doi:10.1016/j.jmb.2010.08.058 J. Mol. Biol. (2011) 405, 1079–1100

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

Probing Designability via a Generalized Model of Helical Bundle Geometry

Gevorg Grigoryan*† and William F. DeGrado*

Department of Biochemistry and Biophysics, School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Received 2 April 2010; received in revised form 26 August 2010; accepted 31 August 2010
Available online 7 October 2010

Edited by M. Sternberg

Keywords:
designability;
protein design;
structural parameterization;
coiled coils;
Crick parameterization

Because the space of folded protein structures is highly degenerate, with recurring secondary and tertiary motifs, methods for representing protein structure in terms of collective physically relevant coordinates are of great interest. By collapsing structural diversity to a handful of parameters, such methods can be used to delineate the space of designable structures (i.e., conformations that can be stabilized with a large number of sequences)—a crucial task for *de novo* protein design. We first demonstrate this on natural α-helical coiled coils using the Crick parameterization. We show that over 95% of known coiled-coil structures are within 1-Å C_α root mean square deviation of a Crick-ideal backbone. Derived parameters show that natural geometric space of coiled coils is highly restricted and can be represented by "allowed" conformations amidst a potential continuum of conformers. Allowed structures have (1) restricted axial offsets between helices, which differ starkly between parallel and anti-parallel structures; (2) preferred superhelical radii, which depend linearly on the oligomerization state; (3) pronounced radius-dependent **a**- and **d**-position amino acid propensities; and (4) discrete angles of rotation of helices about their axes, which are surprisingly independent of oligomerization state or orientation. In all, we estimate the space of designable coiled-coil structures to be reduced at least 160-fold relative to the space of geometrically feasible structures. To extend the benefits of structural parameterization to other systems, we developed a general mathematical framework for parameterizing arbitrary helical structures, which reduces to the Crick parameterization as a special case. The method is successfully validated on a set of non-coiled-coil helical bundles, frequent in channels and transporter proteins, which show significant helix bending but not supercoiling. Programs for coiled-coil parameter fitting and structure generation are provided via a web interface at http://www.gevorggrigoryan.com/cccp/, and code for generalized helical parameterization is available upon request.

Introduction

The number of conformational states accessible to even small proteins is astronomically large. Nevertheless, the space of natively folded protein structures is rather limited in comparison, with highly recurring secondary and tertiary motifs. Clearly, simple physical principles, such as volume exclusion or electrostatic repulsion, preclude many structures.

*Corresponding authors. E-mail addresses: gevorg@alum.mit.edu; wdegrado@mail.med.upenn.edu.

† To whom software-related correspondence should be addressed.

Abbreviations used: TM, transmembrane; CCCP, coiled-coil Crick parameterization.

However, beyond this simple filter, natively folded structures are much more restricted due to other requirements imposed by biology, including robustness of the fold in sequence space.[1,2] In the context of folding and inverse folding problems, structure robustness in sequence space has been referred to as designability.[3–5] Natural abundance of structures and their designability are related. Structural motifs that recur frequently in nature must be able to accommodate a reasonably wide ensemble of sequences and are thus designable. Koehl and Levitt have shown that the designability of a protein structure, measured as the number of sequences compatible with its backbone in an atomistic-level *in silico* protein design experiment, correlates well with its evolutionary designability, as assessed with sequence bioinformatic tools.[6] Kuhlman and Baker demonstrated that the evolutionary sequence profile of the SH3 domain can be recapitulated from a small collection of SH3-domain backbones via computational protein design.[7]

Finding the relationship between structure and designability is a fundamental problem in protein design, as one would like to *a priori* limit oneself to engineering only reasonably designable structures. However, such a relationship is difficult to synthesize because of the very large number of parameters required to exactly define the geometry of a protein. Thus, approaches for representing folded structures via a reduced set of effective parameters can greatly facilitate this process. One class of such methods, which we will refer to as ideal backbone parameterization, are particularly well suited for this task. These approaches model the overall shape into which secondary-structure elements fold with a few parameters, and can often capture the majority of observed structural variability, producing deviations between ideal and real structures within 1 Å. Examples of this include the Crick parameterization of coiled coils,[8,9] mathematical description of β-barrel structures,[10,11] statistical parameterization of the structure of collagen,[12] and parameterization of di-iron helical bundles[13] and of transmembrane helix interaction geometry.[14,15] Such methods have been quite useful in *de novo* protein design,[16–21] modeling of thermodynamic consequence of mutations,[21] protein–protein interaction prediction,[22,23] structure prediction and modeling,[14,15,24,25] and X-ray crystallography.[26]

Recent studies of designability, using lattice models and native protein structures, suggest that more designable folds are those that have more contacts per residue (contact density),[1,4–6,27,28] and that more designable structures also tend to be more thermodynamically stable.[1,4] These important findings further our understanding of designability of different folds and choice of fold by evolution. Such studies can be thought of as an instance of structural parameterization with the goal of relating it to designability. Contact density and on-lattice models provide a coarse view that is appropriate for differentiating folds from one another in the global structure space. Here, we are concerned with a more fine-grained description of variability within a given fold, and thus a higher-resolution, but nevertheless, low-parameter description of designability is sought. For *de novo* protein design, such a detailed description of designability is often needed, as this information is critical for the choice of target template coordinates.

Here, we aim to quantify the constraint of designability on the space of natural helical bundles, starting with the coiled coil. Others have shown that the Crick parameterization can closely fit natural coiled-coil backbones.[9,16,23,29,30] In fact, we demonstrate that it can reproduce to within 1- Å C_α root mean square deviation (RMSD) over 95% of all available coiled-coil structures, with a median RMSD of 0.44 Å. Further, sharp distributions of Crick parameters indicate significant biases in the space of naturally designable structures. For example, axial offsets between helices are restricted to a small range relative to the space of geometrically possible offsets, and these allowed regions differ strongly between parallel and anti-parallel helix pairs. Overall, we estimate that the space of designable structures is at least 160-fold smaller in relation to the space of geometrically reasonable structures, illustrating the importance of considering designability *a priori* in protein engineering applications.

The Crick parameterization has been difficult to extend to helical bundles in general due to their less regular structure, in which each chain is not necessarily identical and the implied symmetry is difficult to observe. While a trained eye may be able to discover residual symmetry underlying considerable complexity,[13,31] it has been difficult to do this in an objective and automated manner. Moreover, there is a continuum of helical bundles, ranging from the very regular coiled coils to ones that defy simple canonical categorization. On the other hand, helical bundles do form with relatively well-defined and recurring helix-crossing angles that reflect the restraints associated with achieving good side-chain packing at the local level[32–36] (see Fig. 1). Early work by Murzin and Finkelstein showed that helices in globular helical proteins tend to lie approximately along the ribs of certain close-to-spherical polyhedra.[34] Due to the nonintegral nature of the α-helix, good packing generally does not occur when helices are packed with their axes precisely parallel or anti-parallel with one another. Instead, side-chain packing tends to be optimized when the helical axes are at an angle to one another, resulting in the divergence of straight helices. Supercoiling, as observed in coiled coils, reflects the resolution of the frustration between optimizing interactions locally

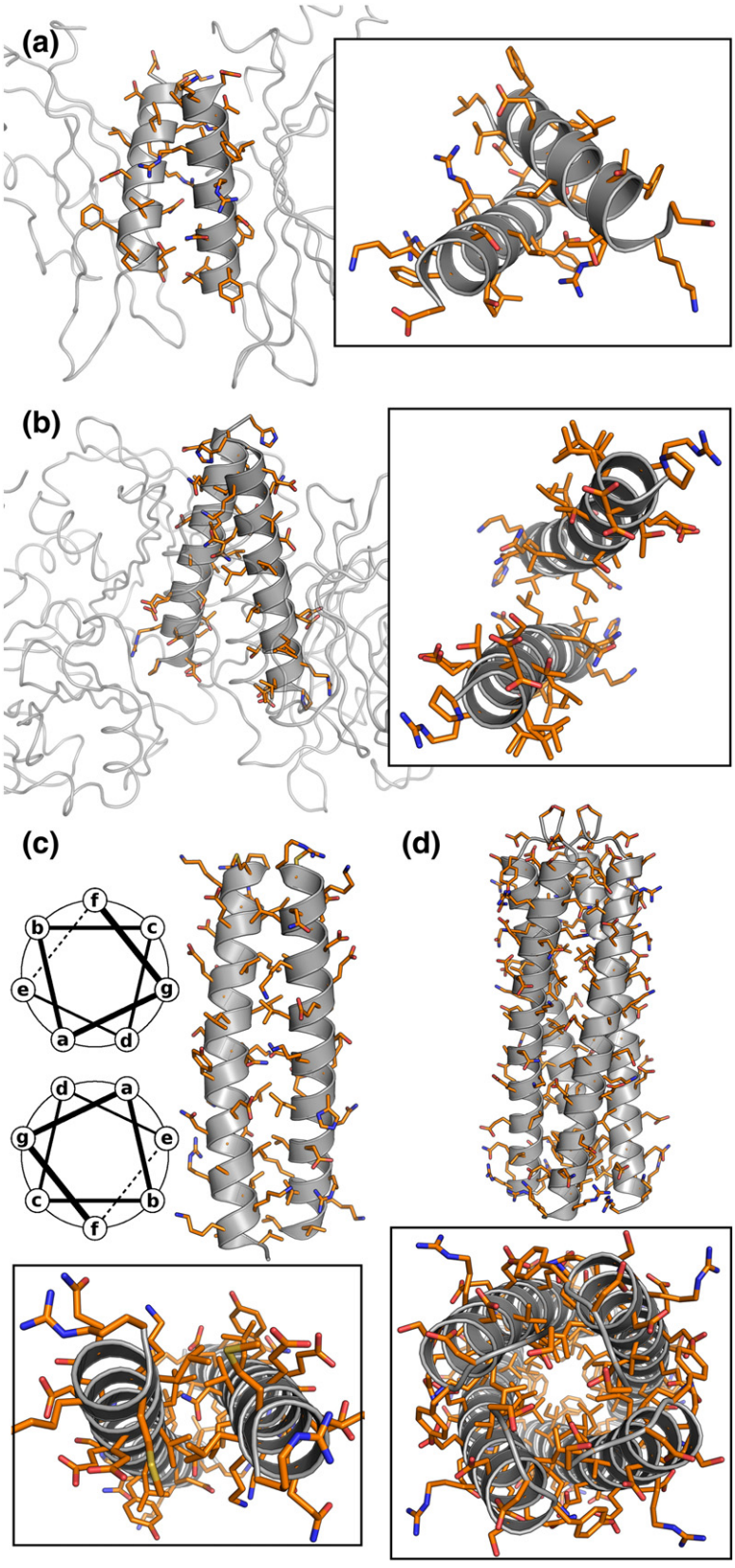

Fig. 1. Common types of helical bundles. Generally, helical bundles form via relatively well-defined and recurring motifs. The classical diverging straight-helix bundle shown in (a) is from the structure *Escherichia coli* PhoQ sensor domain (PDB code 3BQ8). The bundle shown in (b) (from bacteriophytochrome chromophore binding domain, PDB code 2O9C) also diverges, but the individual helices are bent into circular arcs and do not wrap around each other like in coiled coils. This motif can be thought of as a first-order generalization of the straight-helix diverging bundle. Shown in (c) is the left-handed coiled-coil domain from the yeast transcription factor GCN4 (PDB code 2ZTA). The helical wheel diagram on the left illustrates the seven-residue heptad repeat commonly characteristic of left-handed coiled coils. Positions **a** and **d** are in the core and often occupied with hydrophobic amino acids; **e** and **g** are along the interface, often forming salt bridges and polar interactions; and **b**, **c**, and **f** adjust to the environment. Although significantly less common, right-handed coiled coils also occur, such as the example shown in (d) from the human VASP tetramerization domain (PDB code 1USE). Boxed insets represent an alternative view of each bundle.

versus globally over the entire structure. Shorter helical bundles, however, do not necessarily form coiled coils. Indeed, natural divergence of straight helices can form important functional pockets, or the helices can bend or kink in more complex manners. These articulation points can serve as hinges for functional motions involved in signaling, gating, and ion conduction.[37]

As a second goal of this work, we develop a general ideal backbone parameterization approach for arbitrary helical structures. This mathematical framework describes α-helices forming any superstructure defined via a three-dimensional parametric curve, and reduces to the Crick parameterization in the special case when the parametric curve is itself a helix.

The generalized framework provides significant simplification of structural representation while still capturing natural structural variability. We illustrate this using a helical bundle motif common in membrane channels and signaling proteins. This motif consists of helices that curve into nearly perfect circular arcs to form a bundle but do not wrap around each other like coiled coils. Such structures appear as inner and outer bundles lining channel pores (in fact, they are present in virtually all channel families, as defined by the OPM database[38]) and oligomerization interfaces of bacterial signaling proteins. Our parameterization is able to fit such native bundles with C_α RMSD of better than 1.0 Å in 52 of the 55 representative examples analyzed. These findings are particularly significant in light of the functionally important regions of structure these bundles are found in, introducing the potential that functionally relevant structural changes may be modeled and understood in terms of simple geometric parameters.

Results

Coiled-coil structures from the CC+ database[39] longer than 11 residues (see Materials and Methods for details) were analyzed using our automated program CCCP (coiled-coil Crick parameterization) for fitting coiled-coil backbones. The modified Crick equations used in CCCP allow for anti-parallel and mixed orientations and for helical sliding with respect to the interfacial axis (see Materials and Methods for details and Fig. 2 for explanation of Crick parameters). For 95.7% of the analyzed structures, C_α RMSD between ideal and native structures was below 1.0 Å, and the performance did not significantly depend on the orientation or oligomerization state (see Fig. 2 for error distributions). The error did, however, depend on structure size, increasing roughly linearly with the total number of residues fit up to ~0.9 Å. This is not surprising because small local deviations from ideality can lead to large deviations over long distances.

