Bioinformatics, 36(22-23), 2020, 5368–5376
doi: 10.1093/bioinformatics/btaa1041
Advance Access Publication Date: 16 December 2020
Original Paper

OXFORD

Structural bioinformatics

A library of coiled-coil domains: from regular bundles to peculiar twists

Krzysztof Szczepaniak[1], Adriana Bukala[1], Antonio Marinho da Silva Neto[2], Jan Ludwiczak [1,3] and Stanislaw Dunin-Horkawicz [1,*]

[1]Laboratory of Structural Bioinformatics, Centre of New Technologies, University of Warsaw, 02-097 Warsaw, Poland, [2]Molecular Prospecting and Bioinformatics Group, Laboratory of Immunopathology Keizo Asami, Federal University of Pernambuco, 50670-901 Recife, Brazil and [3]Laboratory of Bioinformatics, Nencki Institute of Experimental Biology, 02-093 Warsaw, Poland

*To whom correspondence should be addressed.
Associate Editor: Arne Elofsson

Received on May 12, 2020; revised on October 30, 2020; editorial decision on December 4, 2020; accepted on December 7, 2020

Abstract

Motivation: Coiled coils are widespread protein domains involved in diverse processes ranging from providing structural rigidity to the transduction of conformational changes. They comprise two or more α-helices that are wound around each other to form a regular supercoiled bundle. Owing to this regularity, coiled-coil structures can be described with parametric equations, thus enabling the numerical representation of their properties, such as the degree and handedness of supercoiling, rotational state of the helices, and the offset between them. These descriptors are invaluable in understanding the function of coiled coils and designing new structures of this type. The existing tools for such calculations require manual preparation of input and are therefore not suitable for the high-throughput analyses.

Results: To address this problem, we developed SamCC-Turbo, a software for fully automated, per-residue measurement of coiled coils. By surveying Protein Data Bank with SamCC-Turbo, we generated a comprehensive atlas of ∼50 000 coiled-coil regions. This machine learning-ready dataset features precise measurements as well as decomposes coiled-coil structures into fragments characterized by various degrees of supercoiling. The potential applications of SamCC-Turbo are exemplified by analyses in which we reveal general structural features of coiled coils involved in functions requiring conformational plasticity. Finally, we discuss further directions in the prediction and modeling of coiled coils.

Availability and implementation: SamCC-Turbo is available as a web server (https://lbs.cent.uw.edu.pl/samcc_turbo) and as a Python library (https://github.com/labstructbioinf/samcc_turbo), whereas the results of the Protein Data Bank scan can be browsed and downloaded at https://lbs.cent.uw.edu.pl/ccdb.

Contact: s.dunin-horkawicz@cent.uw.edu.pl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Coiled coils are widespread and diverse domains that can be found in approximately 5% of proteins. They were first described as components of fibrous structural proteins such as keratins and myosins; however, subsequent studies revealed that coiled coils are involved in a much wider range of biological functions, including signal transduction, regulation of gene expression, oligomerization and transport of other molecules (Lupas *et al.*, 2005, 2017). Structurally, coiled coils comprise two or more α-helices in a parallel or antiparallel orientation that are wound around each other to form a regular supercoiled bundle.

The hallmark of coiled-coil structures is a specific mode of interaction between helices, termed *knobs-into-holes* (Walshaw *et al.*, 2001), in which a residue from one helix (knob) packs into a cavity formed by side-chains of residues from the opposing helix (or helices). Such a regular packing requires that side chains occupy periodically equivalent positions along the helix interface. This cannot be achieved with undistorted α-helices that are characterized by a periodicity of 3.63 residues per turn (Arnott *et al*, 1967) and a continuous drift of side-chain positions (Fig. 1). To compensate for the drift and form a coiled-coil structure, the helices need to globally change their periodicity. Since the constraints imposed by the hydrogen bonds

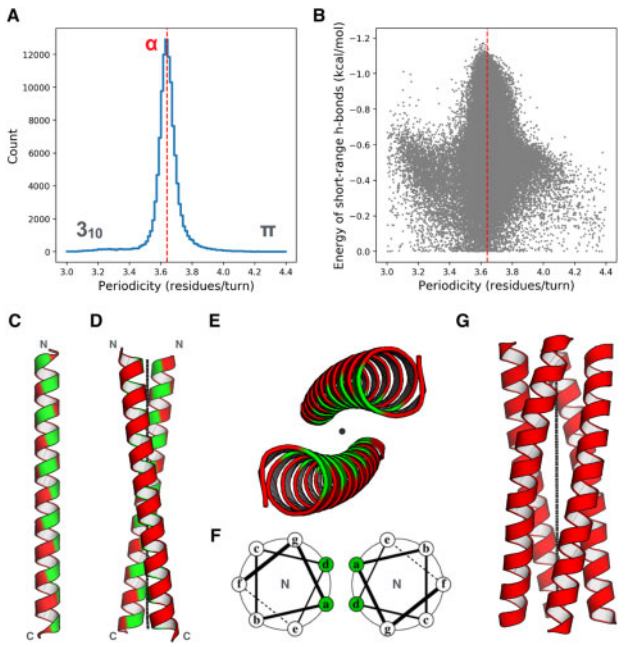

Fig. 1. Periodicity and supercoiling of α-helices. (**A**) Distribution of the average periodicity (residues per turn) of >80 000 helices comprising at least five residues. (**B**) Relationship between the helical periodicity and the short-range hydrogen bonding energy predicted with Rosetta. (**C**) A straight undistorted α-helix (periodicity 3.63 residues per turn). Residues corresponding to the positions *a* and *d* of the heptad repeat are shown in green. A continuous drift of side-chain positions across the face of the helices is visible. (**D**) The same helix as in (C) but shown in the context of a coiled-coil bundle with left-handed supercoiling of the helices. The side-chains in the individual heptad positions occupy equivalent positions after every two helical turns. The bundle axis is shown as a dashed black line. (**E**) The same structure as in (D) but seen from the top. (**F**) A helical wheel diagram showing the coiled-coil heptad repeat for the structure shown in (D) and (E). (**G**) Example of a non-canonical bundle periodicity (15/4) causing a right-handed twist of a bundle

prevent altering the actual helical periodicity, the change is achieved by bending the helices and wrapping them around each other. In such a supercoiled bundle, the component α-helices retain their inherent periodicities, however, when viewed from the perspective of the bundle axis, their periodicity is effectively reduced to 3.5 so the side-chains occupy structurally equivalent positions after every 7 residues (or 2 helical turns; $7/2 = 3.5$). Such an effective periodicity will be hereinafter referred to as bundle periodicity to avoid confusion with the helical periodicity. Importantly, the structural regularity resulting from the supercoiling of helices underlies a repeating seven-residue long sequence pattern, the heptad repeat. The seven positions of heptad repeat are labeled *a–g*; residues in the positions *a* and *d* form the hydrophobic core, whereas residues in the remaining positions (*b, c, e, f, g*) are exposed toward the solvent and are typically occupied by hydrophilic residues.

