segments in FilP and Scy show a very uncommon core residue distribution, as they are enriched in alanine. This residue is typically disfavored in core positions as its small size results in suboptimal knobs-into-holes packing. However, the presence of alanine in

core positions of hendecads might be a necessary feature to support the formation of dimers in FilP and Scy; due to the constrained nature of the knobs-to-knobs packing geometry in hendecads, they tend to either assemble into higher oligomeric states, such as trimers, or feature a small side-chain in their core positions to avoid steric clashes. Except for some short fragments, we did not obtain satisfactory models for the alanine-rich coiled-coil stalks in Scy and FilP (Fig. 10).

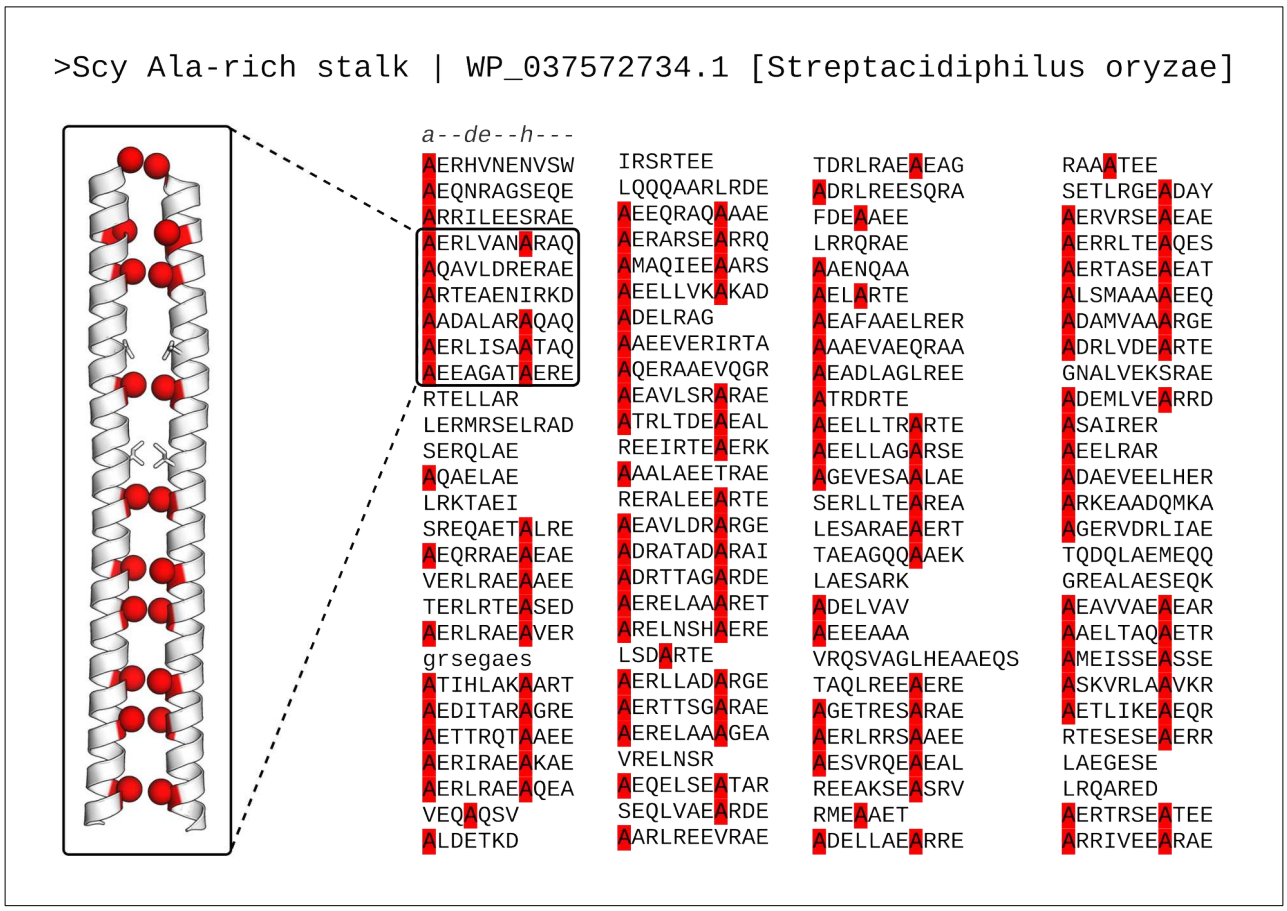


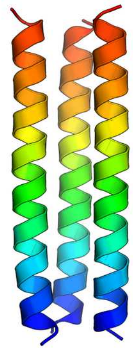
Figure 10: Sequence annotation of the alanine-rich hendecad stalk of the Scy protein (right) and AlphaFold model corresponding to a short fragment (left)

During the analysis of our initial, more general hendecad cluster map, we also noticed many instances of highly similar, adjacent hendecad repeats, pointing to recent amplification events. Some of these occurred within already existing coiled-coil domains, as inferred by the divergent nature of the surrounding repeats; these illustrate the known propensity of coiled-coil domains to gain or lose repeats during evolution[16]. Other amplification events resulted in sequences that consist almost entirely of identical 11-residue repeats, suggesting that they are not real proteins, but rather open reading frames that were erroneously detected as genes during the automatic annotation of genomic sequences; in our cluster map, these behave as *singletons*, this is, sequences without any similarities to any other sequence in the map, which are typically pushed to the outer edge (Fig.

4D). Some of these consisted of sequences compatible with coiled-coil structure, as judged by AlphaFold models. Analysis of the genomic sequences corresponding to these highly-identical 11-residue repeats revealed that some of them had diverging repeats at the DNA level, indicating a possibly ongoing evolutionary process (Fig. 11). We hypothesized that these illustrate a mechanism by which new coiled-coil domains could potentially arise *de novo* from non-coding sequence, by I) random sampling of a single repeat unit compatible with coiled-coil structure, II) amplification of said repeat, and III) acquisition of genomic features that allow for the gene expression and translation.

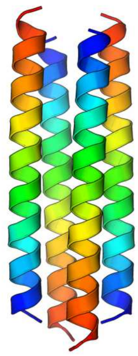
A

```
>GAO15575.1 [Ustilaginoidea virens]
ATGGCACCGACCAACGAAAAACCAGATTGCATC      MAPTNEKPDCI
ACTTTGGCAAATTATTTTCGAAAAGTAGGAGAC      TLANYFRKVGD
GAGGTTGAACTTTTCCAAAGTCTTCCTGCTCTG      EVELFQSLPAL
GATATTGGCGCAGCACTACTGGAACGCATGGAC      DIGAALLERMD
CGATTAATGCTCGAG                        RLMLE
ACGGCATCGTTCCGCCGCGAGGTTCAGTCAGAG      TASFRREVQSE
CTTACGTCGTTCCGCCGCGAGGTTCAGTCAGAG      LTSFRREVQSE
TTTATGTCGTTCCGCCGCGAGGTTCAGTCAGAG      FMSFRREVQSE
TTCACGTCGTTCCGCCGCGAGGTTCAGTCAGAG      FTSFRREVQSE
TCCACGTCGTTCCGCCAAGAATTTGACATAAAG      STSFRQEFDIK
CTCCGAGCAATGTAA                        LRAM
```



B

```
>EFO99190.1 [Caenorhabditis remanei]
ATGATACCCCATCGTCTCAAACTGAGGGAACTC      MIPHRLKLREL
CTGGAGCAACATTGTCTTAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTTAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTTAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTTAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC      LEQHCLKLREL
AAAATTGGTTAA                           KIG
```



C

```
>OWF39119.1 [Mizuhopecten yessoensis]
                           ATGAAT                MN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT      VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT      VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT      VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT      VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT      VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT      VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT      VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT      VQKVLSKMSDN
GTACAAAATGGTTCACAATGA                  VQNGSQ
```

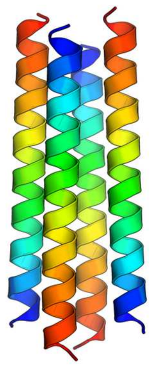


Figure 11: Examples of recently amplified sequences compatible with coiled-coil structure; DNA sequence, protein sequence, and AlphaFold model of the highlighted fragment are shown for (A) a fragment with non-identical DNA and protein repeats, (B) non-identical DNA but identical protein repeats, and (C) identical DNA and protein repeats.

4.3 Discussion

In our survey for hendecad proteins, we aimed for broadness and confidence, not completion. This lead us to choose RepWin as a repeat detection routine, because based on our experience, it is able to detect unambiguous repeat periodicities which nevertheless are formed by reasonably divergent repeats. The likelihood of RepWin detecting a repetitive segment in a sequence is proportional to both the number of contiguous repeats and to the identity between them. This highlights a potential issue in our survey, this being that there might be hendecad proteins that we cannot detect, due to them being too short or divergent. Detecting these instances of hendecad coiled coils might not be possible with our approach, and would require the development of an automatic hendecad prediction tool similar to the ones that exist for heptads.

Going into this project, our initial aim was to develop such a hendecad prediction tool, following the blueprint of COILS, whose scoring matrices were derived from a representative set of annotated coiled-coil proteins. Part of the reason for the success of COILS was that said representative set, which mainly consisted of long dimeric parallel dimers, captured the sequence preferences of the majority of coiled-coil domains in nature. This turned out not to be the case for hendecads, as the sequence preferences of long hendecad proteins do not seem to be representative of shorter, much more common instances of hendecad motifs, often found interspersed between heptad repeats. This divergence between the sequence properties of long and short hendecad domains can be rationalized in various ways for the three longest hendecad coiled-coil domains that are currently known, the phage tail TMPs, the MACH family, and the Scy proteins.

Phage tail TMP proteins are thought to determine the length of the bacteriophage tails by assembling into a hollow tube that forms the interior of the tail. As such, it would be reasonable to assume that their hendecads would be bifaceted, this is, featuring two hydrophobic seams, each pointing to each of the two flanking helices instead of to the center. Bifaceted coiled-coil repeats are classified on the basis of how much spacing there is between their hydrophobic seams, and thus, for heptads, the maximum spacing is 1 residue, also known as type III bifaceted (see **Background** section). Theoretically, the maximum angle between hydrophobic seams that such a coiled-coil helix would allow is 154 degrees (360 / 7 x 3), imposing an upper limit on the size accessible to heptad-based coiled-coil barrels (i.e., limiting the number of helices). This constraint could be aleviated in hendecad-based coiled-coil barrels, as the maximum angle between their hydrophobic

seams is 163 degrees (360 / 11 x 5). This wider aperture would allow the incorporation of additional helices, thereby facilitating the creation of larger cavities within a barrel structure. It is tempting to hypothesize that the fact that TMP proteins are enriched in hendecads is related to their need of assembling into large enough tubes with which to permit the transit of genomic material, but currently, the details of their exact role in genome injection are not clear (for instance, it has been proposed that the TMP proteins are injected before the DNA, and then undergo a conformational change to form a stable channel[59]), and the stoichiometry of the assembly is yet to be elucidated. A stronger case for how hendecads allow larger assemblies to be formed, owing to the increased separation between hydrophobic seams, is found in the SPFH family of proteins, which includes Stomatin, Prohibitin, HlfKC, and Flotillin[60], and in the vault protein[61]. All these assemble into large coiled-coil basket-like structures, the largest of which comprises 44 helices. Instead of parallel to the length of the structure however, these massive complexes align their coiled-coil helices at an angle, which likely increases their structural stability.

The 11-residue repeat of Scy, as well as its alanine-rich core, which makes possible a stable dimeric arrangement, had already been observed before we published our work regarding the DivIVA-like domain[62]. Our sequence analyses of this superfamily suggest an evolutionary scenario whereby a duplication of DivIVA resulted into FilP, and a subsequent duplication of this resulted into Scy. This evolutionary transition from short dimeric heptad into a fibrous hendecads implies that, after the initial duplication from DivIVA, FilP was under pressure to maintain its dimeric topology. This would explain why the hendecad repeats that it acquired had to feature small residue in core positions, but it does not inform whether a hendecad periodicity was necessary or not. Currently it is not known whether the FilP and Scy hendecad stalks provide a functional feature that a heptad stalk would not be able to provide, as the biophysics of the DivIVA-like superfamily are not well understood. Thus, it is possible that the initial acquisition of a short hendecad segment was a fortuitous event, and that by amplification, it expanded to form the long hendecad stalks that we observe today.

The MACH proteins were conspicuous among other clusters in our map due to their long segments of hendecad repeats. While the phage tail TMPs and the Scy proteins featured specialized sequence features, clearly fitting their particular topology, the hendecad segments of MACH would seem more representative of a "standard", more common type of hendecad coiled coils; furthermore, AlphaFold models of the N-terminal transmembrane region, as well as analysis of the coiled-coil stalk, support the formation of a trimer, the preponderant form of hendecads in nature. In spite of

this, the MACH hendecad stalks are enriched in glycine relative to the average coiled coil, which might confer the fiber additional flexibility. Based on the length of their coiled-coil stalk and their membrane anchor, it is likely that these proteins are involved in a form of intracellular compartmentalization or provide structural support.

More generally, our results illustrate the topological diversity that preponderantly hendecad coiled coils exhibit in nature, as opposed to heptad fibers of a comparable size, which tend to be dimeric. This divergence could mean that it is not possible to develop a broadly-applicable hendecad scoring matrix, akin to that for heptads in COILS. How to move forward, towards a more general coiled-coil prediction model? The usability of COILS and the subsequent coiled-coil prediction programs stemmed from their ability to predict the presence and register of coiled-coil domains from sequence. A straightforward way to extend the register assignment to any coiled-coil compatible sequence would be to depart from the idea of a heptad or hendecad register, and instead adopt a framework where coiled coils would be formed by blocks of 3- and 4- residues. Training such a model could prove difficult however, given the diverse sequence preferences that are needed to encode the multitude of coiled-coil topologies. Currently, a formal unifying framework to account for the outstanding structural diversity that coiled coils can adopt seems out of reach. Instead, the field seems to have shifted its focus towards the possibilities that deep-learning based programs such as AlphaFold bring, by producing a three-dimensional structural prediction that incorporates some form of biophysics.

5 On the modeling of coiled coils

5.1 Introduction

The ubiquitous presence of coiled-coil structures in the proteome of life has motivated the development of various coiled-coil modeling programs. These exploit a number of coiled-coil paradigms, such as the Crick parametric equations that describe their backbone, their symmetry, or their packing constraints. Although not strictly modeling tools, routines to predict coiled-coil topology (number and orientation of helices) play a fundamental role in the modeling effort, and connect this to coiled-coil prediction. In a larger framework, these tools represent the ongoing effort in understanding how natural evolution has built coiled-coil structures with specific binding affinities, stabilities, dynamics, and topologies, and how to design synthetic coiled-coil sequences to perform customized molecular roles.

Within this coiled-coil framework, the different coiled-coil modeling programs have attempted to solve a particular task. For example, programs such as ISAMBARD and CCCP exploit the Crick parametric equations to build coiled-coil backbones, while CCfold threads helical fragments that best fit the underlying sequence. A common feature among these programs is that they require the user to set the topology of the bundle (number and orientation of the helices), effectively removing any predictive capability in this regard. The development of AlphaFold and AlphaFold multimer resulted in a general protein-structure prediction routine, capable of predicting the folds of polypeptide chains from sequence, as well as their relative orientation in multimeric complexes. This chapter will summarize our work in benchmarking the accuracy of AlphaFold in modeling coiled-coil domains, as well as our efforts to expand its usability in this context.

5.2 Results

In order to evaluate the performance of AlphaFold[63,64] in the modeling of coiled-coil domains, we developed a benchmark to compare predicted coiled-coil models with their experimental counterparts. While structure similarity measures such as RMSD (Root Mean Square Deviation) and lDDT (Local Distance Difference Test)[65] are typically used for such comparisons, these may not be that informative in the case of coiled coils, due to such measures not reflecting the kinds of features that we would like to see predicted (number, orientation, and axial rotation of helices). Thus, we devised two benchmarks, designed to evaluate the two main aspects of coiled-coil modeling: local geometry, and global topology.

To benchmark the accuracy of AlphaFold in the task of modeling coiled-coil local geometry, we focused on the degree and direction of supercoiling and on the orientation of each side-chain relative to the central axis (also known as the Crick angle). We used these features to compare experimental structures from the PDB and their corresponding predicted AlphaFold models. For this geometry benchmark, we considered dimeric, trimeric, and tetrameric coiled-coil bundles, filtering out structures with resolutions below 3.5 Å or those with complex topologies other than regular bundles. Then, we produced multimeric AlphaFold models with the number of chains displayed by the crystal structures. After using US-Align[66] to check that the overall topology matched between predicted and experimental models (i.e. that all the chains could be confidently aligned), we used SamCC Turbo[35] to detect coiled-coil segments and to compute their structural parameters (supercoiling and helical axial rotation). Our results show that the predicted models that displayed the correct topology accurately matched the features of the corresponding experimental structures, even those that had been deposited after AlphaFold was trained (Fig. 12, left panel). Notably, most of the predicted models that did not match the topology of the crystal structure consisted of antiparallel arrangements. The overall ratio of correctly assigned topologies was of 85%, although this number was reduced to 62% for a manually curated set of antiparallel bundles.

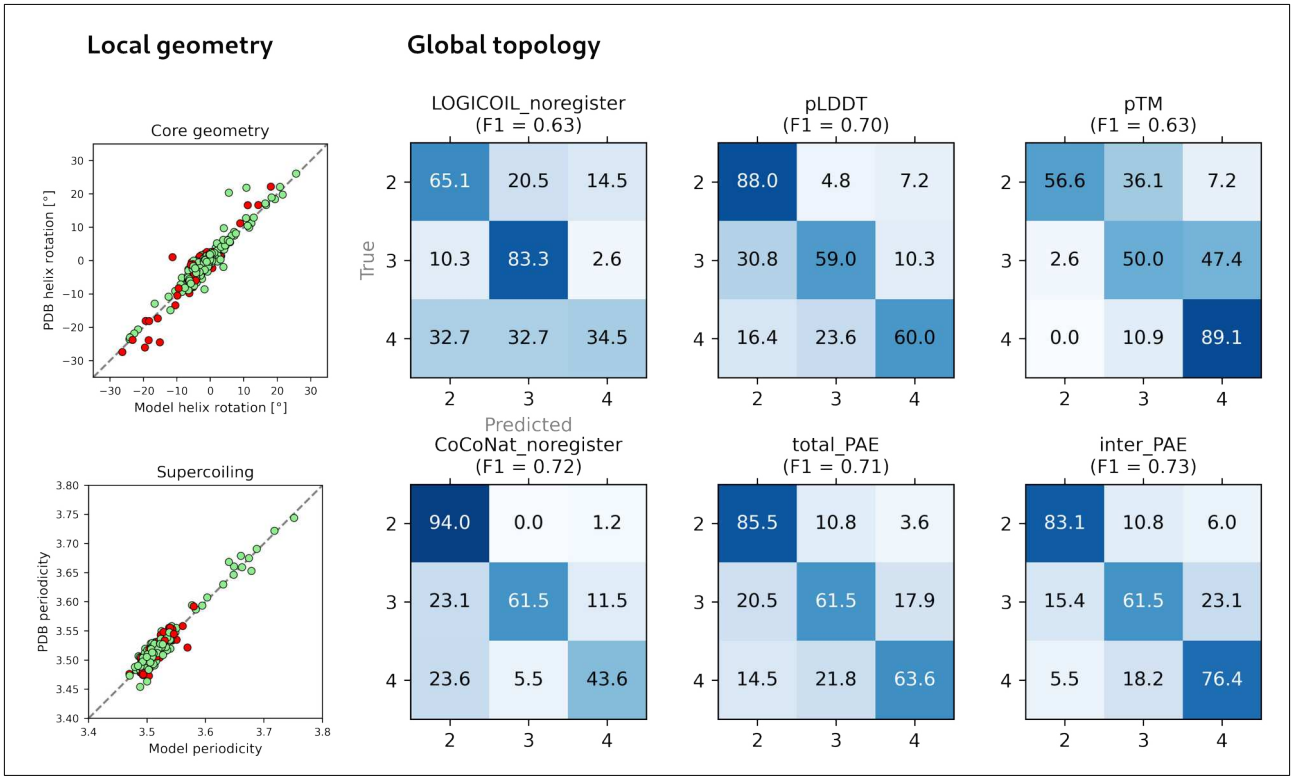


Figure 12: Summary of the AlphaFold coiled-coil modeling benchmarks. (Local geometry) predicted vs experimentally observed core geometry (Crick angle) and supercoiling (bundle periodicity); the diagonal line represents a perfect prediction, and the color shows whether the structure was submitted to PDB before (green) or after (red) AlphaFold was trained. (Global topology) confusion matrices for the predicted vs experimentally observed oligomerization state for the various programs and AlphaFold metrics that we tested; weighted F1 scores are shown for each matrix; the rows in the (_noregister) matrices do not sum up to a 100 because these programs sometimes yielded inconclusive predictions (e.g., 27.3% of the tetramers were predicted as inconclusive by CoCoNat).

The global topology benchmark aimed to assess whether AlphaFold models and their scores could be used with predictive purposes to infer the oligomeric state of coiled-coil domains. For this, we considered a subset of experimental structures from the geometric benchmark set, removing heterooligomers and structures for which the heptad register could not be confidently assigned. We chose to benchmark LOGICOIL[31] and CoCoNat[33] as representatives of state-of-the-art coiled-coil oligomerization state prediction methods. Although these programs can produce predictions based only on sequence, their prediction accuracy can be improved by providing coiled-coil register information; because of this, we tested them in a best-case-scenario, by providing the true register, and in a more realistic scenario, without it (no_register). To test the predictive capabilities of AlphaFold, we modeled each protein in our benchmark set as a dimer, a trimer, and a tetramer, and taking the top-scoring model for each of them, produced oligomeric state predictions based on different AlphaFold quality metrics, such as pLDDT, pTM, and PAE; we further unfolded PAE into the average of the pairwise PAE matrix (total_PAE), and the average of the pairwise PAE scores between different chains (inter_PAE). We computed confusion matrices for all these prediction

schemes by comparing the predicted and the observed oligomerization states (Fig. 12, right panel). Our results showed that, as expected, LOGICOIL and CoCoNat performed quite well when provided with the register (F1 scores of 0.71 and 0.77), albeit they were clearly biased towards predicting dimers and trimers respectively. When register information was not provided (no_register), the prediction accuracy fell significantly (F1 scores of 0.63 and 0.72), while keeping the biases. Among the AlphaFold scores that we tested, pTM performed the worst (F1 = 0.63), while showing a strong bias towards predicting tetramers, and pLDDT featured a better accuracy (F1 = 0.70), with a mild bias towards dimers. PAE, in its inter-chain variant, showed a slightly better accuracy (F1=0.73), with the most balanced distribution of prediction errors across all the AlphaFold scores that we benchmarked.

The fact that the predictive value of AlphaFold scores was comparable to coiled-coil specific methods, even when they make use of the coiled-coil register annotation, motivated us to investigate whether we could develop an even better prediction routine based on AlphaFold. Projecting the AlphaFold internal representation into an observable 2-D map revealed a loose clustering of the three oligomeric states, with a fair number of seemingly out-of-place instances. Encouraged by this observation, we trained a number of machine-learning models to predict the correct oligomeric state from the AlphaFold internal representations, and observed that even a simple logistic regressor was able to yield a significantly improved prediction accuracy (F1 = 0.82), without noticeable biases (Fig. 13).

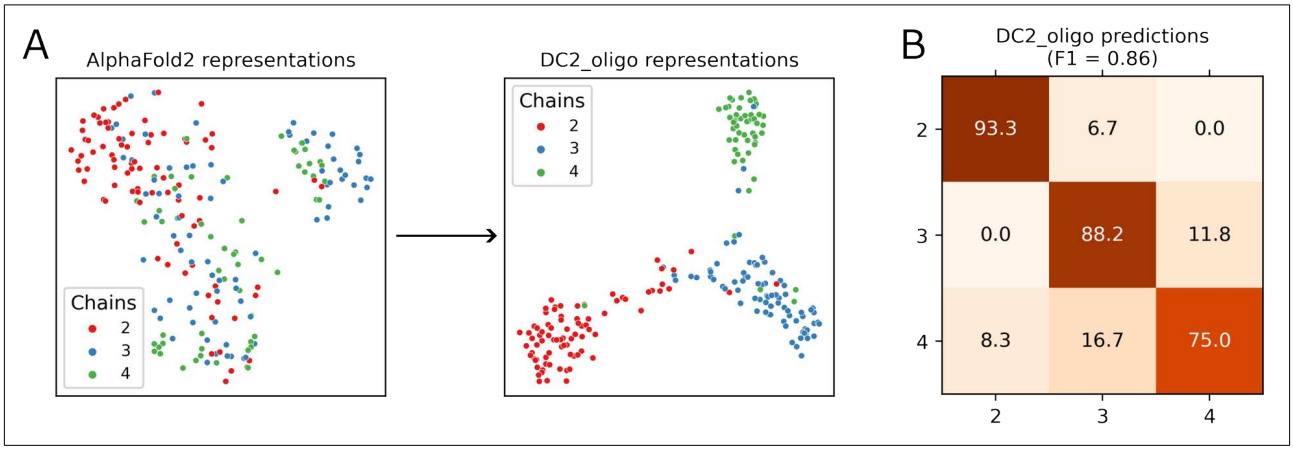


Figure 13: Application of AlphaFold2 representations for the prediction of the oligomeric state. (A) (left) 2D projection of raw AlphaFold2 representations of the benchmark cases and (right) representations obtained from a downstream model; colors represent oligomerization states. (B) A confusion matrix with true and predicted labels on the y and x axes, respectively, showing the performance of the final prediction model trained on AlphaFold2 representations; the weighted F1 score is also shown.

The success of AlphaFold in predicting the oligomeric state of coiled-coil domains suggests that, to some extent, it captures the molecular basis of coiled-coil folding and oligomerization. Understanding these properties is particularly challenging in the case of long coiled coils, due to how folding, stability, and topology determinants are distributed along the length of the structure (e.g. trigger sequences, short motifs that promote coiled-coil formation). This motivated us to develop CCfrag[67], a program to explore the local folding propensities of coiled-coil domains via piece-wise modeling. CCfrag automates the generation of the files necessary to run AlphaFold predictions for a given specification, which is defined by window length, overlap length, and oligomerization state. Once the AlphaFold predictions are produced, CCfrag analyzes the modeled fragments and integrates them into a rich per-residue representation (Fig. 14).

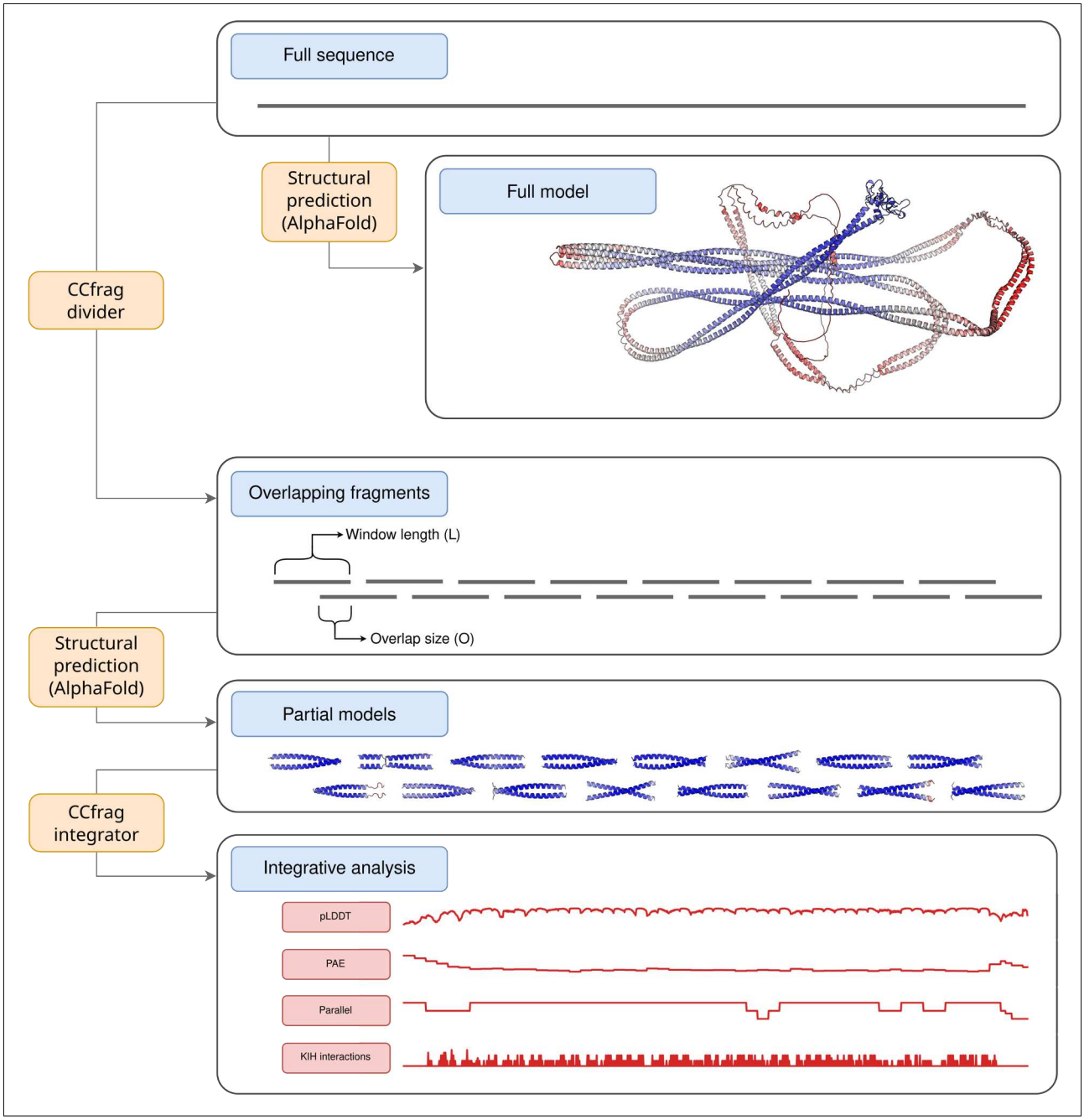


Figure 14: Schematic representation of the CCfrag pipeline. Modeling long coiled-coil domains with AlphaFold generally yields suboptimal models; on top, a full-length model of H. sapiens EEA1 is shown, colored by pLDDT (red-worst to blue-best). By dividing the full-length sequence into fragments, the resulting models are predicted with higher confidence, and can be analyzed for local properties not seen in the full-length model (bottom).

We illustrated some of its potential use-cases with three examples: a) Modeling human EEA1 with different window sizes shows that the confidence of the predictions (plDDT) improves when compared with the full-length model, that AlphaFold is able to detect regions of low coiled-coil forming propensity, and that regions of high coiled-coil forming potential *pull* from their neighboring regions to promote folding (Fig. 15); b) Piece-wise modeling of a member of the MACH family (which is described in a previous section) and subsequent scanning of the models for knobs-into-holes interactions allows the detection of its hendecad stalk, which DeepCoil and COILS fail to detect; c) Scanning the oligomerization landscape of the human coronavirus spike protein via piece-wise modeling shows regions with high coiled-coil forming potential, and recapitulates assemblies that have been experimentally characterized.

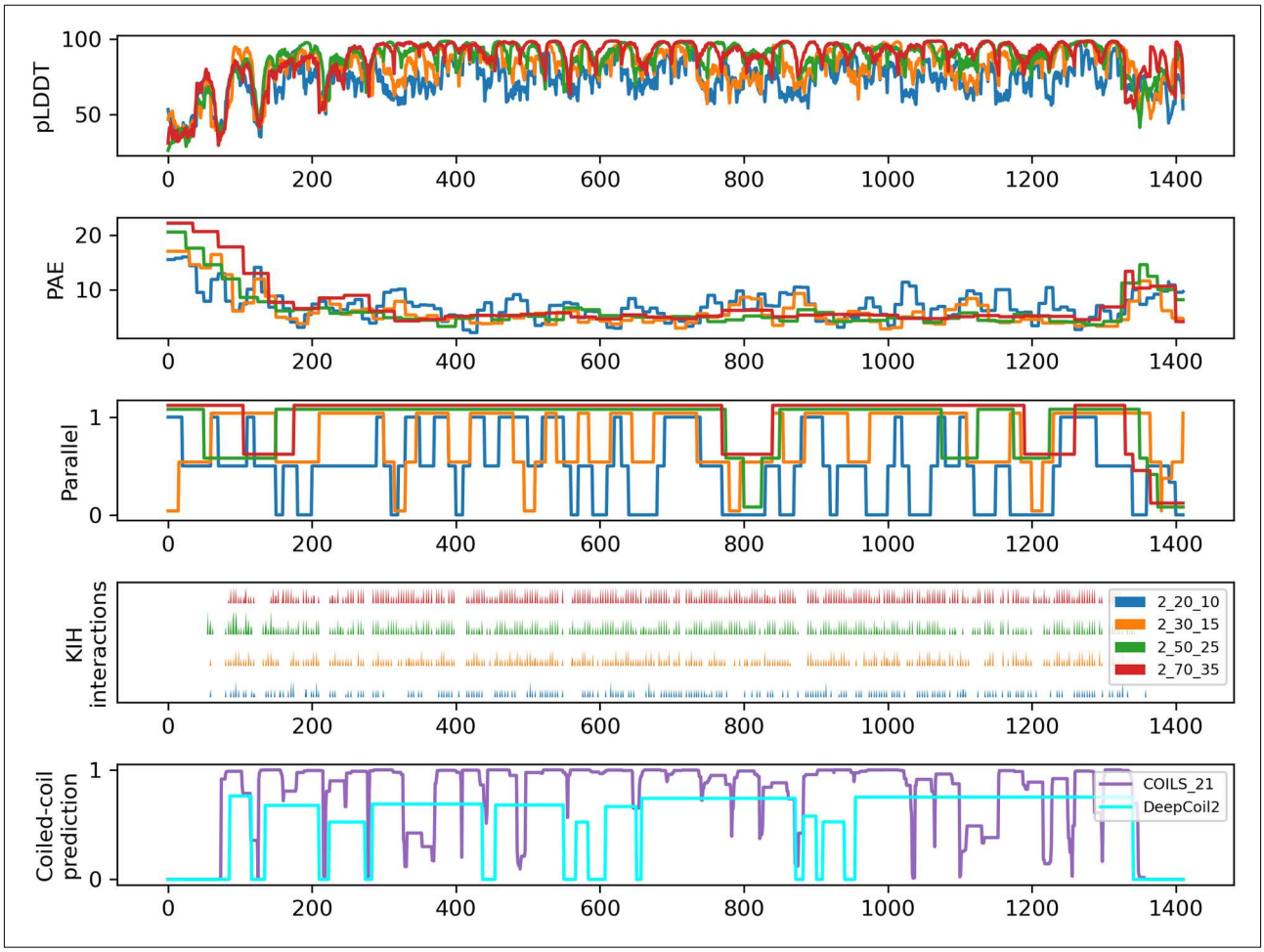


Figure 15: Graphical summary of the CCfrag representation of EEA1 of H. Sapiens. The protein is modeled as a dimer, in four different specifications with window sizes of 20, 30, 50, and 70 residues and an overlap of half the corresponding window size. In the top 4 panels, plots for various features for each specification (color-coded as in the legend in the right of the fourth panel) are shown (pLDDT, PAE, Parallel, KIH interactions), averaged for each residue (since the overlap is half the window size, each residue is covered by two models, resulting in two values for each feature, which are averaged for the visualization). In the bottom panel, coiled-coil prediction probabilities for COILS (window size = 21) and DeepCoil2 are shown.