The set of coiled coils with known structure contains a significant amount of redundancy, whereby some sequence families (or, in some cases, many mutants of the same sequence) are significantly overrepresented. To ensure that our analysis was devoid of such biases, we filtered the data set down to less than 50% sequence redundancy, resulting in 868 structures (see Materials and Methods). All further analysis was performed on this data set. The analysis of error distribution was repeated for this subset and resulted in virtually identical trends (see Supplementary Fig. S2).

Superhelical radius

The distribution of superhelical radii is shown in Fig. 3. Several subpopulations are apparent and these correspond to different oligomerization

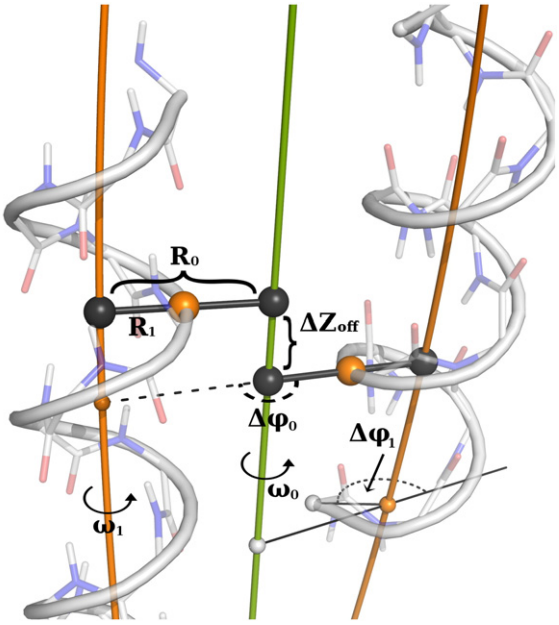

Fig. 2. Visual representation of parameters used in coiled-coil fitting. Geometrical meanings of the superhelical radius (R_0), the helical radius (R_1), the superhelical frequency (ω_0), the helical frequency (ω_1), chain axial offset (ΔZ_{off}), chain superhelical phase offset ($\Delta\phi_0$), and starting helical phase ($\Delta\phi_1$) are shown. The green tube represents the interfacial axis. Orange curves depict local helical axes, which in a coiled coil form a superhelix. The gray tube represents the helical curve, which passes through C_α atoms. Orange balls show the inward-facing points on the helical curves (not necessarily corresponding to locations of atoms) defined as points with helical phase of π. The distance along the interfacial axis between an inward-facing point on one helix and its closest counterpart on the opposite helix is defined as ΔZ_{off}, with sign indicating the order of the two points, relative to N→C of the first helix. The depicted case is an anti-parallel coiled coil with a positive ΔZ_{off} of 2.4 Å.

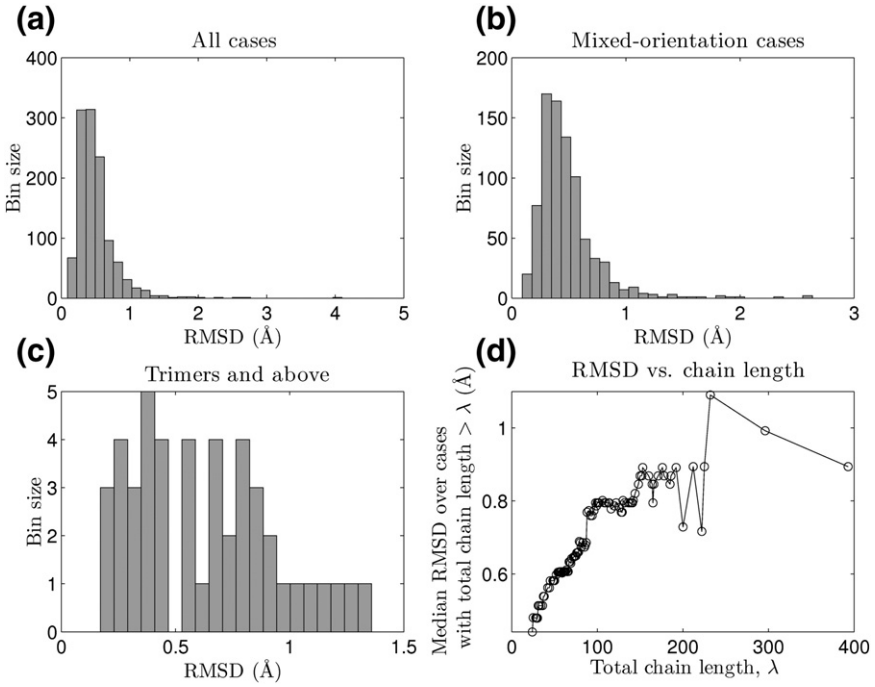

Fig. 3. Distribution of fitting error (C_α RMSD between a natural structure and its Crick parameterized version). (a) Error distribution for the entire dataset. Four structures were fit with RMSD above 2.0 Å, but in 95.7% of cases RMSD was below 1.0 Å. (b) Error distribution for structures with helices of opposing orientation (i.e. anti-parallel dimer and mixed-orientation higher-order oligomers). The performance on these structures is not significantly different from the overall performance. (c) Distribution of fitting error for trimers, tetramers, and pentamers. (d) Fitting error as a function of the total number of residues fit. The Y-axis shows the median C_α RMSD over cases where the total number of residues fit was above the corresponding value on the X-axis.

states (see Fig. 2b–d). The medians of the subpopulations (4.85, 6.36, 7.30, and 8.59 Å for dimers, trimers, tetramers, and pentamers, respectively) correlate nearly perfectly with oligomerization state with the best fit line $R_0 = 1.24 \cdot n + 2.4$ (RMSD 0.1 Å; see Supplementary Fig. S3), where n is the oligomerization state. This indicates that the interhelical distance, which can be expressed as $(1.24 \cdot n + 2.4)\sqrt{2 - 2\cos(\frac{2\pi}{n})}$ (limiting value of $1.24 \cdot 2 \cdot \pi \approx 7.8$ Å), decreases with increasing oligomerization state above the dimer.

One of the restrictions of the Crick parameterization is that the superhelical radius (R_0) is assumed to be constant throughout the structure. We wondered to what degree R_0 actually changes in natural structures. To that end, for dimers longer than three heptads, we compared the value of R_0 obtained by fitting the entire structure to values obtained by fitting individual two-heptad windows scanned across the structure (see Supplementary Fig. S4). In most cases, local and global R_0 coincide closely (median deviation is 0.1 Å); however, exceptions to this do exist, especially for long structures (Supplementary Figs. S4 and S5). The highest observed rates of superhelical radius change are around 0.14 Å per residue. The median autocorrelation length of R_0 is seven residues (defined in Materials and Methods), indicating that the structure has some memory of the superhelical radius for about a heptad.

Even though variations of R_0 within an oligomerization state are small, we wondered whether these correlate with amino acid composition of core **a** and **d** positions. To test this, we considered two simple models for the dependence of R_0 on the core composition. In the first, superhelical radius depends linearly on the composition of **a** and **d** positions, and in the second, amino acids with the largest radius preferences determine the value of R_0 (see Materials and Methods for details). Fitting local R_0 values to either model leads to a weak agreement such that roughly 20% of the observed superhelical radius variability can be attributed to either of these simple expressions (see Supplementary Fig. S6). Sequence context likely plays a significant role in determining which aspects of the local coiled-coil geometry are perturbed upon changes in the identity of core amino acids (e.g., axial sliding of chains, changes in side-chain rotameric states, or adjustments in helical and superhelical phases).

On the other hand, amino acids at **a** and **d** may still have pronounced radius preferences, even though **a**/**d** composition alone may not amount to

a quantitatively predictor of local R_0 variation. To measure these preferences, we considered a subset of the dimeric two-heptad segments described above, in which either position **a** or position **d** was in the center and which had parallel orientation. The superhelical radii of these segments were then studied in relation to the amino acids occupying the central **a** or **d** position. All segments were binned by value of R_0. Next, for each amino acid occurring at either the central **a** or **d** positions, the frequency of its occurrence in each bin was calculated and normalized by the overall frequency of the amino acid in the coiled-coil database, to arrive at amino acid specific propensities for different radii. To minimize noise from statistics of small numbers, for a given amino acid aa_i, bins with very few occurrences of

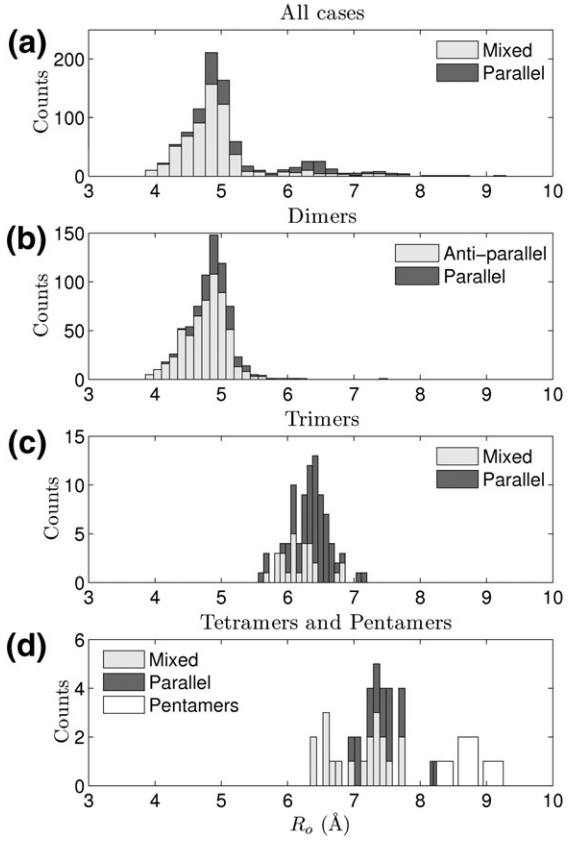

Fig. 4. Distribution of superhelical radius. Stacked bar plots show the overall distribution of R_0 (total histogram) as well as contributions to it from structures of parallel orientation (dark gray bars) and anti-parallel or mixed orientation (light gray bars). (a) The distribution of R_0 is multi-modal with peaks corresponding to different oligomerization states. (b) and (c) show the distributions of R_0 for dimers and trimers, respectively, and (d) combines tetrameric and pentameric cases (of the four pentameric structures, all are of parallel orientation). The median values of the distributions are 4.85 Å for dimers, 6.36 Å for trimers, 7.30 Å for tetramers, and 8.59 Å for pentamers.

aa_i were grouped with neighboring bins and the average number of counts was assigned to each bin of the group when computing propensities for aa_i. Figure 4 shows the resulting propensities for the top 10 most frequent amino acids at either the **a** or **d** position. In these plots, amino acids are ordered, left to right and top to bottom, in the order of decreasing frequency of occurrence at the respective position (e.g., the first four most frequent at **a** are Leu, Ile, Val, and Asn, in that order).

At the **a** position, Leu appears to be the most malleable amino acid as it is widely accommodating of a large range of radii. Leu is also the most common amino acid at this position, which, in part, helps explain why the precise local deviation of the superhelical radius from its characteristic value cannot be easily predicted purely from the composition of core amino acids. Val, Asn, and Ile, by contrast, appear much more specialized for particular radius ranges, especially pronounced in the case of Asn and likely due to its strong preference to pair with another Asn at **a**′. Ala, not surprisingly, shows an extremely strong preference for small radii, and so do Lys and Arg, though with a considerably lower magnitude of the preference. As the number of observations of a particular amino acid decreases, statistical error grows, but the preference of Tyr and Met for larger radii is consistent with their sterics.

Like at the **a** position, at **d** Leu occurs in a wide range of radii and Ala shows strong preference for small radii (see Fig. 4b). The amino acid alphabet is much more restricted at **d**, compared to **a**, with Leu and Ala amounting to 60% of all cases (whereas it takes the top four amino acids at **a** to get to 60%), and thus statistics fall off rapidly after Ala, as indicated by larger error bars in Fig. 4b.