While most coiled coils follow the canonical packing described above, a variant packing mode was observed in some antiparallel structures. It is brought about by the global axial rotation of all helices by up to 26° that shifts the relative position of residues and introduces additional heptad position to the hydrophobic core (*e* or *g*, depending on the rotation direction) (Dunin-Horkawicz *et al.*, 2010b; Lupas *et al.*, 2017; Szczepaniak *et al.*, 2014). Variations in helix axial rotation states were also observed in parallel structures; however, in this case, the extent of rotation is considerably smaller (Szczepaniak *et al.*, 2018). Another type of deviation from the canonical coiled coils is related to the degree of supercoiling. Many coiled-coil structures contain fragments with bundle periodicities different from 7/2 (heptad), such as $11/3 = 3.666$ (hendecad), $15/4 = 3.75$ (pentadecad) or $19/5 = 3.8$ (nonadecad). The bundle periodicity defines the handedness of supercoiling: values above and below 3.63 residues per turn indicate the increasingly tighter right-handed and left-handed twisting, respectively.

Owing to the regular nature of coiled-coil domains, their structures can be fully described with parametric equations (Crick, 1953a, 1953b), making them an ideal model system for protein design (Woolfson, 2017) and studying sequence-structure relationships (Szczepaniak *et al.*, 2018, 2014). For a given structure, the parametrization permits quantifying features such as degree of helix axial rotation and periodicity. These features characterize the geometry of the hydrophobic core and the degree of supercoiling, respectively. Among other parameters that can be traced numerically are horizontal shift between the helices, angular displacement between them, and the bundle radius. All these parameters can be determined at the per-residue resolution, enabling the detection of local fluctuations and transitions between various coiled-coil types. A variety of bioinformatics tools were developed to perform parameterization-based measurements of coiled coils based on their backbone structures (Lupas *et al.*, 2017). One of the first was TWISTER (Strelkov *et al.*, 2002), an algorithm for the determination of local, per-residue coiled-coil parameters of parallel and symmetric bundles. Its successor, SamCC (Dunin-Horkawicz *et al.*, 2010b), initially developed to aid the analysis of coiled-coil domains of prokaryotic signal transduction proteins (Ferris *et al.*, 2011), can be used to measure antiparallel and asymmetric coiled coils with four or more helices. In contrast to these methods, CCCP (Grigoryan *et al.*, 2011) globally fits Crick parameters to a given backbone structure and cannot provide detailed per-residue readouts. Recent work presented an alternative approach that relies on a minimal set of independent parameters derived using principal component analysis (Guzenko *et al.*, 2018b). The Crick parameterization can also be used to generate coordinates for the main chains of coiled coils based on specified parameters. Such backbone models can be obtained with CCCP (Grigoryan *et al.*, 2011), BeammotifCC (Offer *et al.*, 2002) and CCBuilder (Wood *et al.*, 2018). Moreover, the coiled-coil parametric equations were also introduced to the Rosetta modeling package, in the *BundleGridSampler* module (Dang *et al.*, 2017).

The abundance of coiled-coil domains in protein structures triggered the development of SOCKET, an automatic method for coiled-coil detection based on the presence of *knobs-into-holes* packing (Walshaw *et al.*, 2001). By applying SOCKET to all Protein Data Bank structures, Woolfson and coworkers have created the CC+ database that catalogs coiled-coil bundles according to their architecture (number and orientation of helices) (Testa *et al.*, 2009). More recently, the SOCKET data was also used to construct Atlas of coiled coils (Heal *et al.*, 2018) in which the bundles are represented and classified in a form of graphs.

The wealth of data stored in the CC+ database was analyzed with CCCP to obtain general statistics on parameters such as bundle radius, degree of supercoiling, relative offset and rotation of helices (Grigoryan *et al.*, 2011). However, this analysis lacks residue-level resolution because, as mentioned above, CCCP works by the iterative fitting of an idealized model to a given structure, and thus provides only averaged values and misses the local variations of the parameters. Such a high-throughput analysis could not have also been made with SamCC or TWISTER since these methods require a precise definition of a bundle under consideration that is neither provided by SOCKET nor could be obtained automatically with existing software. Consequently, our view on the variability of coiled-coil parameters at the per-residue level is based only on the anecdotal analysis of particular proteins.

In this study, we present SamCC-Turbo, a computational tool for the high-throughput and automatic measurement and classification of coiled-coil structures. By surveying Protein Data Bank with SamCC-Turbo, we have built a database of ∼50 000 coiled-coil bundles. We show how this machine learning-ready dataset can be used in focused and global analyses in several examples. Finally, we discuss the possible application of the SamCC-Turbo in developing tools for the detection and modeling of coiled-coil domains.

2 Materials and methods

2.1 Calculation of helical periodicity

The PDB clustered at a 70% pairwise sequence identity was used as a representative set of structures for the analyses. Structures were minimized using the Rosetta relax protocol (Park *et al.*, 2016) (five independent runs, each involving five cycles of minimization), and for each structure, the lowest-energy model was retained. In these models, helices were defined using DSSP (Kabsch *et al.*, 1983) as any set of consecutive helical residues (DSSP codes H, G and I). The residues per turn values were calculated using the Python implementation (Gowers *et al.*, 2016; Michaud-Agrawal *et al.*, 2011) of the HELANAL algorithm (Bansal *et al.*, 2000) and averaged for each helix. To evaluate the energy cost associated with the helix periodicity variation (Fig. 1), the Rosetta hydrogen bonding (*hbond_sr_bb*) energy term was used.

2.2 Bundle measurement

The measurement of a coiled-coil bundle comprising m helices of length n residues each involves the following steps: first, the bundle is cut into n layers, which are roughly perpendicular to the bundle axis and parallel to each other (Fig. 2). Every layer consists of all atoms defining m residues originating from the m helices. Then, helical axes and the bundle axis are defined as described in (Dunin-Horkawicz *et al.*, 2010b; Strelkov *et al.*, 2002). Briefly, the axis of each helix is described as a spline line defined by points O_n, so that each O_n corresponds to the n-th residue of a given helix. As the procedure cannot define O_n for the first and last residue, these are excluded from further analysis. Finally, C_n points defining the bundle axis are calculated as the geometric average of O_n points (Fig. 2C).

Once the helical axes are traced, the bundle's local parameters can be determined as a function of the residue number n. The helical phase yield per residue ($\Delta\omega_1$) is defined as the mean of the two dihedral angles A_n O_n O_{n+1} A_{n+1} and A_n O_n O_{n-1} A_{n-1}, where A_n denotes C_α atom. Similarly, the coiled-coil phase yield per residue ($\Delta\omega_o$) is calculated as the mean of the two dihedral angles O_n C_n C_{n+1} O_{n+1} and O_n C_n C_{n-1} O_{n-1}. In both cases, for the first and the last layer, only one angle value is considered. The local helix periodicity p is calculated as $360/\Delta\omega_1$, whereas the local bundle periodicity P is calculated as $p_\alpha/(1 - p_\alpha{}^*\Delta\omega_o/360)$, where $p_\alpha = 3.63$ is the number of residues per turn in an undistorted helix. The distances between the O_n points and C_n points define local radius values, and the angle between the vectors O_nC_n and O_nA_n correspond to the positional orientation angle (Crick angle; φ_n), which gives the location of a residue n relative to the supercoil axis (Fig. 2).