5.3 Discussion

It is somewhat disappointing to observe that, after so many advances in coiled-coil bioinformatics driven by our understanding of this motif, the most powerful coiled-coil modeling tool has been borne out of a deep-learning model for general protein structure prediction. For all of its merits though, AlphaFold presents two outstanding issues in the modeling of coiled-coil domains: it tends to fold back long coiled coils, creating spurious contacts between side-chains that are distant in sequence, and it fails to model more *exotic* coiled-coil families, such as the hendecad stalks that we described in the previous section. The bent models could be explained by the repetitive nature of coiled-coil sequences, which may affect AlphaFold's pairwise distance prediction between residues, or by a bias in the training set that drives AlphaFold towards increasing the number of contacts (long coiled-coil fibers, which feature a comparatively low number of contacts compared to the average protein structure, are hardly represented in the PDB). The inability of AlphaFold to model certain coiled-coil families could also be attributed to a lack of representation of those in the PDB. Attempting to produce models of such sequences, such as MACH or SCY proteins, typically yields helical structures with poor coiled-coil packing and low pLDDT. This suggests that AlphaFold is more driven by inductive reasoning (i.e. structures that are similar in sequence) than by the biophysical principles that it has implicitly learned.

Similar issues occur when attempting to model long canonical coiled coils (i.e., typical heptads, robustly predicted by coiled-coil prediction programs), such as EEA1; here, extensive segments which are confidently predicted at the sequence level to be coiled coils are misfolded in the structural model. This suggests that the failure to fold some coiled coils is not only due to lack of examples in the training set, but also to a fundamental issue with the internal representation of AlphaFold. A possibly related issue occurs when attempting to produce a multimeric model with a high number of chains, where, for some sequences, AlphaFold consistently produces structures with clashes and sometimes even completely intersecting polypeptide chains (a phenomenon we name "bowls of spaghetti", due to the thin tubes that PyMol uses to represent such structures).

The issue of the spuriously bent AlphaFold models inspired us to develop CCfrag, which uses AlphaFold to probe local coiled-coil folding potential via piece-wise modeling. Not surprisingly, the windowing procedure gets rid of the tendency of long models to bend back and create spurious contacts. More interestingly, the procedure significantly improves the pLDDT of the predictions,

probably due to the fact that modeling multiple short windows is an “easier” task for AlphaFold than modeling full-length models, where the coordinates of one residue depend on a larger number of previous and subsequent residues, in other words, a larger context. This might explain why modeling non-heptad coiled-coil domains yields better results when AlphaFold is run on shorter windows, instead on the full sequence. In some cases, this improvement is dramatic, as illustrated by EEA1 and MACH, where some misfolded segments in the full-length model become confidently-modeled coiled coils.

Removing protein sequences from their larger sequence context opens up a number of questions. It is immediately obvious that the model of a fragment might be non-physical, in the sense that it is not bound by the same constraints as when in the native protein. On the other hand, such non-physical models might be a bridge between what AlphaFold has learned (to predict the atomic coordinates of crystallized folded proteins) and what we want it to teach us (the biologically relevant conformation that proteins acquire in living systems). For instance, the fact that modeling a coiled-coil stalk in short windows shows hotspots of coiled-coil folding that get *propagated* to neighboring residues in larger windows (Fig. 15) is suggestive of AlphaFold’s ability to detect the localized coiled-coil forming potential within given sequence. However, the applicability of such method is harder to gauge when the resulting models adopt an antiparallel orientation. This can be understood, within a coiled-coil framework, as a fragment that does not possess any topological specificity, as this feature is encoded in another part of the stalk; this makes sense specially if we consider the fact that the energy landscape of coiled-coil structures tend to be isoenergetic, and that coiled-coil domains evolve fast, by amplification and deletions of repeats, leading to regions that do not necessarily encode topological information, and only need to be compatible with that which is encoded by the rest of the sequence.

6 Conclusions

Coiled-coil bioinformatics have come a long way since Francis Crick surmised the periodic occurrence of hydrophobic residues in coiled-coil sequences. A rich set of coiled-coil examples in the form of both sequences and structures has allowed the development of many tools for their prediction, design, and modeling. In spite of this, a unitary model for coiled coils has remained elusive, due to the lack of non-canonical coiled-coil representation, and universally applicable coiled-coil modeling routines. In this dissertation, advancements have been made on both fronts by providing new examples of hendecad coiled-coil families and by assessing the coiled-coil modeling capabilities of AlphaFold and their potential extensions.

Our initial motivation for conducting a survey for hendecad coiled coils was to gather a set of reliable examples with which to build a hendecad prediction program. However, globally hendecad coiled coils turned out to be more divergent in sequence than we anticipated, likely due to their topological diversity, exemplified by the large oligomers that bifaceted hendecads able to form. This suggests that hendecads increase the accessible structural space of coiled coils, and provide them with specialized features for their particular function. Although these findings make it more difficult to think of a general coiled-coil prediction routine, they also represent a new benchmark with which to measure and advance our understanding of these versatile protein structure motifs.

It is clear that AlphaFold represents a qualitative leap in protein structure prediction, and that it is poised to be at the core of new tools. Even though it shows considerable potential to advance our understanding of how coiled-coil structures fold and evolve, its successful application is contingent on its ability to generalize the biophysics that govern such processes, something which, by the non-canonical examples that we have outlined, clearly has yet to achieve. There are many challenges still to be tackled. The generalized adoption of deep-learning in biological sciences has brought a problem that tends to follow fast progress, which is that our understanding is lagging behind our ability to generate predictions. Moving forward, the field of bioinformatics must push forward models and routines that promote the understanding of the governing principles of folding and function. In this endeavor, coiled coils might prove, once again, an attractive research target.

7 References

1. Polypeptide chain configurations in crystalline proteins. https://royalsocietypublishing.org/doi/epdf/10.1098/rspa.1950.0142 doi:10.1098/rspa.1950.0142.
2. Pauling, L., Corey, R. B. & Branson, H. R. The Structure of Proteins. *Proc Natl Acad Sci U S A* **37**, 205–211 (1951).
3. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**, 95–99 (1963).
4. Crick, F. H. C. Is α-Keratin a Coiled Coil? *Nature* **170**, 882–883 (1952).
5. Crick. The Fourier transform of a coiled-coil. *Acta Cryst* **6**, 685–689 (1953b).
6. Crick. The Packing of a-Helices: Simple Coiled-Coils. *Acta Cryst.* **6**, 689–697 (1953a).
7. Parry, D. A. D. Analysis of the primary sequence of α-tropomyosin from rabbit skeletal muscle. *Journal of Molecular Biology* **98**, 519–535 (1975).
8. Wilson, I. A., Skehel, J. J. & Wiley, D. C. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. *Nature* **289**, 366–373 (1981).
9. Gruber, M. & Lupas, A. Historical review: Another 50th anniversary – new periodicities in coiled coils. *Trends in Biochemical Sciences* **28**, 679–685 (2003).
10. Harbury, P. B., Zhang, T., Kim, P. S. & Alber, T. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401–1407 (1993).
11. Hartmann, M. D. *et al.* A coiled-coil motif that sequesters ions to the hydrophobic core. *Proc Natl Acad Sci U S A* **106**, 16950–16955 (2009).
12. Testa, O. D., Moutevelis, E. & Woolfson, D. N. CC+: a relational database of coiled-coil structures. *Nucleic Acids Res* **37**, D315–D322 (2009).
13. Lupas, A. N., Bassler, J. & Dunin-Horkawicz, S. The Structure and Topology of α-Helical Coiled Coils. *Fibrous Proteins: Structures and Mechanisms* **82**, 95–129 (2017).
14. Dawson, W. M. *et al.* Coiled coils 9-to-5: rational de novo design of α-helical barrels with tunable oligomeric states. *Chem. Sci.* **12**, 6923–6928 (2021).

15. Koronakis, V., Sharff, A., Koronakis, E., Luisi, B. & Hughes, C. Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. *Nature* **405**, 914–919 (2000).
16. Lupas, A. N. & Gruber, M. THE STRUCTURE OF a-HELICAL COILED COILS. in (2005).
17. Steinmetz, M. O. *et al.* A distinct 14 residue site triggers coiled-coil formation in cortexillin I. *EMBO J* **17**, 1883–1891 (1998).
18. Kammerer, R. A. *et al.* An autonomous folding unit mediates the assembly of two-stranded coiled coils. *Proceedings of the National Academy of Sciences* **95**, 13419–13424 (1998).
19. Peters, J., Baumeister, W. & Lupas, A. Hyperthermostable Surface Layer Protein Tetrabrachion from the ArchaebacteriumStaphylothermus marinus: Evidence for the Presence of a Right-handed Coiled Coil Derived from the Primary Structure. *Journal of Molecular Biology* **257**, 1031–1041 (1996).
20. Murray, D. H. *et al.* An endosomal tether undergoes an entropic collapse to bring vesicles together. *Nature* **537**, 107–111 (2016).
21. Adamczyk, M. *et al.* Revealing biophysical properties of KfrA-type proteins as a novel class of cytoskeletal, coiled-coil plasmid-encoded proteins. *BMC Microbiol* **21**, 32 (2021).
22. Cavini, I. A. *et al.* X-ray structure of the metastable SEPT14–SEPT7 coiled coil reveals a hendecad region crucial for heterodimerization. *Acta Cryst D* **79**, 881–894 (2023).
23. Kreitler, D. F. *et al.* A Hendecad Motif is Preferred for Heterochiral Coiled-Coil Formation. *J Am Chem Soc* **141**, 1583–1592 (2019).
24. Leo, J. C. *et al.* The structure of E. coli IgG-binding protein D suggests a general model for bending and binding in trimeric autotransporter adhesins. *Structure* **19**, 1021–1030 (2011).
25. Sun, L. *et al.* Icosahedral bacteriophage ΦX174 forms a tail for DNA transport during infection. *Nature* **505**, 432–435 (2014).
26. Lupas, A., Van Dyke, M. & Stock, J. Predicting Coiled Coils from Protein Sequences. *Science* **252**, 1162–1164 (1991).
27. Berger, B. *et al.* Predicting coiled coils by use of pairwise residue correlations. *Proceedings of the National Academy of Sciences* **92**, 8259–8263 (1995).

28. Wolf, E., Kim, P. S. & Berger, B. MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci* **6**, 1179–1189 (1997).

29. Delorenzi, M. & Speed, T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics* **18**, 617–625 (2002).

30. Gruber, M., Söding, J. & Lupas, A. N. Comparative analysis of coiled-coil prediction methods. *Journal of Structural Biology* **155**, 140–145 (2006).

31. Vincent, T. L., Green, P. J. & Woolfson, D. N. LOGICOIL--multi-state prediction of coiled-coil oligomeric state. *Bioinformatics* **29**, 69–76 (2013).

32. Ludwiczak, J., Winski, A., Szczepaniak, K., Alva, V. & Dunin-Horkawicz, S. DeepCoil—a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics* **35**, 2790–2795 (2019).

33. Madeo, G., Savojardo, C., Manfredi, M., Martelli, P. L. & Casadio, R. CoCoNat: a novel method based on deep learning for coiled-coil prediction. *Bioinformatics* **39**, btad495 (2023).

34. Kumar, P. & Woolfson, D. N. Socket2: a program for locating, visualizing and analyzing coiled-coil interfaces in protein structures. *Bioinformatics* **37**, 4575–4577 (2021).

35. Szczepaniak, K., Bukala, A., da Silva Neto, A. M., Ludwiczak, J. & Dunin-Horkawicz, S. A library of coiled-coil domains: from regular bundles to peculiar twists. *Bioinformatics* **36**, 5368–5376 (2021).

36. Kumar, P. *et al.* CC+: A searchable database of validated coiled coils in PDB structures and AlphaFold2 models. *Protein Science* **32**, e4789 (2023).

37. Strelkov, S. V. & Burkhard, P. Analysis of alpha-helical coiled coils with the program TWISTER reveals a structural mechanism for stutter compensation. *J Struct Biol* **137**, 54–64 (2002).

38. Dunin-Horkawicz, S. & Lupas, A. N. Measuring the conformational space of square four-helical bundles with the program samCC. *J Struct Biol* **170**, 226–235 (2010).

39. Grigoryan, G. & DeGrado, W. F. Probing Designability via a Generalized Model of Helical Bundle Geometry. *J Mol Biol* **405**, 1079–1100 (2011).

40. Offer, G., Hicks, M. R. & Woolfson, D. N. Generalized Crick equations for modeling noncanonical coiled coils. *J Struct Biol* **137**, 41–53 (2002).

41. Wood, C. W. *et al.* ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design. *Bioinformatics* **33**, 3043–3050 (2017).

42. Guzenko, D. & Strelkov, S. V. CCFold: rapid and accurate prediction of coiled-coil structures and application to modelling intermediate filaments. *Bioinformatics* **34**, 215–222 (2018).

43. Gruber, M., Soding, J. & Lupas, A. N. REPPER--repeats and their periodicities in fibrous proteins. *Nucleic Acids Research* **33**, W239–W243 (2005).

44. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195–202 (1999).

45. Frickey, T. & Lupas, A. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702–3704 (2004).

46. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

47. Mahony, J. *et al.* Functional and structural dissection of the tape measure protein of lactococcal phage TP901-1. *Sci Rep* **6**, 36667 (2016).

48. Kizziah, J. L., Manning, K. A., Dearborn, A. D. & Dokland, T. Structure of the host cell recognition and penetration machinery of a Staphylococcus aureus bacteriophage. *PLoS Pathog* **16**, e1008314 (2020).

49. Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).

50. Nishikino, T. *et al.* Structure of MotA, a flagellar stator protein, from hyperthermophile. *Biochemical and Biophysical Research Communications* **631**, 78–85 (2022).

51. Holmes, N. A. *et al.* Coiled-coil protein Scy is a key component of a multiprotein assembly controlling polarized growth in *Streptomyces*. *Proc. Natl. Acad. Sci. U.S.A.* **110**, (2013).

52. Javadi, A., Söderholm, N., Olofsson, A., Flärdh, K. & Sandblad, L. Assembly mechanisms of the bacterial cytoskeletal protein FilP. *Life Sci. Alliance* **2**, e201800290 (2019).

53. Hammond, L. R., White, M. L. & Eswara, P. J. ¡vIVA la DivIVA! *J Bacteriol* **201**, (2019).

54. Cleverley, R. M. *et al.* The cell cycle regulator GpsB functions as cytosolic adaptor for multiple cell wall enzymes. *Nat Commun* **10**, 261 (2019).

55. Fröjd, M. J. & Flärdh, K. Apical assemblies of intermediate filament-like protein FilP are highly dynamic and affect polar growth determinant DivIVA in *Streptomyces venezuelae*. *Mol Microbiol* **112**, 47–61 (2019).

56. Halbedel, S. & Lewis, R. J. Structural basis for interaction of DivIVA/GpsB proteins with their ligands. *Mol Microbiol* **111**, 1404–1415 (2019).

57. Bailey, T. L. & Elkan, C. Fitting a Mixture Model By Expectation Maximization To Discover Motifs In Biopolymer. 9.

58. Martinez-Goikoetxea, M. & Lupas, A. N. New protein families with hendecad coiled coils in the proteome of life. *Journal of Structural Biology* **215**, 108007 (2023).

59. Cumby, N., Reimer, K., Mengin-Lecreulx, D., Davidson, A. R. & Maxwell, K. L. The phage tail tape measure protein, an inner membrane protein and a periplasmic chaperone play connected roles in the genome injection process of E. coli phage HK97. *Molecular Microbiology* **96**, 437–447 (2015).

60. Fu, Z. & MacKinnon, R. Structure of the Flotillin Complex in a Native Membrane Environment. 2024.05.09.593390 Preprint at https://doi.org/10.1101/2024.05.09.593390 (2024).

61. Tanaka, H. *et al.* The structure of rat liver vault at 3.5 angstrom resolution. *Science* **323**, 384–388 (2009).

62. Walshaw, J., Gillespie, M. D. & Kelemen, G. H. A novel coiled-coil repeat variant in a class of bacterial cytoskeletal proteins. *Journal of Structural Biology* **170**, 202–215 (2010).

63. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

64. Evans, R. *et al. Protein Complex Prediction with AlphaFold-Multimer.* (2021) doi:10.1101/2021.10.04.463034.

65. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).

66. Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat Methods* **19**, 1109–1115 (2022).

67. Martinez-Goikoetxea, M. CCfrag: Scanning folding potential of coiled-coil fragments with AlphaFold. 2024.05.24.595610 Preprint at https://doi.org/10.1101/2024.05.24.595610 (2024).
68. Martinez-Goikoetxea, M. & Lupas, A. N. A conserved motif suggests a common origin for a group of proteins involved in the cell division of Gram-positive bacteria. *PLoS One* **18**, e0273136 (2023).
69. Madaj, R., Martinez-Goikoetxea, M., Kaminski, K., Ludwiczak, J. & Dunin-Horkawicz, S. Applicability of AlphaFold2 in the modelling of coiled-coil domains. 2024.03.07.583852 Preprint at https://doi.org/10.1101/2024.03.07.583852 (2024).

8 Contributions

In this section, the individual contributions of all the colleagues that co-authored the papers presented in this dissertation are detailed.

- MMG - Mikel Martinez Goikoetxea
- AL - Andrei Lupas
- RM - Rafal Madaj
- KK - Kamil Kaminski
- JL - Jan Ludwiczak
- SDH - Stanislaw Dunin-Horkawicz

Paper one: **New protein families with hendecad coiled coils in the proteome of life**[58]

- AL: Conceptualization, supervision, data collection, bioinformatic analyses, interpretation of results, writing-draft, writing-review; MMG: Data collection, bioinformatic analyses, interpretation of results, visualization, writing-draft, writing-review.

Paper two: **A conserved motif suggests a common origin for a group of proteins involved in the cell-division of Gram+ bacteria**[68]

- AL: Conceptualization, supervision, interpretation of results, writing-draft, writing-review; MMG: Data collection, bioinformatic analyses, interpretation of results, visualization, writing-draft, writing-review.

Paper three: **Applicability of AlphaFold2 in the modelling of coiled-coil domains**[69]

- SDH conceptualized the project and provided supervision. All authors contributed bioinformatic analyses. KK and JL provided creative input. SDH and MMG generated figures. RM, SDH, and MMG wrote the manuscript with contributions from the rest of the authors.

Paper four: **CCfrag: Scanning folding potential of coiled-coil fragments with AlphaFold**[67]

- MMG: code development, bioinformatic analyses, visualization, and writing, with significant input from AL

9 Appendix

Paper one: **New protein families with hendecad coiled coils in the proteome of life**[58]

- Status: Published

Paper two: **A conserved motif suggests a common origin for a group of proteins involved in the cell-division of Gram+ bacteria**[68]

- Status: Published

Paper three: **Applicability of AlphaFold2 in the modelling of coiled-coil domains**[69]

- Status: In revision

Paper four: **CCfrag: Scanning folding potential of coiled-coil fragments with AlphaFold**[67]

- Status: Ready for submission

Journal of Structural Biology 215 (2023) 108007

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

New protein families with hendecad coiled coils in the proteome of life

Mikel Martinez-Goikoetxea, Andrei N. Lupas *

Department of Protein Evolution, Max Planck Institute for Biology, 72076 Tübingen, Germany

ARTICLE INFO

Edited by Andrey KAJAVA

Keywords:
Coiled coils
Protein evolution
Hendecad repeats
De novo proteins
TMP
ZorA
SCY
FilP

ABSTRACT

Coiled coils are a widespread and well understood protein fold. Their short and simple repeats underpin considerable structural and functional diversity. The vast majority of coiled coils consist of 7-residue (heptad) sequence repeats, but in essence most combinations of 3- and 4-residue segments, each starting with a residue of the hydrophobic core, are compatible with coiled-coil structure. The most frequent among these other repeat patterns are 11-residue (hendecad, 3 + 4 + 4) repeats. Hendecads are frequently found in low copy number, interspersed between heptads, but some proteins consist largely or entirely of hendecad repeats. Here we describe the first large-scale survey of these proteins in the proteome of life. For this, we scanned the protein sequence database for sequences with 11-residue periodicity that lacked β-strand prediction. We then clustered these by pairwise similarity to construct a map of potential hendecad coiled-coil families. Here we discuss these according to their structural properties, their potential cellular roles, and the evolutionary mechanisms shaping their diversity. We note in particular the continuous amplification of hendecads, both within existing proteins and *de novo* from previously non-coding sequence, as a powerful mechanism in the genesis of new coiled-coil forms.

1. Introduction

Coiled coils are formed by two or more α-helices that wind around a central axis and interlock their side-chains systematically along the core of the structure (Lupas and Bassler, 2017). The geometry of this interaction, widely considered to be the hallmark of coiled coils, is referred to as knobs-into-holes packing (Crick, 1953a). This geometry places the side-chain of a core residue of one helix (knob) into a cavity formed by four side-chains of a facing helix (hole), next to its symmetry-related core residue on the other helix. Sequence repeats of 7 residues (heptads) project two residues per repeat into the core of the helical bundle. If the 7 recurring positions of a heptad are labeled *a-g*, the two core residues are located in positions *a* and *d* (Fig. 1, left panel). These positions are usually occupied by hydrophobic residues, whose burial in the core drives the folding of the structure. Heptads can thus be considered to consist of alternating elements of 3 and 4 residues, each starting with a core residue (3 + 4 = 7), resulting in an average interval of 3.5 residues (7/2) between core residues. Since straight α-helices feature a periodicity of 3.63 residues per turn on average, orienting the core residues towards the central axis requires the helices of heptad coiled coils to wind with the opposite handedness from the constituent α-helices. With respect to the central axis, the structural periodicity of the coiled coil is thus reduced to 3.5 residues per turn, ensuring the recurrence of the knobs-into-holes packing geometry over the length of the structure, which could potentially be extended indefinitely. It is noteworthy that the coiled coil model, essentially as described here, was proposed by Francis Crick in 1953 as an explanation for discrepancies between the expected and observed fiber diffraction data for keratin and other proteins of the α form (Crick, 1953a). His account of coiled coils built of heptads, for which he derived a set of parametric equations describing the backbone structure (Crick, 1953b), was immediately convincing and became the canonical description of coiled coils, decades before the first sequences and structures confirmed the accuracy of his model.

An important extension of this model resulted from the realization that, in higher oligomers, coiled-coil helices show knobs-into-holes packing along two seams of core residues, whose separation on the surface of the helices increases with the size of the oligomer; these helices are called bifaceted (Walshaw and Woolfson, 2001; Woolfson et al., 2012; Zaccai et al., 2011). A further important extension resulted from the observation that coiled coils can have other sequence periodicities than the heptad (Gruber and Lupas, 2003). Although the heptad repeat is by far the most abundant, possibly because it is the only one that allows for continuous knobs-into-holes packing, other combinations of

* Corresponding author.
E-mail address: andrei.lupas@tuebingen.mpg.de (A.N. Lupas).

https://doi.org/10.1016/j.jsb.2023.108007
Received 28 April 2023; Received in revised form 30 June 2023; Accepted 28 July 2023
Available online 29 July 2023

elements of 3 and 4 residues – each starting with a core residue – produce coiled-coil structures (Hicks et al., 1997), albeit at the cost of local departures from knobs-into-holes packing (Lupas et al., 1995). The limiting factor for the sequence repeats of coiled coils is the difference between the average spacing of their core residues and the periodicity of an undistorted α-helix. If this difference is larger than about 0.23, the coiled-coil helix has to wind beyond its breaking point in order to project the core residues towards the central axis. Within this constraint, a number of repeat periodicities are possible and have been observed in nature. The degree of winding they induce in their helices (generally referred to as supercoiling) assumes different values, depending on the difference to an undistorted helix. Thus, periodicities of 3.4 (17/5), 3.5 (7/2), or 3.57 (25/7) are left-handed, because they are smaller than 3.63, periodicities of 3.6 (18/5) or 3.67 (11/3) are essentially straight, and periodicities of 3.75 (15/4) or 3.8 (19/5) are right-handed.

Among the non-heptad periodicities, the hendecad ($3 + 4 + 4 = 11$) repeat is the most frequent and is often found in single copy between heptad repeats. If its 11 positions are labeled *a-k* (Fig. 1, right panel), a hendecad coiled coil either projects residues *a-d-h* into the core, or residues *a-d-e-h*, depending on the degree of axial rotation of the constituent helices relative to the central axis. Historically, hendecads interspersed between heptads have been considered discontinuities and rationalized in terms of an insertion of 4 residues into a continuous heptad pattern (called a *stutter*). Structurally, the addition of 4 residues to the repeat changes the degree of supercoiling and introduces a core layer of residues in which side-chains are projected towards each other and not towards a hole (called knobs-to-knobs packing). More generally, any coiled-coil repeat with consecutive elements of the same type ($3 + 3$ or $4 + 4$), introduces a layer of knobs-to-knobs interactions, which imposes a steric constraint, as the residues engaging in such interactions must either have enough available space or a small side-chain (Gly, Ala), in order to avoid clashing with other knobs. One way in which coiled coils can increase the available space for core packing is by increasing the oligomerization state, i.e. the number of helices in the structure; this is why hendecad coiled coils usually form trimers or tetramers, rather than dimers, which are the preponderant form of heptad coiled coils.

Most long coiled coils are built of heptad repeats and contain small numbers of hendecads, but some are in fact composed predominantly or even entirely of hendecads. The conspicuous presence of 11-residue repeats with the potential to fold into amphipathic helices has been noted in a number of protein families, among them α-synucleins (Bussell and Eliezer, 2003), LEA proteins (Dure et al., 1989), and apolipoproteins (Boguski et al., 1985), but the limited structural information for synucleins and LEA proteins does not support the formation of hendecad coiled coils, and the structures of apolipoproteins show a great diversity of helical bundles with irregular interactions. However, there are also a few examples of structurally well-characterized hendecad coiled coils, such as tetrabrachion (Peters et al., 1996; Stetefeld et al., 2000), the trimeric autotransporter adhesin EibD (Leo et al., 2011), or the H protein of bacteriophage phiX174 (Sun et al., 2014). The very limited number of such examples is contrasted by the fairly widespread presence of 11-residue repeats in putatively fibrous proteins, prompting us to analyze the protein database more broadly for families likely to form hendecad coiled coils. Here, we describe the results of our analysis and discuss the structural properties of several new families of coiled coils.

2. Methods

2.1. Gathering a dataset of potential hendecad coiled coils

We searched for sequences compatible with hendecad coiled-coil structure in the non-redundant protein sequence database from NCBI, filtered to a maximum pairwise identity of 50% (nr50, version August 2022). Because our goal was to identify reliable new instances of proteins assuming this structure, we asked for clear sequence repetition and for predicted secondary structure compatible with a coiled coil. Specifically, we identified sequences that showed 11-residue periodicity over at least 33 residues, as detected by REPwin (Gruber et al., 2005) at a score threshold of 2, and eliminated those whose repeats contained β-strand predictions in PSIPRED (Jones, 1999), run in single-sequence

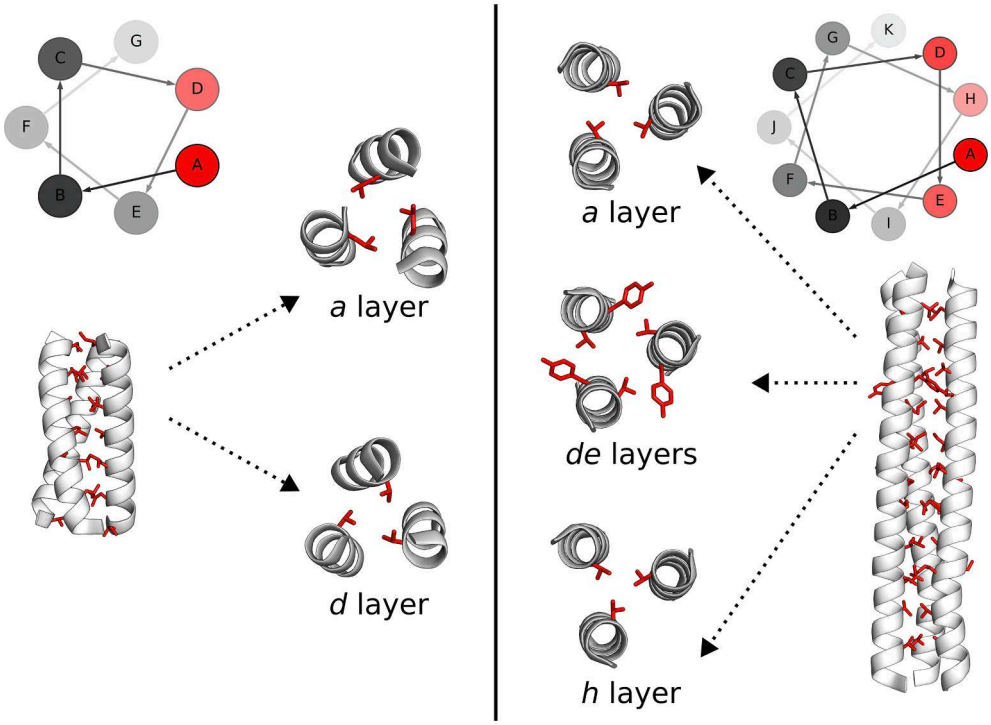


Fig 1. Graphical summary of heptad and hendecad coiled coils. For each structure, we show a helical wheel representation, a side view, and top-down views of the core layers. The heptad trimer (left panel) shows a coiled-coil segment from a voltage-gated channel (PDB 3HFE) and the hendecad trimer (right panel) a segment from the trimeric autotransporter adhesin EibD (PDB 2XQH).

mode (version 4.01). The final dataset consisted of 36,455 sequences (0.03% of nr50).

2.2. Creation and equilibration of the cluster map

We then explored the relationships between the sequences in the dataset by clustering them according to their pairwise BLAST p-values (Camacho et al., 2009) in CLANS (Frickey and Lupas, 2004) (https://toolkit.tuebingen.mpg.de/tools/clans), after masking repetitive segments in the sequences, as detected by REPwin, in order to decrease the number of spurious matches. We performed the clustering until convergence at a p-value threshold 1E-14. We chose this threshold because it was the one that maximized the number of unconnected clusters with a minimum size of 10 sequences (Fig S1). In order to define clusters in the map, we used the Python package NetworkX (https://networkx.org) to construct an undirected graph representation of BLAST p-values at a threshold of 1E-14 and then detected unconnected subnetworks within it, obtaining 136 clusters. The cluster map and its constituent groups of sequences are available as a CLANS file in a Mendeley repository (https://doi.org/10.17632/mbth74w7wy.1).

2.3. Interactive cluster sequence analyses

We analyzed each sequence cluster for taxonomic spectrum (using NCBI taxonomic assignments), the presence of known domains with HMMER 3.3 (Finn et al., 2011) and the PFAM database version 34.0 (Mistry et al., 2021), conserved residues by multiple sequence alignment with ClustalO (Sievers and Higgins, 2018), secondary structure with PSIPRED (Jones, 1999), transmembrane helices with TMHMM (Krogh et al., 2001), intrinsically unstructured regions with IUPRED2A (Erdős and Dosztányi, 2020), tandem repeats with REPwin (Gruber et al., 2005) and coiled-coil segments with COILS (Lupas et al., 1991) and DeepCoil2 (Ludwiczak et al., 2019). Additionally, we selected a number of clusters for in-depth analysis, based on cluster size, length of their 11-residue periodic segments, taxonomic breadth, and the presence of unusual sequence patterns. This analysis was based substantially on tools available in the MPI Bioinformatics Toolkit (Zimmermann et al., 2018), accessible at https://toolkit.tuebingen.mpg.de, and focused on the potential coiled-coil nature of the 11-residue periodic segments, because, being trained mainly on heptad coiled coils, none of the available prediction methods is able to reliably identify hendecad coiled coils. Specifically, the analysis typically started with a PSI-Blast search against the alphafold_uniprot_Aug22 database, which gave us access to precomputed AlphaFold models, and was further extended by secondary structure prediction in Quick2D, domain identification in HHpred, searches for residue patterns against various databases in PatternSearch, and oligomer modelling as detailed below.