Superhelical frequency and pitch angle

Superhelical frequency (ω_0) determines the degree of twist of the superhelix. That is, for every residue, the angle by which the superhelix turns around the coiled-coil axis is ω_0. Values around −3.6 °/res are considered canonical (the negative sign indicates a left-handed superhelix) such that the superhelix completes a full turn in 100 residues. This parameter is tied to the superhelical radius via the relationship:

$$R_0 \cdot \omega_0 = d \sin(\alpha) \quad (1)$$

where ω_0 is in radians, α is the pitch angle, and d is the rise per residue. The latter is the arc length along the superhelical curve from one residue to the next. In a straight helix that is aligned along the Z-axis, d is just the displacement in the Z direction from one residue to the next. This is roughly 1.51 Å for an

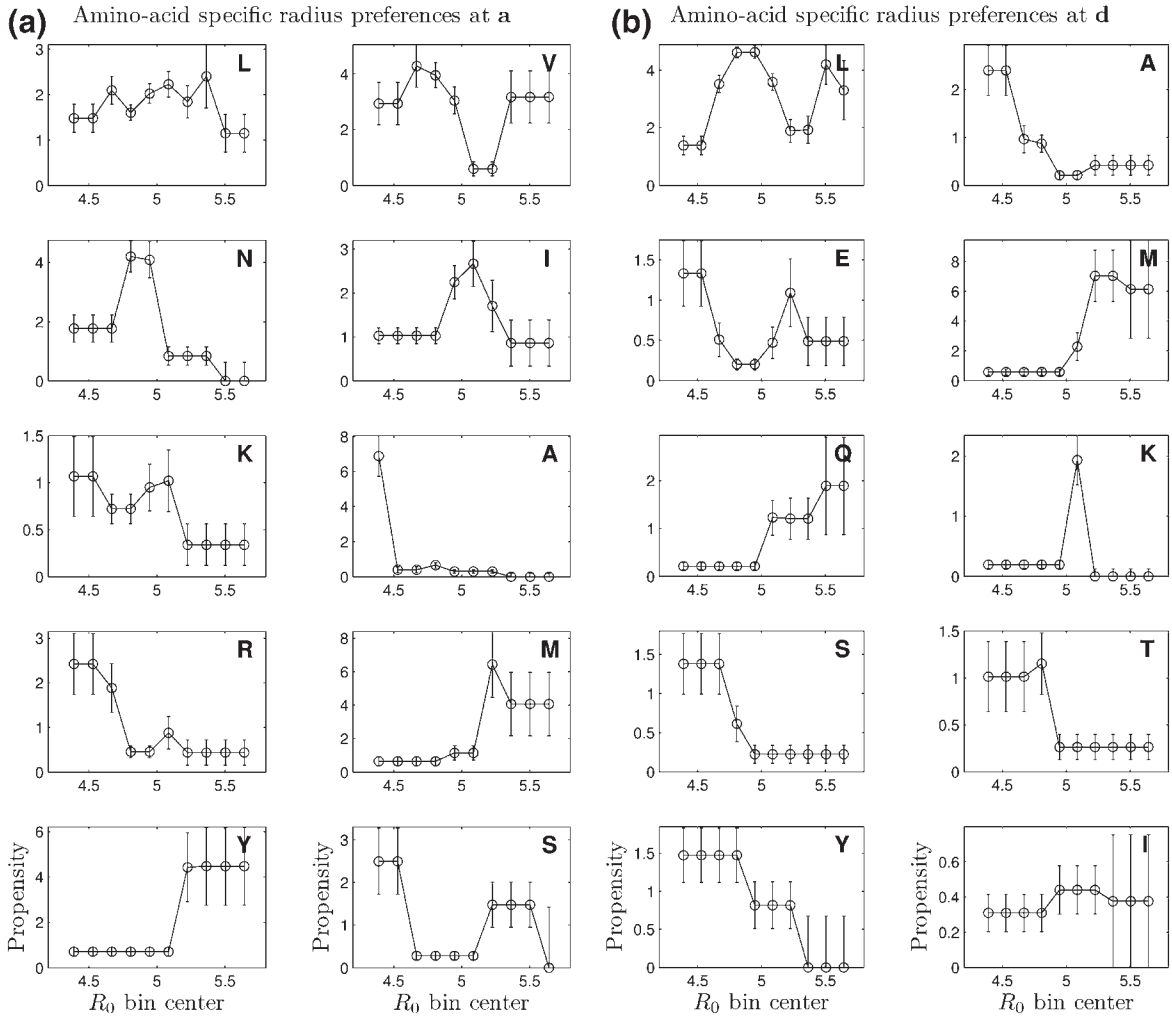

Fig. 5. **a**/**d**-position amino acid superhelical radius propensities. Two-heptad windows of dimeric structures were considered in which **a** (in panel a) or **d** (in panel b) positions were in the center. Values of R_0 obtained by Crick parameter fitting were divided into 10 equally spaced bins. For every amino acid aa_i, values plotted as a function of superhelical bin b_j are $\frac{N(aa_i,b_j)}{N_{\text{tot}} \cdot f_{b_j} \cdot f_{aa_i}}$, where $N(aa_i, b_j)$ is the number of occurrences of aa_i in bin b_j, N_{tot} is the total number of data points, f_{aa_i} is the frequency of aa_i in the entire data set, and f_{b_j} is the fraction of structures with a superhelical radius within bin b_j. To correct for low counts, bins with the number of counts of a particular amino acid below 3 were grouped with neighboring bins until the total number of counts became above 3, and the average number of counts over the group was assigned to each bin within the group. Error bars designate the error of determination of each propensity and were computed by considering all possible binomial distributions that could have given rise to the observed number of successes $k = N(aa_i, b_j)$ out of the number of trials $N = N_{\text{tot}} \cdot f_{b_j}$ and by calculating the error expectation as $\sqrt{\int_0^1 (p-\bar{p})^2 P_b(k,N,p)dp}$, where p is the probability of success of the considered binomial distribution, $P_b(k, N, p)$ is the binomial probability density of k successes from N trials and a success probability of p, and $\bar{p}$ is the expectation of p given k successes and N trials, or $\int_0^1 p \cdot P_b(k,N,p)dp = \frac{k+1}{N+2}$.

ideal α-helix. Superhelical parameters that result in d significantly different from its ideal value correspond to cases where the α-helix is locally strained (e.g., stretched or compressed). This places restraints on the product $R_0\omega_0$, such that for any given radius, the absolute value of ω_0 can vary between 0 (corresponding to straight aligned helices) and roughly 1.51 R_0^{-1} (in radians per residue).

Figure 5a shows the distribution of ω_0 observed in the structural data set. The range of absolute values is between 0 and 8.7 °/res and the median of the distribution is −3.5 °/res, meaning that the most tightly wound coiled coil has a period of about 41 residues (since the period is $2\pi\omega_0^{-1}$ residues) and the average coiled coil completes its period in roughly 102 residues. As expected, the overwhelming majority of annotated coiled coils are left-handed (negative ω_0). In fact, only seven unique structures in the CC+ database showed a significant overall right-handed character, among which are RH4 and

RH4B—coiled coils designed to be right-handed.[18,40] It was formally possible that the number of right-handed structures in CC+ was unfairly reduced due to a possible bias in the program SOCKET[41] for recognizing the type of knobs-into-holes packing more characteristic of canonical left-handed structures. We therefore performed C_α distance matrix-based structural similarity searches of the entire Protein Data Bank (PDB) (see Materials and Methods for details) as well as keyword-based searches of the PDB and the literature to look for additional right-handed coiled-coil structures. In so doing, we uncovered five more clear examples of right-handed bundles with supercoiling and included these in all of our analyses. Although none of the above methods likely recover all available examples of right-handed coiled coils, the significantly lower rate of discovering right-handed *versus* left-handed coiled coils in the structural database argues that right-handed coiled coils are significantly less common and likely less designable. Within the structures annotated in CC+, there are examples of local right-handed character in coiled-coil packing (such as in a fragment of the *Staphylothermus marinus* surface layer protein tetrabrachion[42]), whereas Fig. 5a shows values of ω_0 optimal for entire coiled-coil regions. In a separate set of fits, performed to study the local variation of parameters, we considered two-heptad windows that were scanned along all dimeric coiled coils. This resulted in over 3000 separate two-heptad fragments, out of which 4.8% best fit with a positive value of ω_0.

Figure 5b shows the observed distribution of pitch angle α, which is the angle between a tangent to the superhelical curve and the coiled-coil axis. Once again, this is not a completely free parameter and is related to R_0 and ω_0 via Eq. 1. The peak of the distribution is at around 12 °. For both pitch angle and superhelical frequency, the peak values are very close to the values that best fit the coiled-coil region of the yeast transcription factor GCN4[43]—the first high-resolution structure of a parallel dimeric coiled coil as well as one that has long served as the prototypical coiled-coil example. It is interesting that in Crick parametric terms, this structure is indeed prototypical.

Helical axial shift

As outlined in Materials and Methods, we modified the classic Crick parameterization to allow for a helical sliding along the interfacial axis. Thus, in a given structure, each helix after the first one is assigned a value of ΔZ_{off} relative to the first helix. ΔZ_{off} between two helices is defined as the distance along the interfacial axis between the most inward-facing points on the helical curves of the two helices (not necessarily representing C_α atoms; see Fig. 6 and Materials and Methods). This definition makes ΔZ_{off} independent of helical phase, rendering it a convenient parameter for sampling and fitting. Interestingly, the distribution of ΔZ_{off} strongly depends on whether the helix pairing in question is parallel or anti-parallel (see Supplementary Fig. S7). Whereas for parallel alignments ΔZ_{off} tends to cluster tightly around 0, corresponding to helices with turns *Z*-aligned at the interface, for anti-parallel alignments, the distribution is much more broad and has two maxima at around ± 2.0 to 2.5 Å, corresponding to helices with interdigitated turns (the sign of ΔZ_{off} indicates direction of the axial shift, with the positive sign corresponding to the direction shown in Fig. 6).

To allow for an easier interpretation of axial offset in terms of coiled-coil geometry, CCCP additionally calculates the parameter $\Delta Z_{aa'}$ (not used in the fitting), defined as the axial offset between an **a** position on one helix and the closest **a**′ position on the interacting helix. Sign of $\Delta Z_{aa'}$ indicates whether the closest **a**′ is *downstream* (e.g., in the N→C direction of the first chain, corresponding to a positive sign) or *upstream* (e.g., opposite to N→C direction of the first chain, corresponding to a negative sign). Given its definition, $\Delta Z_{aa'}$ can vary

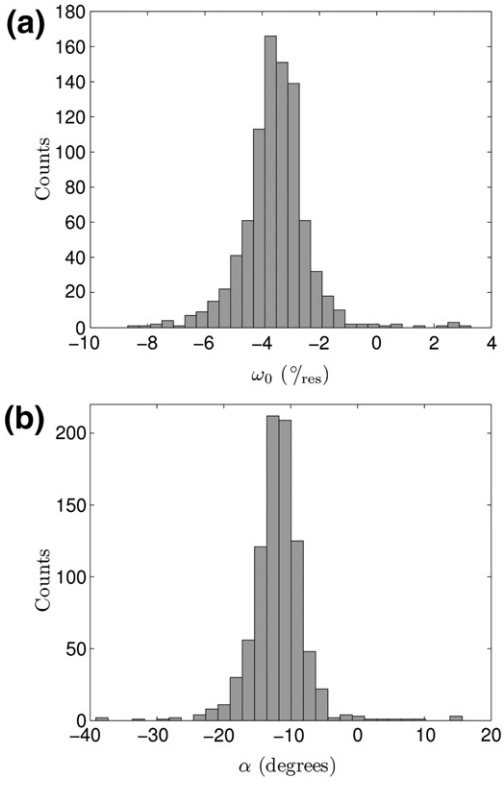

Fig. 6. (a and b) Distributions of superhelical frequency (ω_0) and pitch angle (α), respectively. The median and mean values for ω_0 are −3.5°/res and −3.6°/res, respectively, corresponding to a superhelical repeat of about 100 residues or pitch of approximately 150 Å. The median and mean values for α are −11.9° and −12.0° , respectively.

between plus and minus half a helical pitch, roughly $\left[-\frac{2d\cdot\pi}{\omega_1}, \frac{2d\cdot\pi}{\omega_1}\right]$ (where ω_1 and d are the helical frequency and rise per residue, respectively) or [−5.3,5.3]Å for a typical coiled coil. Figure 7 shows that the distribution of $\Delta Z_{aa'}$ strongly depends on the mutual orientation of the pair of helices. Parallel coiled-coil helix pairs prefer to line up opposing **a** positions with a roughly zero axial offset, whereas anti-parallel pairs prefer to interdigitate. Parameter $\Delta Z_{aa'}$ further tells us that anti-parallel coiled coils interdigitate in an asymmetric manner with respect to the heptad, such that a given **a** position is usually closer to the **a′** residue downstream of it rather than the upstream **a′** residue. In fact, a very large range of $\Delta Z_{aa'}$ values, roughly from −5 to 1 Å is unpopulated, narrowing the designable space of anti-parallel coiled-coil backbones. It is common to think of anti-parallel coiled coils as roughly aligning **a** positions of one chain with **d′** positions of the opposing chain. The distribution of the axial offset between **a** and the closest **d′** is shown in Supplementary Fig. S8. It is clear that in anti-parallel pairs, **a** and the opposing **d′** positions are rarely exactly aligned, and the closest **d′** tends to be upstream of **a**, with only a minority of structures displaying the opposite alignment.

Minor helical phase φ_1 and frequency ω_1

Helical frequency ω_1 characterizes the angular rotation of the α-helix around its local axis with each residue. For an ideal straight helix, this value is 100 °/res, such that a full turn of the α-helix is completed in 3.6 residues. In a canonical coiled coil, ω_1 is around 102.8°/res ; thus, a complete turn is made in 3.5 residues and two turns in 7 residues. This gives rise to the well-known heptad repeat, commonly denoted with positions **a** through **d** (see Fig. 1c). In agreement with this, the distribution of ω_1 within the coiled-coil structures analyzed amounts to a normal distribution with a mean of 102.8 °/res and an SD of only 1.1 ° (data not shown). Variations in parameter ω_1 to a significant extent reflect changes in ω_0 (i.e., changes in the reference frame defining a complete helical turn), as opposed to reflecting changes in α-helical geometry (ω_1 and ω_0 anti-correlate with a correlation coefficient of $R=-0.7$).