The per residue helix axial rotation values can be then calculated as the difference between observed φ_n values and the values expected for the individual positions. For example, in a canonical coiled coil ($P = 7/2$), the consecutive expected Crick angles are separated by $360°/P = 102{,}8°$, and thus their values for the heptad positions a to g are $i + 0.0°$, $i + 102.9°$, i+-154.3°, i+-51.4°, $i + 51.4°$, $i + 154.3°$ and i+-102.9°. The variable i denotes an empirically chosen shift value defining the reference rotation of the helices. It can be set to a value averaged over a panel of two-, three- and four-helical coiled coils (Dunin-Horkawicz *et al.*, 2010b; Offer *et al.*, 2002); however, it is advisable to use a shift value defined specifically for a given bundle type (e.g. 14.5° and 19.5° for antiparallel and parallel four-helical bundles, respectively; Szczepaniak *et al.*, 2014, 2018).

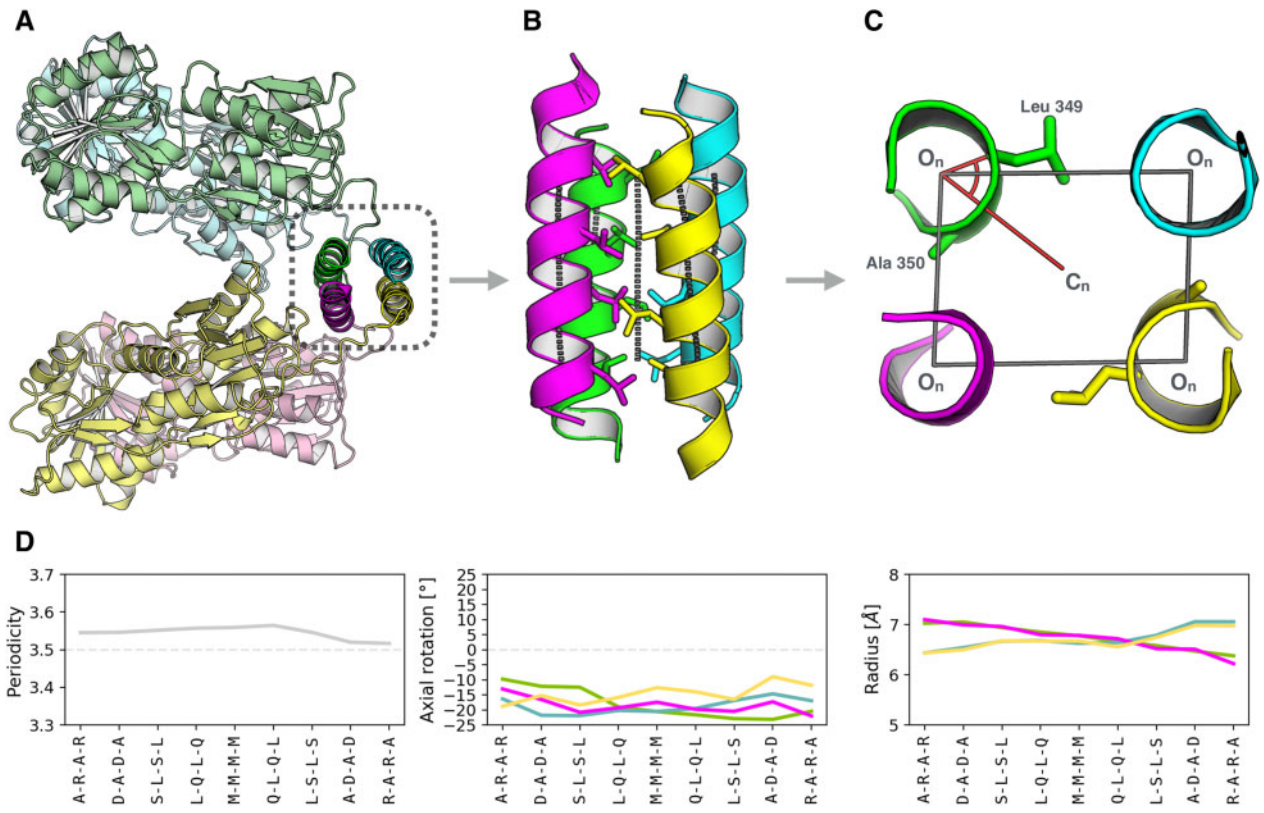

Fig. 2. Measurement of a coiled-coil bundle. (**A**) An example input, a tetrameric structure of lactose operon repressor (PDB: 1LBI). Each chain is shown in a different color. (**B**) A coiled-coil bundle identified in the input structure. Side-chains participating in the formation of the hydrophobic core are shown as sticks. Helices and bundle axes are shown as dashed gray lines. (**C**) The bundle can be 'sliced' into layers roughly perpendicular to the bundle axis and parallel to each other. The depicted layer is composed of residues A (chain): Ser-345, B: Leu-349, C: Leu-349 and D: Ser-345 (for clarity the neighboring residues are also shown). The Crick angle for residue Leu-349 is shown in red. (**D**) Per-layer measurement of the bundle defined as in (B). The four letters on the *X*-axis denote residues participating in the formation of a given layer. In the case of helix axial rotation and radius, all four helices are shown separately

2.3 Manual and automatic mode

In SamCC-Turbo, layers can be defined in automatic or manual mode. In the manual mode, the exact definition of a bundle to be measured must be provided as an input. For instance, the definition of a coiled-coil tetramerization domain of the lactose operon repressor (PDB: 1LBI; Fig. 2) includes: the residue ranges for each of the four helices (A : 342-352, B : 352-342, C : 352-342, D : 342-352), information about the topology (chains B and C are antiparallel and thus their ranges are read backward), and the order of chains in the bundle (A, C, D, B). In the case of the above structure, the first and the last measurable layers are defined by residues A : 343(Ala), C : 351(Arg), D : 343(Ala), B : 351(Arg) and A : 351(Arg), C : 343(Ala), D : 351(Arg), B : 343(Ala), respectively (as mentioned above, the first and the last residues of the user-defined bundle are excluded due to the limitations of the method for the helical axes calculation).

In the automatic mode, the whole input structure is first analyzed with SOCKET (default cut-off = 7.4 Å) (Walshaw *et al.*, 2001) to detect *knobs-into-holes* packing and to roughly annotate possible coiled-coil regions (Fig. 3A). For each detected coiled-coil bundle, SOCKET returns its oligomerization (dimer, trimer, tetramer, etc.) and residue ranges of helices along with their relative orientation. The residue ranges provided by SOCKET do not imply any interhelical interactions and cannot be used to define layers. For this reason, all the detected bundles are passed to the layer detection procedure described below and in Figure 3.

First, to discard possible dangling ends (i.e. helical fragments defined by SOCKET but not participating in the bundle formation), all O_n points that do not possess any other O_n points from the remaining helices within a radius of 20 Å are removed. Next, in each helix axis, three sets of r (by default $r = 5$) consecutive O_n points are selected, one at the beginning, second in the middle and third at the end (Fig. 3B). All the O_n points from the corresponding sets, i.e. start, middle and end are grouped. Then, within each group, all possible simple n-sided polygons (where n is the number of helices in a bundle) are constructed (Fig. 3C). Each such polygon comprises one O_n point from each helix (in a tetrameric bundle and $r = 5$, this approach yields 5^4 polygons in each of the three groups). The obtained polygons define the possible initial layers from which the optimal one is selected. To this end, all polygons from the three sets are gathered together, and a few of them (by default 9) with the lowest perimeter are selected. Then, a median of perimeters of the selected layers is calculated, and these with perimeter higher than the median are discarded. The final list of initial layers is used to define the remaining layers by traversing up- and/or downstream the bundle up to the point where one of the helices ends. From these settings, the one exhibiting minimal median of angles between the neighboring layers is selected as the best one and assigned to a given bundle (Fig. 3D). Since the O_n points forming a single layer typically do not lay on the same plane, they are projected onto a plane before calculating the angle. Moreover, in dimers, which contain only two points per layer, the best setting is selected based on the minimal average distance between points forming a layer.