2.4. Modeling

We computed structural models with both AlphaFold version 2.3.0 (Evans et al., 2021; Jumper et al., 2021) and its implementation in Colabfold version 1.5.1 (Mirdita et al., 2022). Unless otherwise mentioned in the figure legend, each model was computed from the full sequence. In some cases, we computed models for varying oligomeric states, and compared the respective structural scores (pTM, pLDDT) for information on the preferred oligomeric state. All the models that we show are available at a Mendeley repository (https://doi.org/10.17632/mbth74w7wy.1).

2.5. Detection of recently amplified hendecads

In order to detect proteins with tandem repeats of high pairwise sequence identity in our dataset, indicative of recent DNA amplification, we aligned the repeats and computed the average of the pairwise sequence identities (we refer to this as the Column Identity Average, or CIA, which can assume values between 0 and 1). We then scanned sequences with a window of 33 residues and looked for CIA values above 0.9, as indicators of recent amplification events. We were particularly interested in sequences with CIA values of 1.0 (complete identity) as potential instances of new hendecads in the process of emergence from non-coding DNA. In Fig. 2D, we show the cluster map colored by the maximum CIA value of each sequence.

3. Results and discussion

3.1. A cluster map of proteins containing 11-residue repeats

We searched for new hendecad coiled-coil families in the nr50 database by scanning for sequences with 11-residue repeats that lacked β-strand prediction (see Methods). This scan yielded a set of approximately 40 k sequences, whose relationships we explored by cluster analysis with CLANS. This program represents protein sequences as points in a 2D or 3D space, and lets them attract or repel each other in proportion to the statistical significance (p-value) of their all-against-all pairwise comparison. CLANS then generates a representation of the map by iteratively minimizing the energy of the system at a given p-value threshold. By this procedure, related sequences group together towards the center of the map in connected clusters and the unrelated ones (singletons) drift to the periphery. These singletons may be sequences that do not have matches specifically within the given set of sequences, *true* singletons that lack any detectable homologs in databases (orphan genes), or open reading frames that result from automated genome annotation, but do not actually encode proteins.

The similarity between sequences may stem from a shared ancestry (homology), which could be global, and thus reflect a family of proteins, or local, and reflect the presence of common domains in otherwise unrelated proteins. It is also possible that similarities between unrelated sequences may have arisen by convergent evolution. Such is the case with many coiled coils, which typically match one another in sequence searches even when they are unrelated (Mistry et al., 2013), due to their common sequence periodicity and residue distribution. Therefore, in order to reduce the number of spurious matches in our cluster map, we masked the repetitive segments before making the all-against-all pairwise comparison. At a p-value threshold of 1E-14 and a minimum number of 10 sequences for a cluster, we detected 136 clusters with no connections between them, the most salient of which we will present in this paper.

The largest cluster in the map contains sequences of phage and bacteria (Fig. 2A), many of which are annotated as phage tail length Tape Measure Protein (TMP, Fig. 2B); we will discuss these in the next section. The second and third largest clusters (Fig. 2 A,B) contain bacterial ankyrin and eukaryotic HEAT repeats, respectively. Both these motifs are helical hairpins with a similar hydrophobic profile and packing geometry as coiled coils, and, more importantly, a repeat size that is a multiple of 11, which is why they were detected in our scan for helical proteins with 11-residue sequence periodicity. We note the presence of many further proteins containing putative ankyrin or HEAT repeats in the unclustered periphery of the map, which do not make connections at the 1E-14 threshold. We did not consider ankyrin or HEAT repeat-containing proteins further in our search for new coiled-coil families.

To our surprise, apolipoproteins, which represent the most frequent PFAM domain annotation in our dataset (Fig. 2B) did not form any larger clusters and are mostly found in the unconnected periphery of the map. Apolipoproteins are helical proteins with an underlying degenerate 11-residue sequence periodicity (Boguski et al., 1985), but they form fairly irregular structures that combine knobs-into-holes packing with ridges-into-grooves and could be considered to form a structural bridge between regular coiled coils and irregular helical bundles (Lupas et al., 2017). We therefore also omitted them from our search for new coiled coils. Among the sequences with frequent PFAM domain annotations

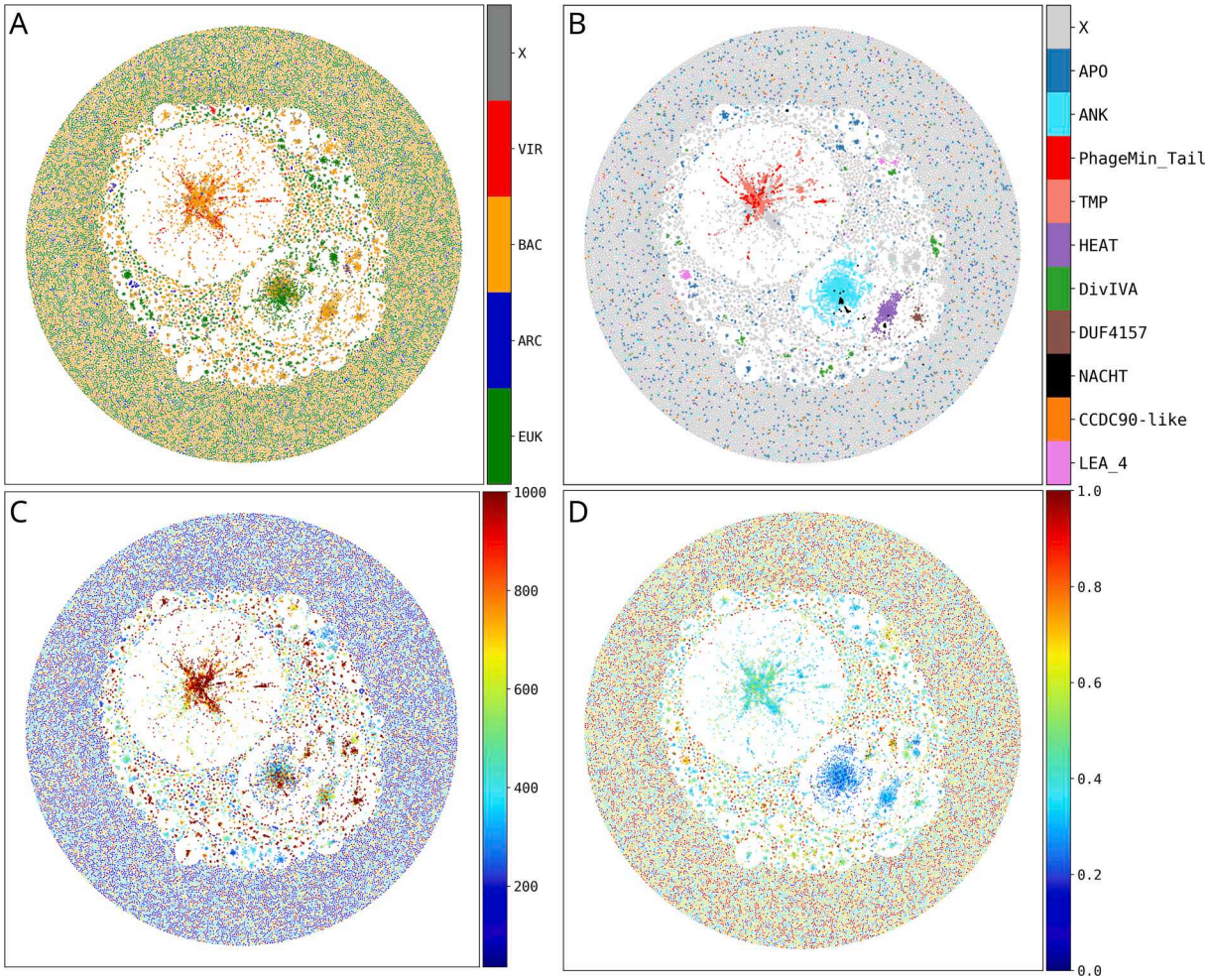


Fig 2. The CLANS cluster map of our dataset, at a p-value threshold of 1E-14. The four panels show the map colored by (A) taxonomy, (B) detected PFAM domains (top 10), (C) sequence length and (D) maximum column identity average.

(Fig. 2B) we also omitted two additional groups from further consideration. One comprises LEA proteins of plants (Late Embryogenesis Abundant), which have 11-residue sequence periodicity and have been proposed to form hendecad coiled coils (Dure et al., 1989), and the other proteins predicted to contain the DUF4157 PFAM domain. A potential coiled-coil structure for these proteins is neither supported by our analysis nor by recent structure prediction methods. Two groups of proteins that we expected to find were those with the PFAM domains DivIVA (Martinez-Goikoetxea and Lupas, 2023) and CCDC90-like, which we described in detail elsewhere (Adlakha et al., 2019). However, both of these groups failed to aggregate into connected clusters, which may be due to the fact that at our p-value threshold, their short hallmark N-terminal domains are not sufficient to bring them together after the repeat-masking step.

3.2. *Phage tail length Tape Measure proteins (TMP)*

The largest cluster in the map consists mainly of proteins annotated as TMPs. Taxonomically, many of these belong to caudoviruses (tailed phages), but most are annotated as bacterial (from firmicutes and actinobacteria), presumably because they are part of prophages. Although little is known about the structure of these proteins, they clearly extend along the length of the phage tail, whose dimensions they appear to determine. In addition to extended hendecad repeats, TMPs often also contain smaller numbers of heptads and pentadecads (Fig. 3A).

Further clustering at a more stringent p-value threshold of 1E-45 yielded 42 smaller subclusters. We ran BLAST searches for one representative per subcluster against the nr database with a taxonomic filter for viruses, and found that all of their top matches were to Siphoviridae and to a lesser extent Myoviridae, two of the three families of Caudoviruses (the third, Podoviridae, has only very short tails (Ackermann, 1998)). The sequence identity between different subclusters is very low (in the 'midnight zone' of Rost (Rost, 1999)). Within every subcluster, however, there is a highly conserved N-terminal region, which, despite sharing little sequence similarity between subclusters, is consistently predicted as a solenoid fold, with a structurally conserved α-helical repeat unit (Fig. 3B). This domain is the only part of these proteins that is predicted confidently in the monomeric state (Fig. 3C).

The hendecad region forms the C-terminal part of these proteins, has a clear helical propensity, and is presumably responsible for the elongated structure of the protein. Although it has been noted before that many TMP proteins show 11-residue repeats (Mahony et al., 2016), their coiled-coil nature has not yet been suggested. This may be due to them having many features that are usually disfavoured in coiled coils, such as an enrichment of proline and glycine residues. However, imagining these proteins in the context of an assembled phage immediately suggests that they might form an oligomeric α-helical barrel, in which their sequence properties could easily be accommodated. Particularly the long helices of SPFH-family proteins come to mind here, which have similar residue preferences and assemble into large helical complexes via bifaceted interactions between hendecad repeats (but too divergent to be detected by the approach we took here). Nevertheless, despite the

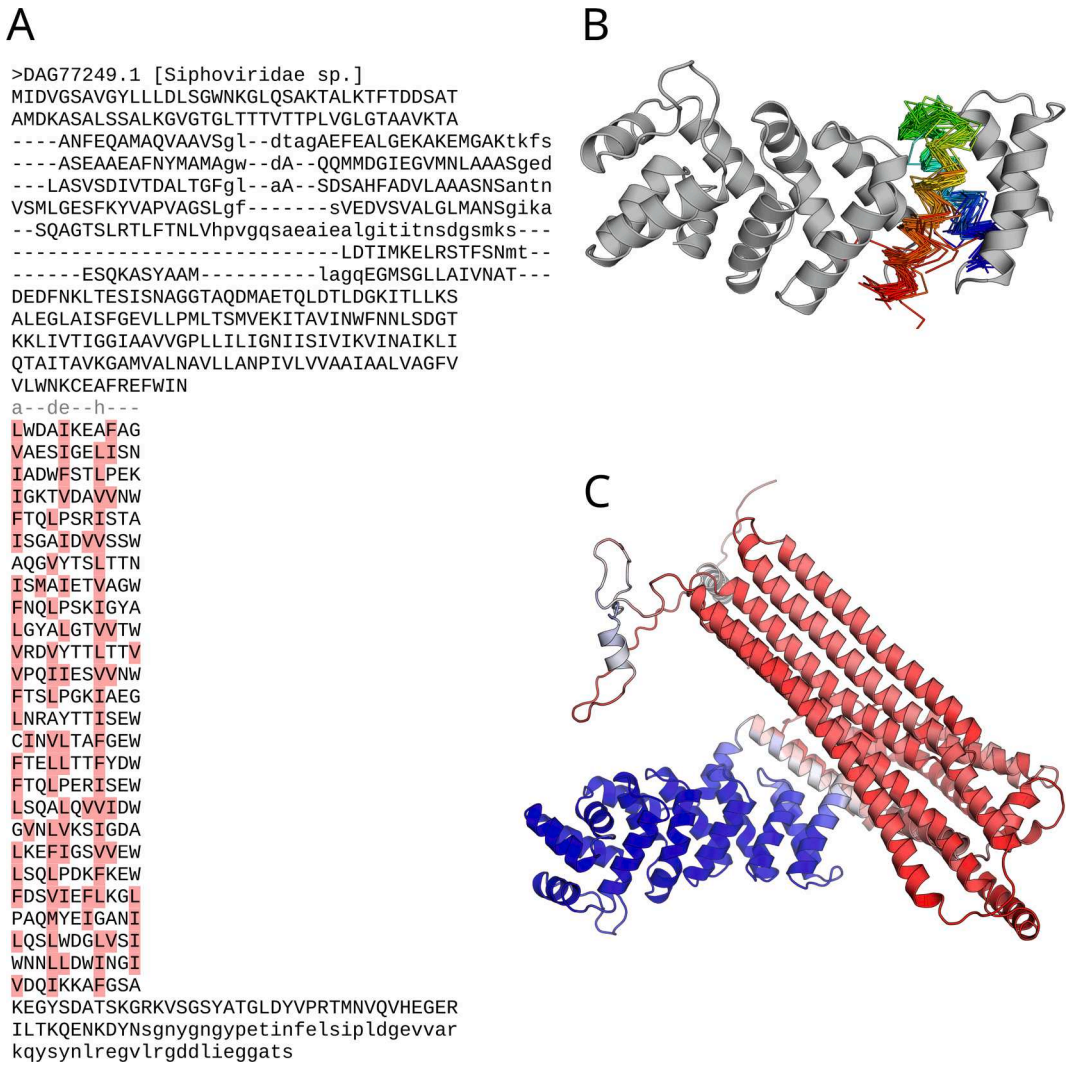


Fig 3. Graphical summary of the viral TMP proteins in our cluster map. (A) Sequence annotation of a representative of the largest subcluster of TMP proteins (DAG77249.1) with, N-terminally, an alignment of the helical hairpins that form the solenoid and, C-terminally, the hendecad stalk annotation. (B) Detail of the N-terminal alpha-helical solenoid domain; in gray, the AlphaFold model of DAG77249.1; in superimposed rainbow-colored ribbons, model hairpins of representatives from all 42 subclusters. (C) AlphaFold monomer prediction of DAG77249.1 (full sequence), colored by pLDDT, from lowest (red) to highest (blue).

similarity between the long helices in SPFH and TMP proteins, we think that TMPs are more likely to form a long tubular structure, akin to the phage phiX174 H protein (PDB: 4JPP), than to form the large basket structures of SPFH proteins (PDB: 7WI3, 4HL8). In an attempt to evaluate this possibility we tried modeling TMPs with AlphaFold, but they are too large to produce higher oligomeric models, and limiting the models to parts of the hendecad region only yielded inconclusive, low-confidence results.

3.3. MACH, a new family of membrane-bound coiled coils in alphaproteobacteria

We found the longest segments of hendecad repeats in a tight cluster of 142 proteins, primarily from Hyphomicrobiales (a suborder of alphaproteobacteria). It includes proteins from many genera relevant to human welfare, such as *Agrobacterium* (e.g. WP_234632661.1), *Mesorhizobium* (e.g. PZO80479.1), and *Bartonella* (e.g. WP_008039692.1). The proteins feature conserved domains at the N- and C-termini, connected by a long helical stalk which appears to be trimeric, based on sequence properties and structure prediction. This helical stalk is the reason for which many of these proteins are annotated as kinesins, apolipoproteins, or SMC-like proteins, even though these annotations are clearly incorrect. The N-terminal domain is anchored to the membrane by two transmembrane segments that form an antiparallel helical hairpin, projecting almost the entire bulk of the protein into the cytosol (Fig. 4). The N-terminal domain is linked to the coiled coil by a short, highly conserved, non-helical sequence, and the coiled coil is linked to the C-terminal domain by a very long, poorly conserved sequence predicted to be intrinsically unstructured. Whereas the N-terminal domain and the first part of the coiled coil could be modeled reliably as a homotrimer in AlphaFold, we only obtained inconclusive oligomeric models for the last part of the coiled coil and the C-terminal domain (Fig. 4C). None of our search programs found reliable matches to annotated domains or experimental structures for either the N- or the C-terminal domains. Based on the features presented here, we propose to call this family MACH, for Membrane-Anchored Coiled coils in Hyphomicrobiales.

3.4. Proteobacterial ZorA proteins

A conspicuous cluster in our map was formed by sequences annotated variously as hypothetical, ZorA, or MotA/TolQ/ExbB proton channel family proteins. They share a common domain architecture, consisting of three N-terminal transmembrane helices followed by a long

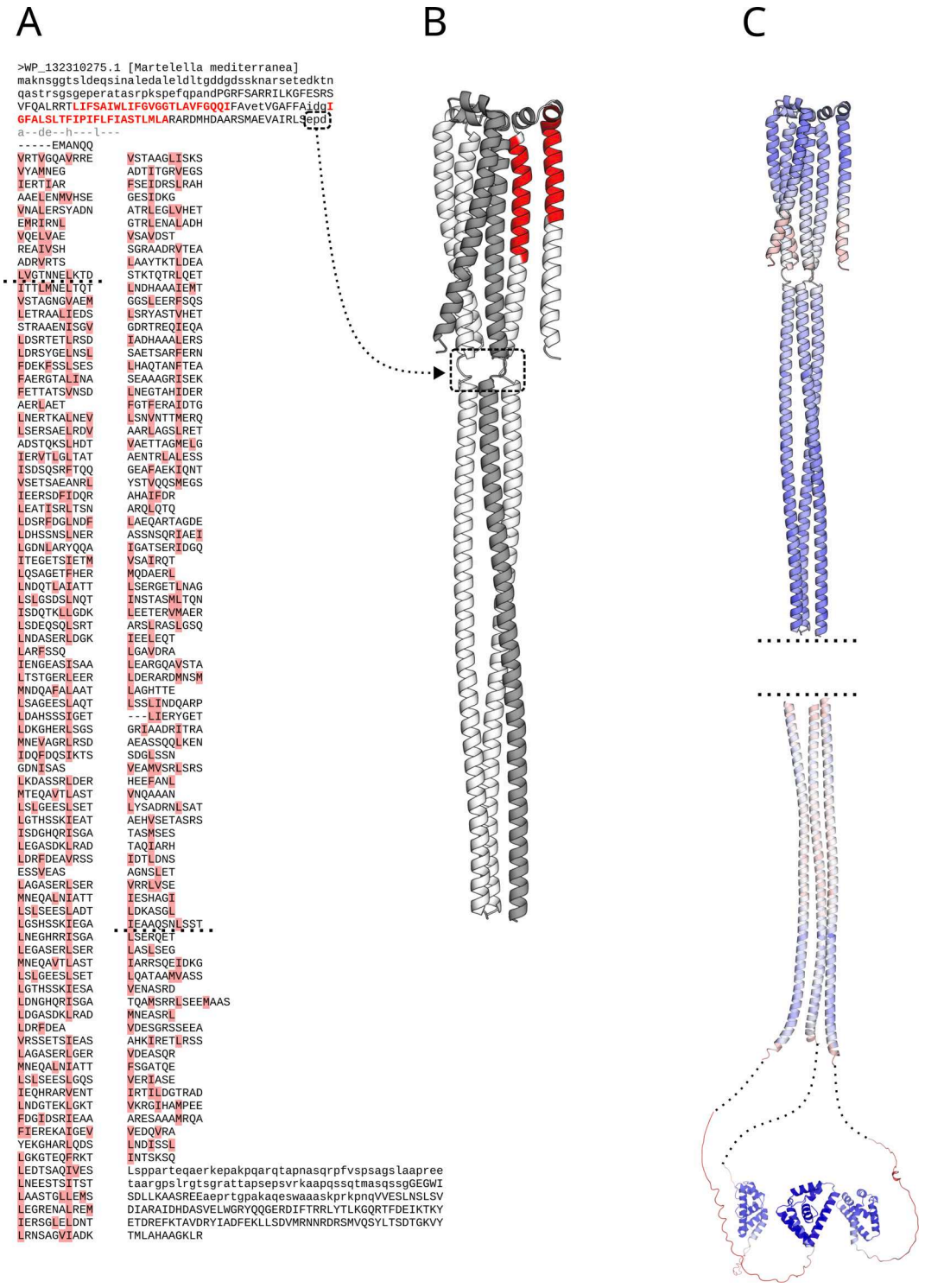


Fig 4. Graphical summary of the MACH proteins. (A) Sequence annotation of the member of the cluster with the longest uninterrupted hendecad stalk according to REPwin (WP_132310275.1), showing N-terminally the two predicted transmembrane helices and C-terminally the hendecad stalk annotation. (B) Trimeric AlphaFold model of the N-terminal part of the protein; the predicted transmembrane helices are coloured red in one subunit. (C) Trimeric AlphaFold models of the N- and C-terminal parts of WP_132310275.1, colored by pLDDT, from lowest (red) to highest (blue); the C-terminal loops have been truncated for clarity and replaced with dotted lines.

stalk, whose C-terminal half is composed of between 10 and 20 hendecad repeats (Fig. 5). As noted previously, the transmembrane part of the protein resembles the MotA/TolQ/ExbB family and most likely forms a proton channel (Doron et al., 2018). This assignment is supported by the presence of a protein homologous to MotB/TolR/ExbD in the same operon and by AlphaFold predictions, which consistently identify the pentamer as the preferred oligomeric state for these proteins (Fig S2), in accordance to the oligomeric state of MotA (PDB: 6YKM, 8GQY).

Although the stalk domains clearly have the features of coiled coils, assigning core residues by sequence analysis was difficult, due in part to the heterogeneous combinations of sequence periodicities in different

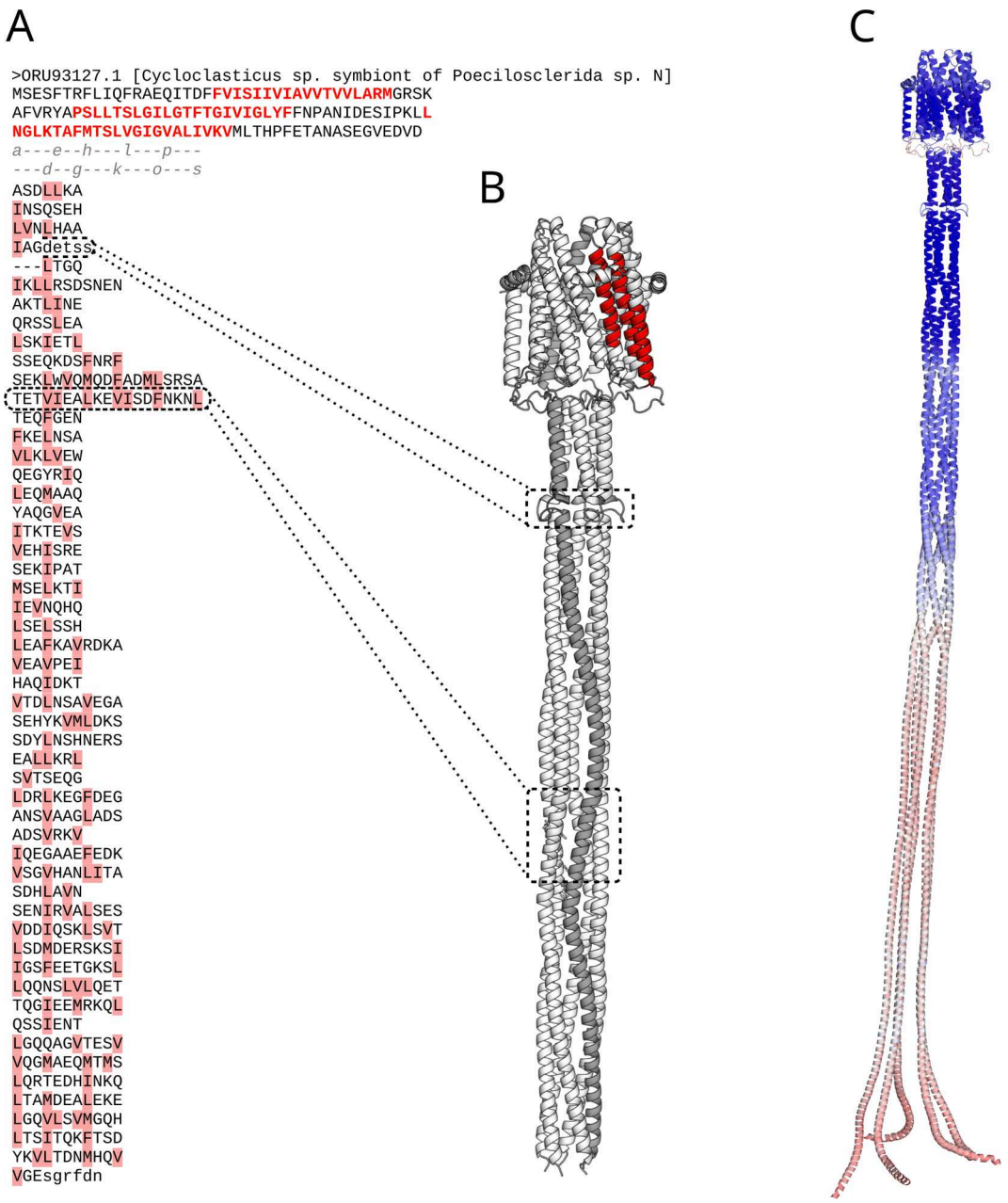


Fig 5. Graphical summary of the ZorA proteins in our cluster map. (A) Sequence annotation of a representative of the ZorA cluster, ORU93127.1, showing the three predicted transmembrane helices (red) and the periodicity of the hendecad stalk, with the two seams of interacting residues marked above the stalk sequence. (B) Pentameric AlphaFold model of the N-terminal part of the protein, the predicted transmembrane helices are coloured red in one subunit. The presence of two 19-residue repeats in a segment of the stalk that has primarily heptad periodicity leads to a strong local perturbation of the supercoil. (C) Pentameric AlphaFold model of ORU93127.1 assembled from two overlapping predictions and colored by pLDDT, from lowest (red) to highest (blue).

proteins of this family, and in part to the fact that, as pentamers, the helices would be bifaceted (Fig. 5A). We therefore used AlphaFold models to assist us in the annotation effort, as shown for example in Fig. 5. Although ZorA proteins were described as a component of the Zorya antiphage defense system (Doron et al., 2018), there are no hypotheses at present regarding the functional need for a long stalk in these proteins, as opposed to all other members of the MotA/TolQ/ExbB family.

3.5. Hendecad coiled coils with a biased residue distribution

We have already pointed out in the context of phage TMPs that some of the sequences in our cluster map were unusual for coiled coils, in that case because of an elevated incidence of glycine and proline. In this section, we will look at some other proteins from our cluster map, which deviate from the canonical coiled-coil residue distribution along at least a segment of their hendecads. For example, a group of long archaeal proteins is enriched in serine, most conspicuously in protein DRP01_01960 from an Archeoglobales archeon (GenBank RLI87417.1), in which a third of the residues over 20 hendecads are serine. These proteins reminded us strongly of coiled-coil segments in trimeric autotransporter adhesins of Burkholderia (e.g. Bcep18194_B0441, GenBank WP_011354054.1), in which two thirds of the residues are serine or threonine (Lupas et al., 2017), albeit in an 18-residue periodic background.

Among the proteins with unusual residue distribution we found two families most striking, because the unexpected residues were found in the core of the coiled coils.

The first family comprises the Scy and FilP cytoskeletal proteins, in which over half of core residues are alanine. We recently described these proteins as part of the DivIVA-like superfamily (Martinez-Goikoetxea and Lupas, 2023); however, in our cluster map they (Fig. 1B, DivIVA PFAM domain) do not form a tight cluster, as mentioned earlier. Both have been characterized as fibrous and feature long stalks formed primarily of hendecads, but they are obligate dimers, an oligomeric state that is generally disfavoured by hendecad repeats because of the constrained nature of knobs-to-knobs packing. The way in which Scy and FilP appear to circumvent this limitation is by greatly enriching for alanine residues in their core positions *a* and *h* (Fig. 6A), allowing for a continuous dimeric coiled coil without steric clashes. As with most of the long coiled coils we attempted to model with AlphaFold, these also mostly yielded inconclusive results (as judged by packing interactions and the continuity of the helices), but short segments occasionally yielded structures at reasonable levels of confidence (Fig. 6A). Although the role of Scy and FilP in the cell division of Gram-positive bacteria has been the object of considerable research interest, there is no explanation for why these proteins have evolved coiled coils with hendecad periodicity, as opposed to the comparatively much more common heptad periodicity in long cytosolic dimers (e.g. kinesins, myosins).

Another striking case of hendecad repeats with unexpected core residues is found in a cluster of fungal kinesins. These eukaryotic motor proteins typically assemble into parallel dimers, and feature an N-terminal ATPase motor domain followed by an extended coiled-coil stalk. The group of fungal kinesins in our cluster map featured hendecad repeats conspicuously enriched for histidine at position *h* (His@h). Histidine is generally disfavoured in core positions of coiled coils due to its polar nature, but there are examples of histidine zippers formed by antiparallel coiled-coil dimers with heptad periodicity (Alexander et al., 2021; Nostadt et al., 2020) (PDB 5LOS). In hendecad repeats however, the geometry of the side-chains precludes similar interactions. We therefore built AlphaFold models for the longest His@h stalk in our dataset (*Tilletia walkeri* KAE8209746.1). Models of the full stalk in different oligomeric states were consistently parallel structures, of which the dimer was favoured by pTM and pLDDT scores. As seen in Fig. 6B, symmetry-related histidines of the core assumed different side-chain rotamers, leading to a stacking of their rings at the interface. We looked for other instances of this motif with pattern searches over the nr database and found a broad range of proteins with this motif. Those for which an oligomeric state is known (such as the histidine kinase RLP27126.1, or the golgin XP_032827899.1) are parallel dimers, supporting the AlphaFold results. In many cases, the His@h hendecads were clearly the result of recent amplification within an existing coiled-coil stalk (Fig S3), indicating that His@h is actually favourable in parallel dimeric hendecad coiled coils. We also found open reading frames that consisted entirely of (nearly) identical His@h hendecads, one of which we will present in the next section.

3.6. New hendecads by genetic amplification

During the analysis of sequences in our cluster map, we noticed many instances of highly similar, adjacent hendecads, pointing to a recent amplification event within an existing coiled coil. For example, in Fig. 6B, the three hendecads at the center of the kinesin fragment shown have > 50% pairwise sequence identity; in a ninein-like protein of a damselfish species, two consecutive hendecads have been amplified into 15 mostly identical repeats (Fig S3); and in a protein family of Ascomycete fungi the stalk of a representative from *Ustilaginoidea virens* has undergone a recent hendecad amplification (Fig. 7A), in contrast to the

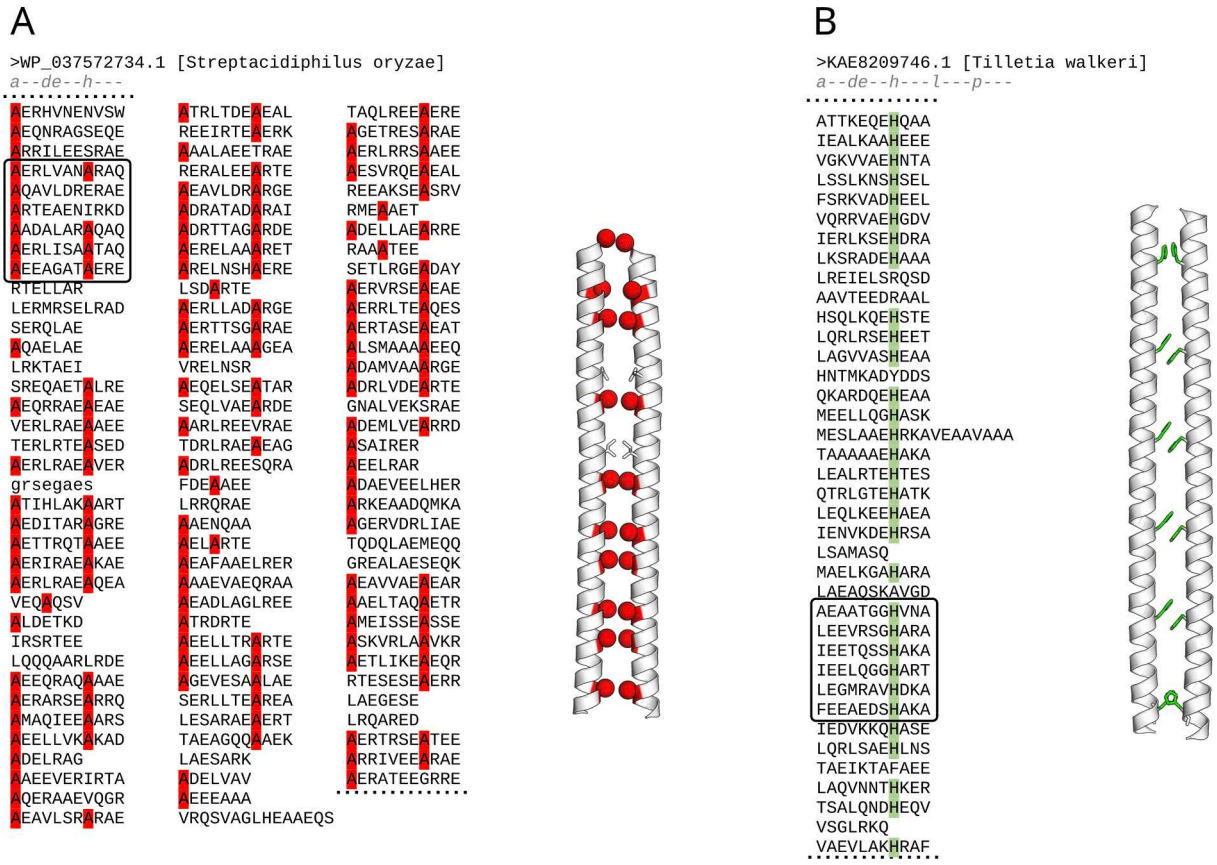
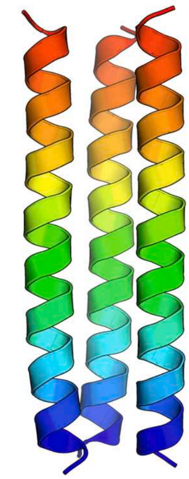


Fig 6. Two examples of hendecad coiled coils with unusual residues in their core positions. (A) hendecad stalk of Scy (Streptomyces CYtoskeletal protein), enriched in alanine at positions *a* and *h*, and a fragment of its AlphaFold dimeric model. (B) Hendecad stalk of a fungal kinesin, conspicuously enriched in histidine at position *h*, and a fragment of its AlphaFold dimeric model.