Parameter ϕ_1 in Crick equations (see Materials and Methods) measures the starting angular register, or helical phase, of the α-helix, and the helical phase of the i-th residue is given by $\phi_1+(i-1)\,\omega_1$. Phase values of 0 and π mean that the C_α atom of the residue points directed away from or towards the interfacial axis, respectively. Phase can be used to determine the heptad position assignment of each residue and CCCP does so automatically. Figure 8 shows the distribution of helical phases for residues in all dimeric coiled coils considered in this study. Empty and filled circles show residue phases as the angular coordinate in the polar-coordinate plot, with each

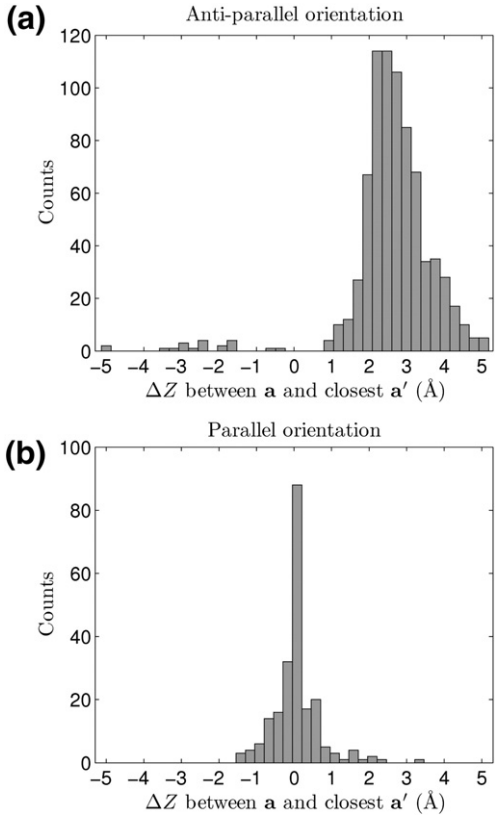

Fig. 7. Distribution of $\Delta Z_{aa'}$. Helix pairs from complexes of all oligomerization states and orientations were considered. The distribution of $\Delta Z_{aa'}$ is shown for (a) anti-parallel pairs and (b) parallel pairs.

point corresponding to a single residue within a single structure. Empty circles represent anti-parallel structures, while filled circles represent parallel structures.

The continuous line in Fig. 8 shows the histogram of helical phase distribution. The seven clearly visible peaks are centered around 41°, 95°, 146°, 197°, 249°, 300°, and 351°, corresponding to heptad positions **c**, **g**, **d**, **a**, **e**, **b**, and **f**, respectively (the helix in the figure is shown in the C→N projection). Although actual phase values can vary around these numbers, the peaks of the distribution are fairly narrow with SDs around 8 °, indicating that heptad positions correspond to well-pronounced minima in the coiled-coil energy landscape.

Overall restriction of designability

The requirements for achieving physically allowed coiled-coil backbone geometries are a reasonable superhelical radius and superhelical phase offsets, such that main-chain clashes do not occur, and the validity of Eq. (1) with a near-native rise per residue. The distributions of native Crick parameters are clearly significantly restricted beyond this filter, and we argue that this remaining

restriction reflects varying designabilities of different structures. To capture this restriction quantitatively, we compared the observed distributions of Crick parameters with uniform random distributions over the same ranges and derived the effective reduction in the number of states in these distributions (see Materials and Methods for the procedure). Considering parameters R_0, α, ϕ_1, and ΔZ_{off} in dimers, we found that the number of states in the corresponding distributions was restricted by factors of 3.8, 4.7, 2.3, and 4.4, respectively. Differences in state-space reduction between parallel and anti-parallel dimers were small, except for ΔZ_{off}, where the reduction factor was 4.7 for parallel dimers and 4.1 for anti-parallel dimers. Therefore, we estimate that the overall reduction of naturally designable space of dimeric coiled coils, relative to physically reasonable geometry space, is at least 160- to 200-fold. This is likely an underestimation of the true restriction because (1) we are ignoring the existing correlations between the analyzed parameters in the space of designable structures; (2) the true range of physically feasible structures extends beyond the limits of naturally observed spread of parameter values; and (3) we disregarded the superhelical phase offset parameter, while it is nearly constant for dimers and extremely restricted for higher-order oligomers.

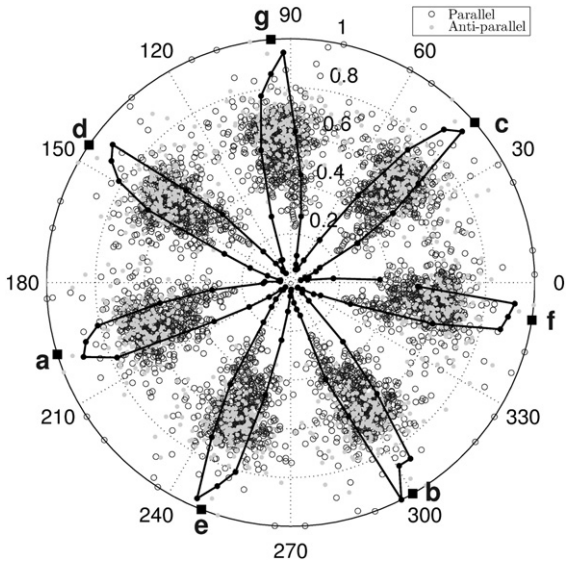

Fig. 8. Distribution of minor-helical phase ϕ_1 in polar coordinates for dimers. Open and filled circles represent parallel and anti-parallel structures, respectively, with each point corresponding to a single residue within a single structure (to avoid redundancy, only the first seven residues in each chain of each structure were considered). The helical phase is denoted by the angular coordinate and the radial coordinate is proportional to the superhelical radius R_0 (normalized for ease of viewing). No detectable correlation between R_0 and phase exists. The continuous line is a histogram of ϕ_1 distribution (for parallel and anti-parallel orientations combined) constructed with bins of width 3.6° . In this case, the radial coordinate represents frequency of each bin. Filled squares denote the means of the seven peaks of the distribution, and the corresponding heptad positions are indicated in bold letters.

Other parametric structures

The Crick parameterization describes an ideal helical backbone wrapping into a superhelix. However, mathematically, any superstructure is possible. In Materials and Methods, we detail a mathematical framework that describes a helix wrapping around an arbitrary parametric curve. This framework can be thought of as a generalization of the Crick parameterization and it reduces to the latter in the case when the parametric curve is itself a helix. The Crick parameterization is useful because many helices in natural structures do curve into superhelices. However, there are other ways in which helices bend in structures, and whenever one observes a common mode, either within a family of structures or more generally, a parameterization can be quite helpful in modeling of natural proteins, design of novel structures, and understanding structural aspects of function.

There are many helical bundles in transmembrane (TM) and near-TM regions of membrane proteins, yet not many of them form coiled coils. We found another common mode of helical bundle formation here, wherein helices that are bent into essentially perfect circular arcs form bundles but do not wrap around each other (see Fig. 1b). We used our generalized parameterization approach to derive a mathematical formulation for such bundles (see Materials and Methods). The parameters of the formulation and their geometric meaning are shown in Fig. 9.

This motif is particularly common in TM and near-TM regions of channel proteins and bacterial signaling proteins. We manually inspected all representative structures of channel proteins in the OPM database[38] and found one or several such arched bundles in all channel superfamilies and in virtually all families (see Table 1 for a list of representative structures). As with the Crick parameterization, arched bundles can be fit with their ideal counterparts rather well, in most instances producing C_α RMSD of less than 1 Å (see Table 1).

Some general trends are already apparent from the limited data on arched-bundle preferred parameters. The degree of helix bending in these bundles can be rather severe, with radii of curvature as low as 23 Å and an average helix diverging by ~23° from its initial direction over 20 residues, the length of a typical TM segment. Such range of bending is similar to what has been observed in water-soluble proteins.[44,45] At the same time, it has been argued from physical considerations that, in general,

membrane-inserted helices should be more regular than their soluble counterparts, and NMR-based mapping of torsion angles of the influenza A M2 proton channel has been used to provide experimental evidence for this.[46] The instances of significant bending we observe in arched bundles may thus be unusual, and this may be of significance in light of the functionally important regions these bundles are found in.

Although we observe a large range of bundle radii in arched bundles, even within a single oligomerization state (average SD within an oligomerization state ~5.3 Å), mean radius, like with coiled coils, follows a roughly linear relationship with the number of chains (see Supplementary Fig. S9). Arc turn angle, which characterizes the rotation of the plane of the arc relative to the bundle axis, is distributed around 90° (see Supplementary Fig. S10), indicating that, most often, the helices bend in a direction roughly tangential to the channel pore. The sign of the pitch angle parameter (see Fig. 9) indicates whether the helix crossing is left-handed or right-handed, corresponding to negative and positive pitch angles, respectively. Based on this, 18 of the analyzed bundles are left-handed and 37 right-handed (see Table 1), an interesting contrast with the coiled coil.

Discussion

Effective low-dimensionality descriptions of protein structure are very desirable as they simplify the task of establishing relationships between structure and function or other properties. For example, understanding biological signaling as it relates to changes in structural states of proteins has been effective in terms of gross structural rearrangements.[47–49] In structure prediction and design, low-complexity representations of structure limit the space of potential templates and can help elucidate sequence-based preferences for various structural environments.[50,51] Simple classification schemes such as hydrophobicity scales or α-helix/β-sheet propensities represent examples of this, and so do more complex approaches such as identification of sequence motifs encoding specific geometries of secondary-structure packing.[32,51]

Another task simplified by the use of reduced structural representations is that of relating structure to designability—the size of sequence space compatible with a given structure. This is a task of great significance in protein design and particularly *de novo* design, where backbone templates cannot be borrowed from existing structures and one must

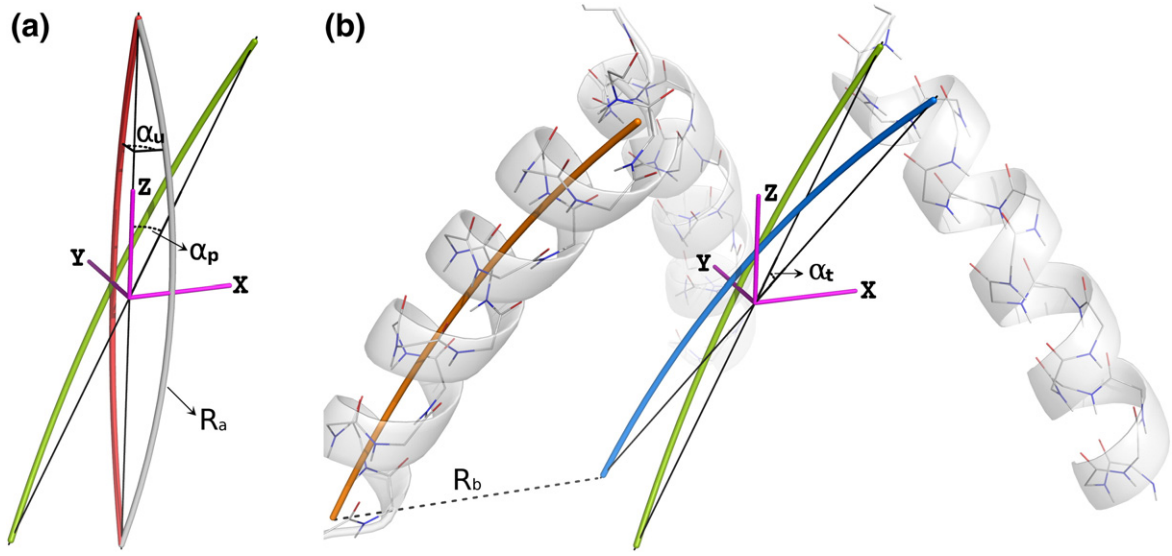

Fig. 9. Parameterization of arched bundles. The geometric meaning of the parameters involved in the arched bundle parameterization can be best understood by considering the transformations necessary to build such a bundle. (a) The tube shown in gray is the initial bent helical curve with a radius of curvature of R_a around which an α-helix is to wind. The *Z*-axis is the principal bundle axis, and the first chain is taken to lie in the *XZ* plane. α_u, the arc turn angle, is the angle by which the helix arc planes of individual monomers are turned with respect to the primary axis in the eventual bundle. Thus, to dial in this parameter, the initial curve is rotated around the *Z*-axis by α_u, resulting in the orange curve. Bundle pitch is established by rotating the new helical curve around the *Y*-axis by the pitch angle α_p, resulting in the curve shown as a green tube. (b) Bundle tilt is created by rotating the helical curve around the *X*-axis by the tilt angle α_t, resulting in the helical curve show as a blue tube. The helical curve is translated along the *X*-axis by an amount equal to the bundle radius R_b and a helix is wound around the resulting curve given a particular helix starting phase ϕ_1 (orange tube and gray helix). Other members of the bundle are either generated using ideal phase offsets of $\frac{2\pi}{n}$, where *n* is the number of chains or, alternatively, phase offsets are assigned individually for each chain after the first one.