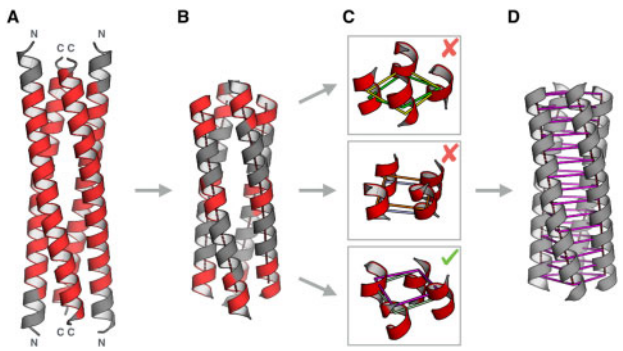

Fig. 3. Automatic layer detection procedure on an example of the proteasomal ATPase Mpa coiled-coil structure (PDB: 3M9H). (**A**) SOCKET is used to detect regions participating in *knobs-into-holes* interactions (shown in red). (**B**) The structure is trimmed to remove dangling ends and three, equidistant regions are defined (shown in red). (**C**) In each region, all possible polygons connecting the helical axes are constructed. For clarity, only two polygons per region are shown, each in a different color. (**D**) The optimal polygon is used to define an initial layer which is then used to define all the remaining layers in a bundle. In this particular example, the purple polygon from the lower region (indicated with a checkmark) was found to be the optimal one. Only odd layers are drawn for clarity of the picture

2.4 Protein Data Bank scan

The Protein Data Bank (version 2020-07-17) was searched with the *localpdb* Python library (Ludwiczak *et al.*, https://github.com/labstructbioinf/localpdb, manuscript in preparation) to obtain 161 684 structures. All of them were processed with the adapted version of the MakeMultimer.py script (Michael Palmer, http://watcut.uwaterloo.ca/tools/makemultimer/) to generate biological assemblies, which were then subjected to the analysis with SOCKET (cut-off = 7.4 Å) (Walshaw *et al.*, 2001). 17 493 structures were found to contain at least one region interacting via *knobs-into-holes* packing (53 325 coiled-coiled regions in total). Each detected bundle was assigned one of the topology classes (for dimers: ↑↑ and ↑↓, for trimers: ↑↑↑, ↑↑↓ and ↑↓↓, etc.) based on the relative orientation of helices provided by SOCKET. Finally, the redundancy in the dataset was removed similarly as in the CC+ database (Testa *et al.*, 2009) by pairwise comparisons of all coiled-coil regions and assuming that two regions are identical when they possess identical sequences, oligomerization and orientation. For each of the redundant groups, the coiled-coil region from the structure with the highest resolution was selected as a representative. Altogether 8747 structures (13 671 coiled-coil regions) were retained in a non-redundant set.

2.5 Identification of periodicity segments

Some coiled coils comprise segments of different bundle periodicity. To detect such segmentation of a bundle, the following procedure was used: first, the bundle periodicity values were averaged in every layer and smoothed with the Savitzky-Golay algorithm (Savitzky *et al.*, 1964) (window length 7, polynomial degree 1). Then, polynomial functions P_n with an increasing degree of n were fitted until Root Mean Square Error (RMSE) was equal or lower than 0.015, or the difference between RMSE for n and n-1 was equal or lower than 75%. Finally, for the best-fitting P_n, the stationary points were calculated and used to define intervals dividing the bundle into segments (Fig. 5). To facilitate the classification of the periodicity segments, each was fitted to a linear function. As a result, each segment was represented by three numbers: a and b coefficients of the linear function and the length.

2.6 Visualization of the coiled-coil sequence space

27 segments fulfilling the following conditions were picked: stable bundle periodicity ($a < 0.0006$; see above), length of at least seven layers and parallel tetrameric topology. In each segment, all residues were assigned based on the Crick angle value into one of the seven bins (bins' ranges were defined as in Fig. 4). In each bin, the average biophysical properties (the average side-chain volume and average hydrophobicity) of contained residues were calculated as described in (Szczepaniak *et al.*, 2014). Moreover, in each bin, the average Crick angle value was calculated and subtracted from a reference value. Each bin's reference value was defined as an average of Crick angles in a non-redundant set of parallel, tetrameric bundles. As a result, each segment has been described with a 14-element 'sequence' vector (average side-chain sizes and volumes in each bin) and a single 'structural' descriptor calculated as an average of Crick angle deviations in the seven bins (differences between observed and reference values). The 'sequence' vectors were scaled to 0-1 values and mapped onto a 2-dimensional space using UMAP (*n_neighbors* = 26) (McInnes *et al.*, 2018) (Fig. 6).

2.7 Database implementation

All SamCC-Turbo readouts, along with segment annotations, were collected into a publicly accessible database CCdb (https://lbs.cent.

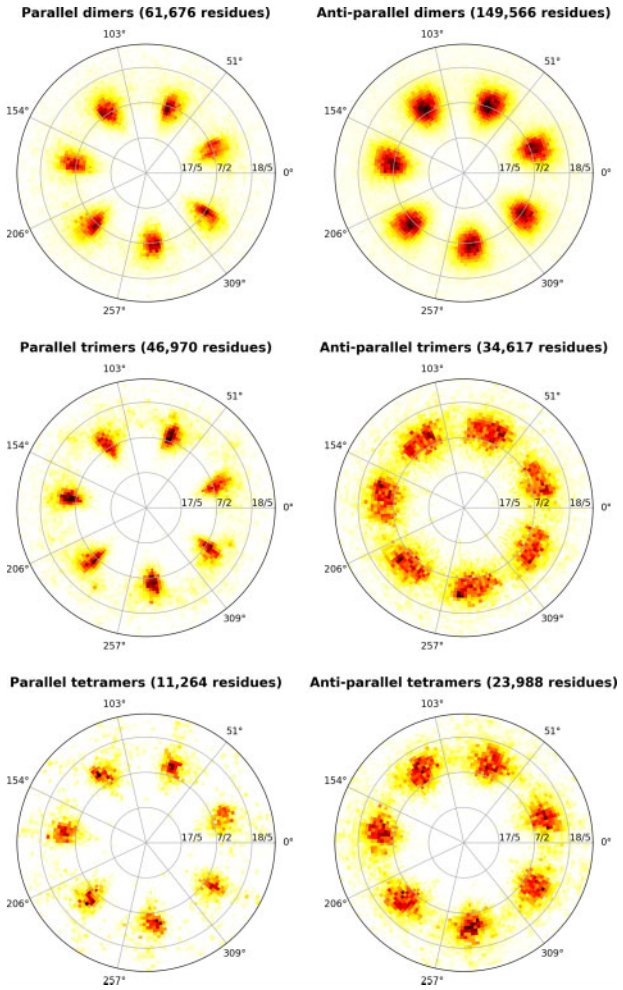

Fig. 4. Distribution of Crick angles in polar coordinates for various coiled-coil topologies. Crick angle is denoted by the angular coordinate (0°/360° indicate residues pointing directly toward the hydrophobic core of a bundle) and the radial coordinate is proportional to the bundle periodicity

uw.edu.pl/samcc_turbo). Web interface for CCdb was created with Django 2.2 and Bootstrap 3.3.7 frameworks.