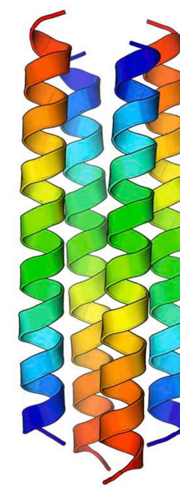
A

```
>GAO15575.1 [Ustilaginoidea virens]
ATGGCACCGACCAACGAAAAACCAGATTGCATC     MAPTNEKPDCI
ACTTTGGCAAATTATTTTCGAAAAGTAGGAGAC     TLANYFRKVGD
GAGGTTGAACTTTTCCAAAGTCTTCCTGCTCTG     EVELFQSLPAL
GATATTGGCGCAGCACTACTGGAACGCATGGAC     DIGAALLERMD
CGATTAATGCTCGAG                       RLMLE
ACGGCATCGTTCCGCCGCGAGGTTCAGTCAGAG     TASFRREVQSE
CTTACGTCGTTCCGCCGCGAGGTTCAGTCAGAG     LTSFRREVQSE
TTTATGTCGTTCCGCCGCGAGGTTCAGTCAGAG     FMSFRREVQSE
TTCACGTCGTTCCGCCGCGAGGTTCAGTCAGAG     FTSFRREVQSE
TCCACGTCGTTCCGCCAAGAATTTGACATAAAG     STSFRQEFDIK
CTCCGAGCAATGTAA                       LRAM
```



B

```
>EFO99190.1 [Caenorhabditis remanei]
ATGATACCCCATCGTCTCAAACTGAGGGAACTC     MIPHRLKLREL
CTGGAGCAACATTGTCTTAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTTAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTTAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTTAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
CTGGAGCAACATTGTCTCAAACTGAGGGAACTC     LEQHCLKLREL
AAAATTGGTTAA                          KIG
```



C

```
>OWF39119.1 [Mizuhopecten yessoensis]
                           ATGAAT              MN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT     VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT     VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT     VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT     VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT     VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT     VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT     VQKVLSKMSDN
GTACAAAAGGTCCTGTCCAAGATGTCTGACAAT     VQKVLSKMSDN
GTACAAAATGGTTCACAATGA                 VQNGSQ
```

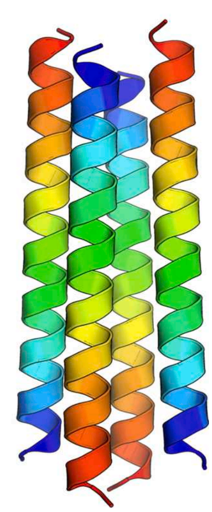


Fig 7. Examples of recently amplified sequences compatible with coiled-coil structure; DNA sequence, protein sequence, and AlphaFold model of the highlighted fragment are shown for (A) a fragment with non-identical DNA and protein repeats, (B) non-identical DNA but identical protein repeats, and (C) identical DNA and protein repeats.

heptad stalks of other representatives of this family, many of which have also undergone recent amplification, albeit with a different periodicity (Fig S4). These examples are in no way unusual and we could list many more, because, even in coiled coils whose repeats have undergone evolutionary differentiation, the strong protein sequence pattern is more likely to lead to local DNA sequence identity, which favours slipped strand mispairing during replication, and results in the amplification and deletion of repeats.

We also observed numerous examples of sequences composed entirely of identical or nearly identical 11-residue repeats, which in our cluster map behave like singletons, and thus tend to be found in the unconnected periphery (Fig. 2D). In contrast to the examples that we have provided for localized amplification within existing proteins, these globally repetitive sequences may either represent recently evolved proteins or ORFs identified by automated genome annotation, which are not currently expressed. Collectively, these sequences chart a pathway for the *de novo* emergence of new coiled coils. Whereas new, folded proteins from non-repetitive sequence are highly unlikely (Lupas et al., 2001), repetition has been identified as a dominant mechanism in the evolution of folded proteins (Eck and Dayhoff, 1966; McLachlan, 1987; Alva and Lupas, 2018). Thus, in essence, every repetitive DNA fragment may encode a new coiled coil if it (I) has a periodicity compatible with coiled-coil structure, (II) lacks stop codons in at least one frame, (III) has a repeating unit that encodes residues favourable to coiled-coil formation at the appropriate positions in that frame, and (IV) acquires the sequence motifs needed for transcription and translation, for example by fusion to an existing protein-coding gene.

We searched our dataset for highly repetitive sequences with clear coiled-coil forming potential, as judged by elevated scores in coiled-coil predictors and confident models in AlphaFold. Fig. 7 shows three examples spanning the range from highly repetitive, but non-identical nucleotide and protein sequence (Fig. 7A), over non-identical nucleotide and identical protein sequence (Fig. 7B), to identical nucleotide and protein sequence (Fig. 7C), in essence tracking key steps in the origin of a new hendecad coiled coil. The first example (Fig. 7A) is clearly a protein, conserved in ascomycetes, which we already pointed out in the first paragraph of this section. The second example (Fig. 7B) may be a protein, judged by the synonymous mutations in its nucleotide sequence, but it has no homologs in other species of its lineage, and would thus be of very recent evolutionary origin. We noted this sequence while searching for hendecads with His@h in our cluster map, but the AlphaFold model placed the histidines into position *d* of an antiparallel tetramer, such that they stack their rings at two interfaces, while the cysteines in position *e* are arrayed along the two other interfaces in a spacing compatible with disulfide bond formation (Fig S5). The confidence of this model (pLDDT ~ 70) is surprisingly high for a sequence without homologs. The third example (Fig. 7C), finally, is the least likely to have already become a protein and is probably the result of open reading frames in a repetitive genomic region of bivalve molluscs. By chance, the repeats have the potential to encode hendecads with a favourable distribution of hydrophobic and hydrophilic residues, as also reflected in the AlphaFold model, which returns an antiparallel tetramer with a confidence that is good for a sequence without homologs (pLDDT ~ 60).

4. Discussion

In this article, we describe the first large-scale survey of hendecad coiled coils in the proteome of life, motivated by the discrepancy between their widespread occurrence in putatively fibrous proteins and their low representation in known protein families. In the absence of prediction programs for hendecad coiled coils, we attempted to identify reliable instances by scanning for proteins with relevant sequence features, i.e. 11-residue periodicity and lack of β-strand prediction, clustering the resulting dataset by sequence similarity, and analyzing the clusters individually.

As our analyses progressed, we realized that the major clusters of long hendecad coiled coils showed a larger diversity in sequence and structure than heptad coiled coils in the same size range. Whereas the latter are predominantly parallel dimers and to a lesser extent trimers, our major groups of hendecad coiled coils went from dimers (Scy/FilP, His@h stalks) to trimers (MACH), pentamers (ZorA), and larger, as yet not fully defined, assemblies (viral TMPs). This range of oligomers is further extended by prominent hendecad coiled coils whose sequence periodicity is too low for REPwin: tetrabrachion (1FE6, tetramer), phage phiX174 pilot protein H (4JPP, decamer), HflKC (7WI3, dodecamer of heterodimers), and the vault protein (4HL8, 39-mer). Although heptad coiled coils can also form higher oligomers, they tend do so at short chain lengths, and the largest currently known oligomer is a barrel of 12 helices in antiparallel orientation (TolC, 1TQQ). The formation of coiled-coil oligomers relies on bifaceted helices (i.e., helices that can interact with other helices along two separate seams of residues) (Walshaw and Woolfson, 2001), and we note that in heptads the largest separation of the seams is 154°, whereas in hendecads it is 164°. If we call oligomers of between 6 and 15 helices *largermers*, following the nomenclature of Woolfson (Woolfson et al., 2012), then the stronger separation of seams possible in hendecads may hold the key to accessing the *largestmers* above 20 helices, such as in HflKC or the vault protein.

The structural diversity of hendecads is very likely the reason for their divergent sequence preferences, since the separation of the two seams in bifaceted helices has a major effect on residue preferences at each position of the repeat unit, and most hendecad helices are bifaceted. The only exception to this is in dimeric structures, where a single seam of residues mediates the interaction. We initially did not anticipate to encounter many examples for hendecad dimers, since the only case we knew of was that of the Scy stalk and we assumed it to be an oddity, but analysis of our cluster map showed the presence of many Scy paralogs, which all shared the preference of Scy for small residues in core positions (Martinez-Goikoetxea and Lupas, 2023). We assumed that the small core residues were essential for the formation of a dimeric structure, but to our surprise we encountered another sequence pattern that appears to lead to dimer formation. Instead of small residues in the core, this pattern showed a preference for the usual core residues of coiled coils (I, V, L, M) in positions *a* and *d*, and a strong preference for histidine in position *h* (His@h). We encountered this pattern in a diverse set of dimeric proteins, including kinesins, histidine kinases, golgins and nineins, often in recently amplified sequence repeats, suggesting that this is a favourable motif for dimer formation. AlphaFold models showed pairwise stacking of the histidine side-chains along the dimer interface, a conjecture supported by the observation that the residue most frequently substituting for histidine in this motif is tyrosine.

The many instances of recently amplified repeats in our dataset, of which His@h hendecads are but one example, led us to look in more detail at sequences with a high degree of internal sequence symmetry. What we observed is that, irrespective of periodicity, coiled coils undergo a constant process of amplification and contraction (e.g. Fig S3), and that this occasionally leads to a global change in periodicity (e.g. Fig S4). Furthermore we encountered many examples of sequences that consisted almost entirely of identical hendecads, made no connections in our cluster map, and had no recognizable homologs in BLAST searches. While some of these seemed likely to represent recently evolved hendecad coiled coils, others were probably just open reading frames in a repetitive part of the genome and their compatibility with coiled-coil structure was entirely fortuitous. We take these sequences as markers for a path that leads from non-coding DNA repeats to new structured proteins.

A question that was on our mind throughout this study was why nature evolved so many coiled coils with hendecad periodicity, given that the packing of their core residues is structurally more constrained than the continuous knobs-into-holes packing of heptad coiled coils. One possible reason could be that bifaceted helices in hendecad coiled coils can achieve greater separation between the two seams of interacting

residues than heptad coiled coils, giving access to higher oligomeric states (*largestmers*). This seems an attractive possibility, but we note that dimeric hendecads appear to have emerged in many protein families, for which this reason would not apply. For example, the comparatively recent evolutionary origin of hendecads in fungal kinesin stalks or the more ancient elaboration of a hendecad stalk in Scy and FilP proteins from an ancestor with heptad periodicity (Martinez-Goikoetxea and Lupas, 2023) suggest that there are other reasons for coiled coils to prefer hendecad periodicity, which we do not currently understand.

Our survey has shown us many coiled coils with hendecad periodicity, with which we might be able to train a predictor that can identify reliably hendecad coiled coils. We are aware that this goal faces a number of obstacles, which we have outlined in this paper. Among them are: (I) the irregular appearance of hendecads and heptads in complex stalks, (II) the presence of unexpected residues in the core of hendecad coiled coils, (III) the bifaceted nature of most hendecad helices, (IV) low availability of experimental hendecad structures needed for validation, and (V) the mostly poor performance of current structure predictors on hendecad sequences. Taken together these obstacles make us expect that a hendecad predictor will be a major challenge, but we have started work towards developing one.

CRediT authorship contribution statement

Mikel Martinez-Goikoetxea: Conceptualization, Investigation, Visualization. **Andrei N. Lupas:** Conceptualization, Investigation, Visualization, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have shared our data in a Mendeley repository (https://doi.org/10.17632/mbth74w7wy.1).

Acknowledgements

We thank Oliver Kohlbacher (Tuebingen University) and John Weir (Friedrich Miescher Laboratory, Tuebingen) for their advice in the early stages of this project. This work was supported by institutional funds of the Max Planck Society.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jsb.2023.108007.

References

Ackermann, H.W., 1998. Tailed bacteriophages: the order caudovirales. Adv. Virus. Res 51, 135–201. https://doi.org/10.1016/s0065-3527(08)60785-x.

Adlakha, J., Karamichali, I., Sangwallek, J., Deiss, S., Bär, K., Coles, M., Hartmann, M.D., Lupas, A.N., Hernandez Alvarez, B., 2019. Characterization of MCU-Binding Proteins MCUR1 and CCDC90B - Representatives of a Protein Family Conserved in Prokaryotes and Eukaryotic Organelles. Structure 27, 464–475.e6. https://doi.org/10.1016/j.str.2018.11.004.

Alexander, L.T., Lepore, R., Kryshtafovych, A., Adamopoulos, A., Alahuhta, M., Arvin, A. M., Bomble, Y.J., Böttcher, B., Breyton, C., Chiarini, V., Chinnam, N. babu, Chiu, W., Fidelis, K., Grinter, R., Gupta, G.D., Hartmann, M.D., Hayes, C.S., Heidebrecht, T., Ilari, A., Joachimiak, A., Kim, Y., Linares, R., Lovering, A.L., Lunin, V.V., Lupas, A.N., Makbul, C., Michalska, K., Moult, J., Mukherjee, P.K., Nutt, W. (Sam), Oliver, S.L., Perrakis, A., Stols, L., Tainer, J.A., Topf, M., Tsutakawa, S.E., Valdivia-Delgado, M., Schwede, T., 2021. Target highlights in CASP14: Analysis of models by structure providers. Proteins 89, 1647–1672. Doi:10.1002/prot.26247.

Alva, V., Lupas, A.N., 2018. From ancestral peptides to designed proteins. Curr. Opin. Struct. Biol. 48, 103–109. https://doi.org/10.1016/j.sbi.2017.11.006.

Boguski, M.S., Elshourbagy, N., Taylor, J.M., Gordon, J.I., 1985. Comparative analysis of repeated sequences in rat apolipoproteins A-I, A-IV, and E. Proc. Natl. Acad. Sci. USA 82, 992–996. https://doi.org/10.1073/pnas.82.4.992.

Bussell, R., Eliezer, D., 2003. A structural and functional role for 11-mer repeats in alpha-synuclein and other exchangeable lipid binding proteins. J. Mol. Biol 329, 763–778. https://doi.org/10.1016/s0022-2836(03)00520-5.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC. Bioinf. 10, 421. https://doi.org/10.1186/1471-2105-10-421.

Crick,, 1953a. The Packing of a-Helices: Simple Coiled-Coils. Acta. Cryst 6, 689–697. https://doi.org/10.1107/S0365110X53001964.

Crick,, 1953b. The Fourier transform of a coiled-coil. Acta. Cryst 6, 685–689. https://doi.org/10.1107/S0365110X53001952.

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., Sorek, R., 2018. Systematic discovery of antiphage defense systems in the microbial pangenome. Science 359, eaar4120. https://doi.org/10.1126/science.aar4120.

Dure, L., Crouch, M., Harada, J., Ho, T.-H.-D., Mundy, J., Quatrano, R., Thomas, T., Sung, Z.R., 1989. Common amino acid sequence domains among the LEA proteins of higher plants. Plant. Mol. Biol 12, 475–486. https://doi.org/10.1007/BF00036962.

Eck, R.V., Dayhoff, M.O., 1966. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science,. New. Series 152, 363–366.

Erdős, G., Dosztányi, Z., 2020. Analyzing Protein Disorder with IUPred2A. Curr. Protoc. Bioinformatics 70. https://doi.org/10.1002/cpbi.99.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., Hassabis, D., 2021. Protein complex prediction with AlphaFold-Multimer (preprint). Bioinformatics. https://doi.org/10.1101/2021.10.04.463034.

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic. Acids. Res. 39, W29–W37. https://doi.org/10.1093/nar/gkr367.

Frickey, T., Lupas, A., 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics 20, 3702–3704. https://doi.org/10.1093/bioinformatics/bth444.

Gruber, M., Lupas, A., 2003. Historical review: Another 50th anniversary – new periodicities in coiled coils. Trends. Biochem. Sci 28, 679–685. https://doi.org/10.1016/j.tibs.2003.10.008.

Gruber, M., Soding, J., Lupas, A.N., 2005. REPPER–repeats and their periodicities in fibrous proteins. Nucleic. Acids. Res. 33, W239–W243. https://doi.org/10.1093/nar/gki405.

Hicks, M.R., Holberton, D.V., Kowalczyk, C., Woolfson, D.N., 1997. Coiled-coil assembly by peptides with non-heptad sequence motifs. Fold. Des. 2, 149–158. https://doi.org/10.1016/S1359-0278(97)00021-7.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol 292, 195–202. https://doi.org/10.1006/jmbi.1999.3091.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol 305, 567–580. https://doi.org/10.1006/jmbi.2000.4315.

Leo, J.C., Lyskowski, A., Hattula, K., Hartmann, M.D., Schwarz, H., Butcher, S.J., Linke, D., Lupas, A.N., Goldman, A., 2011. The structure of E. coli IgG-binding protein D suggests a general model for bending and binding in trimeric autotransporter adhesins. Structure 19, 1021–1030. https://doi.org/10.1016/j.str.2011.03.021.

Ludwiczak, J., Winski, A., Szczepaniak, K., Alva, V., Dunin-Horkawicz, S., 2019. DeepCoil—a fast and accurate prediction of coiled-coil domains in protein sequences. Bioinformatics 35, 2790–2795. https://doi.org/10.1093/bioinformatics/bty1062.

Lupas, A.N., Bassler, J., 2017. Coiled Coils – A Model System for the 21st Century. Trends. Biochem. Sci 42, 130–140. https://doi.org/10.1016/j.tibs.2016.10.007.

Lupas, A.N., Bassler, J., Dunin-Horkawicz, S., 2017. The Structure and Topology of α-Helical Coiled Coils. Fibrous. Prot:. Struct. Mech. 82, 95–129. https://doi.org/10.1007/978-3-319-49674-0_4.

Lupas, A., Van Dyke, M., Stock, J., 1991. Predicting Coiled Coils from Protein Sequences. Science 252, 1162–1164. https://doi.org/10.1126/science.252.5009.1162.

Lupas, A., Müller, S., Goldie, K., Engel, A.M., Engel, A., Baumeister, W., 1995. Model structure of the Omp alpha rod, a parallel four-stranded coiled coil from the hyperthermophilic eubacterium Thermotoga maritima. J. Mol. Biol 248, 180–189. https://doi.org/10.1006/jmbi.1995.0210.

Lupas, A.N., Ponting, C.P., Russell, R.B., 2001. On the Evolution of Protein Folds: Are Similar Motifs in Different Protein Folds the Result of Convergence, Insertion, or Relics of an Ancient Peptide World? J. Struct. Biol. 134, 191–203. https://doi.org/10.1006/jsbi.2001.4393.

Mahony, J., Alqarni, M., Stockdale, S., Spinelli, S., Feyereisen, M., Cambillau, C., van Sinderen, D., 2016. Functional and structural dissection of the tape measure protein

of lactococcal phage TP901-1. Sci. Rep 6, 36667. https://doi.org/10.1038/srep36667.
Martinez-Goikoetxea, M., Lupas, A.N., 2023. A conserved motif suggests a common origin for a group of proteins involved in the cell division of Gram-positive bacteria. PLoS. One 18, e0273136.
McLachlan, A.D., 1987. Gene duplication and the origin of repetitive protein structures. Cold. Spring. Harb. Symp. Quant. Biol 52, 411–420. https://doi.org/10.1101/sqb.1987.052.01.048.
Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., Steinegger, M., 2022. ColabFold: making protein folding accessible to all. Nat. Methods 19, 679–682. https://doi.org/10.1038/s41592-022-01488-1.
Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., Punta, M., 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic. Acids. Res. 41, e121–e. https://doi.org/10.1093/nar/gkt263.
Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., Bateman, A., 2021. Pfam: The protein families database in 2021. Nucleic. Acids. Res. 49, D412–D419. https://doi.org/10.1093/nar/gkaa913.
Nostadt, R., Hilbert, M., Nizam, S., Rovenich, H., Wawra, S., Martin, J., Küpper, H., Mijovilovich, A., Ursinus, A., Langen, G., Hartmann, M.D., Lupas, A.N., Zuccaro, A., 2020. A secreted fungal histidine- and alanine-rich protein regulates metal ion homeostasis and oxidative stress. New. Phytol. 227, 1174–1188. https://doi.org/10.1111/nph.16606.
Peters, J., Baumeister, W., Lupas, A., 1996. Hyperthermostable surface layer protein tetrabrachion from the archaebacterium Staphylothermus marinus: evidence for the presence of a right-handed coiled coil derived from the primary structure. J. Mol. Biol 257, 1031–1041. https://doi.org/10.1006/jmbi.1996.0221.
Rost, B., 1999. Twilight zone of protein sequence alignments. Protein. Eng. Des. Sel. 12, 85–94. https://doi.org/10.1093/protein/12.2.85.
Sievers, F., Higgins, D.G., 2018. Clustal Omega for making accurate alignments of many protein sequences: Clustal Omega for Many Protein Sequences. Protein. Sci. 27, 135–145. https://doi.org/10.1002/pro.3290.
Stetefeld, J., Jenny, M., Schulthess, T., Landwehr, R., Engel, J., Kammerer, R.A., 2000. Crystal structure of a naturally occurring parallel right-handed coiled coil tetramer. Nat. Struct. Biol 7, 772–776. https://doi.org/10.1038/79006.
Sun, L., Young, L.N., Zhang, X., Boudko, S.P., Fokine, A., Zbornik, E., Roznowski, A.P., Molineux, I.J., Rossmann, M.G., Fane, B.A., 2014. Icosahedral bacteriophage ΦX174 forms a tail for DNA transport during infection. Nature 505, 432–435. https://doi.org/10.1038/nature12816.
Walshaw, J., Woolfson, D.N., 2001. Open-and-shut cases in coiled-coil assembly: α-sheets and α-cylinders. Protein. Sci 10, 668–673.
Woolfson, D.N., Bartlett, G.J., Bruning, M., Thomson, A.R., 2012. New currency for old rope: from coiled-coil assemblies to α-helical barrels. Curr. Opin. Struct. Biol. 22, 432–441. https://doi.org/10.1016/j.sbi.2012.03.002.
Zaccai, N.R., Chi, B., Thomson, A.R., Boyle, A.L., Bartlett, G.J., Bruning, M., Linden, N., Sessions, R.B., Booth, P.J., Brady, R.L., Woolfson, D.N., 2011. A de novo peptide hexamer with a mutable channel. Nat. Chem. Biol 7, 935–941. https://doi.org/10.1038/nchembio.692.
Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., Alva, V., 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J. Mol. Biol. 430, 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007.

Supplementary Information

New protein families with hendecad coiled coils in the proteome of life

Mikel Martinez-Goikoetxea, Andrei N. Lupas*

Department of Protein Evolution, Max Planck Institute for Biology, 72076 Tübingen, Germany

* To whom correspondence should be addressed:

Andrei N. Lupas

Tel.: +49 7071 601 341

Fax: +49 7071 601 349

E-mail: andrei.lupas@tuebingen.mpg.de

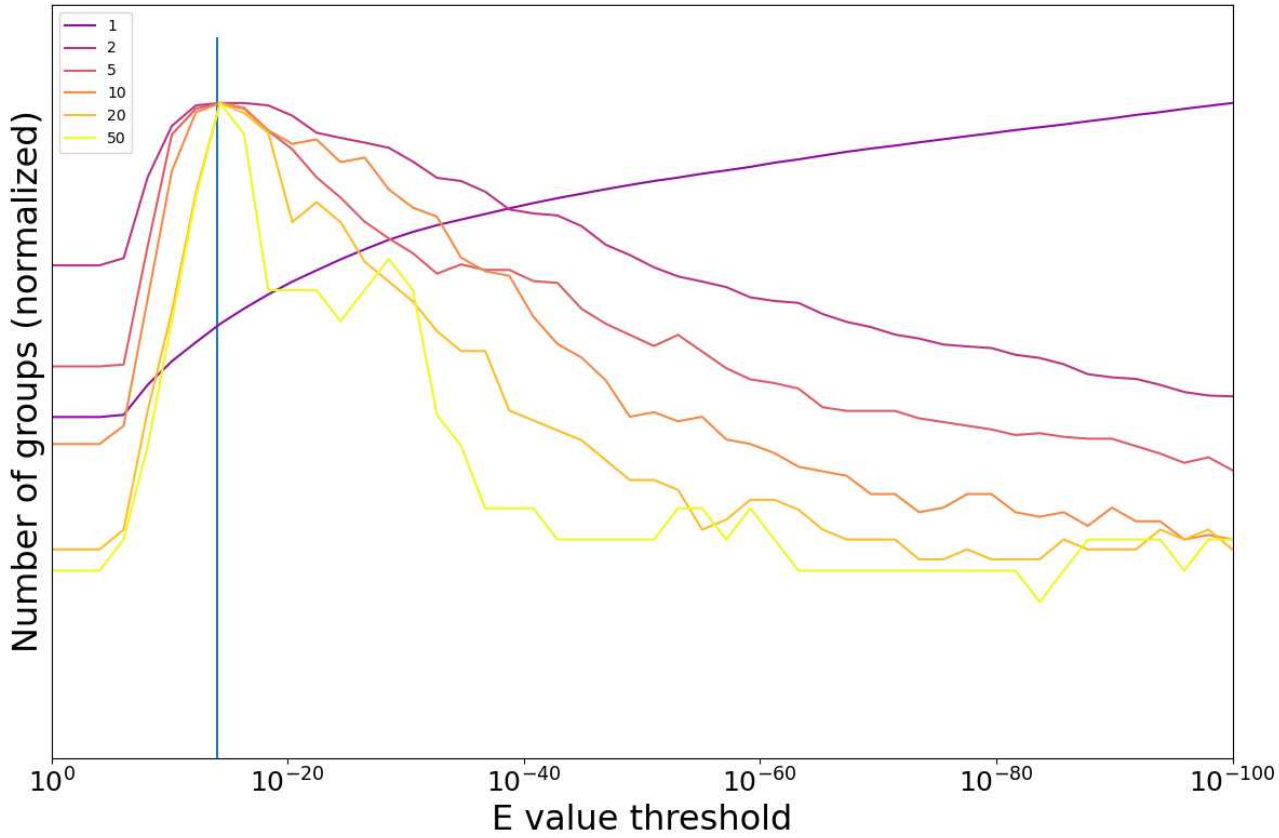


Fig S1. A plot of the number of unconnected clusters in our dataset as a function of the p-value threshold. As the threshold is made more stringent, more connections are removed and new clusters are created. The colored lines reflect different minimum cluster sizes. Thus, if a minimum of 1 is used (implying that singletons are smallest possible cluster), the number of clusters increases continuously. With a minimum cluster size >1, small clusters will not be considered and thus their number will start decreasing after a point. For any minimum cluster size above 1, the threshold that consistently yields the maximum number of separate clusters is 1E-14 (blue vertical line).

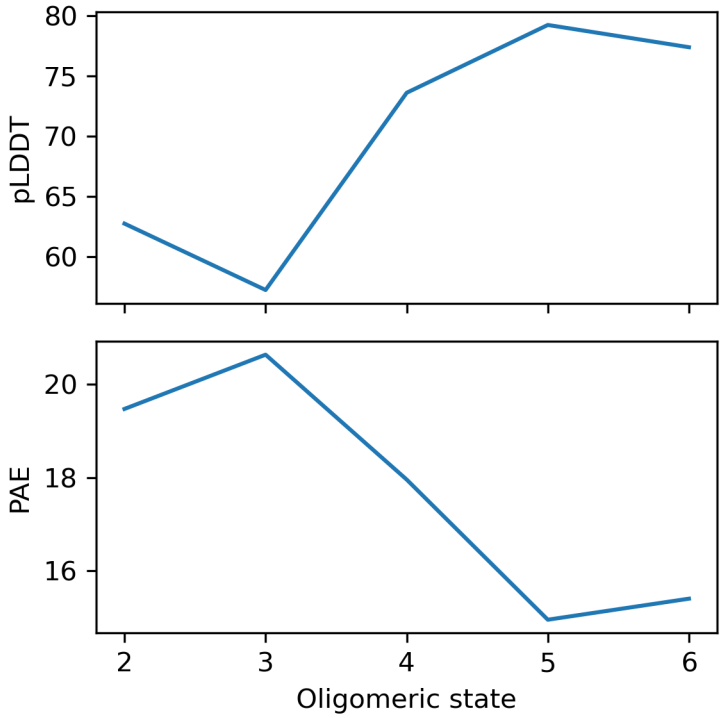


Fig S2. A plot of the AlphaFold model scores for a ZorA protein. The pLDDT (higher is better) and PAE (lower is better) plots strongly suggest that the preferred oligomeric state is the pentamer.

```
>XP_051815946.1 ninein-like protein
[Acanthochromis polyacanthus]
a--d---h---l---
...
SSEQQRRMQWE
RAELEQNFARE
IGNLVQRLSAE
KEQLEAELKLK
MDQELLLVRDESQQQ
LAQMKLQHTES
QHRLLHQLHRQ
RGFWEQQ
VIQSEEQHQEK
VRHLEEQHQEE
VQHLEKQHQEE
VQHLEEQHQEE
VQHLEEQHQEE
VQHLEEQHQEK
VQHLEKQHQEE
VQHLEKQHQEE
VQHLEKQHQEE
VQHLEEQHQEE
VQHLEEQHQEE
VQHLEEQHQEE
VQHLEKQHQEK
VQHLEKQHQEE
VQHLEEQRSRA
EEAGRSE
LEACWGR
SSELEAR
LEEVCTQ
LQESIGF
LESHEVL
NRRLAAE
KSSVEER
...
```

```
>XP_054874357.1 ninein-like protein
[Amphiprion ocellaris]
a--d---h---l---
...
SSEQQRRMQRE
RAELEQNFARE
IGNLVQKLSAE
KEQLEAELKLK
MDQELLLVRDESQQQ
LAQMKLHHAEL
QRRLLHQLHRQ
RGFWENR
VVQSEEQHQEE
----------
----------
----------
----------
----------
----------
----------
----------
----------
----------
----------
----------
----------
VRRLEEQLSRP
EEAGRSE
LEACRGR
SSELEAR
LEEACTQ
-------
LQSHEVL
NRHLAAA
KSSVEEQ
   ...
```

Fig S3. Coiled-coil segments from two ninein-like proteins with histidine in *h*, the first from the damselfish *Acanthochromis polyacanthus*, the second from the clownfish *Amphiprion ocellaris*. Consensus residues in the amplified region of the damselfish protein are highlighted in yellow, residues of the clownfish sequence differing from the damselfish one are highlighted in grey.

```
>XP_042998175.1 uncharacterized protein
UV8b_04743 [Ustilaginoidea virens]
...
KPDCITLANYFRKVGDEVELFQSLPALDIGAALLER
MDRLMLE
-------
-------
-------
-------
TASFRREVQSE
LTSFRREVQSE
FMSFRREVQSE
FTSFRREVQSE
STSFRQEFDIK
-------
LRAMNKNISSRLVNQWALSPEVSLSPMYNVSTGDEI
ANCPKT
```

```
>KAG8412782.1 hypothetical protein
J3458_013219 [Metarhizium acridum]
...
QPDFGRLASNFKDIGEQVQRFENIPAFEGGTQLMRK
MDRIFEA
VTEMRHE
INGFRTE
MTDMRRE
MTDMRRE
LDGFGLN
MTDMRRE
MTDMRRE
MTDMRRE
MTDMRRE
LDGLGLK
MMITDKNFQARMANSIVVSGEMTLSPLYNVTTGEEL
SHCPET
...
```

Fig S4. Independently amplified coiled-coil segments from two representatives of a protein family in ascomycete fungi. The stalk of the protein from *Ustilaginoidea virens* was amplified with a hendecad periodicity, the one from *Metarhizium acridum* with a heptad periodicity. Consensus residues in the two amplified regions are highlighted in yellow. Residues of the *Metarhizium* protein differing from the *Ustilaginoidea* protein are highlighted in grey.