Table 1. List of representative structure with arched bundles

PDB ID and Region[a]	N[b]	R_a, Å (ε,°)[c]	R_b, Å[d]	α_t,°[e]	α_p,°[f]	α_u,°[g]	ϕ_{eff},°[h]	Err, Å[i]	Annotation
1NKZ, outer bundle	9	93.9 (14.7)	30.6	−8	13	−115	84	0.34	light-harvesting complex LH2
1NKZ, inner bundle	9	93.8 (11.1)	16.9	−1	4	−150	−106	0.29	light-harvesting complex LH2
1IJD, outer bundle	9	83.4 (15.5)	30.7	−8	13	−116	−176	0.36	light-harvesting complex LH3
1IJD, inner bundle	9	94.5 (10.1)	17.4	−1	5	−136	−13	0.28	light-harvesting complex LH3
3BEH, inner bundle	4	87.5 (11.9)	11.3	29	22	53	60	0.37	bacterial cyclic nucleotide-regulated channel MlotiK1
1R3J, inner bundle	4	234.9 (4.9)	10.2	27	23	−85	74	0.36	potassium channel KcsA
1R3J, outer bundle	4	51.3 (20.1)	17.0	−14	15	118	−24	0.24	potassium channel KcsA
1S5H, outer bundle	4	50.4 (20.3)	16.9	−14	16	118	−26	0.23	potassium channel KcsA
1S5H, inner bundle	4	240.7 (4.8)	10.1	27	23	−88	74	0.36	potassium channel KcsA
1S5H, selectivity filter bundle	4	53.5 (10.4)	13.4	47	−8	−95	145	0.18	potassium channel KcsA
3EFF, outer bundle	4	96.6 (10.5)	19.1	−5	19	88	−83	1.26	potassium channel KcsA, full-length, closed state
3EFF, inner bundle	4	374.1 (2.7)	12.0	27	22	−13	74	0.56	potassium channel KcsA, full-length, closed state
2A79, inner bundle	4	42.5 (18.2)	13.2	26	26	70	−179	0.38	potassium channel Kv1.2
2A79, outer bundle	4	64.2 (10.6)	18.4	−11	38	−164	73	0.29	potassium channel Kv1.2
2A79, selectivity filter bundle	4	41.7 (13.7)	13.8	49	−8	−57	150	0.32	potassium channel Kv1.2
2R9R, outer bundle	4	66.1 (13.6)	17.9	−9	34	−150	−25	0.23	K+ channel Kv1.2, membrane-like environment
2R9R, inner bundle	4	53.6 (15.3)	13.4	25	25	64	73	0.31	K+ channel Kv1.2, membrane-like environment
2R9R, selectivity filter bundle	4	23.9 (24.8)	14.1	51	−9	−38	136	0.52	K+ channel Kv1.2, membrane-like environment
1ORQ, inner bundle	4	41.7 (35.0)	8.0	−5	52	99	20	0.81	K+ channel KvAP
2A0L, outer bundle	4	70.2 (14.8)	19.5	−2	35	−78	−89	0.43	K+ channel KvAP
2A0L, inner bundle	4	36.4 (34.0)	9.1	9	47	105	36	0.65	K+ channel KvAP
3E86, outer bundle	4	138.7 (5.9)	19.5	−1	21	−118	133	0.21	K+ channel, open state
3E86, inner bundle	4	23.3 (40.1)	12.2	17	36	96	164	0.54	K+ channel, open state
3E86, selectivity filter bundle	4	189.1 (2.9)	12.9	45	−1	−118	161	0.17	K+ channel, open state
2AHY, outer bundle	4	149.6 (5.3)	19.0	−5	13	107	123	0.24	K+ channel, closed state
2AHY, inner bundle	4	99.6 (9.1)	12.7	21	16	94	173	0.28	K+ channel, closed state
2AHY, selectivity filter bundle	4	104.0 (5.3)	12.9	44	−17	−84	166	0.19	K+ channel, closed state
2QKS, outer bundle	4	169.7 (6.1)	17.7	−6	40	−131	86	0.24	Kir3.1-prokaryotic Kir channel chimera
2QKS, inner bundle	4	95.4 (13.4)	10.0	17	28	77	61	0.26	Kir3.1-prokaryotic Kir channel chimer
1P7B, outer bundle	4	396.8 (2.4)	17.9	−4	41	−177	−10	0.33	K+ channel Kirbac1.1
1P7B, inner bundle	4	115.4 (9.2)	11.4	25	22	68	71	0.45	K+ channel Kirbac1.1
1XL6, outer bundle	4	237.2 (4.2)	18.4	−5	40	−48	91	0.28	potassium channel Kirbac3.1
1XL6, inner bundle	4	96.0 (13.2)	10.5	18	28	40	59	0.38	potassium channel Kirbac3.1

2OAU, TM3 bundle	7	73.7 (15.5)	25.8	−26	25	136	−108	1.92	mechanosensitive channel MscS, expanded state
2OAU, TM2 bundle	7	131.3 (7.7)	24.5	32	19	64	158	2.01	mechanosensitive channel MscS, expanded state
2OAU, TM1 bundle	7	116.7 (5.9)	10.6	1	26	76	72	0.82	mechanosensitive channel MscS, expanded state
2VV5, TM3 bundle	7	91.0 (12.8)	27.7	−11	49	149	−125	0.34	mechanosensitive channel MscS, open state
2VV5, TM2 bundle	7	122.8 (8.4)	23.3	18	32	61	175	0.37	mechanosensitive channel MscS, open state
2VV5, TM1 bundle	7	58.3 (11.6)	13.4	−8	6	109	87	0.24	mechanosensitive channel MscS, open state
3EAM, inner bundle	5	194.2 (5.1)	9.6	−6	−1	43	−139	0.39	ligand-gated ion channel, open conformation
2RLF, TM bundle	4	109.6 (7.9)	7.6	−4	−16	32	−162	0.35	M2 channel TM domain, closed state
3C9J, TM bundle	4	662.0 (1.3)	8.5	−20	−31	91	179	0.47	M2 channel TM domain, open state, complex with amantadine
3H9V, inner bundle	3	42.8 (27.0)	9.7	−33	−42	−20	−107	0.61	ATP-gated P2X4 ion channel, the closed
2O9B, GAF domain helices α8	2	121.7 (8.3)	6.3	7	−11	−61	−170	0.90	bacteriophytochrome chromophore binding domain
2O9C, GAF domain helices α8	2	113.4 (8.9)	6.3	7	−10	−65	−170	0.90	bacteriophytochrome chromophore binding domain
2O9C, GAF domain helices α4	2	38.9 (16.8)	8.0	−23	3	158	26	0.51	bacteriophytochrome chromophore binding domain
3E98, dimer interface helices	2	37.1 (22.1)	10.0	6	−6	−40	−155	0.66	GAF domain of unknown function
3E98, dimer interface helices	2	121.2 (6.0)	6.6	−3	−3	78	−34	0.30	GAF domain of unknown function
2HGV, dimer interface helices	2	131.6 (5.8)	7.0	3	4	−3	−77	0.24	GAF domain of transcriptional repressor CodY
2HGV, dimer interface helices	2	58.9 (11.7)	8.6	−9	−2	−112	−30	0.30	GAF domain of transcriptional repressor CodY
1IXM, dimer interface helices	2	77.2 (13.5)	5.9	−3	7	−88	−8	0.36	*B. subtilis* Spo0B phosphotransferase
1LIH, dimer interface helices	2	63.3 (17.8)	3.9	−1	−9	154	155	0.28	Bacterial aspartate receptor ligand-binding domain
2LIG, dimer interface helices, Asp-bound form	2	64.4 (17.6)	4.0	−1	−8	157	150	0.29	Bacterial aspartate receptor ligand-binding domain
1LIH, dimer interface outer helices	2	170.0 (5.8)	10.9	4	−11	−172	49	0.19	Bacterial aspartate receptor ligand-binding domain
2LIG, dimer interface outer helices, Asp-bound form	2	171.9 (5.7)	11.1	4	−9	180	49	0.32	Bacterial aspartate receptor ligand-binding domain

[a] Supplementary Table S3 lists the regions fit within each OPM entry.
[b] Number of chains.
[c] Arc radius with arc incident angle given in parentheses.
[d] Bundle radius.
[e] Tilt angle.
[f] Pitch angle.
[g] Arc turn angle.
[h] Effective helix phase ($\phi_1 + \alpha_u$).
[i] Fitting C_α RMSD (Å).

limit oneself to targeting only templates of reasonable designability. Here, we demonstrate the use of structural parameterization for describing designability, starting with the Crick parameterization of the α-helical coiled coil. We show that for over 95% of annotated coiled-coil structures, an ideal parameterized backbone can be found within 1. 0-Å C_α RMSD of the native structure. In fact, we find that instead of the very large space of possible parameter value combinations, natural structures populate only very restricted areas in parameter space. For example, characteristic values of the superhelical radius grow linearly with the number of chains, and the amount of variation within each oligomerization state is around ±1 Å (an SD of ~0.5 Å; see Supplementary Fig. S3). This means that given the desired oligomerization state to be targeted in design, reasonably designable templates have superhelical radii in a narrow range. Further, within this range, certain amino acids, when occurring at core **a** and **d** positions, have well-defined preferences for certain radius values, whereas others are fairly forgiving (see Fig. 4). In the latter category is Leu at the **a** position, the most common residue in this position of coiled coils. This makes Leu a very important amino acid for the designability of coiled coils, without which the space of designable structures would be much more restricted. Superhelical radius can also change significantly within a single coiled coil (see Supplementary Figs. S5 and S4). In fact, the radius can vary by as much as 0.14 Å per residue, amounting to a contraction or expansion of 0.7 Å over a single heptad. The ability of coiled coils to adopt different local parameters can be a useful feature in design.

It is well known that amino acid preferences at **a** and **d** are different,[52] and we have now also shown that the ability of amino acids to accommodate different-sized supercoils also differs at these two positions (compare Fig. 4a and b). Indeed, **a** and **d** are also nonsymmetric in the sense of helical phase, which is apparent from Fig. 8. On average, the C_α atom of an **a** position is closer to the interhelical axis than that of a **d** position, with the vectors connecting these atoms to the helical center offset by 17° relative to the interhelical vector (see Fig. 8). The distribution of helical phases shows seven clear peaks, corresponding to the heptad repeat. Notably, these peaks are rather narrow with SDs ~8°. Interestingly, we found no correlation between the locations of these peaks and orientation, oligomerization state, or axial offset. Alber and co-workers have studied the packing differences of **a** and **d** residues in different oligomerization states by comparing core structures of dimeric, trimeric, and tetrameric variants of the GCN4 coiled coil.[53] The authors point out that the angle between the C_α–C_β vector of the **a**/**d** residue, defining the "knob," and the opposing C_α–C_α vector, defining the "hole" into which the knob packs, change in going from dimer to trimer to tetramer. These findings are entirely consistent with our observation of helical phase invariance with respect to oligomerization state. As helices move around the superhelical circle to make room for more monomers, the α-helical phase of each monomer (defined with respect to the interfacial axis) stays roughly unchanged, whereas the orientation of the C_α–C_α vector defining the "hole" changes with respect to partnering helices, giving precisely the trend observed by Alber and co-workers (see Supplementary Fig. S11).

We found that the axial alignment at the helix–helix interface is very restricted in natural backbones, but in different ways for parallel and anti-parallel helix pairs. As can be seen from the distribution of parameter $\Delta Z_{aa'}$ in Fig. 7, parallel coiled-coil pairings tend to align helical turns of adjacent helices at the same level along the interface axis, corresponding to $\Delta Z_{aa'}$ of zero, whereas in anti-parallel orientations, helical turns tend to interdigitate at the interface. Further, anti-parallel coiled coils interdigitate with considerable asymmetry, with **a** positions preferring to pack between the upstream **d**′ (e.g., in the C→N direction of the first helix) and the downstream **a**′ (N→C direction of the first helix; see Figs. 7 and 10 and Supplementary Fig. S8).

The origin of this difference can be rationalized by considering the chirality of the C_α atom and the fact that in an α-helix, the C_α–C_β vector has a positive component in the C→N direction. Adjacent **a** and **d** positions of a coiled coil are on opposite sides of the interface, separated along the interfacial axis by ~4.5 Å from **a** to the next **d** and by ~6.0 Å from **d** to the next **a**. In a parallel alignment, where the C_α–C_β vectors of both chains point in the same direction along the helix, this makes it possible for the **a** and **d** positions of opposite chains to situate across from one another, forming **a**–**a**′ and **d**–**d**′ interactions (see Fig. 10a). In an anti-parallel alignment, however, because C_α–C_β vectors of opposite chains point in opposite directions, potential steric interactions arise between core residues of opposite chains located on the same side of the interface (i.e., **a**–**a**′ and **d**–**d**′), as shown Fig. 10b. Indeed, it is easy to see that unless core residues are small, an on-level alignment (e.g., $\Delta Z_{off}=0$) will not be preferred, but rather, an **a** residue of one chain will pack between adjacent **d**′ and **a**′ residues of the other chain and *vice versa*. The detailed alignment, of course, depends on the exact sequence, but the overall trends of ΔZ_{off} are consistent with these simple geometric considerations. These findings are also consistent with the description of the anti-parallel Alacoil by Richardsons and co-workers.[54] By analyzing seven unrelated proteins, the authors identified an anti-parallel dimeric coiled-coil motif, the Alacoil, in which small residues at **a** and **d**

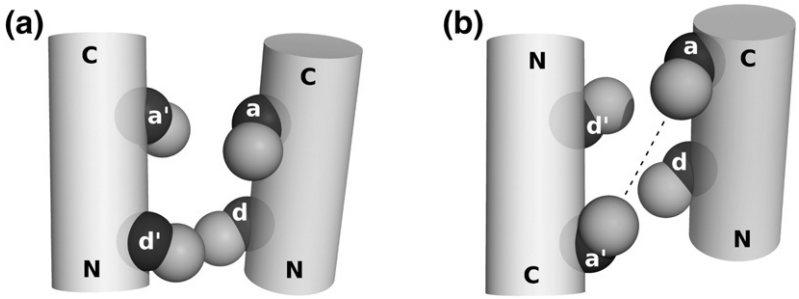

Fig. 10. Structural reasoning behind different ΔZ_{off} preferences between parallel and anti-parallel alignments. (a) and (b) illustrate parallel and anti-parallel dimeric coiled coils, respectively (images generated with coiled coils taken from PDB entries 2ZTA and 2NOV, respectively). Cylinders represent local orientation of helices, black spheres show C_α atoms, and gray spheres depict C_β atoms. Only **a** and **d** positions are shown. Given the direction of C_α–C_β vectors on opposite chains, in anti-parallel alignments, a potential steric repulsion between **a** and **a′** residues in adjacent layers (depicted with a dotted line) pushes the preference of ΔZ_{off} away from zero.

positions allow for a very close approach of helices. In addition to the on-level helical alignment akin to the one seen in parallel coiled coils, the authors observed another type of axial offset, corresponding precisely to the interdigitated alignment described above. In the case of Alacoil, geometric preferences can be understood in terms of the close helical packing and small residues at **a** and **d**.[54] However, we observed the median superhelical radius for anti-parallel dimers to be only marginally smaller than that for parallel dimers (4.82 and 4.92 Å, respectively), with corresponding distributions overlapping very significantly (see Fig. 3b). On the other hand, the difference between distributions of helical offsets for parallel and anti-parallel alignments is striking, indicating that closer helical approach alone cannot account for this effect. The difference between axial alignment of parallel and anti-parallel coiled coils is also in agreement with the recent results reported by Gellman and co-workers, which demonstrate significant thermodynamic coupling between residues at the **a** position of one chain and the **a′** position of the opposite chain in anti-parallel dimeric coiled coils.[55]

All of the analyzed coiled-coil parameters are strongly restricted, reflecting a well-defined and narrow space of designable structures. In fact, based on analyzing the degree of restriction in parameters R_0, α, $\Delta Z_{aa'}$, and ϕ_1, we estimate that the space of designable structures is at least 160- to 200-fold reduced relative to the space of physically reasonable backbones. In the context of *de novo* design, this means that a template selected purely on the basis of geometric feasibility is very unlikely to have significant designability. This is likely to be the case for protein structures in general, illustrating the need for effective low-dimensionality parameterization schemes in other systems.