3 Results

3.1 Manual analysis of coiled-coil structures

Before a coiled-coil can be measured, boundaries (beginnings and ends) and topology (orientation and order) of its constituent helices have to be determined (Fig. 2). Such a defined bundle can be then 'sliced' into parallel layers, each of which is formed by one residue from each helix. All the subsequent measurements are performed in a per-layer fashion, and the obtained values are either assigned to the individual residues or averaged within the layer. The per-layer and the per-residue measurements are represented as curves where fluctuations in the individual parameters can be traced. For example, measurement of the tetramerization domain of the lactose operon repressor (PDB: 1LBI) reveals an antiparallel bundle with canonical periodicity ~3.5 (7 residues per 2 helical turns) and non-canonical geometry of the hydrophobic core caused by rotation of the helices by approximately -20° away from values typically observed in coiled coils of this type (Fig. 2) (Dunin-Horkawicz *et al.*, 2010b; Szczepaniak *et al.*, 2014). Analogous measurements performed based on the manual layer assignment were the basis of many studies, such as those aiming at understanding the sequence features defining structural parameters of coiled coils (Szczepaniak *et al.*, 2018), revealing the conformational changes occurring in signal transduction proteins (Duclert-Savatier *et al.*, 2018; Ferris *et al.*, 2011, 2012), describing designed (Xu *et al.*, 2013; Zhang *et al.*, 2018), fusion (Deiss *et al.*, 2014) and natural structures (Dunin-Horkawicz *et al.*, 2010b).

3.2 Automatic analysis of coiled-coil structures

The manual assignment of bundles is a time-consuming process, frequently requiring expert knowledge. Bearing this in mind, we developed SamCC-Turbo, a coiled coils analysis tool that can automatically detect bundles and define their parameters solely based on the input structure. Although such an approach is fast, we deemed it necessary to assess its accuracy and applicability to high-throughput scans. To this end, we compared manual and automatic assignments in 35 bundles and found that despite slight differences, the automatic assignments are of high quality and the structural parameters calculated based on them do not deviate from those obtained manually (Supplementary Fig. S1). These results indicate that SamCC-Turbo reproduces the results of manual measurements reasonably well and can be used to perform automatized analyses of coiled coils.

To assess the abundance of coiled-coil domains, we performed a scan of the National Center for Biotechnology Information (NCBI) protein database filtered to 70% maximum pairwise sequence identity with DeepCoil (Ludwiczak *et al.*, 2019), a deep-learning tool for the prediction of coiled coils. We found that 8% of the sequences contain at least a single predicted coiled-coil segment comprising 14 or more residues. A similar estimate was obtained by scanning the Protein Data Bank filtered to 70% maximum pairwise sequence identity with SOCKET (Walshaw *et al.*, 2001). In this case, we found that 5% of structures contain at least a single coiled-coil region of 14 or more residues. The possibility of automatic measurements with SamCC-Turbo enables investigating this area of 'protein universe' in terms of sequences, structures and the relations between the two. Among 161 684 structures deposited in Protein Data Bank, we identified 53 325 coiled-coil bundles in 17 493 structures. After removing redundancy, we obtained a set of 8747 structures containing 13 671 unique coiled-coil regions (Supplementary Fig. S2). Upon removing structures with resolution below 3 Å, the resulting non-redundant set of 6555 structures containing 9690 bundles was used in the three exemplary analyses described below.

In the first analysis, we performed a global survey of Crick angles in bundles of various topology. The Crick angle (Fig. 2C) defines a residue's position relative to the bundle axis; the values of 0/360° and 180° denote residues pointing directly toward and outwards the hydrophobic core, respectively. We found that regardless of the oligomerization state, Crick angles in parallel bundles are more constrained than in bundles containing one or more antiparallel helices (Fig. 4). In parallel dimers, trimers and tetramers, Crick angles fall into seven focused groups, corresponding to the seven heptad positions. In the case of antiparallel structures, the seven groups are considerably less focused, reflecting the greater rotational freedom of their helices. These general observations agree with our previous studies in which we analyzed small, hand-picked sets of parallel (Szczepaniak *et al.*, 2018) and antiparallel (Szczepaniak *et al.*, 2014) four-helix bundles. However, as shown in Figure 4 (background signal in shades of yellow), there is a minor fraction of parallel structures that assume unusual Crick angle values. Among them, we found right-handed bundles with non-canonical periodicity (e.g. tetrabrachion stalk domain; PDB codes 1YBK and 6CRD), which is expected, since such structures adopt an alternative variant of coiled-coil interhelical packing (knobs-to-knobs; (Lupas *et al.*, 2017)) characterized by the presence of residues pointing directly to the core, so-called *x* layers. Other examples include M2 proton transporters (3LBW, 6BMZ and 6NV1) and hemagglutinins (4FQI, 4MHI, 5E2Z, 5BNY and 5BQY) of influenza A viruses, spike protein (5ZHY) of human coronavirus 229E, HAMP domain (2Y20, 3ZRX and 4CQ4), a signal-transducing module adopting multiple conformational states (Dunin-Horkawicz *et al.*, 2010a; Ferris *et al.*, 2011; Hulko *et al.*, 2006), and structures with odd sequence composition such as the designed 'Phe-zipper' (2GUS) containing phenylalanine residues at the core (Liu *et al.*, 2006). It is