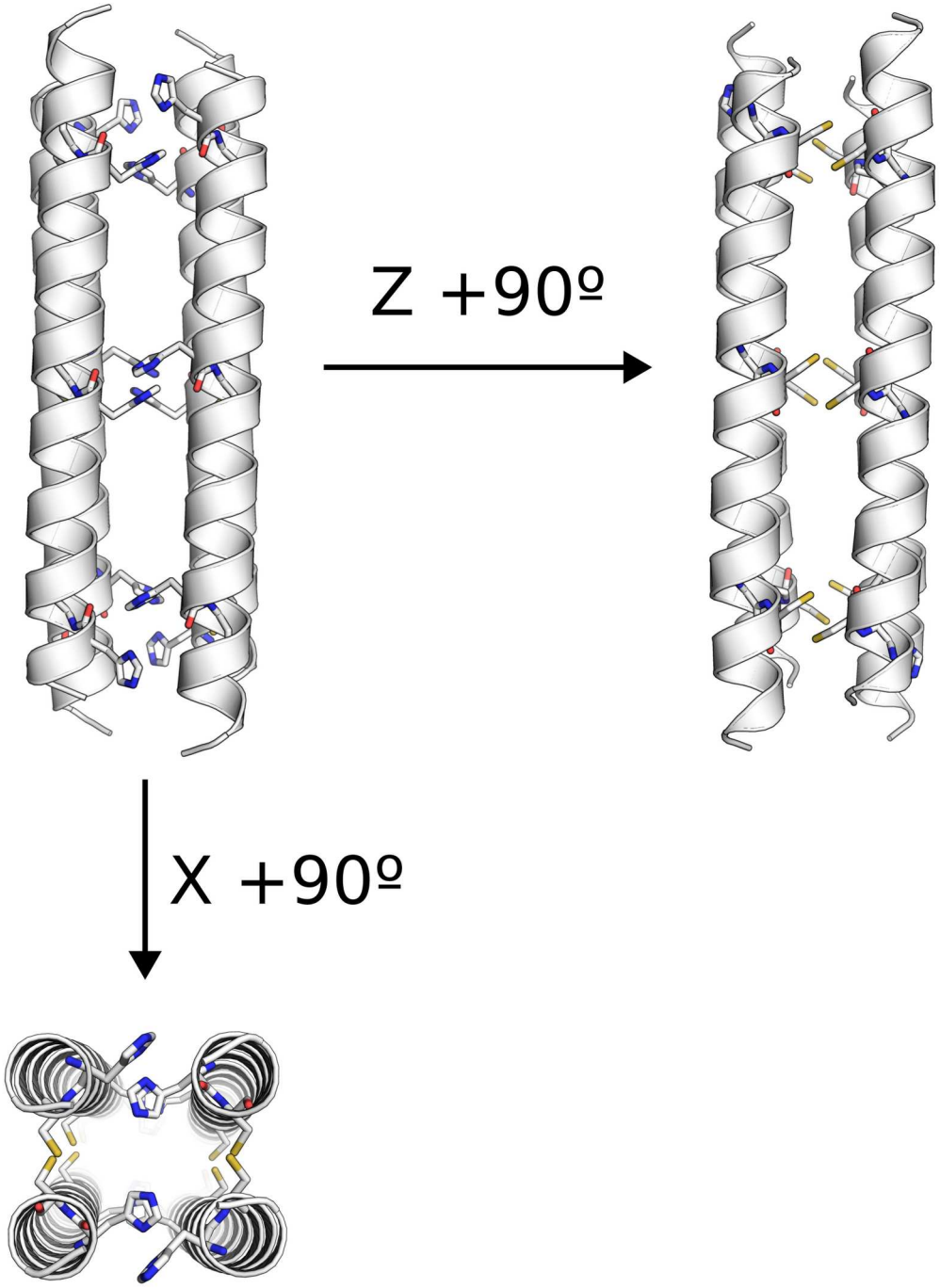


Fig S5. Detail of the AlphaFold model of a recently amplified hendecad fragment (EFO99190.1 of *Caenorhabditis remanei*, Fig 7B). As opposed to the expected His@h arrangement, AlphaFold predicted the histidines to be in position *d*.

PLOS ONE

RESEARCH ARTICLE

A conserved motif suggests a common origin for a group of proteins involved in the cell division of Gram-positive bacteria

Mikel Martinez-Goikoetxea, Andrei N. Lupas*

Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany

* andrei.lupas@tuebingen.mpg.de

Citation: Martinez-Goikoetxea M, Lupas AN (2023) A conserved motif suggests a common origin for a group of proteins involved in the cell division of Gram-positive bacteria. PLoS ONE 18(1): e0273136. https://doi.org/10.1371/journal.pone.0273136

Editor: Matteo De March, University of Nova Gorica, SLOVENIA

Received: August 2, 2022

Accepted: December 29, 2022

Published: January 20, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0273136

Data Availability Statement: All relevant data are within the paper, its Supporting information file,

Abstract

DivIVA, GpsB, FilP, and Scy are all involved in bacterial cell division. They have been reported to interact with each other, and although they have been the subject of considerable research interest, not much is known about the molecular basis for their biological activity. Although they show great variability in taxonomic occurrence, phenotypic profile, and molecular properties, we find that they nevertheless share a conserved N-terminal sequence motif, which points to a common evolutionary origin. The motif always occurs N-terminally to a coiled-coil helix that mediates dimerization. We define the motif and coiled coil jointly as a new domain, which we name DivIVA-like. In a large-scale survey of this domain in the protein sequence database, we identify a new family of proteins potentially involved in cell division, whose members, unlike all other DivIVA-like proteins, have between 2 and 8 copies of the domain in tandem. AlphaFold models indicate that the domains in these proteins assemble within a single chain, therefore not mediating dimerization.

Introduction

DivIVA proteins localize to negatively-curved membranes, such as the cell poles and the division septum, to which they recruit various proteins, depending on the organism and the physiological conditions [1]. Homologs of this protein are found predominantly in Gram-positive bacteria, but have also been reported in other major clades of bacteria. GpsB, a paralog of DivIVA in Firmicutes, is thought to complement the role of DivIVA in coordinating peptidoglycan synthesis at sites of cell division and elongation [2, 3]. In contrast to DivIVA and GpsB, FilP is considered to be primarily a cytoskeletal protein, but has also been implicated in cell division [4]. Thus, in Streptomyces species, FilP forms a concentration gradient that increases towards the cell poles, and is thought to contribute to the polar localization of DivIVA [5]. Scy (Streptomyces CYtoskeletal protein) is a paralog of FilP [6], found in Streptomyces species only. It is the largest protein in this set at about 4 times the size of FilP and has been reported to interact with both FilP and DivIVA [7]. All these proteins form dimeric coiled coils, differing in the length and periodicity of their coiled-coil segments.

and a Mendeley repository available at https://data.mendeley.com/datasets/bn627zbymx.

Funding: The authors were supported by institutional funds of the Max Planck Society. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Coiled coils are a widespread supersecondary structure motif that consists of two or more α-helices that wind around a central axis and interlock their side-chains systematically along the core of the structure [8]. Their packing geometry, widely considered the hallmark of coiled coils, is known as knobs-into-holes and is achieved by packing a given core residue (knob) into a cavity formed by 4 residues on a symmetry-related α-helix (hole) [9]. Because of their different solvent accesibility, core residues tend to be hydrophobic, while all other residues of the coiled coil are typically polar. In nature, coiled coils are primarily built by sequence repeats of seven residues (heptads), featuring hydrophobic residues at the first and fourth positions. Using the general nomenclature of coiled coils, which denotes the seven positions of the repeat as *a*-*g*, the hydrophobic residues tend to be found in positions *a* and *d*. While heptads are the only repeat fully compatible with knobs-into-holes packing, hence, possibly, their dominance in nature, other sequence repeats may form coiled-coil structures. The most frequent of these are eleven-residue repeats (hendecads, labeled *a*-*k*, with hydrophobics predominantly in *a*, *d*, and *h*) and fifteen-residue repeats (pentadecads, labeled *a*-*o*, with hydrophobics predominantly in *a*, *d*, *h* and *l*). In all these proteins, the degree to which individual helices wind around the central axis of the bundle (supercoiling) is given by the difference between the sequence periodicity of hydrophobic residues and the structural periodicity of an undistorted α-helix. Thus, for example, the sequence periodicity of hydrophobic positions in heptads is 7/2 = 3.5, while the structural periodicity of an undistorted α-helix is 3.63 residues/turn. Therefore, the α-helix being a right-handed spiral, the difference of -0.13 specifies a left-handed supercoil. Correspondingly, the equivalent difference for hendecads is 0.03, specifying a very minor degree of supercoiling, and the difference for pentadecads is 0.12, specifying a right-handed supercoil of the same magnitude as the left-handed supercoil in heptads. Besides changing the degree of supercoiling, departures from heptad periodicity also affect the packing geometry. Particularly in hendecads, the different structural periodicity leads to one core residue per repeat pointing directly towards the central axis of the bundle, a packing mode referred to as knobs-to-knobs. The reduced distance between the side-chains results in a steric constraint, and, to avoid clashes, the distance must be increased by means of assembling into higher oligomeric states or with the core position featuring a small side-chain, particularly Alanine. Correspondingly, FilP and Scy, which are dimers but contain extended stretches of hendecad coiled coils, show a high proportion of Alanine residues in their hydrophobic cores [6].

Whereas heptad repeats have been studied extensively and can be predicted effectively from protein sequences, hendecads have hitherto remained largely unexplored. In order to produce a database of reliable hendecad sequences, we therefore searched for this periodicity by tandem repeat detection over the non-redundant protein sequence database (NR). In the process, we also encountered Scy and FilP, which, to our surprise, bore a remarkable similarity to DivIVA and GpsB in their N-terminal part. Here, we define the common motif in these four protein families and describe a fifth family that uniquely bears between 2 and 8 repetitions of this DivIVA-like motif. We propose that this family is yet another component of the Gram-positive cell-division machinery.

Materials and methods

1 Discovery of the similarities between DivIVA, GpsB, FilP and Scy

As part of the sequence analysis of proteins detected in our survey for hendecads, we also analyzed representatives of Scy and FilP (WP_066885126.1 and WP_172158822.1) with HMMER (version 3.3, http://hmmer.org) [10] against the PFAM domain database (release 34.0, https://pfam.xfam.org) [11], and obtained intriguing matches between their N-terminal regions and the corresponding region in the DivIVA profile. Further HMM-HMM (Hidden Markov

Model) searches with HHpred against the PDB (Protein Data Bank) database in the MPI Bioinformatics Toolkit (https://toolkit.tuebingen.mpg.de/tools/hhpred, PDB_mmCIF70_14_Apr database using default parameters) [12] gave strong matches to DivIVA (PDB identifier 2WUJ), GpsB (4UG3), and Wag31 (6LFA). All matches were anchored by a set of highly conserved residues N-terminal to a dimeric heptad coiled coil.

2 Finding the conserved motif in DivIVA-like proteins

We collected homologs of DivIVA, GpsB, FilP and Scy using BLAST at NCBI (https://blast.ncbi.nlm.nih.gov) against the NR database (version May 2022) at the default E-value cutoff (0.05). In order to obtain a representative coverage of the protein families, we chose as starting sequences DivIVA from Bacillus (WP_163131312.1) and Streptomyces (WP_190537660.1), GpsB from Bacillus (NP_390100.1), and FilP and Scy from Streptomyces (GAU65504.1, MYV96918.1). We obtained 21171 sequences, which we then filtered with MMSeqs2 (https://toolkit.tuebingen.mpg.de/tools/mmseqs2) [13] to a maximum pairwise sequence identity of 80% and 80% minimum coverage, which yielded a total of 1816 sequences. In order to find conserved motifs in this dataset, we used MEME (version 5.4.1, https://meme-suite.org/meme/tools/meme) [14], whose highest scoring motif coincided with the conserved residues already identified in the HHpred searches. In order to ascertain the presence and conservation of this motif, we computed multiple sequence alignments (MSAs) for every representative sequence and their BLAST matches with ClustalOmega (version 1.2.4, http://www.clustal.org/omega) [15], and constructed logos using WebLogo (version 3.7.11, http://weblogo.threeplusone.com) [16]. Consensus sequences from the BLAST searches are shown in Fig 1C, and the logo produced by MEME for the highest scoring motif is shown in Fig 1D.

3 Construction of a cluster map of sequences with a DivIVA-like motif

In order to broadly collect sequences that feature this motif, we used the pattern GY[DN]xx[QE]V[DN] for searches of the NR30 database (NR database filtered to a maximum pairwise identity of 30%). In particular, we omitted the arginine in the MEME motif because we had noticed that some clades systematically contained a different residue at that position (E in actinobacterial DivIVA), and wanted to see whether other clades showed further diversity (indeed, the new family of DivIVA-like proteins we describe here shows a wide range of residues at this position, including a substantial proportion of Glycine residues). The pattern search yielded 5552 sequences. The relationships between these sequences were explored by clustering them according to their pairwise BLAST P-values in CLANS (https://toolkit.tuebingen.mpg.de/tools/clans) [17]. Clustering was done in default settings (attract = 10, repulse = 5, exponents = 1), and the map was imaged at the P-values given in the figure legends. In order to define clusters in the map, we constructed an undirected graph representation of the BLAST P-values at a threshold of 1E-15 and then used the Girvan-Newman algorithm as implemented in the NetworkX Python package (https://networkx.org) to automatically detect densely connected groups of sequences.

4 Detection and analysis of protein families containing DivIVA-like domains

In order to identify which of the sequences satisfying the pattern were part of protein families containing a DivIVA-like domain, we extended the pattern to include hydrophobic residues of the coiled-coil segment (GY[DN]xx[QE]V[DN]xx[ILV]xx[ILV]), and selected clusters in which at least half of the sequences had a match.

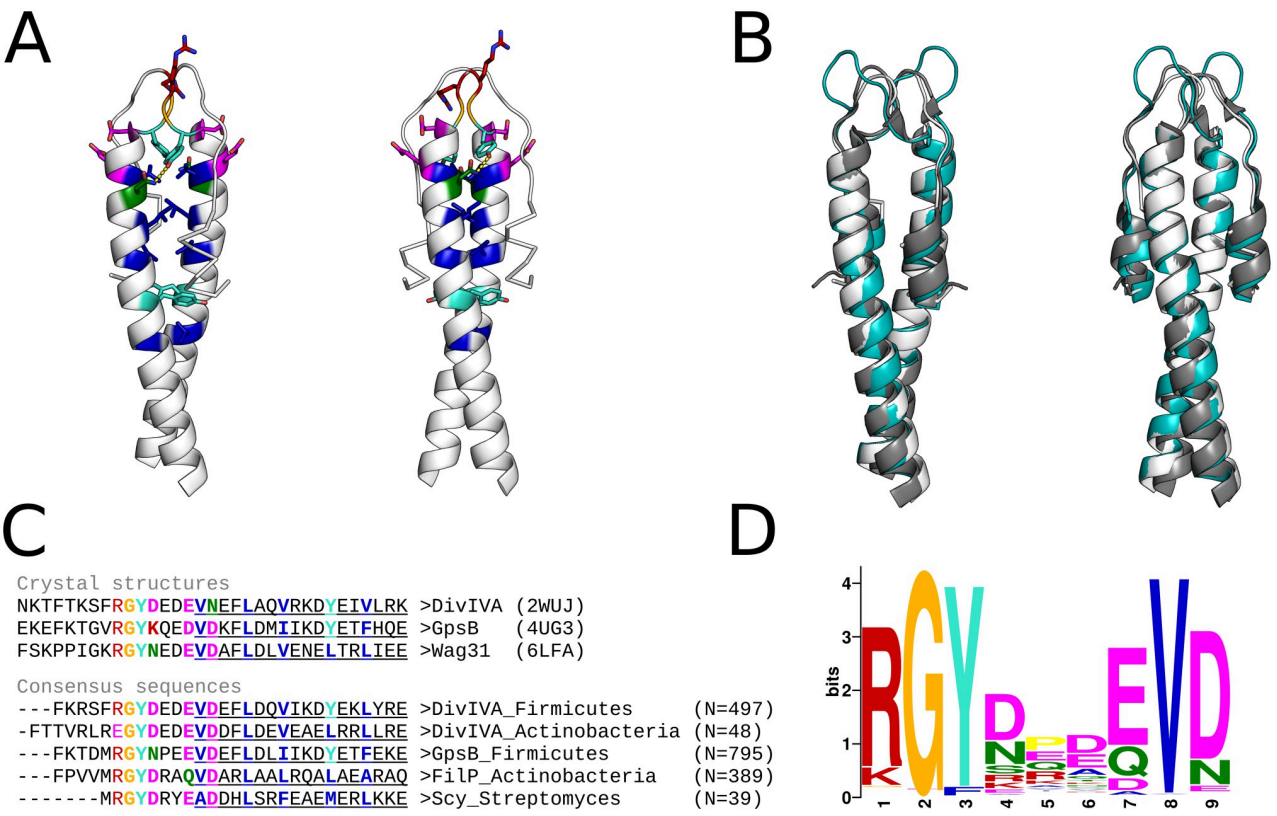


Fig 1. Summary of the DivIVA-like domain. (A) Cartoon renders of the DivIVA N-terminal domain (2WUJ); color-coded to match panel C, are the residues of the conserved motif as well as the core residues of the two following heptad coiled coils. (B) Superimposition of DivIVA (2WUJ, white), GpsB (4UG3, gray) and Wag31 (6LFA, teal); RMSD of the superimposition to 2WUJ is 1.0 and 2.2 Angstrom respectively. (C) Alignment of the sequences for the structures shown in panel B to the consensus sequences (N = number of sequences) of representatives for the major groups of DivIVA homologs (see Methods); the conserved motif and the core residues of the two following heptad coiled coils are highlighted in colors, and the Quick2D α-helical prediction is shown as underlined characters. (D) Logo representation of the top scoring motif found by MEME in the set of 1816 DivIVA homologs obtained as described in the Methods.

https://doi.org/10.1371/journal.pone.0273136.g001

We analyzed these clusters for their taxonomic spectrum and sequence features, including interactive coiled-coil annotation with PCOILS and REPPER (https://toolkit.tuebingen.mpg.de/tools/repper) [18], disorder prediction with IUPRED2 [19], secondary structure prediction with Quick2D (https://toolkit.tuebingen.mpg.de/tools/quick2d) [12], and genomic context with GCsnap [20].

5 Structural predictions

We computed structural predictions for the PolyDIV proteins with AlphaFold (version 2.1.1) [21, 22]. First, we predicted monomers and homodimers for representative sequences of the PolyDIV clusters. Then, we produced artificial PolyDIV sequences by replacing the DivIVA-like domains of three PolyDIV proteins with DivIVA-like domains from either DivIVA, or GpsB, or FilP, for a total of 9 constructs (see S1 File). We predicted these both as monomers and as homodimers as well. Examples of the different topologies obtained are shown in Fig 3, and the complete set of sequences and models is provided in a Mendeley repository.

Results

In a search for protein sequences with helical potential and hendecad periodicity, we encountered the Scy and FilP proteins, which have previously been described to contain hendecads [7]. In a scan for PFAM domains in these sequences (see Methods), we were surprised to obtain matches to the DivIVA domain in the N-terminal part of the proteins. Further, analysis of the N-terminal sequences by HMM-HMM comparison gave strong matches to the structures of DivIVA and its close homologs (Fig 1A and 1B), hinging on a pattern of conserved residues whose broad conservation we ascertained through multiple sequence alignments of the respective protein families. The consensus sequence for the conserved motif is RGYDxxEVD. In the crystal structures, residues RGYD are part of a loop that connects a short N-terminal helix to the coiled-coil helix mediating dimerization, which starts with EVD (Fig 1A–1C). In the motif, the G allows for a close contact between the symmetry-related loops of the dimer, the Y of one chain forms a hydrogen bond with the second D in the other chain, the first D is the N-terminal capping residue of the coiled coil, and the V is the first residue of its hydrophobic core. Based on these observations, we define the DivIVA-like domain as the conserved motif in the context of a coiled coil with heptad periodicity.

In order to collect as broadly as possible sequences featuring the DivIVA-like domain, we used pattern searches against sequence databases, as described in the Methods. We clustered the approximately 5500 sequences obtained in these searches by pairwise sequence similarity using CLANS. As is apparent from Fig 2A, the resulting cluster map consists of groups of well-connected sequences surrounded by a large halo of sequences that make few or no connections in the map. In order to separate the densely-connected groups into individual clusters, we used the Girvan-Newman algorithm for community detection. Finally, we checked the clusters for the presence of a hydrophobic pattern indicative of heptad coiled coils. This left us with the clusters highlighted in color in Fig 2A as the ones containing a DivIVA-like domain. At a P-value threshold of 1, these clusters came together at the center of the map (S1 Fig in S1 File). As we used more stringent thresholds, three main groups of clusters emerged. Mapping the known DivIVA-like proteins into the map, we identified one group as Scy/FilP (red group), and another one as DivIVA/GpsB (blue group). The third (purple) group did not have sequences that could be reliably mapped to any known DivIVA-like proteins.

In the DivIVA/GpsB group of clusters, the center is taken by DivIVA representatives from Firmicutes (lactobacillales, eubacteriales, clostridiales). Radiating outwards from the center of the map is GpsB (as expected, found only in Firmicutes), and two clusters of Firmicute proteins which are uncharacterized at present. Radiating inwards is the cluster of actinobacterial DivIVA proteins (from micrococcales, corynebacteriales and acidimicrobiales). All these sequences show essentially the same domain organization (Fig 2C), with variations in length and in the position and number of unstructured regions, as well as different preferences for the first residue of the motif ([R]GYD is the most common, followed by [E] and [G]). In the PDB structures of DivIVA-like proteins (Fig 1A), this residue protrudes from the top of the structure and may be part of an interacting surface, modulating binding partner specificity. In terms of genomic context, DivIVA proteins are found close to other proteins involved in cell division, such as SepF, FtsZ and FtsA, as well as YggT and YggS homologs. In turn, GpsB proteins are found close to RecU (Holliday junction resolvase), DnaD (component of the PriA primosome) and the penicillin-binding protein PBP1A.

The FilP/Scy group of clusters, as expected, contains exclusively actinobacterial sequences. The center is taken by FilP representatives (mainly from micrococcales, micromonosporales, and pseudonocardiales), as well as a few Scy sequences from Streptomyces (in Fig 2A and 2B, label 2). Radiating outwards are three clusters of uncharacterized proteins (from

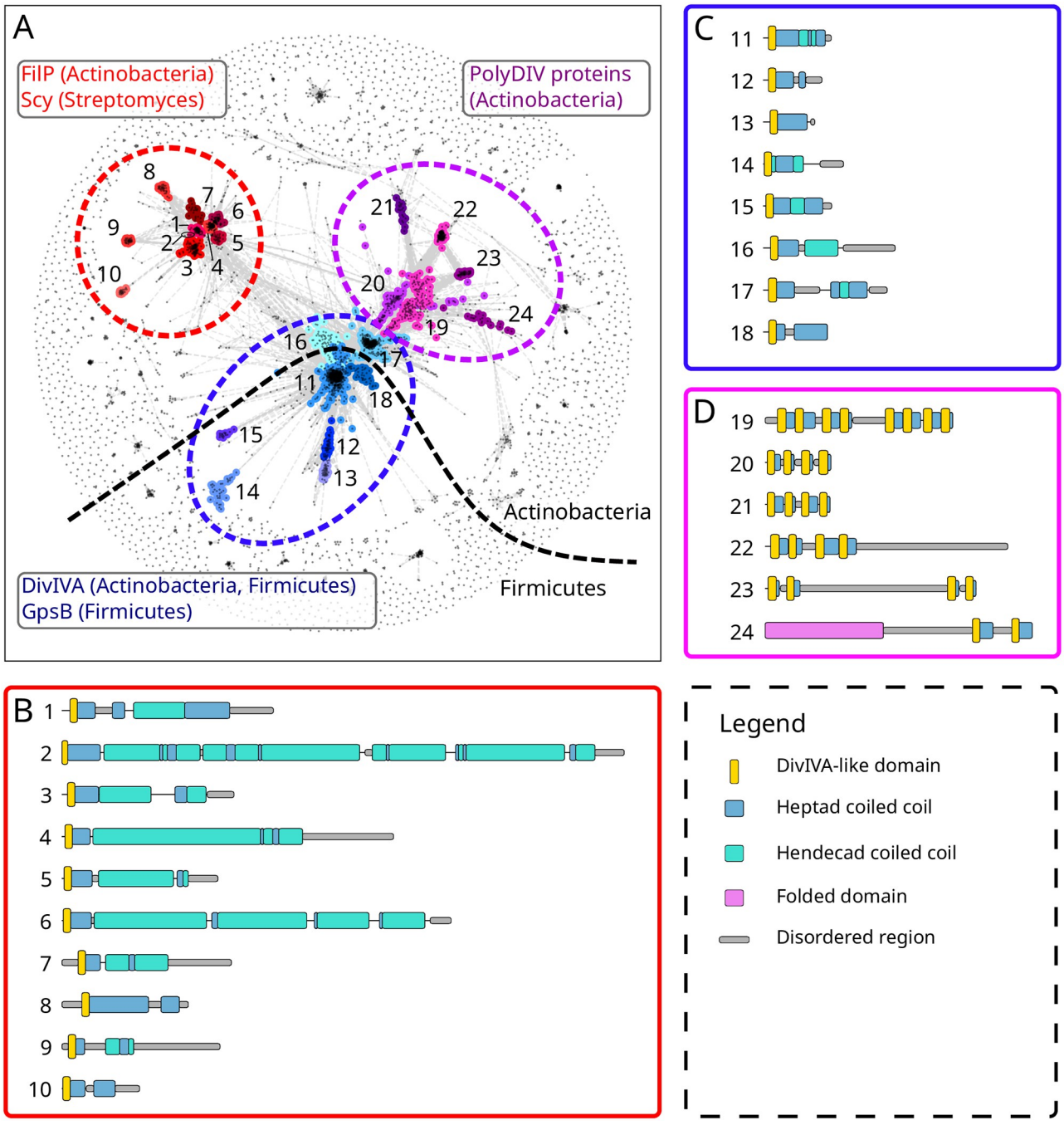


Fig 2. A broad survey of DivIVA-like proteins. (A) Cluster map of sequences with a match to the conserved motif in DivIVA-like proteins (see Methods). Clustering was done in 2D until equilibrium at a BLAST P-value of 1E-10. Connections represent similarities up to a P-value of 1E–10. The three groups of clusters that feature a DivIVA-like domain are colored and labeled. Cartoon representation of the sequence features for representative sequences of Scy/FilP (B), DivIVA/GpsB (C), and PolyDIV groups (D).

https://doi.org/10.1371/journal.pone.0273136.g002

geodermatophilales, micromonosporales and pseudonocardiales). All the sequences in the FilP/Scy group show the same domain organization, but differ substantially in length, from under 200 residues to over 1000 (Fig 2B). All clusters show a preference for R before GYD, except for one of the uncharacterized clusters, where F or T are preferred. Surprisingly, Scy

sequences, which are closely embedded in the central FilP cluster, are N-terminally truncated and start right before the RGYD motif, thus lacking the N-terminal buttressing helix (Fig 1A). In terms of genomic context, there is a greater diversity than in the DivIVA/GpsB group. The Scy proteins as well as some FilP clusters are found close to crotonyl-CoA carboxylase/reductase and methylmalonyl-CoA epimerase genes, and in the case of Scy, also to a cellulose-binding protein, which by BLAST searches can be reliably identified as the FilP of Streptomyces species.

The third group of clusters also contains exclusively actinobacterial proteins (from micrococcales, streptosporangiales, pseudonocardiales, micromonosporales, corynebacteriales and bifidobacteriales). Its constituent sequences are distinctive in that they do not feature only one instance of the DivIVA-like domain, but rather several, ranging from a minimum of 2 to a maximum of 8, while the majority of the sequences feature 4 such domains. We have named this new, hitherto unknown group of proteins PolyDIV. Uniquely among DivIVA-like proteins, one cluster of PolyDIV proteins (no. 24 in Fig 2A and 2D) contains an additional folded domain, which is a Ser/Thr protein kinase. Most of the clusters in the PolyDIV group have a UMP kinase in their genomic vicinity.

The DivIVA-like domains in PolyDIV proteins tend to group in pairs, with connectors within the pairs being typically shorter than between pairs (Fig 2D). Although it is entirely possible that these proteins form dimers through the consecutive association of their DivIVA-like domains, their invariably even number and the length difference in their connectors suggested to us that the PolyDIV proteins might actually be monomeric, with consecutive DivIVA-like domains folding into DivIVA-like structures. Indeed, AlphaFold models predicted all representative PolyDIV proteins as monomers, with consecutive DivIVA-like domains assembling into DivIVA-like structures (first with second and third with fourth, as shown in Fig 3A). As a control, we also predicted the same proteins as homodimers; most models (7 of 9) retained the monomer topology, i.e. folding into DivIVA-like structures within a single chain (Fig 3C), while the rest showed a more complex topology with mixtures of intra- and inter-chain DivIVA-like structures (Fig 3D).

In order to probe the plausibility of intragenic amplification as a mechanism for the origin of PolyDIVs, we constructed synthetic PolyDIVs by amplification of DivIVA-like domains that are optimized for homodimeric interaction. To this end, we inserted the DivIVA-like domains of either DivIVA, or GpsB, or FilP into the linker frames of wild-type PolyDIV proteins (see Methods). When predicted as monomers, the majority of these synthetic PolyDIVs (8 of 9) displayed a topology identical to the wild-type PolyDIV monomers (Fig 3A), but when predicted as dimers, the majority (7 of 9) featured a more complex topology, combining intra- and inter-chain assembly (Fig 3D). These models show that intragenic amplification is sufficient to yield a monomeric PolyDIV topology from a precursor that is obligately homodimeric, but that further adaptations are needed to produce a structurally-specific variant.