To begin addressing this need, we developed a general method for parameterizing arbitrary helical structures. Further, we used it to characterize an α-helical motif we found to be common in inner linings of channel and transporter proteins, the surrounding secondary bundles, as well as in dimerization interfaces of bacterial signaling proteins. These bundles consist of helices that are bent into nearly perfectly circular arcs. Using our generalized method, we describe these structures parametrically with a handful of simple geometric variables (see Fig. 9 and Table 1). We find that the degree of helix curvature in these bundles can be rather significant, with radii of curvature as low as 23 Å (for comparison, a typical coiled coil has a radius of curvature of ~100 Å). In their classical work on the geometry of helices in proteins, Barlow and Thornton found similarly curved helices in water-soluble proteins.[44] In fact, among all helices characterized as curved in their work, the lowest and the median radii of curvature were 29 and 62 Å, respectively, whereas for our set of bundles these are 23 and 64 Å, respectively (for the latter, we defined as curved helices with a radius of curvature below 100 Å). Barlow and Thornton showed that the center of helix curvature of regularly curved (amphipathic) α-helices tended to lie on the hydrophobic side of the helix due to backbone hydrogen bonds being on average shorter in hydrophobic environments and longer in a solvent-exposed environments. Is this also the reason for the curvature we observe in arched bundles? In some instances, this is clearly the case, as there are crystallographically resolved water molecules that form hydrogen bonds with the backbone amide hydrogen and carbonyl oxygen atoms, allowing one side of the helix to stretch out and create a bend. For example, this can be seen in the 1.5-Å resolution structure of an NaK channel from *Bacillus cereus*[56] (PDB code 3E86; see Supplementary Fig. S12A). However, in other cases, such water molecules are not explicitly seen, and a single structural reason for the presence of the bend is difficult to identify (e.g., in the structure of the voltage-dependent potassium channel KvAP,[57] PDB code 1ORQ; see Supplementary Fig. S12B). One might expect that helices in channel-forming bundles may curve radially from the channel central axis, as the extent of solvent exposure

may be quite different for the pore-facing side of the helix and the membrane/protein-facing side. However, the distribution of arc turn angles from our parameterization (see Fig. 9) demonstrates that, most often, bending occurs in the plane approximately tangential to the channel pore (see Supplementary Fig. S10). Likely, membrane helical curvature arises as a consequence of a combination of effects, reconciling different extents of solvent exposure (with potentially specific water molecules present in the membrane), amino acid bilayer solvation preferences, and the energetic preference for a particular crossing angle imposed by the full protein structure.

Conclusions

Backbone-parameterized models of protein structure provide a powerful means of describing the diversity within a structural family and relating it to designability. We have used this principle to demonstrate that the space of naturally designable coiled coils is restricted at least 160- to 200-fold relative to the space of geometrically reasonable structures. We have also provided a general framework for parameterization of arbitrary helical structures and have used it to parametrically describe pore-forming bundles of TM channel proteins and subunit interfacial bundles of bacterial signaling proteins.

Rearrangements along coordinates of a well-chosen parameterization may also be functionally relevant. Structurally well-populated modes of helix–helix interaction occur in specific discrete crossing geometries,[32] such that tilting and pitching of helices with respect to each other in a complex bundle would appear to be a convenient way of establishing a discrete set of local minima. Arching of helices also appears to be well suited for establishing alternate states and transferring information. Because arching corresponds to low-frequency bending modes of the α-helix, differently bent conformations would be expected to interconvert at a low frequency in the background, while the presence of various conditions, such as signaling ligands, could stabilize some conformations relative to others.

Materials and Methods

Database

Coiled-coil structures were obtained from the CC+ database[39] as of August 20, 2009. This database was generated by Woolfson and colleagues using the structure-based coiled-coil recognition program SOCKET[41] applied exhaustively to the entire PDB. Our initial set of 3902 structures was obtained by searching CC+ for all entries longer than 11 amino acids using the web-based dynamic search interface. For later analysis, a sequence redundancy filter of 50% was added in the CC+ search interface, which resulted in 868 structures returned by CC+. Sequence redundancy between two coiled coils in CC+ is considered only for structures of identical topology (e.g., the same orientation and oligomerization state). It is calculated as the average percent identity between matching chains of the two complexes, where each chain in one complex is matched up against its closest sequence counterpart in the other complex (see Ref. 39 for details). Additionally, five structures of right-handed coiled coils were manually added to the database as a result of distance map-based searches of the PDB for such topologies (see next section for a description of the search method). The resulting database contained structures ranging from 12 to 148 residues per chain (or 24 to 393 residues per structure), although shorter structures were more common (see Supplementary Fig. S1). Supplementary Table S2 summarizes the distribution of structures among different topologies within the final data set.

Crick parameterization equations

The basic Crick equations describing Cartesian coordinates of a backbone atom type in the α-helix of a coiled coil are as follows (assuming the interfacial axis is aligned with the Z-axis)[8]:

$$\begin{aligned} x &= R_0\cos(\omega_0 t+\phi_0) + R_1\cos(\omega_0 t+\phi_0)\cos(\omega_1 t+\phi_1) \\ &\quad - R_1\cos(\alpha)\sin(\omega_0 t+\phi_0)\sin(\omega_1 t+\phi_1) \\ y &= R_0\sin(\omega_0 t+\phi_0) + R_1\sin(\omega_0 t+\phi_0)\cos(\omega_1 t+\phi_1) \\ &\quad + R_1\cos(\alpha)\cos(\omega_0 t+\phi_0)\sin(\omega_1 t+\phi_1) \\ z &= \frac{\omega_0 R_0}{\tan(\alpha)}t - R_1\sin(\alpha)\sin(\omega_1 t+\phi_1) \end{aligned}$$

where R_0, ω_0, and ϕ_0 are the superhelical radius, frequency, and phase, respectively; R_1, ω_1, and ϕ_1 are the helical radius, frequency, and phase, respectively; α is the pitch angle; and t is residue index. In order to accommodate the ability of helices to shift with respect to one another along the Z-axis, these equations were modified to include an additional degree of freedom Δz. Thus, the equations used in this study were:

$$\begin{aligned} x &= R_0\cos\left(\omega_0 t+\phi_0'\right) + R_1\cos\left(\omega_0 t+\phi_0'\right)\cos(\omega_1 t+\phi_1) \\ &\quad - R_1\cos(\alpha)\sin\left(\omega_0 t+\phi_0'\right)\sin(\omega_1 t+\phi_1) \end{aligned} \quad (2)$$

$$\begin{aligned} y &= R_0\sin\left(\omega_0 t+\phi_0'\right) + R_1\sin\left(\omega_0 t+\phi_0'\right)\cos(\omega_1 t+\phi_1) \\ &\quad + R_1\cos(\alpha)\cos\left(\omega_0 t+\phi_0'\right)\sin(\omega_1 t+\phi_1) \end{aligned} \quad (3)$$

$$z = \frac{\omega_0 R_0}{\tan(\alpha)}t - R_1\sin(\alpha)\sin(\omega_1 t+\phi_1) + \Delta z \quad (4)$$

where ϕ_0' is $\phi_0+\frac{\Delta z\tan(\alpha)}{R_0}$. This last adjustment was made to decouple axial shift from superhelical phase. Thus, Δz is the helical Z-shift produced by sliding a chain along the superhelical curve. During the fitting procedure, each chain,

except the first one, was assigned its own values of ϕ_0' and Δz, relative to the first chain, and each chain had its own ϕ_1.

An issue with Δz is that it describes the axial offset between chain ends rather than the axial alignment of the interface. To address this, we created two additional derived parameter, ΔZ_{off} and $\Delta Z_{aa'}$. To define ΔZ_{off}, inward-facing points on each helix are identified (see Fig. 6), which are the points along the parametric helical curve that have a phase of π (i.e., they point directly into the interface). The smallest Z-offset between these points on the two chains considered is defined as ΔZ_{off} between the two chains, with sign being positive if the second chain is shifted in the N→C direction of the first one, and negative for the opposite shift direction. The sign is meaningful for anti-parallel helix pairings or for parallel, heterodimeric pairings. Given this definition, the value of ΔZ_{off} can range between plus and minus a quarter of a helical pitch, or $\left[-\frac{d\cdot\pi}{\omega_1}, \frac{d\cdot\pi}{\omega_1}\right]$, where ω_1 and d are the helical frequency and rise per residue, respectively. For a typical helix, this corresponds to the range of [−2.6,2.6 Å].

$\Delta Z_{aa'}$ is defined in a similar manner, except instead of points with a helical phase of π, C_α atoms of **a** positions on either chain are used as reference points (heptad assignment is made upon the completion of the fitting procedure, based on best-fit helical phases, using the canonical values shown in Fig. 8). Thus, $\Delta Z_{aa'}$ can vary between plus and minus half a helical pitch, roughly $\left[-\frac{2d\cdot\pi}{\omega_1}, \frac{2d\cdot\pi}{\omega_1}\right]$, or [−5.3,5.3]Å for a typical coiled coil.

Fitting procedure

The program to fit Crick parameters given a structure, called coiled-coil Crick parameterization (CCCP), was written in Matlab‡, using basic functionality and some functions from the Optimization Toolbox. The code is compatible with GNU Octave§[58] with no significant difference in performance, and the web-based version of CCCP uses Octave.

The entire fitting procedure can be broken down into several steps. First, given the input structure file, the number of chains is determined and C_α coordinates are extracted. For each chain (except the first one), its orientation with respect to the first chain is determined by testing the sign of the dot product between vectors connecting the first and last atoms of the two chains. Next, the order of chains, in a clockwise direction when looking down the positive Z direction, is determined. This is important for oligomeric states above the dimer so that chains of the generated ideal structures are superimposed on their corresponding chains in the native structure. To do this, the C_α center of mass of each chain is identified (e.g., points $\vec{c}_1$ through $\vec{c}_n$ for n chains) and the center of the bundle is taken to be the average of these points (e.g., point $\vec{c}_c$). Chain order is then determined by measuring the angles $[\vec{c}_1, \vec{c}_c, \vec{c}_k]$ for each k between 1 and n. Angles are defined in a clockwise direction, with respect to the positive Z direction and mapped onto the range between 0 and 2π {clockwise *versus* counterclockwise is determined by testing whether the cross product between vectors $\vec{c}_c - \vec{c}_1$ and $\vec{c}_c - \vec{c}_k$ is aligned with the positive Z direction, i.e., whether $[(\vec{c}_c - \vec{c}_1)\times(\vec{c}_c - \vec{c}_k)]\cdot[0;\ 0;\ 1]$ is positive or negative}. Thus, sorting the angles in ascending order produces the order of chains in the clockwise direction.

Next, parameter optimization begins. The objective function to minimize is the RMSD between atoms of the parameterized structure and the input structure (optimal superposition was implemented via the SVD method by Kabsch *et al.*[59] as described by Coutsias *et al.*[60]). Basic parameters R_0, R_1, ω_0, ω_1, α, and ϕ_1 are always varied in optimization, and CCCP has a variety of options to limit additional variables. For example, one may wish to assume rotational symmetry about the supercoil axis and set individual chain superhelix phase offsets to $\frac{2\pi}{n}\cdot i$, for chain i of an n-mer; alternatively, this parameter may be individually variable for each chain. Similarly, one may wish to force all chains to have the same starting phase ϕ_1 or have this parameter float between different chains. In all of the fits performed here, no assumptions of symmetry were made. Thus, in addition to the six basic parameters, each chain was also assigned its own value of ϕ_1, and all chains after the first one were assigned values for ϕ_0 and Δz relative to the first chain. The optimization proceeded in a loop until convergence (i.e., RMSD change per iteration was less than 10e−5 Å). In each iteration of the loop, first, each parameter (except R_1 and ω_1) was optimized individually using the BFGS quasi-Newton method with a cubic line search procedure, implemented in function *fminunc* of Matlab. The order in which individual parameters are optimized was found to be important for optimal convergence, and the order used was Δz for each chain, ϕ_0 for each chain, ϕ_1 for each chain, R_0, ω_0, and α. This was followed by multivariable optimization of all parameters at once (including R_1 and ω_1) first with nonlinear least-squares (function *lsqnonlin* in Matlab) and then with BFGS quasi-Newton method with a cubic line search procedure (*fminunc* in Matlab).