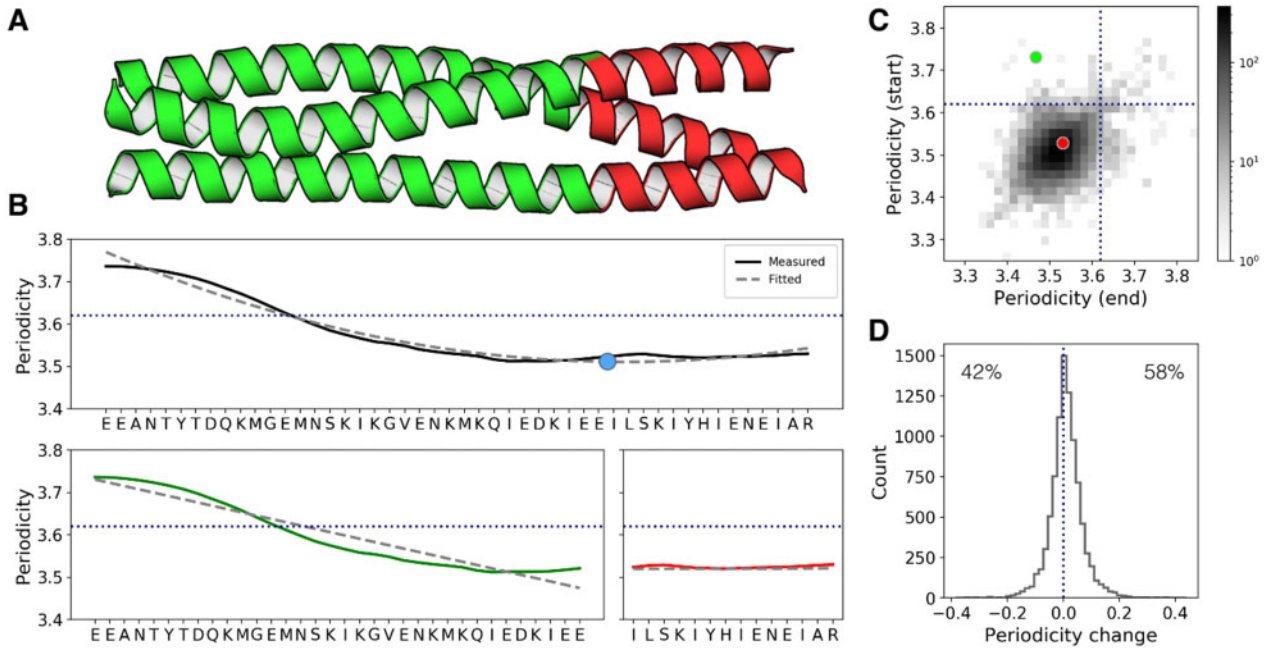

Fig. 5. Definition of bundle periodicity segments. (**A**) Fragment of *Salmonella enterica* SadA trimeric autotransporter structure (PDB: 2YO3). (**B**) The average per-layer bundle periodicity in the whole fragment (upper panel) and two identified segments (two lower panels). Values measured in the structure are shown as solid lines, whereas fitted polynomial and linear functions as dotted lines. The stationary point of the fitted polynomial is indicated with a pale blue dot. (**C**) The 2D histogram in log scale depicting bundle periodicity changes in segments comprising seven or more layers. (**D**) Distribution of bundle periodicity changes shown in (C)

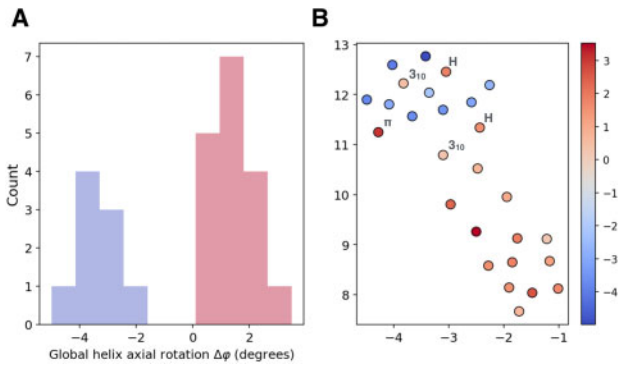

Fig. 6. Structural and sequence descriptors of 27 parallel tetrameric segments. (**A**) The segments were divided into two groups based on their global helix axial rotation. (**B**) UMAP representation of sequence descriptors. Each segment was colored according to the global helix axial rotation value shown in (A). H, 3_{10} and π labels denote the HAMP domain, 3_{10} and π helices, respectively

conspicuous that most of the aforementioned coiled coils are not 'simple' structural scaffolds but are rather parts of dynamic systems involved in processes such as transport, viral binding to host cell and signal transduction. In their work, Grigoryan et al. (2011) used CCCP to perform an analogous analysis; however, they found no difference between parallel and antiparallel bundles in terms of Crick angles distribution, highlighting the advantage of using high-resolution readouts from SamCC-Turbo.

In the second high-throughput analysis, we focused on the observation that some coiled-coil structures comprise segments of different degree of supercoiling. The degree of supercoiling can be described using pitch, i.e. the distance required for the superhelix to complete a full turn or, as described above, bundle periodicity. Because the pitch is very sensitive to small structural changes (Supplementary Fig. S3), we decided to use the latter coefficient and a procedure based on polynomial function fitting (Fig. 5). Briefly, periodicity values in a given bundle are averaged within layers, smoothed and fitted to a polynomial function of degree that provides a reasonable approximation. Once the function is fitted, its stationary points, i.e. points where the change of periodicity switches the sign, are determined. Finally, the stationary points are used to split the bundle into segments. Using the segmentation procedure, we defined 10 005 segments; in most cases, the detected segments corresponded to the whole bundle (96.7%). The remaining bundles (3.3%) have a composite character and comprise two or more segments. To get an overview of the segments, for each, we calculated a periodicity change value, defined as the absolute difference between the bundle periodicity measured in the first and the last layer (for the parallel bundles, the first layer comprises residues closest to the N-termini of the constituent helices, whereas, for the antiparallel bundles, the first and last layers are assigned arbitrarily). We found that 70% of segments are characterized by a stable periodicity (change below 0.05), and among the remaining 30%, only a small fraction featured substantial changes (>0.2), indicating that extreme transitions are rare (Fig. 5D). As in the case of Crick angle deviations, the substantial changes in the periodicity are frequent in structures of dynamic bundles involved in complex functions such as bacterial and viral pathogenicity. The most striking transitions were observed in trimeric autotransporters (TAAs), proteins employed by pathogenic bacteria to adhere to their host cells, YadA (3LT7, 3LT6 and 3H7X) and SadA (2YO3; Fig. 5B) in which supercoiling shifts from strong right-handed (15/4) to canonical left-handed (7/2). A transition in the opposite direction, i.e. from smaller to larger periodicity values, was seen in Hia, a TAA from *Haemophilus influenzae* (7/2 → 11/3). Other examples are coiled-coil domains of fusion glycoproteins from herpes simplex (15/4 → 7/2), pseudorabies (11/3 → 17/5) and respiratory syncytial (7/2 → 15/4) viruses (2GUM, 6ESC and 6Q0S, respectively). We noticed that most of the strong transitions occur from higher to lower periodicities. Interestingly, this bias is also observable when all fragments, regardless of the periodicity difference, are taken into consideration (Fig. 5D). In the context of this analysis, we asked ourselves to what extent the bundle periodicity is defined by its surrounding. To this end, we investigated a non-redundant set of 113 bundles comprising either five or six helices. For each bundle, we calculated its periodicity and