Discussion

We have shown that the DivIVA, GpsB, FilP and Scy proteins share a common N-terminal sequence motif (GYDxxEVD) followed by a short, dimeric heptad coiled coil. We define this combination as the DivIVA-like domain. A survey of proteins with DivIVA-like domains showed that these form three large groups of sequence clusters, the DivIVA/GpsB group, the FilP/Scy group, and a third, hitherto unreported group, which we name PolyDIV, for its most salient feature: unlike all other proteins containing the DivIVA-like domain, PolyDIVs contain the domain in multiple copies (2 to 8) within one chain. Among all these proteins, DivIVA is the one with the broadest taxonomic spectrum, as it is almost universal in Gram-positive

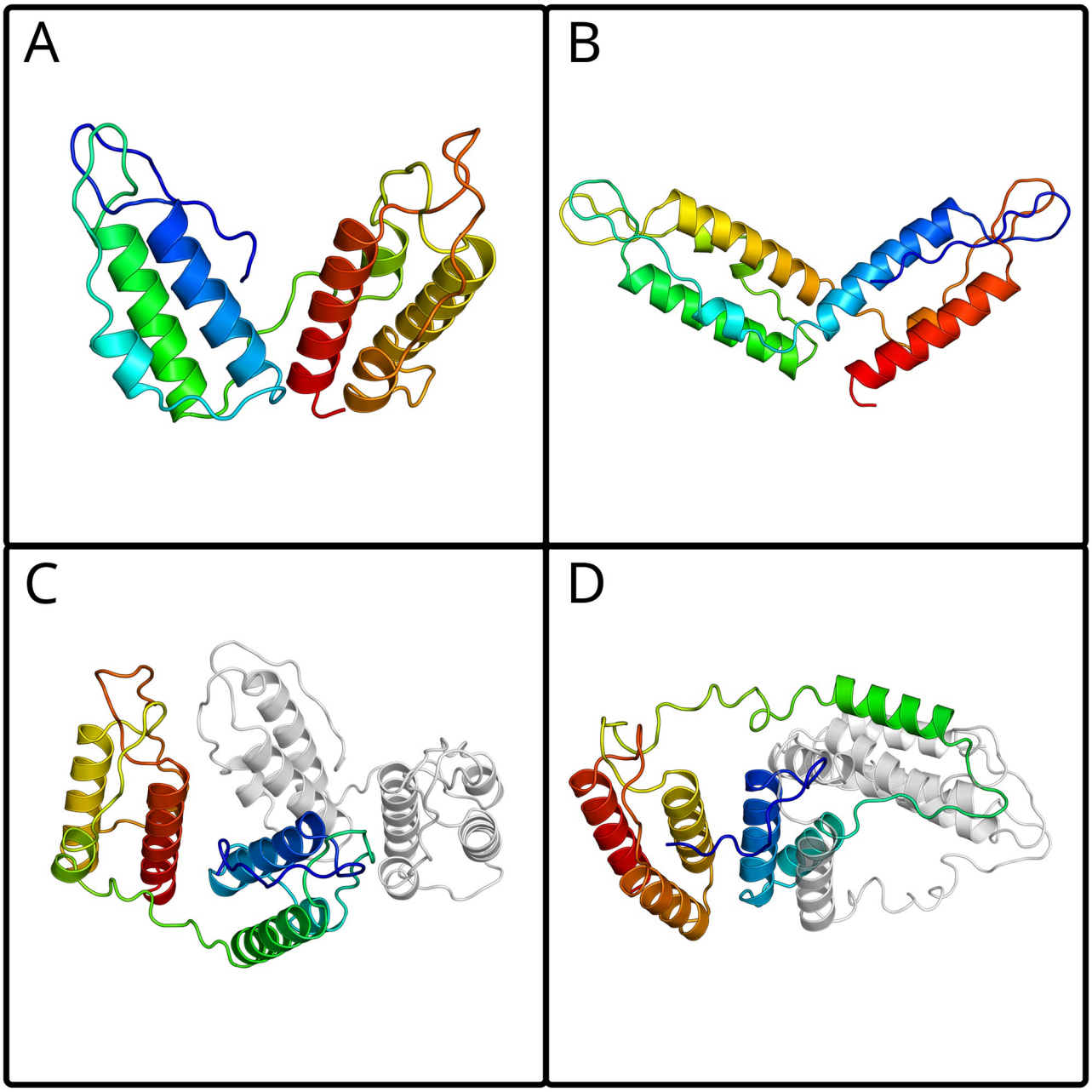


Fig 3. Examples of the topologies observed in the PolyDIV AlphaFold models. Colored from N-terminus (blue) to C-terminus (red). (A) The PURPLE_3_monomer model shows its DivIVA-like domains interacting in a consecutive manner, first with second and third with fourth. (B) The PURPLE_3+DivIVA_monomer shows a non-consecutive DivIVA-like topology. (C) PURPLE_3_dimer shows a topology equivalent to that of (A), where the two chains do not interact via their DivIVA-like domains. (D) PURPLE_1_dimer shows a complex topology where one or more DivIVA-like domains interact with another DivIVA-like domain from another polypeptide chain. All sequences and models are provided in the S1 File.

https://doi.org/10.1371/journal.pone.0273136.g003

bacteria. GpsB is only found in Firmicutes, FilP only in Actinobacteria, and Scy exclusively in Streptomycetales (an order within Actinobacteria). It is well established in the literature that GpsB is a DivIVA paralog [2, 3], the product of an ancient duplication of the gene, and it has been noted that Scy and FilP are probably paralogs of each other as well [6]. Our discovery of a shared domain between all these proteins provides the first evidence for their homologous

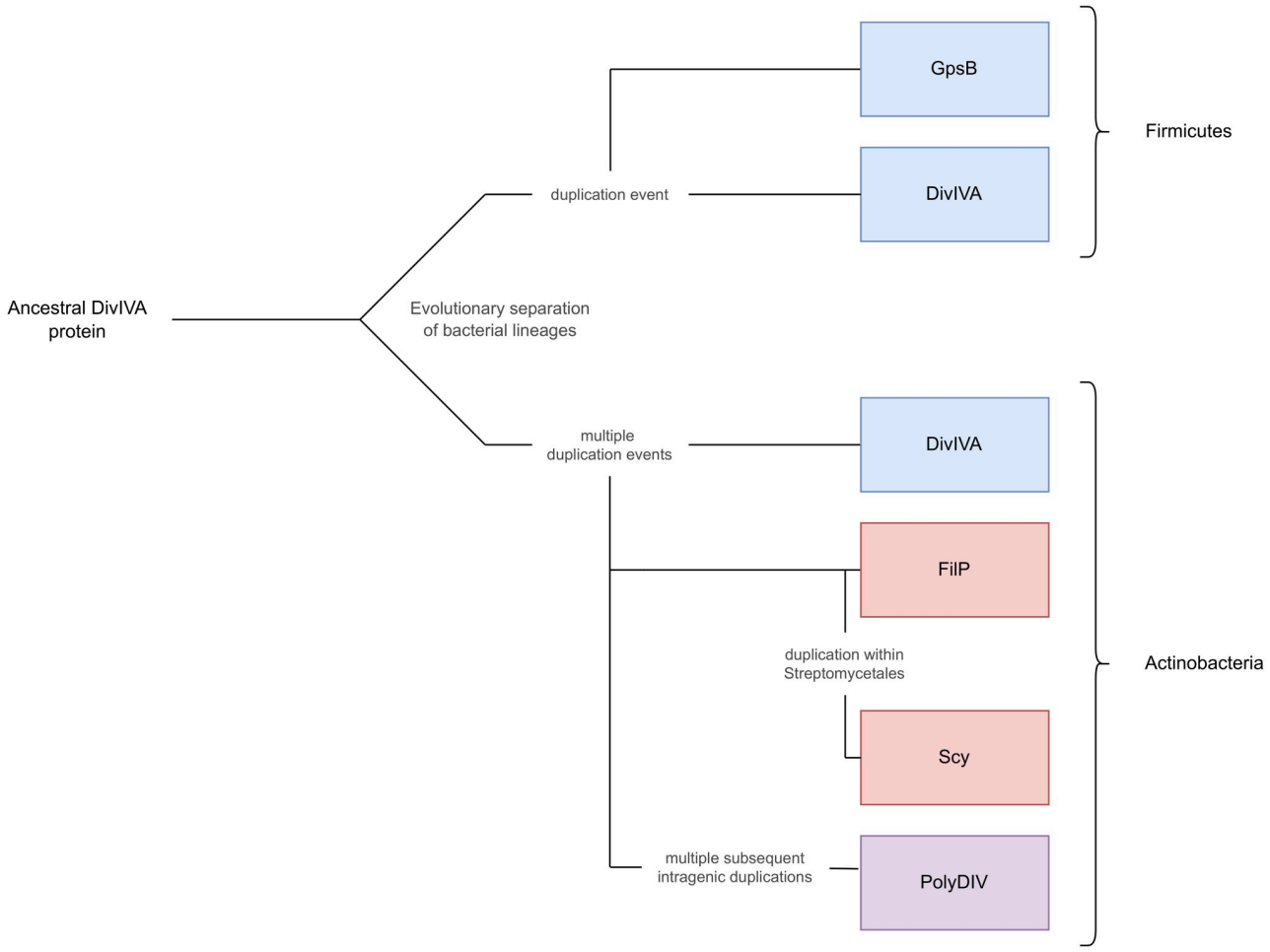


Fig 4. Evolutionary scenario for proteins containing the DivIVA-like domain. The diversity of the DivIVA-like superfamily seems to have arisen mainly through events of duplication (intra- or intergenic) and coiled-coil repeat expansion. The deepest branching point in this tree is an orthologous split due to the separation between Firmicutes and Actinobacteria. All other branch points are due to paralogy.

https://doi.org/10.1371/journal.pone.0273136.g004

origin. We thus define here the new superfamily of DivIVA-like proteins and show that this includes a large group of unannotated proteins, the PolyDIVs. We propose that, like all other members of known function in this superfamily, PolyDIVs are also involved in cell division.

In evolutionary terms, it is clear that DivIVA is at the root of this superfamily, based on its broad taxonomic spectrum and its basic architecture. After the separation of Firmicutes and Actinobacteria within the Gram-positives, independent duplication events led to the origin of GpsB in the Firmicutes, and of FilP and PolyDIV in the Actinobacteria (Fig 4). While GpsB retained the basic architecture of DivIVA, FilP acquired a longer rod segment, in which the N-terminal heptad repeats were gradually expanded by hendecads, and PolyDIV underwent multiple intragenic duplications of the DivIVA domain, which changed the overall topology of the protein from a homodimer to a monomer. At a later stage, after the separation of Streptomycetes from the other Actinobacteria, a further duplication of FilP yielded Scy, which shows the longest coiled-coil rod among DivIVA-like proteins (at over 1000 residues) and is composed almost entirely of hendecads. This outlines the most likely evolutionary events that resulted in the modern complement of DivIVA-like proteins, but BLAST searches within individual proteomes provide many instances of DivIVA-like-related lineage-specific gene duplications (see

S2 Appendix in S1 File). These indicate that having multiple paralogs of one or the other of the DivIVA-like proteins is a widespread phenomenon and that the underlying duplications are an ongoing process.

Structurally, the FilP/Scy and the DivIVA/GpsB groups are obligate dimers. While our analyses suggest that PolyDIVs mostly fold as monomers, it is entirely possible that some oligomerize into more complex topologies. The advantage of the PolyDIV topology over the ancestral DivIVA one is unclear to us, but we note that PolyDIVs are widespread throughout Actinobacteria. Besides the evident increase in binding-site density along the same polypeptide chain and decreased requirements for folding and assembly, PolyDIVs may have the potential to act as hubs by displaying binding sites with different specificity in close proximity to each other. This, in turn, would explain why the DivIVA-like domains of PolyDIVs are more divergent compared to other DivIVA-like proteins.

The DivIVA, GpsB, FilP and Scy proteins are key players in the process of bacterial division. In spite of considerable effort, the mechanisms by which these proteins act within this process are incompletely understood and have not been synthesized into an overarching molecular model. Our identification of a domain conserved between these proteins and our observation that the highest sequence conservation is found in a loop displayed at the top of the domain immediately suggest that this is the main site for protein-protein interactions, providing a unifying feature for these otherwise diverse proteins.

Supporting information

S1 File. We provide S1 Fig (Series of CLANS maps at different P-value thresholds), S2 Appendix (Evidence of continuous genetic duplication events), and S3 Fig (Heatmap of pairwise sequence similarities for representative members of the DivIVA-like superfamily) as Supplementary information. Furthermore, the following data are available at https://data.mendeley.com/datasets/bn627zbymx: (i) a cluster-map file, which can be navigated interactively in CLANS and gives direct access to all the sequences in this study, (ii) the sequence annotations for representatives of DivIVA-like proteins one for each cartoon shown in Fig 2B–2D, and (iii) the sequences and structural models of the PolyDIVs modeled with AlphaFold, both natural and artificial.
(DOCX)

Acknowledgments

We would like to thank Dr Vikram Alva and Dr Stanislaw Dunin-Horkawicz for useful discussions and comments that contributed greatly to improve this manuscript.

The AlphaFold predictions were computed using the module available at the Raven cluster, part of the Max Planck Computing and Data Facility (https://www.mpcdf.mpg.de).

Author Contributions

Conceptualization: Andrei N. Lupas.

Funding acquisition: Andrei N. Lupas.

Investigation: Mikel Martinez-Goikoetxea.

Supervision: Andrei N. Lupas.

Writing – original draft: Mikel Martinez-Goikoetxea.

Writing – review & editing: Mikel Martinez-Goikoetxea, Andrei N. Lupas.

References

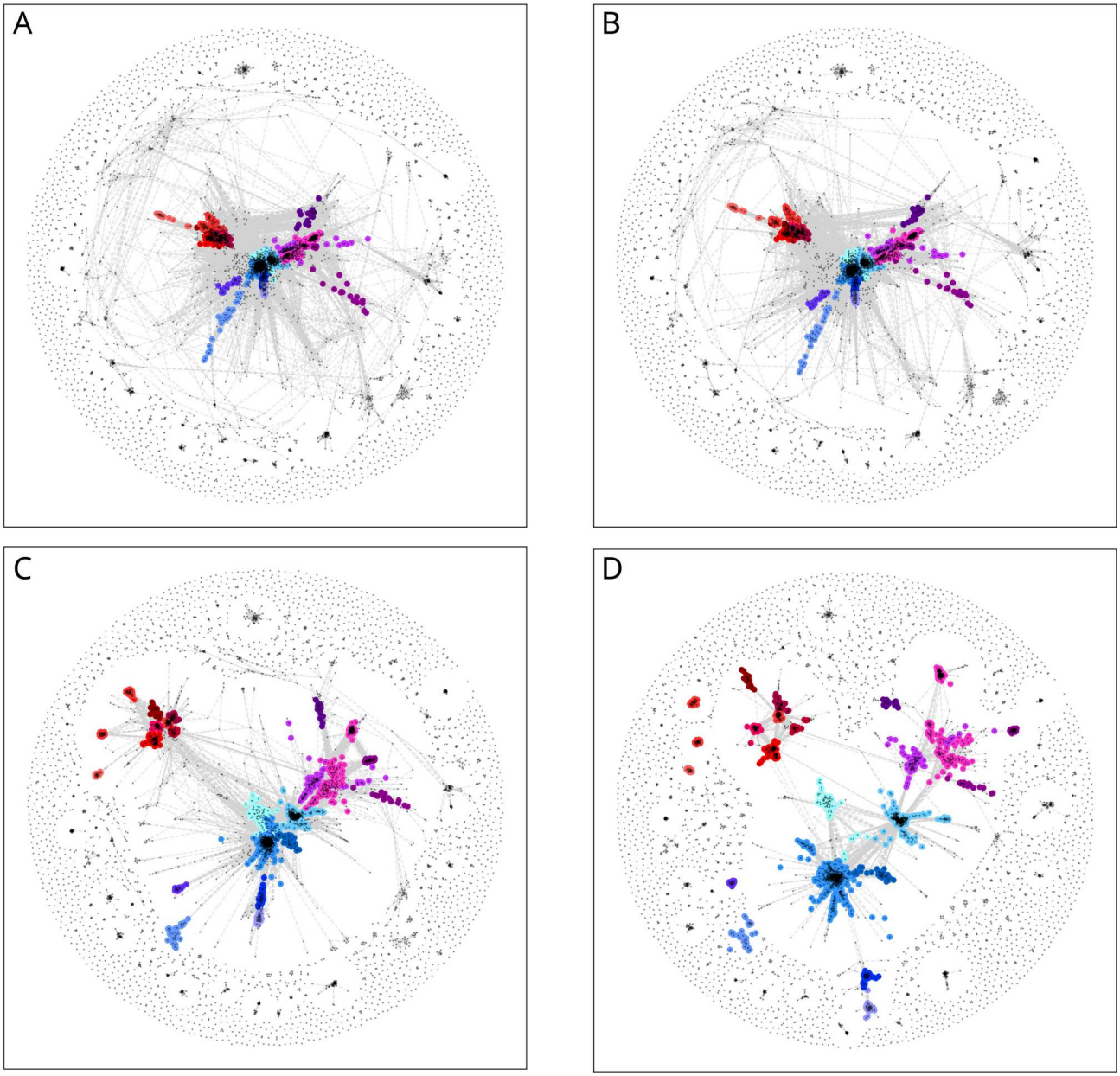
1. Hammond LR, White ML, Eswara PJ. ¡vIVA la DivIVA! Margolin W, editor. J Bacteriol. 2019 Nov; 201 (21).
2. Cleverley RM, Rutter ZJ, Rismondo J, Corona F, Tsui HCT, Alatawi FA, et al. The cell cycle regulator GpsB functions as cytosolic adaptor for multiple cell wall enzymes. Nat Commun. 2019 Dec; 10(1):261. https://doi.org/10.1038/s41467-018-08056-2 PMID: 30651563
3. Halbedel S, Lewis RJ. Structural basis for interaction of DivIVA/GpsB proteins with their ligands. Mol Microbiol. 2019 Jun; 111(6):1404–15. https://doi.org/10.1111/mmi.14244 PMID: 30887576
4. Javadi A, Söderholm N, Olofsson A, Flärdh K, Sandblad L. Assembly mechanisms of the bacterial cytoskeletal protein FilP. Life Sci Alliance. 2019 Jun; 2(3):e201800290. https://doi.org/10.26508/lsa.201800290 PMID: 31243049
5. Fröjd MJ, Flärdh K. Apical assemblies of intermediate filament-like protein FilP are highly dynamic and affect polar growth determinant DivIVA in *Streptomyces venezuelae*. Mol Microbiol. 2019 Jul; 112 (1):47–61.
6. Walshaw J, Gillespie MD, Kelemen GH. A novel coiled-coil repeat variant in a class of bacterial cytoskeletal proteins. J Struct Biol. 2010 May; 170(2):202–15. https://doi.org/10.1016/j.jsb.2010.02.008 PMID: 20178847
7. Holmes NA, Walshaw J, Leggett RM, Thibessard A, Dalton KA, Gillespie MD, et al. Coiled-coil protein Scy is a key component of a multiprotein assembly controlling polarized growth in *Streptomyces*. Proc Natl Acad Sci. 2013 Jan 29; 110(5).
8. Lupas AN, Bassler J. Coiled Coils—A Model System for the 21st Century. Trends Biochem Sci. 2017 Feb; 42(2):130–40. https://doi.org/10.1016/j.tibs.2016.10.007 PMID: 27884598
9. Crick FHC. The Packing of a-Helices: Simple Coiled-Coils. Acta Cryst. 1953; 6:689–97.
10. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011 Jul 1; 39(suppl):W29–37. https://doi.org/10.1093/nar/gkr367 PMID: 21593126
11. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021 Jan 8; 49(D1):D412–9. https://doi.org/10.1093/nar/gkaa913 PMID: 33125078
12. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. 2018 Jul; 430(15):2237–43. https://doi.org/10.1016/j.jmb.2017.12.007 PMID: 29258817
13. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017 Nov; 35(11):1026–8. https://doi.org/10.1038/nbt.3988 PMID: 29035372
14. Bailey TL, Elkan C. Fitting a Mixture Model By Expectation Maximization To Discover Motifs In Biopolymer.: 9.
15. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences: Clustal Omega for Many Protein Sequences. Protein Sci. 2018 Jan; 27(1):135–45.
16. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A Sequence Logo Generator.: 3.
17. Frickey T, Lupas A. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics. 2004 Dec 12; 20(18):3702–4. https://doi.org/10.1093/bioinformatics/bth444 PMID: 15284097
18. Gruber M, Soding J, Lupas AN. REPPER—repeats and their periodicities in fibrous proteins. Nucleic Acids Res. 2005 Jul 1; 33(Web Server):W239–43. https://doi.org/10.1093/nar/gki405 PMID: 15980460
19. Erdős G, Dosztányi Z. Analyzing Protein Disorder with IUPred2A. Curr Protoc Bioinforma. 2020 Jun; 70 (1). https://doi.org/10.1002/cpbi.99 PMID: 32237272
20. Pereira J. GCsnap: Interactive Snapshots for the Comparison of Protein-Coding Genomic Contexts. J Mol Biol. 2021 May; 433(11):166943. https://doi.org/10.1016/j.jmb.2021.166943 PMID: 33737026
21. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug 26; 596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844
22. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. Bioinformatics; 2021 Oct.

Supplementary Information: A conserved motif suggests a common origin for a group of proteins involved in the cell division of Gram+ bacteria

Mikel Martinez-Goikoetxea, Andrei N. Lupas*

Department of Protein Evolution, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

* To whom correspondence should be addressed:
Andrei N. Lupas
Tel.: +49 7071 601 341
Fax: +49 7071 601 349
E-mail: andrei.lupas@tuebingen.mpg.de

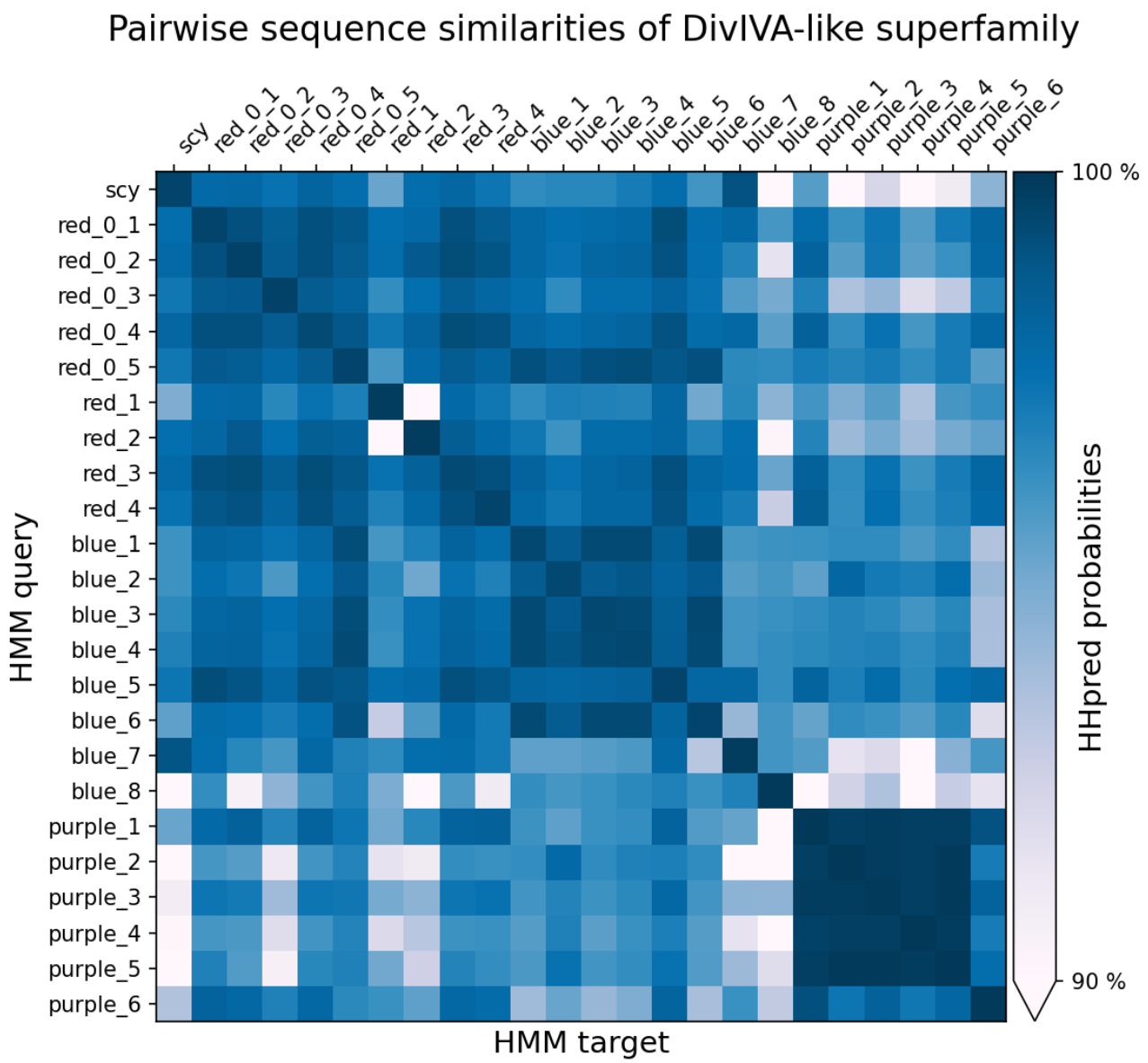


S1 Figure. Series of CLANS maps at different P-value thresholds.

Shown are the CLANS maps, colored to match Fig. 2A in the main manuscript, at the P-value thresholds of (A) p=1, (B) p=1E-5, (C) p=1E-10, (D) p = 1E-15. As the value is decreased, only the most significant matches are taken into account to construct the map, and thus the sub-groups are revealed

S2 Appendix. Evidence of continuous genetic duplication events

While performing BLAST searches of DivIVA-like domains against the proteomes of the seed database (https://pubseed.theseed.org), we found instances of non-streptomyces actinobacteria that featured several additional DivIVA-like proteins besides one copy of DivIVA and one of FilP. For example, in Pseudonocardia dioxanivorans CB1190, we found two sequences (seed identifiers 675635.11.peg. 1714 and 1181) that by BLAST searches against our cluster map could be reliably identified as FilP-like proteins. In Actinoplanes sp. SE50/110, we found two genomically adjacent FilP-like sequences (134676.3.peg. 7340 and 7341), and two sequences that closely resembled DivIVA (134676.3.peg. 2009, 1670). Finally, by the manual analysis of the GCsnap results, we also found that Nocardiopsis chromatogenes has two genomically adjacent proteins (WP_017626108.1, WP_017626109.1) that by BLAST searches can be mapped to different clusters within the PolyDIV group.



S3 Figure. Heatmap of pairwise sequence similarities for representative members of the DivIVA-like superfamily.

The sequences correspond to those annotated in the Mendeley data repository (https://data.mendeley.com/datasets/bn627zbymx). For the computation of the similarities, we only considered the DivIVA-like domains of every sequence, as defined in the main manuscript, and ran the HHpred HMM-HMM comparison tool for every pairwise comparison. For the graphing of the data, we set a lower bound of 90%, which shows that there is a strong similarity signal both within and between groups. The lowest HHred probability score was higher than 80%.

Applicability of AlphaFold2 in the modelling of coiled-coil domains

Rafal Madaj[a,§], Mikel Martinez-Goikoetxea[b,§], Kamil Kaminski[a], Jan Ludwiczak[a,†], and Stanislaw Dunin-Horkawicz[a,b,*]

[a] Institute of Evolutionary Biology, Faculty of Biology, Biological and Chemical Research Centre, University of Warsaw, Zwirki i Wigury 101, 02-089, Warsaw, Poland

[b] Department of Protein Evolution, Max Planck Institute for Biology Tübingen, Max-Planck-Ring 5, 72076 Tübingen, Germany

[†] Present address: Prescient Design, Genentech Research & Early Development, Roche Group, Grenzacherstrasse 124, 4070 Basel, Switzerland.

[§] These two authors contributed equally to this work

[*] To whom correspondence should be addressed

E-mail: stanislaw.dunin-horkawicz@tuebingen.mpg.de

Abstract

Coiled coils are a common protein structural motif involved in cellular functions ranging from mediating protein-protein interactions to facilitating processes such as signal transduction or regulation of gene expression. They are formed by two or more alpha helices that wind around a central axis to form a buried hydrophobic core. Various forms of coiled-coil bundles have been reported, each characterized by the number, orientation, and degree of winding of the constituent helices. This variability is underpinned by short sequence repeats that form coiled coils and whose properties determine both their overall topology and the local geometry of the hydrophobic core. The strikingly repetitive sequence has enabled the development of accurate sequence-based coiled-coil prediction methods; however, the modelling of coiled-coil domains remains a challenging task. In this work, we present the outstanding accuracy of AlphaFold2 in modeling coiled-coil domains, both in modeling local geometry and in predicting global topological properties. Furthermore, we show that the prediction of the oligomeric state of coiled-coil bundles can be improved by using the internal representations of AlphaFold2, with a performance better than any previous state-of-the-art method (code available at https://github.com/labstructbioinf/dc2_oligo).

Introduction

Coiled coils are protein structural motifs consisting of two or more alpha helices, oriented parallel or antiparallel, that wind around a central axis to form rod-like bundles (Lupas *et al.*, 2017). This winding, also known as supercoiling, is caused by periodic interlocking of side chains localized in the hydrophobic core of the bundle. The basis of this interlocking, an interaction known as knobs-into-holes, is considered the hallmark of coiled coils. It places the side chain of a core residue (knob) of one helix into a cavity (hole) formed by residues of the opposite helix. At the sequence level, coiled-coil domains are formed by 7-residue repeats (heptad) in which residues are labeled from *a* through *g*. Residues at two positions, *a* and *d*, are typically hydrophobic and face the bundle axis, forming the hydrophobic core. Since the average spacing between hydrophobic core residues in a heptad is 3.5 (7/2) and the average periodicity of an undistorted alpha helix is 3.63 residues per turn, left-handed supercoiling is required to effectively reduce the helical periodicity to 3.5 and allow the formation of knobs-into-holes interactions along the helices (Figure 1A). Importantly, this effective reduction in periodicity occurs with respect to the bundle axis and does not affect the actual helical periodicity (Szczepaniak *et al.*, 2021).

Although heptads represent the vast majority of known coiled-coil domains, deviations from this canonical periodicity are possible and have been observed, very often as local non-canonical repeats interspersed between arrays of heptads, but also as global non-heptad coiled-coil domains (Martinez-Goikoetxea and Lupas, 2023). For example, one of the most common non-heptad repeats is the hendecad (Figure 1A), which is 11 residues long and contributes three residues to the hydrophobic core. Since the average spacing between core residues (11/3=3.67) is very close to the periodicity of an alpha helix, coiled-coil bundles based on hendecad repeats show almost no supercoiling. Unlike hendecads, pentadecads are 15 residues long, contribute four residues to the core (15/4=3.75), and cause right-handed supercoiling (3.75>3.63). These departures from the heptad periodicity are necessarily accompanied by the partial loss of the knobs-into-holes interactions, as some of them become

knobs-to-knobs, with the side chains of the core residues (knobs) pointing toward the central axis instead of toward a hole. The canonical knobs-into-holes packing can also be disrupted by axial rotation of the constituent helices, which does not alter supercoiling, but can lead to the formation of knobs-to-knobs interactions. For example, counterclockwise rotation of helices (viewed from their N-terminal ends) in an antiparallel 4-helix bundle of 7/2 periodicity by about 26° relative to the canonical packing results in the formation of an *a*-*d*-*e* core, with residues in positions *a* pointing toward the center of the bundle and residues in positions *d* and *e* flanking them (Figure 1B).

Due to their regular nature, coiled-coil structures can be fully described by parametric equations (Wood *et al.*, 2017; Grigoryan and Degrado, 2011; Strelkov and Burkhard, 2003; Szczepaniak *et al.*, 2021; Crick, 1953). Such descriptions allow structural parameters to be traced down to single-residue resolution, highlighting subtle differences within and between coiled-coil domains (Figure 1C). The most important parameters describing coiled-coil structures are their topology, i.e., the number and relative arrangement of helices (parallel vs. antiparallel), the degree of supercoiling (from left-handed to nearly straight to right-handed), and the axial rotation of helices, which defines the architecture of the hydrophobic core. Such a detailed view is important for relating the structural properties of coiled-coil domains to the function of the proteins that contain them. For example, the structural parameters of the HAMP domain, a small signaling 4-helix coiled coil, have been shown to be tightly coupled to the enzymatic activity of downstream domains (Ferris *et al.*, 2011).

The high frequency with which coiled-coil domains are found (Szczepaniak *et al.*, 2021) as well as their well-understood sequence-structure relationship have motivated the development of many computational tools for their detection and modeling. The accuracy of automated coiled-coil prediction tools such as COILS (Lupas *et al.*, 1991) or DeepCoil2 (Ludwiczak *et al.*, 2019) is remarkable, in part due to the convergent nature of coiled-coil sequence repeats (non-homologous coiled coils share similar general sequence patterns due to constraints imposed by

supercoiling). Despite our ability to predict coiled coils from sequence, their structural modeling remains a challenging problem that has been addressed by programs such as CCFold (Guzenko and Strelkov, 2018) and Rosetta (Das *et al.*, 2009), which are based on fragment modeling, and ISAMBARD (Wood *et al.*, 2017), CCBuilder (Wood and Woolfson, 2018), BeammotifCC (Offer *et al.*, 2002), and CCCP (Grigoryan and Degrado, 2011), which use the aforementioned parametric equations. However, none of these programs is universally applicable. Some are limited to certain oligomeric states or require additional data besides a sequence, while others can only model perfectly regular bundles without considering local discontinuities. These discontinuities are often essential for coiled-coil function, as demonstrated in examples such as signal transduction (Ferris *et al.*, 2012) and intracellular trafficking (Murray *et al.*, 2016).

Recent years have seen a number of breakthroughs in protein structure prediction with the development of deep learning-based methods, the most prominent of which is AlphaFold2 (Jumper *et al.*, 2021; Evans *et al.*, 2022). It is implemented as an end-to-end sequence-to-structure model that also exploits evolutionary information provided in the form of a multiple sequence alignment. Benchmarks have demonstrated its superiority over classical homology modeling approaches (Lupas *et al.*, 2021) and its applicability to the modeling of protein complexes (Akdel *et al.*, 2022) and peptides (McDonald *et al.*, 2023). In addition, although not designed for such tasks, AlphaFold2 has been reported to provide insight not only into protein structure but also into its dynamics (Winski *et al.*, 2024; del Alamo *et al.*, 2022; Stein and Mchaourab, 2022; Wayment-Steele *et al.*, 2022).

In this work, we present a systematic benchmark of modeling coiled-coil domains with AlphaFold2. Going into this project, there were a number of features of coiled coils that we thought might be challenging for AlphaFold2. For example, compared to globular proteins, coiled coils have a relatively small number of contacts, which means that their structure is determined by a comparatively small number of residues. This is reflected in the fact that coiled-coil topology (oligomerization state and helical orientation) can be altered by one or a

few mutations, typically near the hydrophobic core (Harbury *et al.*, 1993). Regarding the use of evolutionary information, it is worth noting that the structural features and repetitive nature of coiled coils impose strong constraints on their sequences, which has been identified as a problem because it often leads to false matches between non-homologous coiled-coil sequences (Mistry *et al.*, 2013). This could result in the MSA provided to AlphaFold2 being "contaminated" with non-homologous coiled-coil segments that could potentially have very different structural properties. With these potential challenges in mind, we set out to evaluate the ability of AlphaFold2 to predict coiled-coil features, including global topology and local geometry.

Methods

Datasets

The benchmark was to compare experimentally determined coiled-coil structures from CCdb (Szczepaniak *et al.*, 2021) with their corresponding AlphaFold2 predictions. We focused on structures containing dimeric, trimeric, and tetrameric coiled-coil bundles, filtering out structures with resolutions below 3.5 Å or with complex topologies other than regular bundles. Finally, we retained only structures in which coiled-coil regions, defined as residues involved in knobs-into-holes interactions detected by SOCKET (Walshaw and Woolfson, 2001), accounted for more than 50% of all residues. All the filtering steps were performed using the *localpdb* package (Ludwiczak *et al.*, 2022). This process resulted in an “automatic” benchmark set of 379 coiled-coil structures.

For the oligomer state prediction benchmark, we further refined this data set. We excluded heterooligomers and structures in which the heptad register could not be assigned with TWISTER (Strelkov and Burkhard, 2003) (unambiguous register assignment was critical because some of the benchmarked methods use this feature to improve prediction accuracy). This resulted in a set of 216 structures, with each oligomer class (38% dimers, 36% trimers, and 26%

tetramers) similarly represented. In addition to these sets, we constructed two additional sets based on manually curated examples, consisting of 19 parallel (Szczepaniak *et al.*, 2018) and 21 antiparallel (Szczepaniak *et al.*, 2014) coiled-coil structures, most of which were GCN4 variants. The four benchmark sets are summarized in Supplementary Table 1.

AlphaFold2 modelling

Modeling was performed with ColabFold (version 2.5.2) (Mirdita *et al.*, 2022) using the *alphafold2_multimer_v3* model, 5 rounds of recycling, and Amber minimization. It is important to note that although we modeled the entire sequences (i.e., including the non-coiled coil regions), we performed the structural analyses only on the coiled-coil regions.

Oligomerization state prediction

We selected to benchmark LOGICOL (Vincent *et al.*, 2013) and CoCoNat (Madeo *et al.*, 2023) as representatives of state-of-the-art coiled-coil oligomerization state prediction methods. Unlike AlphaFold2, both programs can be run with coiled-coil register information, which improves the robustness of the predictions. We benchmarked LOGICOL and CoCoNat in two different scenarios, a) with the structurally derived heptad registers obtained with TWISTER (Strelkov and Burkhard, 2003), and b) without explicitly providing register information (indicated by the "noregister" suffix in Figure 2). While CoCoNat implements its own register prediction, we opted to run LOGICOIL in the same way as its web server, with registers predicted by MARCOIL (Delorenzi and Speed, 2002).

To assess the extent to which AlphaFold2 can be used to predict the oligomeric state of coiled-coil domains, we used it to model each of the 216 sequences in the “oligomerization” benchmark set as a dimer, a trimer, and a tetramer, and checked whether the oligomerization state of the best-scoring model matched the oligomeric state observed in the experimental

structures. We tested several metrics, including pLDDT, pTM, and PAE. We further refined the PAE score (total_PAE) into the interchain PAE (inter_PAE), which can be thought of as the uncertainty in predicting the relative position of the helices. Confusion matrices and weighted F1 scores were calculated for each method (Figure 2).