In cases of short dimeric coiled coils with minimal superhelical curvature, multiple combinations of superhelical radius and superhelical phase offset values can be equally fitting in terms of RMSD (for a visual example, see Supplementary Fig. S13). To avoid this problem of underdeterminedness, ϕ_0 was constrained to be around π for dimers. During optimization (and not for the purpose of reporting the final RMSD), the objective function value for dimers was incremented by $0.02\ \frac{|\varphi_0 - \pi|}{\frac{\pi}{3}}\ \frac{L}{21}$ Å, where L was chain length. That is, the score was penalized by 0.02 Å for a deviation of $\frac{\pi}{3}$ from the optimal value of π for every three heptads. ϕ_0 was not simply fixed at π because, in some instances, the most appropriate value of phase offset may truly be different from π (e.g., when the dimer looks more like two chains of a tetramer, as shown in Supplementary Fig. S13) and ϕ_0 should be allowed to adjust for that, provided enough improvement in RMSD results. This small penalty resulted in the overwhelming majority of dimers having a ϕ_0 of nearly π.

Once an optimal C_α trace is obtained, the final step is to build the remaining backbone atoms. These were built using a table of internal coordinates (dihedral angles, angles, and distances) of N and C backbone atoms with respect to known C_α atoms. These internal coordinates and their average values, extracted from the structure of GCN4 (PDB code 2ZTA), are shown in Supplementary Table S1.

‡ http://www.mathworks.com/
§ http://www.octave.org/

Backbone oxygen and hydrogen atoms were then placed using standard bond length, angle, and dihedral angle values from CHARMM.[61] This fast and simple approach proved robust, and backbones thus produced were generally not found to be in need of further adjustment. However, CCCP (including the web-based version) can optionally subject the produced backbone to short molecular mechanics minimization.

Local R_0 as a function of core composition

For all dimers longer than three heptads, local values of R_0 were calculated by CCCP fitting of two-heptad segments scanned across each structure. Superhelical radius autocorrelation length was defined as the lag, in terms of the number of residues, at which the autocorrelation function crossed zero. Two simple models for the dependence of R_0 on the **a**- and **d**-position amino acid composition were considered. The relationship may be linear, whereby each residue contributes a certain fixed context-independent component, whether negative or positive, to the final R_0, relative to some average reference value. Or it may be that each residue has its own radius preference, and the residues with the largest preference dictate the eventual local R_0. To simplify interpretation and to limit the number of parameters, we focused on parallel dimers. In both cases, there are 40 parameters (for 20 naturally occurring amino acids at either the **a** or **d** position), and the two models can be written as:

$$R_0 = \frac{1}{2}\left[\frac{1}{||\vec{a}||}\sum_{i\in\vec{a}}\left(w_a\left(aa_i^A\right) + w_a\left(aa_i^B\right)\right) + \frac{1}{||\vec{b}||}\sum_{i\in\vec{d}}\left(w_d\left(aa_i^A\right) + w_d\left(aa_i^B\right)\right)\right] \quad (5)$$

$$R_0 = \max\left[\max_{i\in\vec{a}}\left[w_a\left(aa_i^A\right) + w_a\left(aa_i^B\right)\right] \max_{i\in\vec{d}}\left[w_d\left(aa_i^A\right) + w_d\left(aa_i^B\right)\right]\right] \quad (6)$$

where $\vec{a}$ and $\vec{d}$ are the indices of **a** and **d** positions in the considered window, respectively; aa_i^A and aa_i^B are the i-th amino acids in either of the two chains of the dimer; and w_a and w_d are amino acid-dependent weights at **a** and **d** positions, respectively. In both models, $w_h(aa_i)$ corresponds to the context-independent contribution of amino acid aa_i at heptad position h (either **a** or **d**) to the final radius. Thus, the contribution of an **a**–**a**′ layer (prime designates the opposite chain) is the sum of contributions of the two amino acids forming the layer or $w_a(aa_i^A) + w_a(aa_i^B)$, and the contribution of a **d**–**d**′ layer is $w_d(aa_i^A) + w_d(aa_i^B)$. What differentiates the models is that in the first one, the final radius preference is dictated by the average of the layer preferences, whereas in the second one, the layer with the largest radius preference dominates. These models were fit to local R_0 data described above, with the results shown in Supplementary Fig. S6.

Quantifying the degree of restriction of natural coiled-coil space

For a given coiled-coil parameter x, for which we observe values ranging from $x_{\min}$ to $x_{\max}$ in natural coiled coils, we compare its observed distribution with a uniform random distribution over the range $x_{\min}$ to $x_{\max}$. To this end, we create a histogram of the observed distribution by binning values into n bins and calculate the entropy of this histogram as $S(x,n) = -\sum_{i=1}^{i=n}\frac{N_i}{N_{\text{obs}}}\cdot\ln\left(\frac{N_i}{N_{\text{obs}}}\right)$, where N_{obs} is the total number of data points and N_i is the number of x values within bin i. Next, we compose a uniform random distribution of N_{obs} values over n bins by apportioning $k = \left\lfloor\frac{N_{\text{obs}}}{n}\right\rfloor$ points into all bins and an additional point into $r = \left(N_{\text{obs}} - \left\lfloor\frac{N_{\text{obs}}}{n}\right\rfloor\cdot n\right)$ bins. The entropy of this distribution thus becomes $S_{\text{ref}}(x,n) = -r\cdot\frac{k+1}{N_{\text{obs}}}\cdot\ln\left(\frac{k+1}{N_{\text{obs}}}\right) - (n-r)\cdot\frac{k}{N_{\text{obs}}}\cdot\ln\left(\frac{k}{N_{\text{obs}}}\right)$. The quantity $\Delta S = S_{\text{ref}}(x,n) - S(x,n)$ is then a measure of information content within the naturally observed distribution of x. If we interpret entropy in the statistical mechanical sense, then $e^{\Delta S}$ is the ratio between the effective "number of states" in the uniform distribution and the native distribution, thereby providing us with a reasonable estimate of the degree of restriction of the natural space.

Given a limited amount of data, the number of bins n has an impact on the calculated information content. With very few bins, the natural and uniform distributions converge as everything eventually collapses to one bin. The same happens when the number of bins is too large, as eventually all bins, in both the natural and uniform distributions, either have one observation or none. To find a value of n that optimally captures the amount of information present in the observed data set, we chose the value that maximizes the quality ΔS.

Generalized parameterization

An α-helix is described by a simple parametric set of equations for a left-handed screw:

$$\begin{aligned} x &= R_1\cos(\omega_1 t + \phi_1) \\ y &= R_1\sin(\omega_1 t + \phi_1) \\ z &= d\cdot t \end{aligned} \quad (7)$$

where t is the parameter (residue index) and d is rise per residue (discussed in Results). The Crick parameterization describes what happens to this screw when it, in turn, is twisted into a larger screw, the superhelix. To address a general version of this problem, suppose that the α-helix is bent into an arbitrary shape described by a generic parametric set of equations:

$$\begin{aligned} x_0 &= f_x(s) \\ y_0 &= f_y(s) \\ z_0 &= f_z(s) \end{aligned}$$

where subscript 0 designates coordinates of the superstructure and s is the new superstructure parameter. For convenience, let us consider the locations of C_α atoms, although identical reasoning can be applied for any backbone atom type. In an ideal Z-aligned helix, the Z coordinates of successive C_α atoms change in increments of d, such that a local coordinate frame can be defined for the i-th C_α atom with an origin $[0, 0, z(t_i)]$ and axes pointing in the same direction as the laboratory frame in which Eq. (7) is defined. Correspondingly, when the helix winds around the parametric curve, these local origins

will translate onto points on the new curved axis, each corresponding to a particular parameter value s_i. For every residue in the final structure, the superstructure ought to sweep out an arc length equal to d, to ensure that the helix is locally neither stretched nor compressed. Thus, for going from residue i to residue $i+1$, we must satisfy:

$$d = \int_{s_i}^{s_{i+1}} \sqrt{\left(\frac{df_x}{ds}\right)^2 + \left(\frac{df_y}{ds}\right)^2 + \left(\frac{df_z}{ds}\right)^2}\, ds \qquad (8)$$

where s_i and s_{i+1} are the values of s corresponding to residue i and $i+1$, respectively. For the case of a superhelix (and thus $f_x(s) = R_0 \cos(\omega_0 s + \phi_0)$, $f_y(s) = R_0 \sin(\omega_0 s + \phi_0)$, and $f_z(s) = \frac{R_0\omega_0}{\tan(\alpha)} S$), we get simply that $d = (s_{i+1} - s_i)\sqrt{(R_{0\omega 0})^2 + \left(\frac{R_{0\omega 0}}{\tan(\alpha)}\right)^2} = \frac{R_{0\omega 0}}{\sin(\alpha)}$ or that $s_{i+1} - s_i = \frac{d\sin(\alpha)}{R_{0\omega 0}}$, meaning that s values along the superhelix, corresponding to C_α atoms, occur at regular intervals. This does not, however, need to be the case in general. For more complicated shapes (e.g., a coiled coil of changing radius), s values from one residue to the next will change differently in different parts of the curve (e.g., more slowly the larger the radius gets).

For many parametric shapes, the integral in Eq. (8) can be analytically solved, in which case values of s_i are easily obtainable. However, even in cases where the integral does not exist in closed form, simple numerical integration methods can be applied to solve Eq. (8) efficiently. In the latter case, values of s_i are built up starting from the known s_1, which defines the starting position along the parametric curve, each time obtaining the value of s_{k+1} that validates Eq. (8), given the known s_k. Finding all s_i values identifies the origins of local coordinate frames associated with each residue in the final structure {e.g., $[f_x(s_i); f_y(s_i); f_z(s_i)]$}, and we must now define the axes. Naturally, the local Z-axis is defined in the direction of the gradient of the parametric curve at s_i, providing a smooth transition from a straight to a curved helix, whereas the X-axis is defined in the direction opposite to the curvature vector of the parametric curve at s_i:

$$\vec{Z}_l = \vec{g} = \frac{1}{\sqrt{(f_x')^2 + (f_y')^2 + (f_z')^2}}\left[f_x'; f_y'; f_z'\right]$$

$$\vec{X}_l = \vec{c} = \frac{1}{\sqrt{(g_x')^2 + (g_y')^2 + (g_z')^2}}\left[g_x'; g_y'; g_z'\right]$$

$$\vec{Y}_l = \vec{Z}_l \times \vec{X}_l$$

where primes designate derivatives with respect to s, and g_x, g_y, and g_z are the components of the normalized gradient vector $\vec{g} = [g_x; g_y; g_z]$. By analogy with the coiled coil, this ensures that the X-axis points away from the center of rotation of the local stretch of the parametric curve. In the case of a superhelical parametric curve, this definition recapitulates exactly that by Crick.[8]

Once the local coordinate frame associated with each residue in the final structure is defined, and given the original local coordinate frame for each residue in the straight helix, two rotations and a translation are applied to transform each residue from the straight helix into its position within the curved helix. Therefore, the generalized helical parameterization framework takes on the form of a numerical procedure, rather than a simple formula, as with the coiled coil, since it is not guaranteed that the integral in Eq. (8) is analytically solvable. If it is, however, a formula can be derived as well, as all terms within the above rotation and translation matrices become analytical. Matlab (or Octave) code implementing the generalized helix bending parameterization is available upon request.

We applied this framework to generate a parameterization of arched bundles. As shown in Fig. 9, individual helices were modeled as being shaped into perfectly circular arcs of radius R_a. The superstructure parametric equations, therefore, were $f_x(s) = R_a \cos(\omega_0 s)$, $f_y(s) = 0$, and $f_z(s) = R_a \sin(\omega_0 s)$, with parameter s varying symmetrically around 0 (e.g., from some $-s_0$ to $+s_0$). This corresponds to a circular arc in the XZ plane, symmetrical about the X-axis (see Fig. 9a). Solving Eq. (8), we get $\Delta s = \frac{R_a\omega_0}{d}$; thus, for simplicity, ω_0 was set to $\frac{d}{R_a}$ (note, this is an arbitrary choice) so that values of s could be varied between $-\frac{L-1}{2}$ and $+\frac{L-1}{2}$ in increments of 1 (L is the number of residues in the chain). The normalized gradient of our superstructure curve is simply $\vec{g} = \frac{1}{R_a\omega_0}[-R_a\omega_0 \sin(\omega_0 s); 0; R_a\omega_0 \cos(\omega_0 s)] = [-\sin(\omega_0 s); 0; \cos(\omega_0 s)]$ (since ω_0 is positive here), and the normalized curvature vector is $\vec{c} = [-\cos(\omega_0 s); 0; -\sin(\omega_0 s)]$. Thus, the rotation and translation matrices for superimposing the local coordinate frame in the ideal helix and the curved superstructure are trivial to find, and the final equations for a helix bent around a circular arc are:

$$\begin{aligned} x(t) &= [R_1 \cos(\omega_1 t + \phi_1) + R_a] \cdot \cos\left[\frac{d}{R_a}\left(t - \frac{L-1}{2}\right)\right] \\ &\quad - R_a \cdot \cos\left[\frac{d \cdot (L-1)}{2 \cdot R_a}\right] \\ y(t) &= R_1 \sin(\omega_1 t + \phi_1) \\ z(t) &= [R_1 \cos(\omega_1 t + \phi_1) + R_a] \cdot \sin\left[\frac{d}{R_a}\left(t - \frac{L-1}{2}\right)\right] \end{aligned} \qquad (9)$$

where t varies from 0 to $L-1$ [the second term in $x(t)$ simply moves the ends of the arc so that they lie on the Z-axis, as shown in Fig. 9a, gray curve].