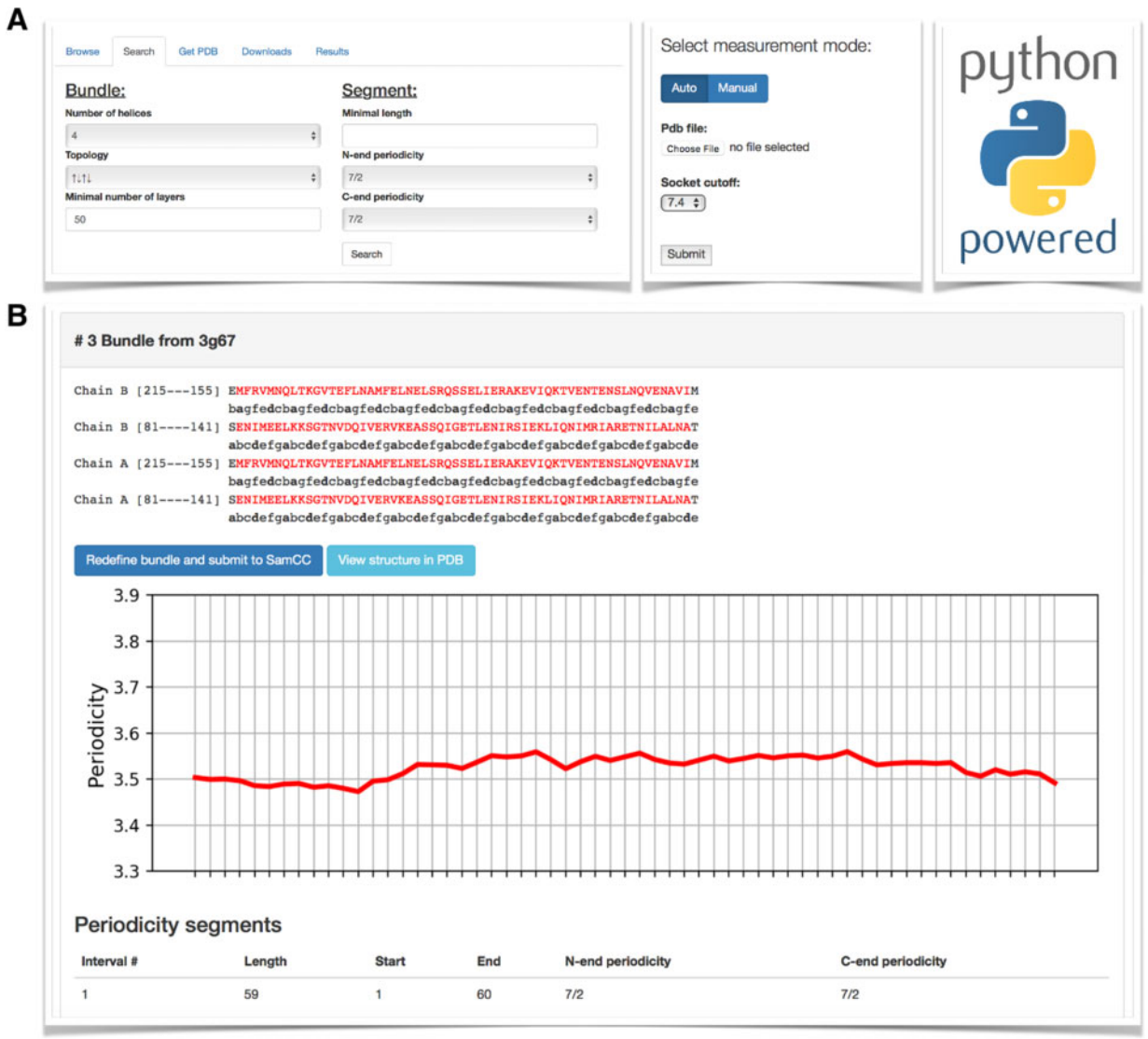

Fig. 7. The web interface of SamCC-Turbo. (**A**) There are three main ways of using SamCC-Turbo: browsing the precalculated database of non-redundant structures, providing own structures for the measurement, and using the Python library. In this example, anti-parallel, four-helix and canonical bundles comprising at least 50 layers were requested. (**B**) The results are presented in the form of a list of bundles that match the query parameters. A coiled-coil structure of chemoreceptor from *Thermotoga maritima* is shown

coverage index defining the percent of the structure corresponding to the coiled-coil domain (100% indicates that the whole structure is a coiled coil). We found (Supplementary Fig. S4) that the periodicity of these bundles appears to be dictated by their structural surrounding: structures of isolated bundles are restrained to a canonical periodicity of 7/2, whereas in bundles that are parts of larger structures (e.g. receptors) the periodicity varies considerably (up to 3.7). Although performed on a relatively small set, this analysis suggests that the supercoiling of a bundle may be defined both by its intrinsic features and the surrounding.

Finally, to exemplify the potential application of the segmentation procedure, we performed an analysis of the sequence-structure relationship in 27 parallel tetrameric fragments of stable periodicity (Supplementary Table S1). Each fragment has been associated with a sequence and structural descriptor (see Methods for details). The sequence descriptor reflects the average biophysical properties of residues localized in the seven heptad positions, whereas the structural descriptor defines global axial rotation of the helices in a bundle. Based on the structural descriptors, the 27 fragments were divided into two groups: 18 'positive' (rotation $>0°$) and 9 'negative' ($<0°$) (Fig. 6A). To investigate whether this structure-based division is also detectable at the sequence level, we mapped the sequence descriptors onto a 2-dimensional space using UMAP (McInnes *et al.*, 2018). The obtained mapping yielded a reasonable separation of the two structural groups, with only five segments being 'misclassified' (see labels in Fig. 6B). Interestingly, such a discrepancy between the structural and sequence descriptors was observed exclusively in structures functionally similar to those found in the analysis of Crick angles and periodicity shifts (see above); namely, the phosphoprotein tetramerization domain (PDB: 4N5B and 6EB8) of Nipah virus, hemagglutinin-neuraminidase stalk domain (PDB: 3TSI) of parainfluenza virus, and aforementioned signal-transducing HAMP domains (PDB: 4I3M and 3ZRV). In the three viral structures, we identified non-canonical helices (a 3_{10} helix in phosphoprotein and a π helix in hemagglutinin-neuraminidase), whose possible function is to accommodate a tension caused by conflicting hydrophobic registers (Hartmann *et al.*, 2016; Schmidt *et al.*, 2017).

The three exemplary analyses described above revealed an association between coiled-coil peculiarities and their role in dynamic systems related to virulence and signal transduction. Since they depend on detecting very subtle differences in the core packing, we

repeated them using a more stringent cut-off of 2 Å for structure resolution (2551 structures containing 3557 bundles). We found that despite using considerably fewer structures, the results do not differ substantially from these obtained using the set generated at the 3 Å cut-off (Supplementary Figs S5 and S6). Finally, we also tested SamCC-Turbo's applicability to other tasks, for example, quantification of coiled-coil parameter fluctuations as a function of the MD simulation progress or clustering of decoys obtained with Rosetta fold-and-dock protocol (Das *et al.*, 2009). Since models resulting from such simulations may vary in terms of bundle definitions, their automatic measurement is especially desirable. Moreover, parameterization-based descriptors are much more informative than RMSD values, typically used to evaluate structural variation.

3.3 Data and method availability

To make the SamCC-Turbo data available to a broad range of researchers, we created a web service at https://lbs.cent.uw.edu.pl/samcc_turbo. The service features access to the up-to-date measurements that can be browsed using various search tools or simply downloaded in a tabular format (Fig. 7). The web server also provides an interface to the SamCC-Turbo tool that can be run in manual or automatic mode. The user has a possibility to switch between the two modes, so for example, a structure under investigation can be first subjected to the automatic mode, and then the obtained definitions can be resubmitted to the manual mode and refined according to the needs. Finally, we provide SamCC-Turbo source code at https://github.com/labstructbioinf/samcc_turbo for performing more advanced tasks.

3.4 Conclusions and perspectives

The crucial step of the automatic measurements with SamCC-Turbo is the detection of *knobs-into-holes* interactions with the SOCKET tool (Walshaw *et al.*, 2001). SOCKET offers a very high specificity in detecting coiled-coil regions; however, it has some limitations. The most important one is a potential bias toward the detection of *knobs-into-holes* interactions typical for canonical 7/2 coiled coils (Grigoryan *et al.*, 2011). For this reason, SOCKET may miss some of the non-canonical bundles characterized by the presence of knobs-to-knobs interactions (Lupas *et al.*, 2017). A partial workaround for this problem has been proposed in a work by Grigoryan et al. (2011), but no uniform solution has been developed so far. The second problem is related to the fact that SOCKET frequently interprets high-order multimers as assemblies of dimers, thus hampering the proper measurement of such structures and leading to the underestimation of their abundance. This issue, important in the context of the growing interest in the design of coiled-coil barrels (Rhys *et al.*, 2018), has been addressed in the Atlas of coiled coils (Heal *et al.*, 2018) in which a new implementation of SOCKET (iSOCKET) was used for proper classification of bundles comprising multiple helices. Considering the above, we see a need for developing new approaches for the detection of coiled-coil structures and incorporating them into the SamCC-Turbo framework.