Given the good accuracy of the AlphaFold2 quality metrics in predicting oligomeric state, and the fact that AlphaFold2 was not directly supervised for oligomeric state prediction (Evans *et al.*, 2022), we hypothesized that accuracy could be further improved by supervised learning on the internal representations of AlphaFold2. To test this, we extracted representations for each benchmark sequence modeled as a monomer (retrieved with the --save-single-representations option). The resulting representations were then averaged over the sequence length dimension, resulting in fixed size vectors of shape 5x256 (5 corresponds to each of the AlphaFold2 pre-trained models and 256 to the embedding size) per sequence. For visualization purposes, the dimensionality of the embeddings was reduced using PacMAP (Wang *et al.*, 2021) (Figure 3A, left panel).

To process these concatenated AlphaFold2 embeddings, we initially trained a simple neural network using fivefold cross-validation. The neural network architecture included multiple dense layers with batch normalization, dropout, and L2 regularization. The model was trained using the Adam optimizer (Kingma and Ba, 2017) and a sparse categorical cross-entropy loss function. Early stopping with a patience of 3 epochs was applied during training to avoid overfitting. The dimensionality of the internal representations (from the last layer) of the model was reduced using PacMAP (Wang *et al.*, 2021) and visualized (Figure 3A, right panel). We also considered simpler classifiers available in the scikit-learn Python package, and based on fivefold cross-validation, we found LogisticRegression to be the best performing. The internal classifier parameters, namely regularization strength and class balance weighting, were optimized, and the final model, available on GitHub, was fitted to all available data without cross-validation (https://github.com/labstructbioinf/dc2_oligo).

Bundle geometry prediction benchmark

We modeled each structure in the "automatic", "parallel", and "antiparallel" benchmark sets with the oligomerization state matching that of the experimental structure. For comparison with the corresponding experimental structures, we took the top-ranked model as representative of each prediction. We then superimposed the predicted and experimental structures using US-Align (Zhang *et al.*, 2022), checking that all chains were aligned in the correct orientation and excluding misaligned cases from further analysis. Finally, we used SamCC Turbo to detect the coiled-coil segments and compute their structural parameters (helical axial rotation and supercoiling). An example of the comparison of these parameters between an experimental and a predicted model is shown in Figure 1C. Details on the calculation of these parameters can be found in (Szczepaniak *et al.*, 2021).

Results and Discussion

Oligomerization state prediction

We began the benchmark by evaluating the performance of current methods for predicting the oligomerization state of coiled-coil domains. In addition to the sequence-based methods LOGICOIL (Vincent *et al.*, 2013) and CoCoNat (Madeo *et al.*, 2023), we also evaluated the predictive power of the AlphaFold2 model quality metrics pLDDT, pTM, and PAE. For predictions based on AlphaFold2 scores, we computed models for oligomerization states ranging from dimer to tetramer for a given sequence. Then, for each metric, the oligomerization of a model with the best score was selected as the prediction. For each approach, we computed a confusion matrix comparing the predicted number of helices with the number of helices in the experimental structures (Figure 2). For AlphaFold2, since it outputs a structural model, we also checked whether the helix orientation was correct (this was the case for 98% of the parallel and 76% of the antiparallel bundles). As expected, both LOGICOIL

and CoCoNat performed well in the oligomerization state prediction task when provided with structurally derived coiled-coil registers, with weighted F1 scores of 0.71 and 0.77, respectively, although LOGICOIL showed a strong bias towards predicting trimers, as did CoCoNat for dimers. When we did not provide the structurally derived heptad register (see Methods), the F1 scores dropped to 0.63 and 0.72, respectively (in addition to a larger number of incorrect predictions, some predictions were ambiguous, marked "?" in Figure 2).

Among the AlphaFold2 metrics tested, the best performing were pLDDT (F1=0.70) and pTM (F1=0.63), with the former showing a strong bias towards dimers and the latter towards tetramers. We also tested the PAE score, which is a 2D matrix representing the predicted error in the pairwise distances between all residues in all chains. In the benchmark, we considered the average of the entire PAE matrix (total_PAE) and the average of the PAE values between residues from different chains (inter_PAE). The average total_PAE (F1=0.71) was comparable to the pLDDT and slightly worse than the inter_PAE (F1=0.73), which additionally showed a much better ability to predict tetrameric bundles (Figure 2).

Intrigued by the fact that AlphaFold2 had comparable accuracy to state-of-the-art methods despite not being trained to predict oligomerization state, we investigated whether such predictions could be made directly from the internal representations of AlphaFold2, rather than from the scores associated with models computed for different oligomerization states. Projecting the AlphaFold2 representations of the “oligomerization” benchmark set cases into 2D space revealed some degree of separation between dimers, trimers, and tetramers (Figure 3A, left panel). Motivated by this observation, we trained a simple neural network to predict a coiled-coil oligomeric state based on its corresponding AlphaFold2 representation. Projection of the embeddings obtained from this neural network showed a clear separation between the groups (Figure 3A, right panel), suggesting the potential applicability of such an approach. In exploring alternative prediction models, we found that even a simple regression model can produce very good results: 5-fold cross-validation on the benchmark set showed significantly

improved performance (F1=0.82) without any noticeable bias in the preferred oligomeric state (Figure 3B). Note, however, that we evaluated this model using 5-fold cross-validation, so the results obtained are not directly comparable to those shown in Figure 2. The implementation, along with the training routines, has been deposited on GitHub (https://github.com/labstructbioinf/dc2_oligo).

Modeling of coiled-coil geometry

To perform a detailed comparison of experimental structures with AlphaFold2 models, we used SamCC Turbo (Szczepaniak *et al.*, 2021), a program designed to automatically detect coiled-coil domains and calculate their structural parameters, including the degree of bundle supercoiling and hydrophobic core geometry (Figure 1). Comparisons were made using three separate sets of benchmarks, each designed to highlight a specific challenge in coiled-coil modeling. The first and second sets focused on packing geometries in parallel and antiparallel 4-helix bundles, respectively, while the third set was more general and included bundles with different oligomerization states and degrees of supercoiling. For each benchmark, all structures were modeled in the correct oligomerization state obtained from the corresponding experimental structure, and the models were then analyzed using SamCC Turbo.

The first set, termed “parallel”, consists of 19 canonical 7/2 parallel 4-helix bundles in which the helices interact by knobs-into-holes packing. Despite their overall similarity, these structures exhibit slight differences in the geometries of their hydrophobic cores. These differences are primarily manifested in the axial rotation of the helices, which is within ±5° (calculated relative to an idealized reference model of a coiled-coil employing knob-into-holes packing; Figure 1). Our previous work showed that these subtle differences are not artifacts of the experimental procedures but can be related to the sequence composition of the hydrophobic core (Szczepaniak *et al.*, 2018). All obtained models showed correct topology and their hydrophobic cores were modeled with high fidelity (Figure 4A). The only exception was the model of the

designed CC-Hex-II structure (PDB: 4h7r), in which the interhelical interfaces were incorrectly predicted.

The "antiparallel" set consisted of 21 four-helix antiparallel bundles with hydrophobic core geometries defined by rotations ranging from -26° to +10°. Structures with rotations up to ±10° are considered to have canonical knobs-in-holes packing (Figure 1B, left panel), whereas those with rotations below -10° are considered to have non-canonical packing (Figure 1B, right panel). Canonical packing is characterized by the interaction of two residues per heptad repeat within the hydrophobic core. In contrast, non-canonical packing involves the co-option of an additional heptad position e into the core (Lupas et al., 2017). In this benchmark, AlphaFold2 correctly modeled only 13 out of 21 structures. In most of the unsuccessful cases, the resulting models had the wrong topology (parallel instead of antiparallel), preventing a meaningful comparison with the experimental structures. However, the models with the correct topology were highly accurate (Figure 4B), except for the coiled-coil domain of the coronavirus S2 transmembrane fusion protein (PDB: 1zv7), where the core geometry was incorrectly predicted.

The two benchmark sets described above contained hand-picked structures, mostly GCN4 variants. To evaluate AlphaFold2 more comprehensively, we generated a third set containing 379 dimers, trimers, and tetramers of various topologies. Unlike the smaller benchmark sets, this set also included non-canonical bundles characterized by right-handed twist and periodicity >3.63 associated with repeats such as hendecads and pentadecads (Figure 1A). As with the two smaller sets, among the correct topology models (323 out of 379), the majority were accurate with respect to both helix packing and bundle periodicity (Figure 4C). Interestingly, the quality of modeling did not depend on the release date of the corresponding experimental structures, and those released after AlphaFold2 training (2018-04-30) were modeled equally well (compare red and green dots in Figure 4C).

Conclusions

While AlphaFold2 provides highly accurate predictions of the oligomeric state (Figures 2 and 3) and fine details such as supercoiling and core geometry (Figure 4), its ability to predict the correct orientation of the helices is limited. In all benchmarks, we observed the same tendency for AlphaFold2 to model antiparallel structures as parallel. For example, in the benchmark focused on predicting the oligomerization state, the helix orientation was correct in 98% of the parallel and 76% of the antiparallel test cases. Similarly, in the benchmark focused on a set of 21 antiparallel structures (Figure 4B), only 13 out of 21 were correctly modeled as antiparallel. The problems in predicting the correct helical arrangement were to be expected, since it is dictated by the long-range contacts, which can be difficult to capture due to the strong convergent evolutionary signal in coiled-coil domains. This problem has less impact on the prediction of the oligomeric state and core geometry, since these are clearly defined by the features of the interhelical interfaces, such as the presence of additional hydrophobic positions (Szczepaniak *et al.*, 2014) or the formation of multiple interfaces by a single helix (Lupas *et al.*, 2017).

As we looked at models that differed from the experimental structures, we realized that some of them might be more accurate. For example, the experimental structure of the human tetherin protein (PDB: 3mqc) is a tetramer, while the corresponding rank 1 AlphaFold2 model takes the form of two separate dimers (Figure 5A). The rank 2 model is a tetramer with the same helical arrangement as the experimental structure, but contains numerous helical distortions that significantly reduce the quality metrics of this model. Although the rank 1 model does not match the experimental structure, it is actually more accurate because the dimeric form of tetherin has been shown to be functional (Yang *et al.*, 2010). Another example is the human CtIP protein (Figure 5B); again, there is a clear discrepancy between the experimental structure (PDB: 7bgf) and the AlphaFold2 rank 1 model. While the experimental structure assumes a straight coiled coil, the AlphaFold2 model appears to be broken in the middle. The observed break in the model results from the presence of a functionally important

region, a zinc hinge, which is essential for the function of the protein (Morton *et al.*, 2021). Such examples highlight the possibility that some of the AlphaFold2 models can provide clues to protein function even when they do not agree with experimental structures.

Finally, it's worth highlighting the unexpected accuracy in predicting the oligomerization state using AlphaFold2 hidden representations. Such an approach proved to be not only effective (Figure 3), but also faster than modeling a given coiled-coil sequence in all possible oligomerization states (Figure 2). Recently, AlphaFold2 embeddings have also been explored in the context of predicting protein-ligand interactions (Gazizov *et al.*, 2023). Prediction of oligomerization states has also been attempted using ESM2 model embeddings (Sledzieski *et al.*, 2023). We also note that new AlphaFold2-based methods for predicting structures of homooligomers (Schweke *et al.*, 2024) and complex heteroligomeric assemblies (Shor and Schneidman-Duhovny, 2024) have recently been proposed.

In this work, we have highlighted the tremendous versatility of AlphaFold2, demonstrated here by its ability to accurately model important details of coiled-coil domains, and presented a tool capable of predicting the oligomeric state of coiled coils from sequence. We expect that our results will be useful to the coiled-coil community, for example to address challenging problems such as understanding the complex oligomerization propensities of some coiled-coil families, and to detect non-canonical coiled coils with divergent sequence properties.

Acknowledgments

S.D-H. and M.M.G. were supported by institutional funds of the Max Planck Society. This work was also supported by the "Excellence Initiative - Research University" program, an internal grant of the University of Warsaw to enhance the research potential of its staff.

Figure Legends

Figure 1. Coiled-coil domains and quantification of their structures. **(A)** Examples of coiled-coil structures with different periodicities. The most typical periodicity is 7/2, which induces the left-handed twist of a bundle, while others such as 11/3 or 15/4 are considered non-canonical. **(B)** Helical wheel diagrams of antiparallel coiled coils with heptad positions labeled *a*-*g*. Letters in the center of the helices indicate whether they are viewed from the N- or C-terminal side. The left diagram corresponds to the canonical *a*-*d* geometry, while the right diagram corresponds to the extended *a*-*d*-*e* geometry caused by a counterclockwise axial rotation of all helices by about 26 degrees, viewed from the N-terminus. **(C)** Experimental structure and corresponding AlphaFold2 model of the tropomyosin fragment. The results of the measurement of two structural parameters, helix axial rotation and periodicity, are shown below in scale. These parameters are defined separately for each individual layer of an input structure (layer is defined as a set of *n* residues, where *n* corresponds to the number of helices in the bundle, roughly localized on a plane perpendicular to the bundle axis).

Figure 2. Accuracy of AlphaFold2, LOGICOIL and CoCoNat in predicting the oligomeric state of coiled coils. LOGICOIL and CoCoNat were run in two modes, with and without ("_noregister" suffix) providing structurally derived heptad registers. AlphaFold2 predictions were made using the model quality metrics pLDDT, pTM and PAE, the latter divided into total_PAE and inter_PAE (see text for details). The accuracy of each method is shown as a confusion matrix with true and predicted labels on the y and x axes, respectively. Weighted F1 coefficients are also provided.

Figure 3. Application of AlphaFold2 representations for the prediction of the oligomeric state. **(A)** 2D plots of raw AlphaFold2 representations of the benchmark cases (left panel) and representations obtained from a downstream model (right panel). Colors represent oligomerization states. **(B)** A confusion matrix with true and predicted labels on the y and x axes, respectively, showing the performance of the final prediction model trained on AlphaFold2 representations. The weighted F1 score is also shown.

Figure 4. Comparison of coiled-coil parameters between experimental (PDB) and AlphaFold2 (model) structures. The scatter plots show the differences in average helix rotation and periodicity, while the accompanying plots (right) show detailed per-layer measurements for exemplary structures (model and experimental structures are shown in red and light blue, respectively). Panels A, B and C show the results for the "parallel", "antiparallel", and "automatic" benchmark sets, respectively.

Figure 5. Examples of successes and failures in modeling coiled-coil structures with AlphaFold2. See the text for details.

References

Akdel,M. *et al.* (2022) A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.*, **29**, 1056–1067.

del Alamo,D. *et al.* (2022) Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife*, **11**, e75751.

Crick,F.H.C. (1953) The Fourier transform of a coiled-coil. *Acta Crystallogr.*, **6**, 685–689.

Das,R. *et al.* (2009) Simultaneous prediction of protein folding and docking at high resolution. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 18978–83.

Delorenzi,M. and Speed,T. (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics*, **18**, 617–25.

Evans,R. *et al.* (2022) Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.10.04.463034.

Ferris,H.U. *et al.* (2012) Mechanism of regulation of receptor histidine kinases. *Structure*, **20**, 56–66.

Ferris,H.U. *et al.* (2011) The mechanisms of HAMP-mediated signaling in transmembrane receptors. *Structure*, **19**, 378–85.

Gazizov,A. *et al.* (2023) AF2BIND: Predicting ligand-binding sites using the pair representation of AlphaFold2. *bioRxiv*, 2023.10.15.562410.

Grigoryan,G. and Degrado,W.F. (2011) Probing designability via a generalized model of helical bundle geometry. *J. Mol. Biol.*, **405**, 1079–100.

Guzenko,D. and Strelkov,S. V (2018) CCFold: rapid and accurate prediction of coiled-coil structures and application to modelling intermediate filaments. *Bioinformatics*, **34**, 215–222.

Harbury,P.B. *et al.* (1993) A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science*, **262**, 1401–7.

Jumper,J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

Kingma,D.P. and Ba,J. (2017) Adam: A Method for Stochastic Optimization.

Ludwiczak,J. *et al.* (2019) DeepCoil-a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics*, **35**, 2790–2795.

Ludwiczak,J. *et al.* (2022) Localpdb- a Python package to manage protein structures and their annotations. *Bioinformatics*.

Lupas,A. *et al.* (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–4.

Lupas,A. *et al.* (2017) The Structure and Topology of α-Helical Coiled Coils. In, *Sub-Cellular Biochemistry*., pp. 95–129.

Lupas,A.N. *et al.* (2021) The breakthrough in protein structure prediction. *Biochem. J.*, **478**, 1885–1890.

Madeo,G. *et al.* (2023) CoCoNat: a novel method based on deep learning for coiled-coil prediction. *Bioinformatics*, **39**.

Martinez-Goikoetxea,M. and Lupas,A.N. (2023) New protein families with hendecad coiled coils in the proteome of life. *J. Struct. Biol.*, **215**, 108007.

McDonald,E.F. *et al.* (2023) Benchmarking AlphaFold2 on peptide structure prediction. *Structure*, **31**, 111-119.e2.

Mirdita,M. *et al.* (2022) ColabFold: making protein folding accessible to all. *Nat. Methods*, **19**, 679–682.

Mistry,J. *et al.* (2013) Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.*, **41**, e121.

Morton,C.R. *et al.* (2021) Structural basis for the coiled-coil architecture of human CtIP. *Open Biol.*, **11**, 210060.

Murray,D.H. *et al.* (2016) An endosomal tether undergoes an entropic collapse to bring vesicles together. *Nature*, **537**, 107–111.

Offer,G. *et al.* (2002) Generalized Crick equations for modeling noncanonical coiled coils. *J. Struct. Biol.*, **137**, 41–53.

Schweke,H. *et al.* (2024) An atlas of protein homo-oligomerization across domains of life. *Cell*.

Shor,B. and Schneidman-Duhovny,D. (2024) CombFold: predicting structures of large protein assemblies using a combinatorial assembly algorithm and AlphaFold2. *Nat. Methods*.

Sledzieski,S. *et al.* (2023) Democratizing Protein Language Models with Parameter-Efficient Fine-Tuning. *bioRxiv*, 2023.11.09.566187.

Stein,R.A. and Mchaourab,H.S. (2022) SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *PLOS Comput. Biol.*, **18**, e1010483.

Strelkov,S. V and Burkhard,P. (2003) Analysis of alpha-helical coiled coils with the program TWISTER reveals a structural mechanism for stutter compensation. *J. Struct. Biol.*, **137**, 54–64.

Szczepaniak,K. *et al.* (2021) A library of coiled-coil domains: from regular bundles to peculiar

twists. *Bioinformatics*, **36**, 5368–5376.

Szczepaniak,K. *et al.* (2014) Designability landscape reveals sequence features that define axial helix rotation in four-helical homo-oligomeric antiparallel coiled-coil structures. *J. Struct. Biol.*, **188**, 123–133.

Szczepaniak,K. *et al.* (2018) Variability of the core geometry in parallel coiled-coil bundles. *J. Struct. Biol.*, **204**, 117–124.

Vincent,T.L. *et al.* (2013) LOGICOIL--multi-state prediction of coiled-coil oligomeric state. *Bioinformatics*, **29**, 69–76.

Walshaw,J. and Woolfson,D.N. (2001) Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *J. Mol. Biol.*, **307**, 1427–50.

Wang,Y. *et al.* (2021) Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization. *J. Mach. Learn. Res.*, **22**, 1–73.

Wayment-Steele,H.K. *et al.* (2022) Prediction of multiple conformational states by combining sequence clustering with AlphaFold2. *bioRxiv*.

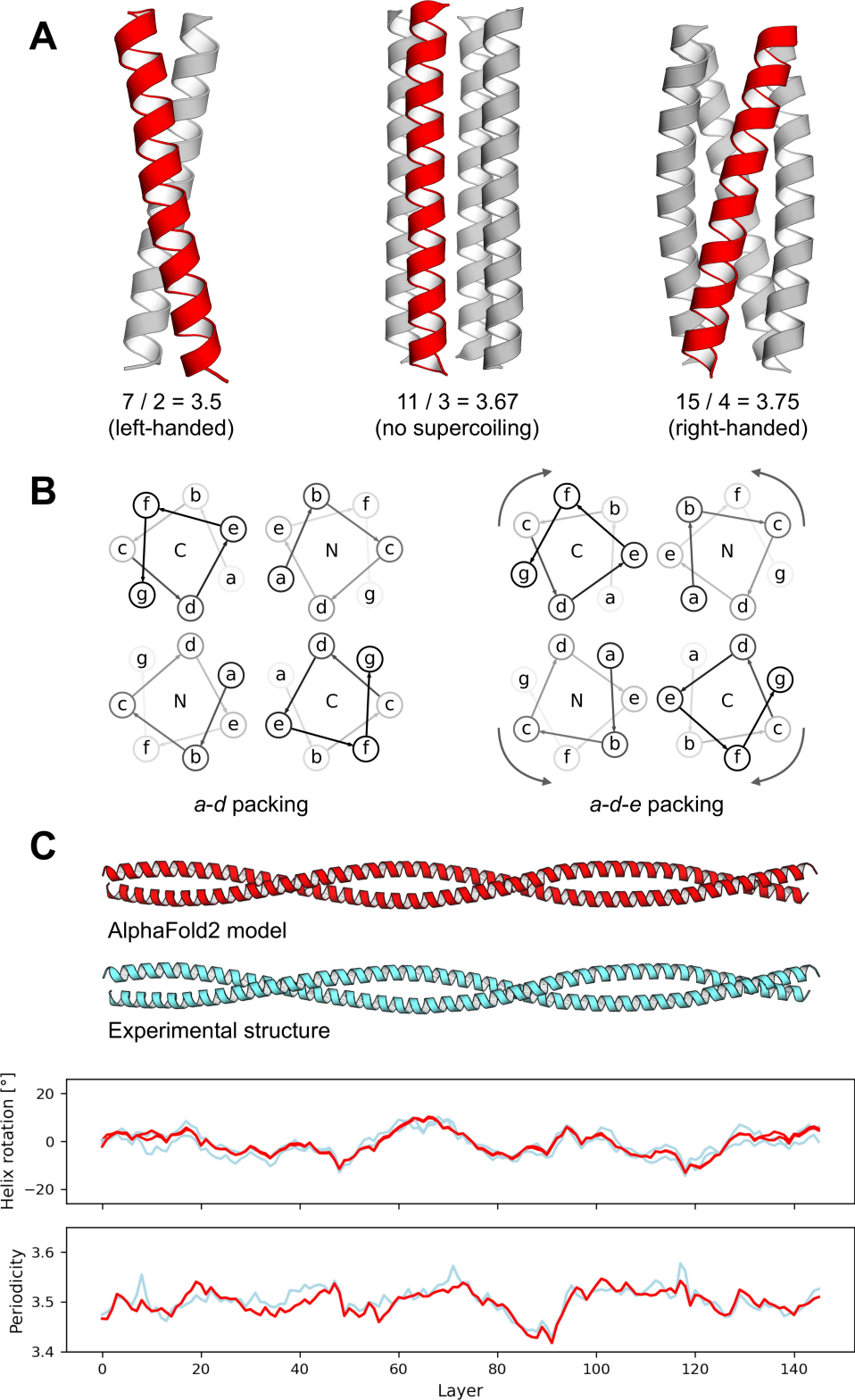
Winski,A. *et al.* (2024) AlphaFold2 captures the conformational landscape of the HAMP signaling domain. *Protein Sci.*, **33**, e4846.

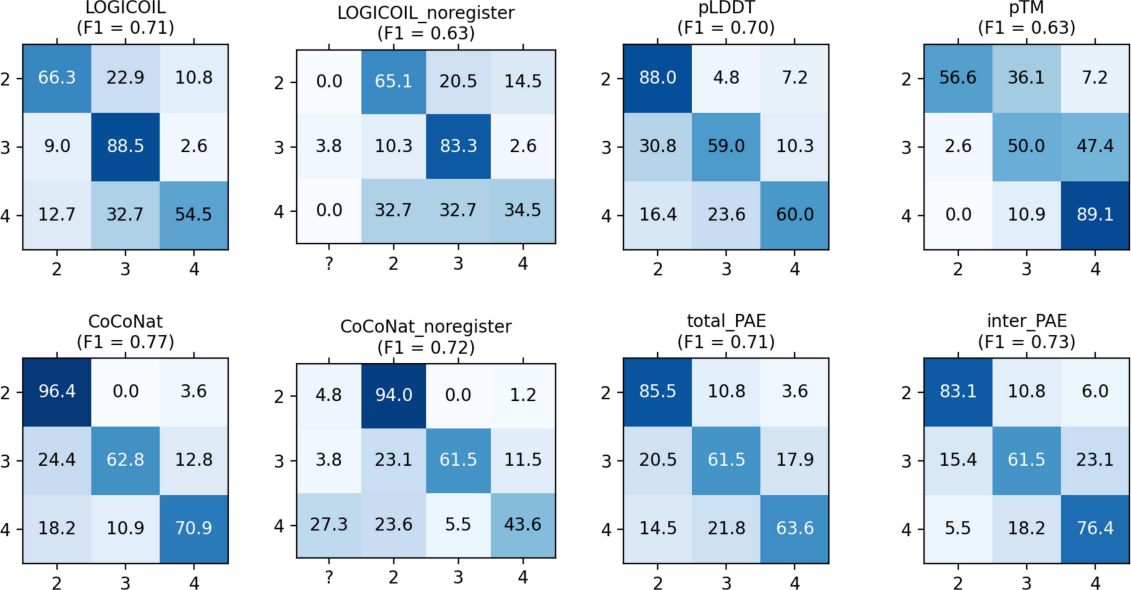
Wood,C.W. *et al.* (2017) ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design. *Bioinformatics*.

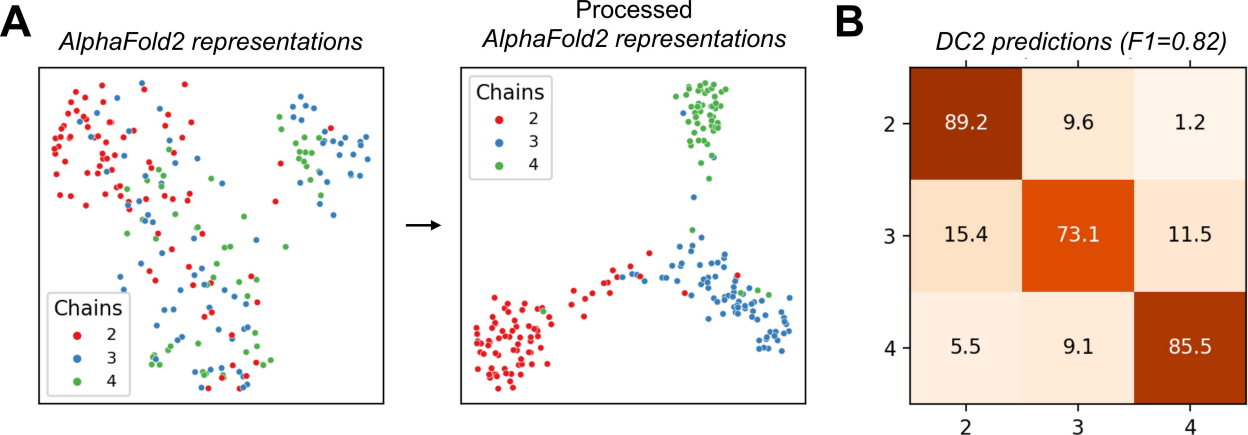
Wood,C.W. and Woolfson,D.N. (2018) CCBuilder 2.0: Powerful and accessible coiled-coil modeling. *Protein Sci.*, **27**.

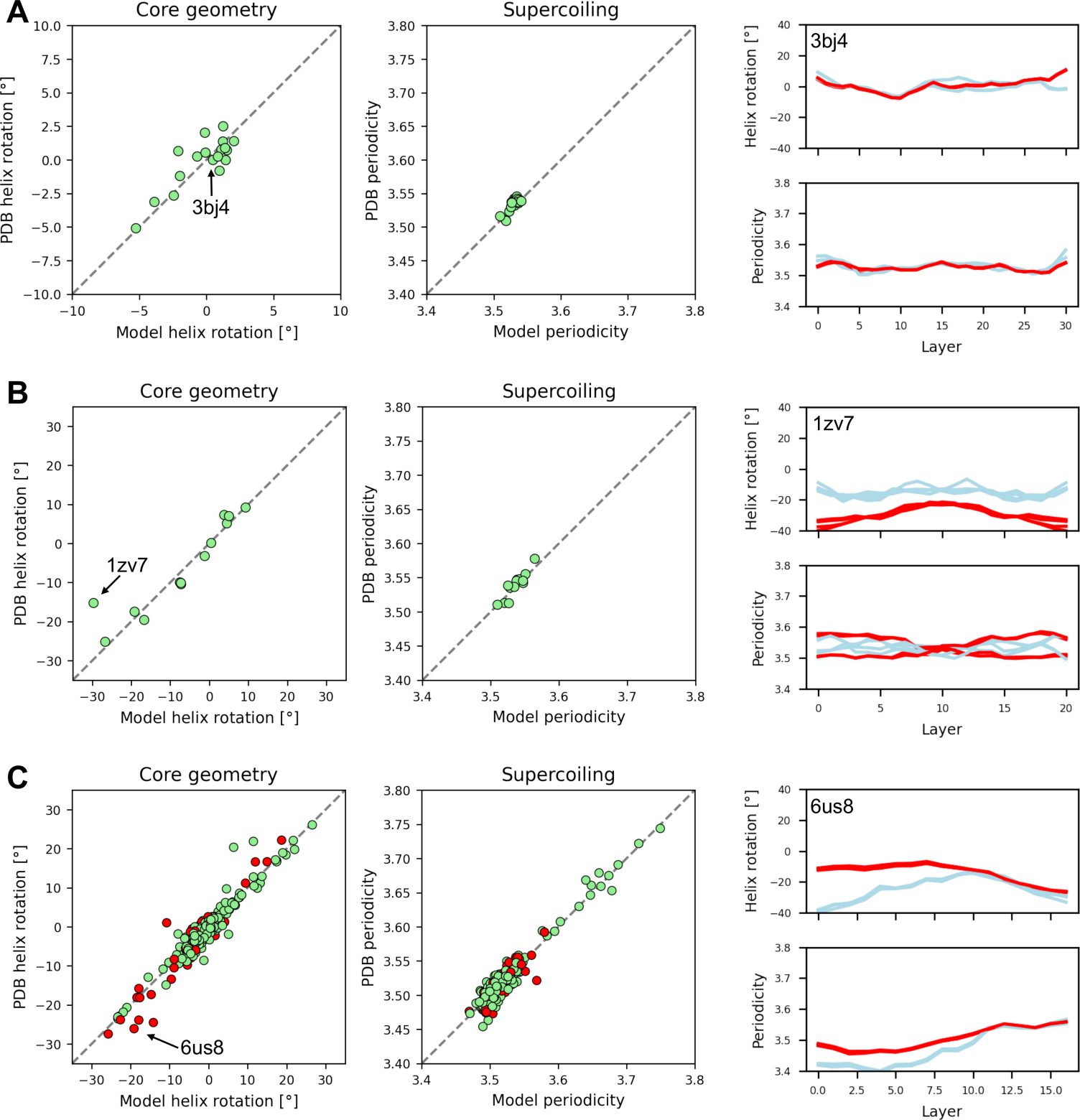
Yang,H. *et al.* (2010) Structural insight into the mechanisms of enveloped virus tethering by tetherin. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 18428–32.

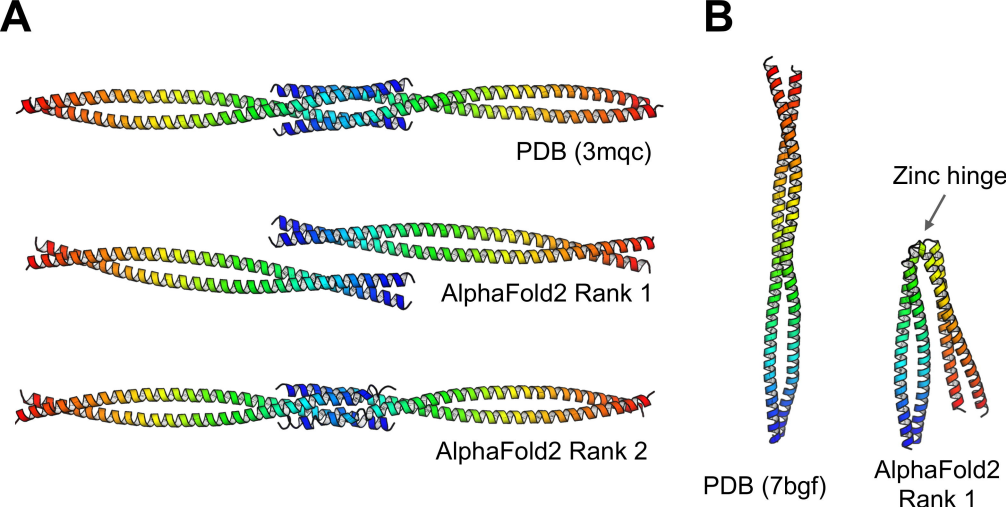
Zhang,C. *et al.* (2022) US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods*, **19**, 1109–1115.