Once the curved helix is generated, it is rotated about the laboratory Z-axis by the arc turn angle α_u, then the laboratory Y-axis by the tilt angle α_t, and then the laboratory X-axis by the pitch angle α_p (see Fig. 9). Finally, the helix is translated along the X-axis by the bundle radius R_b. In our arched-bundle parameterization, we assumed that individual chains are spaced equally around the bundle; thus, upon generating one helix, the rest are generated with rotations about the Z-axis in increments of $\frac{2\pi}{n}$, where n is the number of chains in the bundle.

Table 1 reports parameter values that best fit to a set of representative structures as well as the C_α RMSD of the fits. Note that for convenience of comparing parameters between different structural states, instead of helical phase (variable ϕ_1 in Eq. (7)), Table 1 reports the effective phase that takes into account the arc turn angle (i.e., the actual value reported is $\phi_{eff} = \phi_1 + \alpha_u$). This allows the phase to be reflective solely of the angular position of the first residue in the chain with respect to the interfacial axis, allowing for an independent interpretation of this parameter. Table 1 also reports an additional calculated parameter—the arc

incidence angle, which measures the angle between the tangent to the arc at the initial point and the line connecting arc ends. For a straight helix, this angle is zero; for a semicircle, it is 90 °.

Distance map-based search for right-handed coiled coils

To search for additional examples of right-handed coiled coils in the PDB, we employed a distance map-based search method. Full details of the method are to be published elsewhere, but a short description is provided here. A distance map is an *n*-by-*n* matrix of C_α-to-C_α distances within a protein, where *n* is the number of residues in the protein.[62] We first pre-computed such distance maps for all proteins in the PDB, storing them in a sparse format such that distances above 25 Å were not considered. Next, we used all of the right-handed coiled coils found in the CC+ database as search queries. For each query protein, a distance map corresponding to two helices, two heptads each, was generated. These query maps were then compared to the pre-computed distance maps, looking for optimal agreement between the query map and a sub-map of a potential matching protein, in the least-squares sense. The search was performed taking into account that the two helices will not be contiguous in sequence within the matching structure, and thus a linker of arbitrary length was allowed to locate between the two helices of the query map.

Acknowledgements

We would like to thank Dr. Cinque S. Soto for stimulating discussions and careful reading of the manuscript. Work in this study was supported by NIH grants 1F32GM084631-01A1 and 5F32GM084631-02 to GG.

Supplementary Data

Supplementary data to this article can be found online at doi:10.1016/j.jmb.2010.08.058

References

1. Li, H., Helling, R., Tang, C. & Wingreen, N. (1996). Emergence of preferred structures in a simple model of protein folding. *Science*, **273**, 666–669.
2. Wingreen, N. S., Li, H. & Tang, C. (2004). Designability and thermal stability of protein structures. *Polymer*, **45**, 699–705.
3. Helling, R., Li, H., Melin, R., Miller, J., Wingreen, N., Zeng, C. & Tang, C. (2001). The designability of protein structures. *J. Mol. Graph. Model.* **19**, 157–167.
4. England, J. L., Shakhnovich, B. E. & Shakhnovich, E. I. (2003). Natural selection of more designable folds: a mechanism for thermophilic adaptation. *Proc. Natl Acad. Sci. USA*, **100**, 8727–8731.
5. Govindarajan, S. & Goldstein, R. A. (1996). Why are some proteins structures so common? *Proc. Natl Acad. Sci. USA*, **93**, 3341–3345.
6. Koehl, P. & Levitt, M. (2002). Protein topology and stability define the space of allowed sequences. *Proc. Natl Acad. Sci. USA*, **99**, 1280–1285.
7. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.
8. Crick, F. H. (1953). The Fourier transform of a coiled coil. *Acta Crystallogr.* **6**, 685–689.
9. Harbury, P. B., Tidor, B. & Kim, P. S. (1995). Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc. Natl Acad. Sci. USA*, **92**, 8408–8412.
10. Murzin, A. G., Lesk, A. M. & Chothia, C. (1994). Principles determining the structure of β-sheet barrels in proteins. I. A theoretical analysis. *J. Mol. Biol.* **236**, 1369–1381.
11. Murzin, A. G., Lesk, A. M. & Chothia, C. (1994). Principles determining the structure of β-sheet barrels in proteins. II. The observed structures. *J. Mol. Biol.* **236**, 1382–1400.
12. Rainey, J. K. & Goh, M. C. (2002). A statistically derived parameterization for the collagen triple-helix. *Protein Sci.* **11**, 2748–2754.
13. Lombardi, A., Summa, C. M., Geremia, S., Randaccio, L., Pavone, V. & DeGrado, W. F. (2000). Inaugural article: retrostructural analysis of metalloproteins: application to the design of a minimal model for diiron proteins. *Proc. Natl Acad. Sci. USA*, **97**, 6298–6305.
14. Goldberg, S. D., Soto, C. S., Waldburger, C. D. & DeGrado, W. F. (2008). Determination of the physiological dimer interface of the PhoQ sensor domain. *J. Mol. Biol.* **379**, 656–665.
15. Metcalf, D. G., Law, P. B. & DeGrado, W. F. (2007). Mutagenesis data in the automated prediction of transmembrane helix dimers. *Proteins*, **67**, 375–384.
16. North, B., Summa, C. M., Ghirlanda, G. & DeGrado, W. F. (2001). D(n)-symmetrical tertiary templates for the design of tubular proteins. *J. Mol. Biol.* **311**, 1081–1090.
17. Havranek, J. J. & Harbury, P. B. (2003). Automated design of specificity in molecular recognition. *Nat. Struct. Biol.* **10**(1), 45–52.
18. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science*, **282**(5393), 1462–1467.
19. Grigoryan, G., Reinke, A. W. & Keating, A. E. (2009). Design of protein-interaction specificity gives selective bzip-binding peptides. *Nature*, **458**(7240), 859–864.
20. Offredi, F., Dubail, F., Kischel, P., Sarinski, K., Stern, A. S., de Weerdt, C. V. *et al.* (2003). *De novo* backbone and sequence design of an idealized α/β-barrel protein: Evidence of stable tertiary structure. *J. Mol. Biol.* **325**(1), 163–174, doi:10.1016/S0022-2836(02)01206-8.
21. Keating, A. E., Malashkevich, V. N., Tidor, B. & Kim, P. S. (2001). Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled

coils.*Proc. Natl Acad. Sci. USA*, **98**(26), 14825–14830, doi:10.1073/pnas.261563398; arXiv:. http://www.pnas.org/content/98/26/14825.full.pdf+html.
22. Grigoryan, G. & Keating, A. E. (2006). Structure-based prediction of bzip partnering specificity. *J. Mol. Biol.* **355**, 1125–1142.
23. Apgar, J. R., Gutwin, K. N. & Keating, A. E. (2008). Predicting helix orientation for coiled-coil dimers. *Proteins*, **72**, 1048–1065.
24. Valavanis, I. K., Bagos, P. G. & Emiris, I. Z. (2006). β-Barrel transmembrane proteins: Geometric modelling, detection of transmembrane region, and structural properties. *Comput. Biol. Chem.* **30**(6), 416–424, doi:10.1016/j.compbiolchem.2006.09.001.
25. Furini, S., Domene, C., Rossi, M., Tartagni, M. & Cavalcanti, S. (2008). Model-based prediction of the α-hemolysin structure in the hexameric state. *Biophys. J.* **95**(5), 226–2274, doi:10.1529/biophysj.107.127019.
26. Howard, R. J., Clark, K. A., Holton, J. M. & Jr, D. L. M. (2007). Structural insight into KCNQ (Kv7) channel assembly and channelopathy. *Neuron*, **53**(5), 663–675, doi:10.1016/j.neuron.2007.02.010.
27. Zou, J. & Saven, J. G. (2000). Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. *J. Mol. Biol.* **296**, 281–294.
28. Wolynes, P. G. (1996). Symmetry and the energy landscapes of biomolecules. *Proc. Natl Acad. Sci. USA*, **93**, 14249–14255.
29. North, B. H. (2006). *Application of the Crick parameterization to de novo* protein design. PhD dissertation. University of Pennsylvania.
30. Offer, G., Hicks, M. R. & Woolfson, D. N. (2002). Generalized Crick equations for modeling noncanonical coiled coils. *J. Struct. Biol.* **137**, 41–53.
31. Summa, C. M., Lombardi, A., Lewis, M. & DeGrado, W. F. (1999). Tertiary templates for the design of diiron proteins. *Curr. Opin. Struct. Biol.* **9**, 500–508.
32. Walters, R. F. & DeGrado, W. F. (2006). Helix-packing motifs in membrane proteins. *Proc. Natl Acad. Sci. USA*, **103**, 13658–13663.
33. Engel, D. E. & DeGrado, W. F. (2005). Alpha–alpha linking motifs and interhelical orientations. *Proteins*, **61**, 325–337.
34. Murzin, A. G. & Finkelstein, A. V. (1988). General architecture of the α-helical globule. *J. Mol. Biol.* **204**, 749–769.
35. Bowie, J. U. (1997). Helix packing angle preferences. *Nat. Struct. Biol.* **4**, 915–917.
36. Walther, D., Eisenhaber, F. & Argos, P. (1996). Principles of helix–helix packing in proteins: the helical lattice superposition model. *J. Mol. Biol.* **255**, 536–553.
37. Zhou, H. X. McCammon, J. A. The gates of ion channels and enzymes. *Trends Biochem. Sci.* **35**, 179–185.
38. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. (2006). OPM: orientations of proteins in membranes database. *Bioinformatics*, **22**, 623–625.
39. Testa, O. D., Moutevelis, E. & Woolfson, D. N. (2009). CC+: a relational database of coiled-coil structures. *Nucleic Acids Res*, **37**(database issue), D315–D322.
40. Sales, M., Plecs, J. J., Holton, J. M. & Alber, T. (2007). Structure of a designed, right-handed coiled-coil tetramer containing all biological amino acids. *Protein Sci*, **16**(10), 2224–2232.
41. Walshaw, J. & Woolfson, D. N. (2001). Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *Journal of molecular biology*, **307**(5), 1427–1450, doi:10.1006/jmbi.2001.4545.
42. 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. *Nature*, **7**(9), 772–776.
43. O'Shea, E. K., Klemm, J. D., Kim, P. S. & Alber, T. (1991). X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science*, **254**(5031), 539–544.
44. Barlow, D. J. & Thornton, J. M. (1988). Helix geometry in proteins. *J. Mol. Biol.* **201**, 601–619.
45. Kumar, S. & Bansal, M. (1998). Geometrical and sequence characteristics of α-helices in globular proteins. *Biophys. J.* **75**, 1935–1944.
46. Page, R. C., Kim, S. & Cross, T. A. (2008). Transmembrane helix uniformity examined by spectral mapping of torsion angles. *Structure*, **16**, 787–797.
47. Grottesi, A., Sands, Z. A. & Sansom, M. S. P. (2005). Potassium channels: Complete and undistorted. *Curr. Biol.* **15**(18), R771–R774.
48. Casino, P., Rubio, V. & Marina, A. (2009). Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. *Cell*, **139**, 325–336.
49. Albanesi, D., Martin, M., Trajtenberg, F., Mansilla, M. C., Haouz, A., Alzari, P. M. *et al.* (2009). Structural plasticity and catalysis regulation of a thermosensor histidine kinase. *Proc. Natl Acad. Sci. USA*, **106**, 16185–16190.
50. MacDonald, J. T., Maksimiak, K., Sadowski, M. I. & Taylor, W. R. (2010). *De novo* backbone scaffolds for protein design. *Proteins*, **78**, 1311–1325.
51. Berger, B. W., Kulp, D. W., Span, L. M., DeGrado, J. L., Billings, P. C., Senes, A. *et al.* (2010). Consensus motif for integrin transmembrane helix association. *Proc. Natl Acad. Sci. USA*, **107**, 703–708.
52. Lupas, A., Van Dyke, M. & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
53. Harbury, P. B., Kim, P. S. & Alber, T. (1994). Crystal structure of an isoleucine-zipper trimer. *Nature*, **371**, 80–83.
54. Gernert, K. M., Surles, M. C., Labean, T. H., Richardson, J. S. & Richardson, D. C. (1995). The Alacoil: a very tight, antiparallel coiled-coil of helices. *Protein Sci.* **4**, 2252–2260.
55. Hadley, E. B., Testa, O. D., Woolfson, D. N. & Gellman, S. H. (2008). Preferred side-chain constellations at antiparallel coiled-coil interfaces. *Proc. Natl Acad. Sci. USA*, **105**(2), 530–535.
56. Alam, A. & Jiang, Y. (2009). High-resolution structure of the open NaK channel. *Nat. Struct. Mol. Biol.* **16**, 30–34.
57. Jiang, Y., Lee, A., Chen, J., Ruta, V., Cadene, M., Chait, B. T. & MacKinnon, R. (2003). X-ray structure of a voltage-dependent K+ channel. *Nature*, **423**, 33–41.
58. Eaton, J. W. (2002). *GNU Octave Manual. Network Theory Limited.*
59. Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, **32**(5), 922–923, doi:10.1107/S0567739476001873.

60. Coutsias, E. A., Seok, C. & Dill, K. A. (2004). Using quaternions to calculate rmsd. *J. Comput. Chem.* **25**(15), 1849–1857, doi:10.1002/jcc.20110.
61. Brooks, B. R. R., Brooks, C. L. L., Mackerell, A. D. D., Nilsson, L., Petrella, R. J. J., Roux, B. *et al.* (2009). CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30**(10), 1545–1614, doi:10.1002/jcc.21287.
62. Choi, I. G., Kwon, J. & Kim, S. H. (2004). Local feature frequency profile: a method to measure structural similarity in proteins. *Proc. Natl Acad. Sci. USA*, **101**, 3797–3802.