The survey of the Protein Data Bank with SamCC-Turbo resulted in a comprehensive database of ∼50.000 coiled-coil sequences and structures. As shown in the results section, this dataset can be used in various tasks ranging from picking the desired sub-set of bundles to global analysis of all structures. The SamCC-Turbo data readouts are suitable for machine learning tasks that rely on well-formatted and organized data. For example, DeepCoil, a tool for the prediction of coiled-coil regions in sequences, was trained on raw SOCKET data that lack the information on the local structural parameters (Ludwiczak *et al.*, 2019). We are currently developing a new version of DeepCoil that will be trained not on simple, binary annotations (presence or absence of *knobs-into-holes* interactions) but precise, per-residue structural descriptors. Such a method would provide more detailed predictions which, for example, could be used to pinpoint functional 'hotspots' or to pick fragments suitable for fragment-based modeling tools such as CCFold (Guzenko *et al.*, 2018a) or fold-and-dock (Das *et al.*, 2009; Rämisch *et al.*, 2015), and thus improving their accuracy.

Acknowledgements

The authors thank Vikram Alva for the critical evaluation of the manuscript and numerous valuable comments.

Funding

This work was supported by the Polish National Science Centre [2015/18/E/NZ1/00689 to S.D-.H.] and the First Team programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund [POIR.04.04.00-00-5CF1/18-00 to S.D-.H.]. Computations were carried out with the support of the Interdisciplinary Centre for Mathematical and Computational Modeling (ICM) at the University of Warsaw [GA71-24 to S.D-.H.]. J.L. was additionally supported by the Etiuda scholarship from the Polish National Science Centre [2019/32/T/NZ1/00323 to J.L.].

Conflict of Interest: none declared.

References

Arnott,S. *et al.* (1967) Refinement of bond angles of an alpha-helix. *J. Mol. Biol.*, **30**, 209–212.

Bansal,M. *et al.* (2000) HELANAL: a program to characterize helix geometry in proteins. *J. Biomol. Struct. Dyn.*, **17**, 811–819.

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

Crick,F.H.C. (1953b) The packing of α-helices: simple coiled-coils. *Acta Crystallogr.*, **6**, 689–697.

Dang,B. *et al.* (2017) De novo design of covalently constrained mesosize protein scaffolds with unique tertiary structures. *Proc. Natl. Acad. Sci. USA*, **114**, 10852–10857.

Das,R. *et al.* (2009) Simultaneous prediction of protein folding and docking at high resolution. *Proc. Natl. Acad. Sci. USA*, **106**, 18978–18983.

Deiss,S. *et al.* (2014) Your personalized protein structure: Andrei N. Lupas fused to GCN4 adaptors. *J. Struct. Biol.*, **186**, 380–385.

Duclert-Savatier,N. *et al.* (2018) Conformational sampling of CpxA: connecting HAMP motions to the histidine kinase function. *PLoS One*, **13**, e0207899.

Dunin-Horkawicz,S. *et al.* (2010a) Comprehensive analysis of HAMP domains: implications for transmembrane signal transduction. *J. Mol. Biol.*, **397**, 1156–1174.

Dunin-Horkawicz,S. *et al.* (2010b) Measuring the conformational space of square four-helical bundles with the program samCC. *J. Struct. Biol.*, **170**, 226–235.

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**, 378–385.

Gowers,R. *et al.* (2016) MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In: Benthall, S. and Rostrup, S. (eds.) *Proceedings of the 15th Python in Science Conference*. SciPy, Austin, TX, pp. 98–105.

Grigoryan,G. *et al.* (2011) Probing designability via a generalized model of helical bundle geometry. *J. Mol. Biol.*, **405**, 1079–1100.

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

Guzenko,D. *et al.* (2018b) Optimal data-driven parameterization of coiled coils. *J. Struct. Biol*, **204**, 125–129.

Hartmann,M.D. *et al.* (2016) α/β coiled coils. *Elife*, **5**, e11861.

Heal,J.W. *et al.* (2018) Applying graph theory to protein structures: an Atlas of coiled coils. *Bioinformatics*, **34**, 3316–3323.

Hulko,M. *et al.* (2006) The HAMP domain structure implies helix rotation in transmembrane signaling. *Cell*, **126**, 929–940.

Kabsch,W. *et al.* (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Liu,J. *et al.* (2006) Conformational transition between four and five-stranded phenylalanine zippers determined by a local packing interaction. *J. Mol. Biol.*, **361**, 168–179.

Ludwiczak,J. *et al.* (2019) DeepCoil-a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics*, **35**, 2790–2795.
Lupas,A.N. *et al.* (2017) The structure and topology of α-helical coiled coils. *Sub-Cell. Biochem.*, **82**, 95–129.
Lupas,A.N. *et al.* (2005) The structure of alpha-helical coiled coils. *Adv. Protein Chem.*, **70**, 37–78.
McInnes,L. *et al.* (2018) UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.*, **3**, 861.
Michaud-Agrawal,N. *et al.* (2011) MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.*, **32**, 2319–2327.
Offer,G. *et al.* (2002) Generalized Crick equations for modeling noncanonical coiled coils. *J. Struct. Biol.*, **137**, 41–53.
Park,H. *et al.* (2016) Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.*, **12**, 6201–6212.
Rämisch,S. *et al.* (2015) Exploring alternate states and oligomerization preferences of coiled-coils by de novo structure modeling. *Proteins*, **83**, 235–247.
Rhys,G.G. *et al.* (2018) Maintaining and breaking symmetry in homomeric coiled-coil assemblies. *Nat. Commun.*, **9**, 4132.
Savitzky,A. *et al.* (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, **36**, 1627–1639.
Schmidt,N.W. *et al.* (2017) The accommodation index measures the perturbation associated with insertions and deletions in coiled-coils: application to understand signaling in histidine kinases. *Protein Sci.*, **26**, 414–435.
Strelkov,S.V. *et al.* (2002) 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.* (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.
Testa,O.D. *et al.* (2009) CC+: a relational database of coiled-coil structures. *Nucleic Acids Res.*, **37**, D315–22.
Walshaw,J. *et al.* (2001) Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *J. Mol. Biol.*, **307**, 1427–1450.
Wood,C.W. *et al.* (2018) CCBuilder 2.0: powerful and accessible coiled-coil modeling. *Protein Sci.*, **27**, 103–111.
Woolfson,D.N. (2017) Coiled-coil design: updated and upgraded. *Subcell. Biochem.*, **82**, 35–61.
Xu,C. *et al.* (2013) Rational design of helical nanotubes from self-assembly of coiled-coil lock washers. *J. Am. Chem. Soc.*, **135**, 15565–15578.
Zhang,S.-Q. *et al.* (2018) De novo design of tetranuclear transition metal clusters stabilized by hydrogen-bonded networks in helical bundles. *J. Am. Chem. Soc.*, **140**, 1294–1304.