A
7 / 2 = 3.5
(left-handed)
11 / 3 = 3.67
(no supercoiling)
15 / 4 = 3.75
(right-handed)
B
a-d packing
a-d-e packing
C
AlphaFold2 model
Experimental structure
Helix rotation [°]
20
0
−20
Periodicity
3.6
3.5
3.4
0
20
40
60
80
100
120
140
Layer


LOGICOIL
(F1 = 0.71)
2 66.3 22.9 10.8
3 9.0 88.5 2.6
4 12.7 32.7 54.5
2 3 4
LOGICOIL_noregister
(F1 = 0.63)
2 0.0 65.1 20.5 14.5
3 3.8 10.3 83.3 2.6
4 0.0 32.7 32.7 34.5
? 2 3 4
pLDDT
(F1 = 0.70)
2 88.0 4.8 7.2
3 30.8 59.0 10.3
4 16.4 23.6 60.0
2 3 4
pTM
(F1 = 0.63)
2 56.6 36.1 7.2
3 2.6 50.0 47.4
4 0.0 10.9 89.1
2 3 4
CoCoNat
(F1 = 0.77)
2 96.4 0.0 3.6
3 24.4 62.8 12.8
4 18.2 10.9 70.9
2 3 4
CoCoNat_noregister
(F1 = 0.72)
2 4.8 94.0 0.0 1.2
3 3.8 23.1 61.5 11.5
4 27.3 23.6 5.5 43.6
? 2 3 4
total_PAE
(F1 = 0.71)
2 85.5 10.8 3.6
3 20.5 61.5 17.9
4 14.5 21.8 63.6
2 3 4
inter_PAE
(F1 = 0.73)
2 83.1 10.8 6.0
3 15.4 61.5 23.1
4 5.5 18.2 76.4
2 3 4


A
AlphaFold2 representations
Chains
2
3
4
Processed
AlphaFold2 representations
Chains
2
3
4
B
DC2 predictions (F1=0.82)
2
3
4
89.2
9.6
1.2
15.4
73.1
11.5
5.5
9.1
85.5
2
3
4


A
Core geometry
Supercoiling
3bj4
PDB helix rotation [°]
Model helix rotation [°]
PDB periodicity
Model periodicity
Helix rotation [°]
Periodicity
Layer
B
Core geometry
Supercoiling
1zv7
PDB helix rotation [°]
Model helix rotation [°]
PDB periodicity
Model periodicity
Helix rotation [°]
Periodicity
Layer
C
Core geometry
Supercoiling
6us8
PDB helix rotation [°]
Model helix rotation [°]
PDB periodicity
Model periodicity
Helix rotation [°]
Periodicity
Layer


A
PDB (3mqc)
AlphaFold2 Rank 1
AlphaFold2 Rank 2
B
Zinc hinge
PDB (7bgf)
AlphaFold2
Rank 1

CCfrag: Scanning folding potential of coiled-coil fragments with AlphaFold

Mikel Martinez-Goikoetxea

Department of Protein Evolution, Max Planck Institute for Biology, 72076 Tübingen, Germany

Structured abstract

Motivation: Coiled coils are a widespread structural motif consisting of multiple α-helices that wind around a central axis to bury their hydrophobic core. Although their backbone can be uniquely described by the Crick parametric equations, these have little practical application in structural prediction, given that most coiled coils in nature feature non-canonical repeats that locally distort their geometry. While AlphaFold has emerged as an effective coiled-coil modeling tool, capable of accurately predicting changes in periodicity and core geometry along coiled-coil stalks, it is not without limitations. These include the generation of spuriously bent models and the inability to effectively model globally non-canonical coiled coils. In an effort to overcome these limitations, we investigated whether dividing full-length sequences into fragments would result in better models.

Results: We developed CCfrag to leverage AlphaFold for the piece-wise modeling of coiled coils. The user can create a specification, defined by window size, length of overlap, and oligomerization state, and the program produces the files necessary to run structural predictions with AlphaFold. Then, the structural models and their scores are integrated into a rich per-residue representation defined by sequence- or structure-based features, which can be visualized or employed for further analysis. Our results suggest that removing coiled-coil sequences from their native context can in some case improve the prediction confidence and avoids bent models with spurious contacts. In this paper, we present various use cases of CCfrag, and propose that fragment-based prediction is useful for understanding the properties of long, fibrous coiled coils, by showing local features not seen in full-length models.

Availability and Implementation: The program is implemented as a Python module. The code and its documentation are available at https://github.com/Mikel-MG/CCfrag.

Contact: mikel.martinez@tuebingen.mpg.de

Introduction

Coiled coils consist of multiple α-helices that wind around a central axis to bury their hydrophobic core. They are widespread in proteomes, where they can be found in a variety of forms, depending on the number and orientation of their constituent helices (Lupas and Bassler 2017). This topological diversity is underpinned by the seven-residue heptad repeat, labeled a-g, where the *a* and *d* positions form the core. The geometry of coiled-coil interaction, known as knobs-into-holes, involves the core residue of a helix (knob) packing into a cavity formed by four residues (hole) of an adjacent helix (Crick 1953a). They are the best understood protein fold, as indicated by the number of programs that are able to detect coiled-coil forming propensity from sequence (Lupas, Bassler and Dunin-Horkawicz 2017), and by the existence of the Crick parametric equations that describe their backbone (Crick 1953b). In spite of this, coiled-coil structural prediction has remained a substantial challenge, due to the fact that, although preponderantly repetitive in sequence and structure, coiled-coil domains are rarely without occasional interruptions in the form of non-heptad repeats, which locally alter their packing interactions and geometry.

Even as a general-purpose protein structure model, AlphaFold (Evans *et al.* 2021; Jumper *et al.* 2021) has demonstrated an outstanding ability to accurately model coiled-coil domains, particularly with respect to their supercoiling and core geometry (Madaj *et al.* 2024). Additionally, it has been shown that it can be used to inform of dynamic protein conformations (Wayment-Steele *et al.* 2024; Winski *et al.* 2024). Despite its merits, AlphaFold cannot robustly model some long coiled coils, for which it often outputs oddly bent models that feature spurious contacts or even atomic clashes. We have also observed that it does not generate confident models for non-canonical coiled coils, which are modeled with poor or absent side-chain packing (Martinez-Goikoetxea and Lupas 2023).

Motivated by these observations, we wondered whether computing AlphaFold models of overlapping windows along a sequence would yield better quality models, by simplifying the prediction task. Thus, we developed CCfrag, a pipeline to automate the division of a sequence into fragments and the subsequent integration of the corresponding AlphaFold models into a rich per-residue representation. Our results suggest that not only does this improve the modeling quality of challenging coiled coils, but additionally, it can be used to scan sequences for local structural properties not observed in full-length coiled-coil models. We anticipate that this framework will be of particular interest in the context of understanding long fibrous coiled coils, such as myosins and kinesins.

Implementation and features

CCfrag is implemented as a Python module that contains two main classes, the divider and the integrator. The former is used to divide a sequence according to a user-defined *specification*, and its output consists of the FASTA files necessary to run the AlphaFold predictions (Fig. 1). After running AlphaFold predictions, the integrator module extracts a number of features from the models, and incorporates them into a rich per-residue representation that can be displayed graphically or used for further analysis. CCfrag also provides limited support for ESMfold (Lin *et al.* 2022), increasing speed at the expense of prediction accuracy.

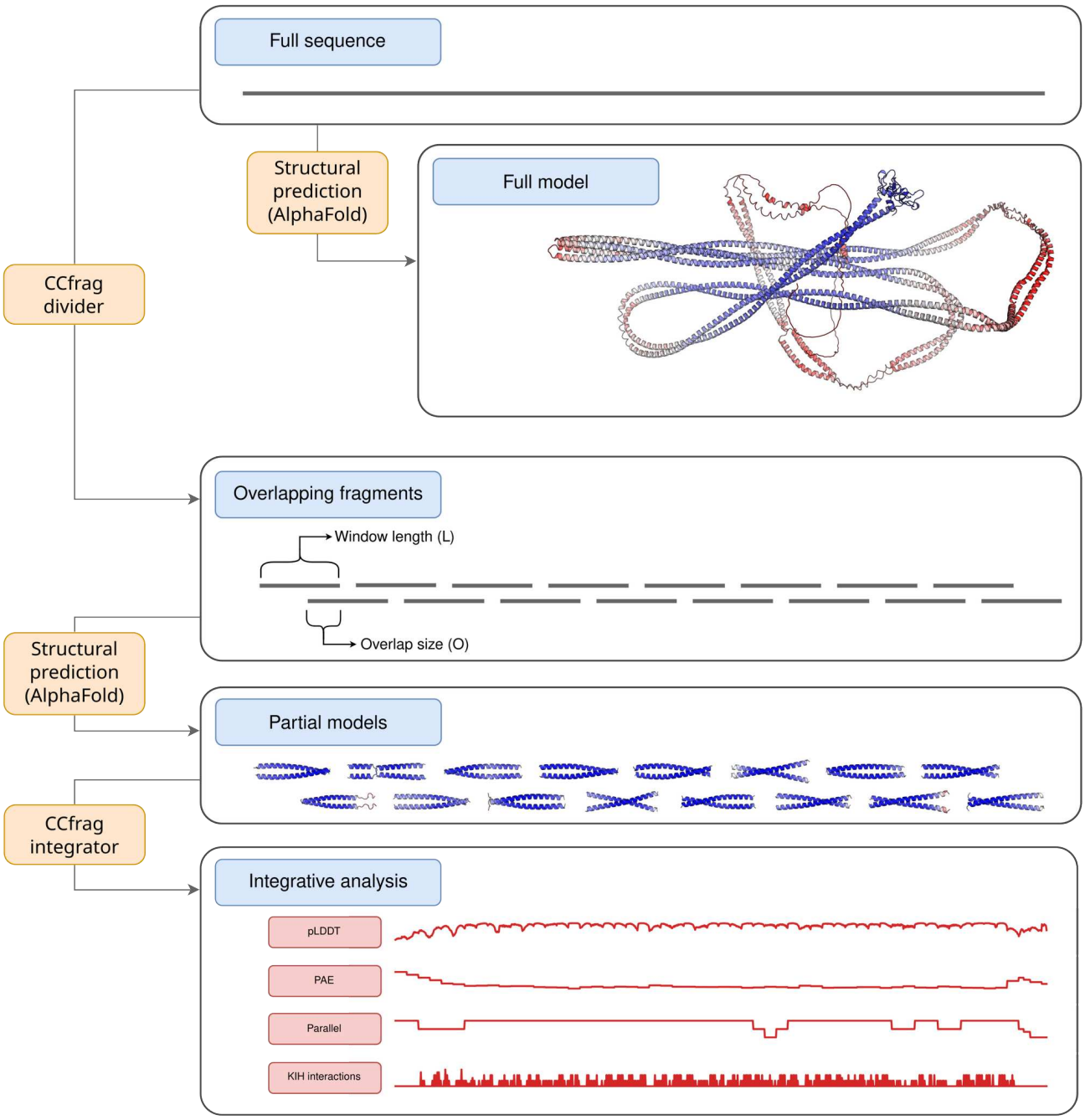


Figure 1. Schematic representation of the CCfrag pipeline. Modeling long coiled-coil domains with AlphaFold generally yields suboptimal models; on top, a full-length model of *H. sapiens* EEA1 is shown, colored by pLDDT (red-worst to blue-best). By dividing the full-length sequence into fragments, the resulting models are predicted with higher confidence, and can be analyzed for local properties not seen in the full-length model (bottom).

The divider module accepts various parameters, the most important of which are the window size (*L*), length of overlap between contiguous fragments (*O*), and oligomeric state (*N*). These define a *specification*, noted in the framework of the program as *N_L_O* (for example, 2_30_15 would be 30-residue windows with 15-residue overlap, modeled as a dimer). The minimum overlap size is 0 (no overlap), and the maximum is the length of the sequence minus one. If during the windowing of the

input sequence, the C-terminal fragment is not long enough to produce a full window-sized fragment, the window size parameter (*L*) is given priority, and the overlap is increased for that last fragment. An additional feature of CCfrag is the possibility of setting a flanking sequence (*flank*), which will be attached N- and C- terminally to each window for the AlphaFold modeling, but will be removed in the assembly step. The addition of a flanking sequence does not contribute in itself to the scores during assembly (see below), but it can be used to promote the folding of the fragments; this feature is inspired by experimental techniques used to study coiled coils, whereby the addition of GCN4 adaptors is routinely used to promote folding and subsequent crystalization (Hernandez Alvarez *et al.* 2008). The output of the divider module consists of a parameter file (*parameters.json*), a table of constructs (*constructs.csv*), and a folder (*queries*) which contains the input FASTA files for AlphaFold.

CCfrag does not implement a wrapper to run AlphaFold. This is a necessary limitation given the variety of ways in which users can run AlphaFold (e.g. local machine, high performance cluster), but also a way to decouple the concept of fragment-based modeling from the prediction software itself. Thus, the program can be easily updated to work with newer and potentially better sequence-to-structure prediction programs.

After the AlphaFold models are generated, the integrator module reads most of its required parameters from the configuration file (*parameters.json*) and the list of constructs (*constructs.cs*v) generated by the divider module. Additionally, the user can specify a list of features that will be extracted or computed from each window. CCfrag includes functions to extract pLDDT and PAE (the average), as well as a function to compute whether the models are parallel or antiparallel, assuming they feature an extended helical conformation. It also includes a wrapper to run SOCKET (Kumar and Woolfson 2021) to detect knobs-into-holes interactions, the hallmark of coiled-coil structures. The addition of arbitrary features (e.g., solvent-accessible surface area) is possible, but requires defining a new function within the source code (an example is provided in the documentation). The output of the integrator module is a table that stores per-residue values for each combination of feature and specification. During the integration process, the features of overlapping windows are flattened via averaging, although this can be customized. This means that, for the case of the parallel/antiparallel feature (numerically encoded as 1 and 0 respectively), some positions of the sequence may have a value of 0.5, meaning that half of the overlapping models showed a parallel arrangement, and the other half an antiparallel one; as illustrated in the next section, this can be interpreted as the lack of topological encoding in the local sequence.

Case studies

In this section, three examples of the use of CCfrag are presented. The code to generate and visualize these examples is included in the GitHub repository in the form of Jupyter notebooks.

Scanning long coiled coils for folding potential

EEA1 (Early Endosome Antigen 1) is a protein that features an N-terminal zinc finger domain, a long parallel dimeric coiled coil, and a C-terminal FYVE domain. It has been extensively studied for its involvement in endosomal trafficking, where its coiled-coil domain is thought to switch between extended and flexible states (Murray *et al.* 2016). Using CCfrag to model EEA1 fragment-wise shows that the pLDDT scores are significantly better than those of the full-length prediction (Fig. 1), with the largest window (L=70) showing the best scores (Fig. S1). On the other hand, it can be observed that shorter windows generally do not encode for a parallel orientation, except for some segments that seem to *pull* from their neighboring residues in larger window sizes; for example, the 400th residue is found in an antiparallel orientation when modeled within a 20-residue context, but the neighboring segments

promote the adoption of a parallel arrangement. As expected, there is significant overlap between the SOCKET knobs-into-holes (KIH) detection and the sequence-based coiled-coil predictors DeepCoil2 (Ludwiczak *et al.* 2019) and COILS (Lupas, Van Dyke and Stock 1991). Notably, the absence of KIH interactions also coincides with weak coiled-coil predictions in one or the other program, suggesting that these correspond to flexible regions or segments with ambiguous topological specificity.

Scanning for non-canonical coiled coils

In recent work we described a number of protein families that predominantly featured hendecad coiled-coil repeats (Martinez-Goikoetxea and Lupas 2023), and pointed out that these are a particularly difficult targets for AlphaFold. These non-canonical coiled coils are underrepresented in natural proteomes, and possibly as a result, are poorly detected by sequence-based coiled-coil predictors. Using CCfrag to model a member of the MACH family illustrates how piece-wise modeling can be applied to essentially *scan* sequences for potential KIH interactions, thus detecting coiled coils that even coiled-coil specific methods fail to predict (Fig. S2).

Multi-state modeling

It has been shown that AlphaFold prediction quality metrics such as pLDDT and PAE can inform of the likely oligomeric states of protein complexes (Madaj *et al.* 2024; Schweke *et al.* 2024). In the context of fragment-based modeling of coiled coils, oligomeric state prediction is particularly challenging due to the fact that coiled-coil sequences can often assemble into various nearly-isoenergetic oligomeric states (Harbury *et al.* 1993). A consequence of this is that when coiled-coil fragments are taken out of their native context, they often crystallize in oligomeric states that do not correspond to the native stoichiometry of the full-length protein. Nevertheless, these fragments inform of the topological specifity that is encoded locally. A great example of this can be found in the spike protein of SARS Coronavirus, whose coiled-coil segments have been crystallized as trimers and tetramers (Deng *et al.* 2006), even though the full-length protein is known to assemble into a trimer. Using CCfrag, we modeled this protein as dimers, trimers, and tetramers, and observed that some of the predicted models matched experimentally-validated coiled-coil domains present in the spike protein (Fig S3). We also observed additional fragments that, even though they do not form coiled coils, were nevertheless predicted as coiled-coil structures. This suggests these fragments feature cryptic coiled-coil folding potentials, and that could fold as such if removed from their native context.

Conclusions

CCfrag extends the functionality of AlphaFold by facilitating the process of dividing a protein sequence into smaller overlapping fragments, running AlphaFold predictions on these, and integrating the resulting models into a rich representation. We find that this piece-wise modeling can improve the robustness of the predicted models in the case of long coiled coils, even non-canonical ones. Additionally, we propose that this representation can reveal local features not seen in full-sequence models, such as oligomerization preference or folding propensity. CCfrag, together with its documentation and additional examples, is available at https://github.com/Mikel-MG/CCfrag.

Acknowledgements

We would like to thank Dr Andrei Lupas, Dr Stanislaw Dunin-Horkawicz, Dr Pedro Escudeiro, and Adrian Dobbelstein for useful discussions and comments that contributed greatly to improve this manuscript.

References

Crick. The Packing of a-Helices: Simple Coiled-Coils. *Acta Cryst* 1953a;**6**:689–97.

Crick. The Fourier transform of a coiled-coil. *Acta Cryst* 1953b;**6**:685–9.

Deng Y, Liu J, Zheng Q *et al.* Structures and Polymorphic Interactions of Two Heptad-Repeat Regions of the SARS Virus S2 Protein. *Structure (London, England : 1993)* 2006;**14**:889.

Evans R, O'Neill M, Pritzel A *et al. Protein Complex Prediction with AlphaFold-Multimer*. Bioinformatics, 2021.

Harbury PB, Zhang T, Kim PS *et al.* A Switch Between Two-, Three-, and Four-stranded Coiled Coils in GCN4 Leucine Zipper Mutants. *Science* 1993;**262**:1401–7.

Hernandez Alvarez B, Hartmann MD, Albrecht R *et al.* A new expression system for protein crystallization using trimeric coiled-coil adaptors. *Protein Engineering, Design and Selection* 2008;**21**:11–8.

Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.

Kumar P, Woolfson DN. Socket2: a program for locating, visualizing and analyzing coiled-coil interfaces in protein structures. *Bioinformatics* 2021;**37**:4575–7.

Lin Z, Akin H, Rao R *et al. Evolutionary-Scale Prediction of Atomic Level Protein Structure with a Language Model*. Synthetic Biology, 2022.

Ludwiczak J, Winski A, Szczepaniak K *et al.* DeepCoil—a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics* 2019;**35**:2790–5.

Lupas A, Van Dyke M, Stock J. Predicting Coiled Coils from Protein Sequences. *Science* 1991;**252**:1162–4.

Lupas AN, Bassler J. Coiled Coils – A Model System for the 21st Century. *Trends in Biochemical Sciences* 2017;**42**:130–40.

Lupas AN, Bassler J, Dunin-Horkawicz S. The Structure and Topology of α-Helical Coiled Coils. *Fibrous Proteins: Structures and Mechanisms* 2017;**82**:95–129.

Madaj R, Martinez-Goikoetxea M, Kaminski K *et al.* Applicability of AlphaFold2 in the modelling of coiled-coil domains. 2024:2024.03.07.583852.

Martinez-Goikoetxea M, Lupas AN. New protein families with hendecad coiled coils in the proteome of life. *Journal of Structural Biology* 2023;**215**:108007.

Murray DH, Jahnel M, Lauer J *et al.* An endosomal tether undergoes an entropic collapse to bring vesicles together. *Nature* 2016;**537**:107–11.

Schweke H, Pacesa M, Levin T *et al.* An atlas of protein homo-oligomerization across domains of life. *Cell* 2024;**187**:999-1010.e15.

Wayment-Steele HK, Ojoawo A, Otten R *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* 2024;**625**:832–9.

Winski A, Ludwiczak J, Orlowska M *et al.* AlphaFold2 captures the conformational landscape of the HAMP signaling domain. *Protein Science* 2024;**33**:e4846.

CCfrag: Scanning folding potential of coiled-coil fragments with AlphaFold

Supplementary Material

Mikel Martinez-Goikoetxea

Department of Protein Evolution, Max Planck Institute for Biology, 72076 Tübingen, Germany

E-mail: mikel.martinez@tuebingen.mpg.de

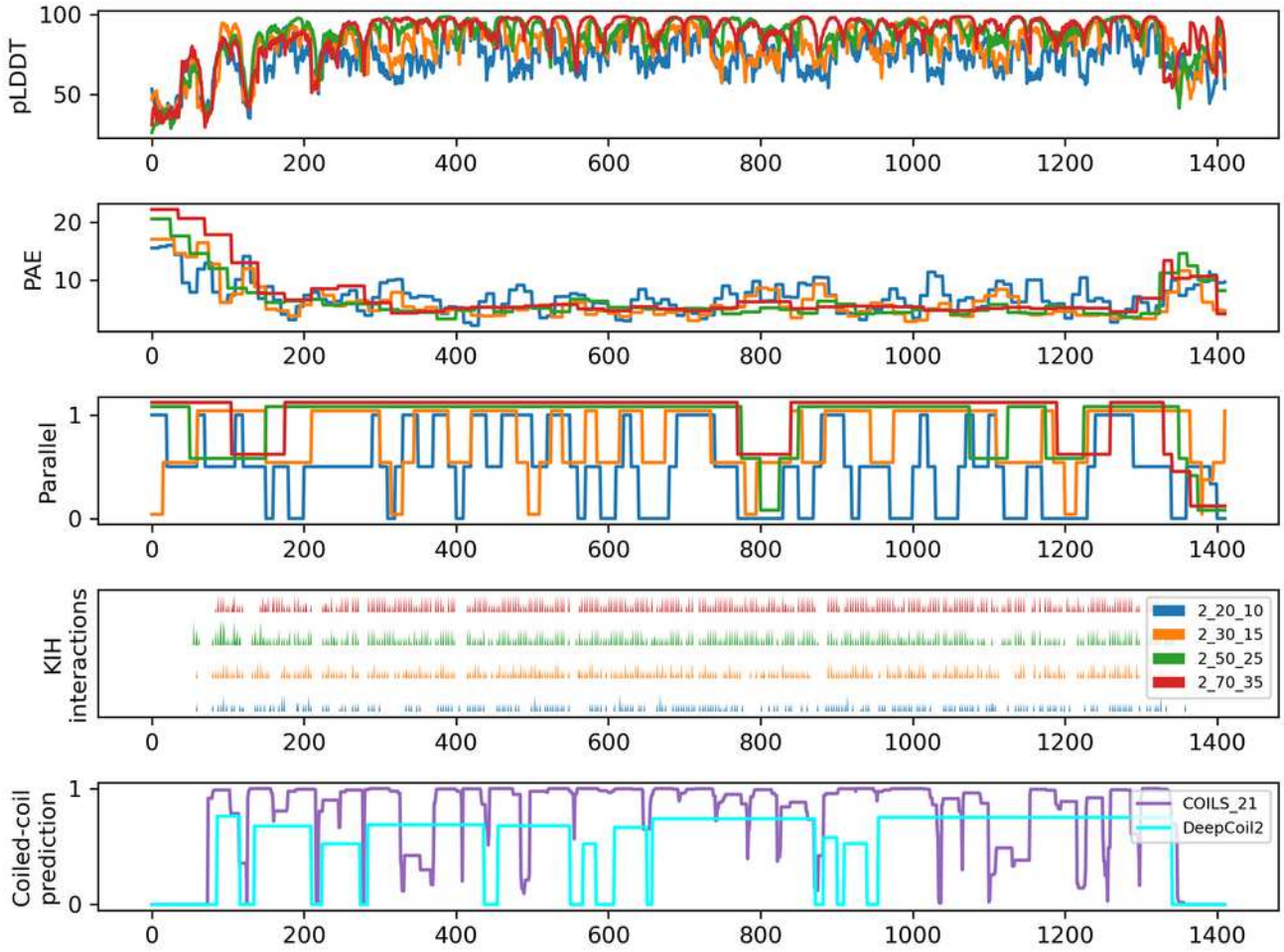


Figure S1. Graphical summary of the CCfrag representation of EEA1 of *H. Sapiens*. The protein is modeled as a dimer, in four different specifications with window sizes of 20, 30, 50, and 70 residues and an overlap of half the corresponding window size. In the top 4 panels, plots for various features for each specification (color-coded as in the legend in the right of the fourth panel) are shown (pLDDT, PAE, Parallel, KIH interactions), averaged for each residue (since the overlap is half the window size, each residue is covered by two models, resulting in two values for each feature). In the bottom panel, coiled-coil prediction probabilities for COILS (window size = 21) and DeepCoil 2 are shown.

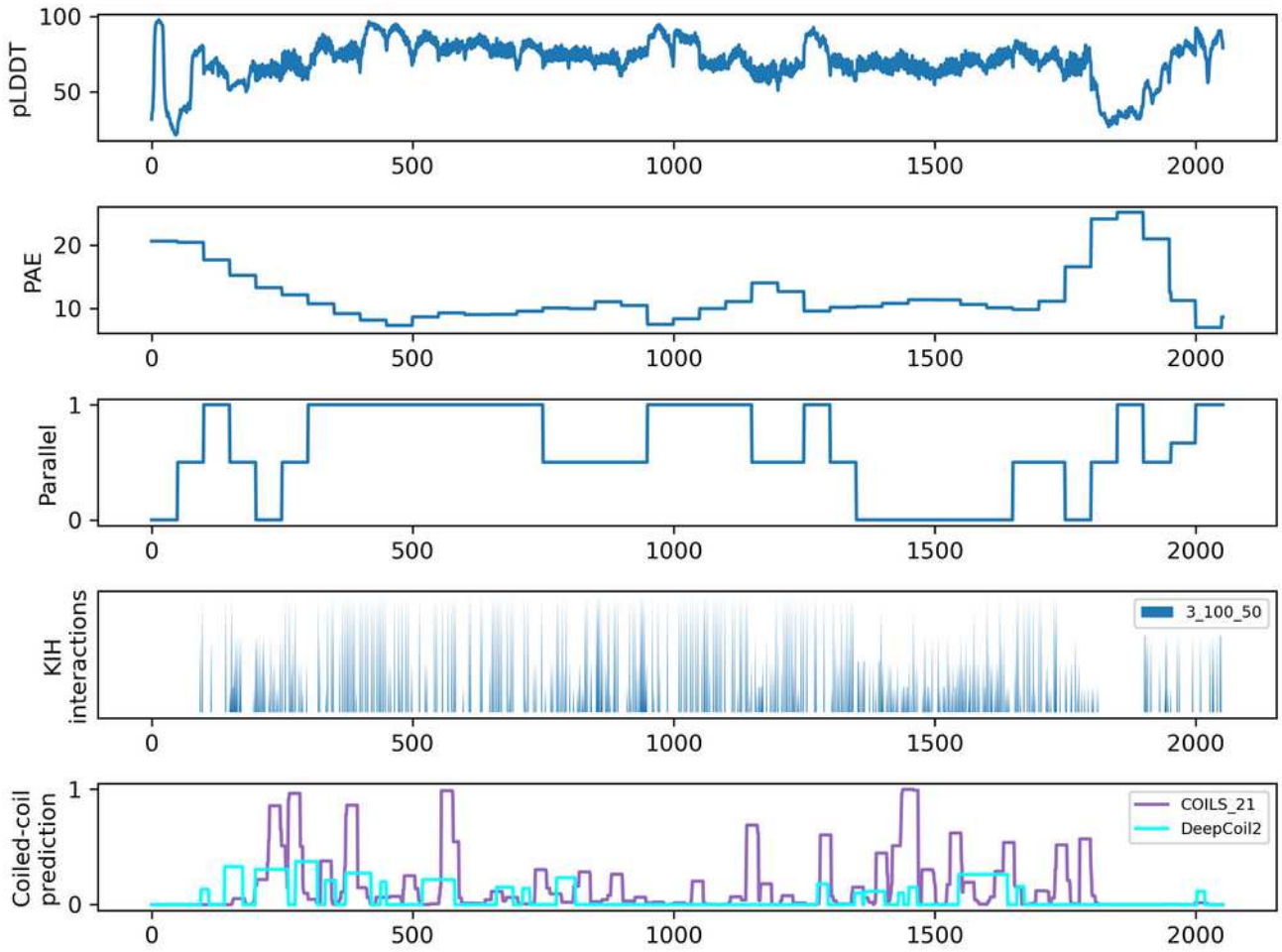


Figure S2. Graphical summary of the CCfrag representation of WP_132310275 from *Martelella mediterranea*, a member of the MACH protein family; according to interactive sequence analyses, these proteins feature an extensive hendecad coiled-coil domain. The protein is modeled as a trimer, in a specification of 100-residue windows with 50-residue overlap. Note that even though the sequence-based coiled-coil prediction is very poor, the knobs-into-holes (KIH) interactions can be detected in the structural models along subsequent models, supporting a long coiled-coil stalk.

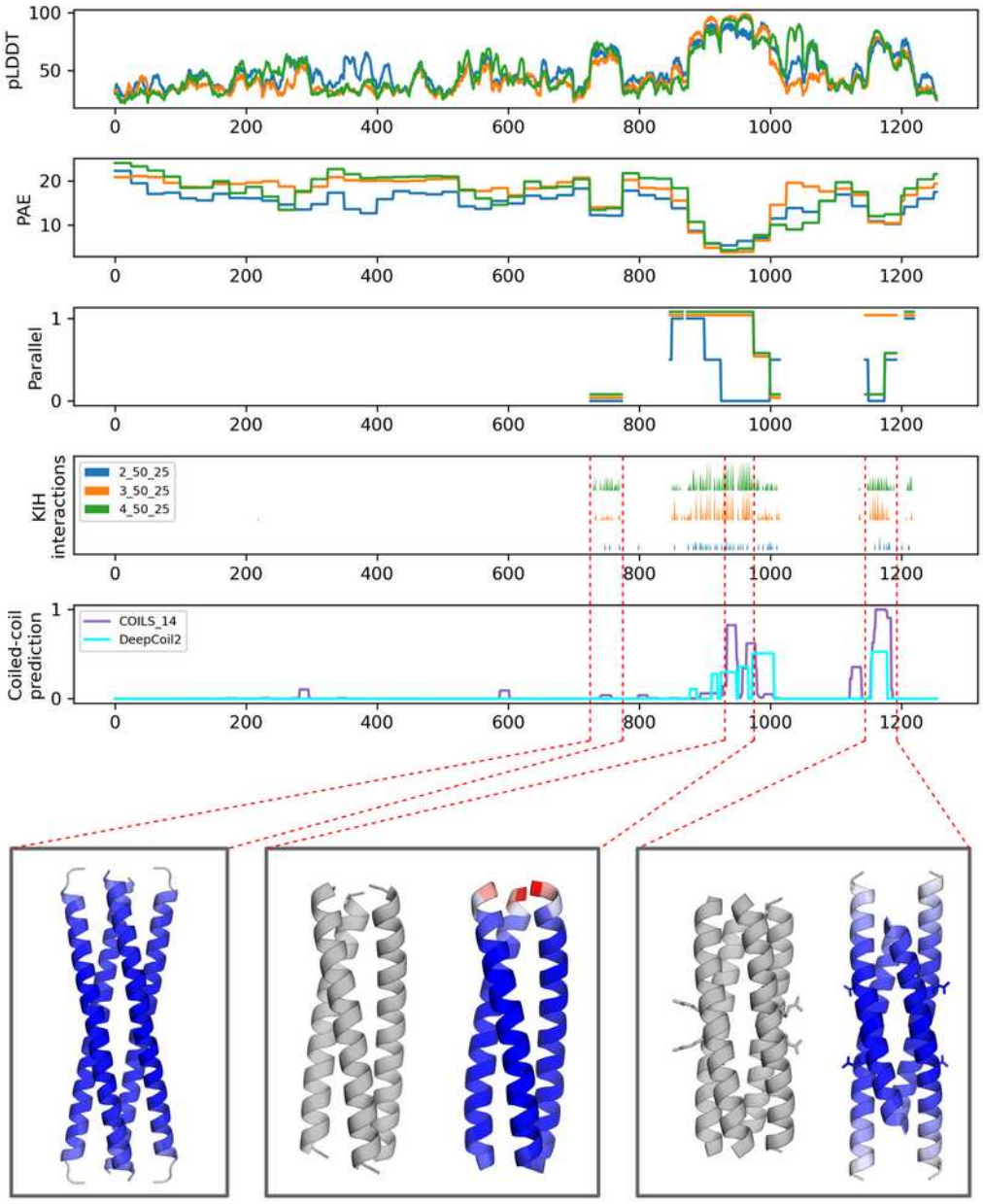


Figure S3. Graphical summary of the CCfrag representation of the spike protein of human SARS coronavirus (UniProt P59594). The spike protein is involved in the fusion between the viral and cellular membranes, and although it is not a fibrous protein, it features two functionally relevant coiled-coil domains, known as HR1 and HR2 (for heptad repeat). The protein is modeled as a dimer, a trimer, and a tetramer, each in 50-residue windows with a 25-residue overlap. There are three regions which are more confidently predicted than their surrounding context (high pLDDT, low PAE), two of which correspond to HR1 and HR2. Bottom panel: Left) Fragment 725-775 of the protein is predicted as an interlocking array of coiled coils; Center) Fragment 925-975 is predicted as a parallel trimer, nearly identical to the experimentally solved structure (1ZVB, in grey; RMSD: 0.4 Angstroms). Right) Fragment 1150-1200 is predicted as an antiparallel tetramer, although with a register different of that of the experimentally solved structure (PDB 1ZV7, in grey). The predicted structures are colored by pLDDT (red-worst to blue-best).