Journal of Structural Biology xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Structural Biology

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

This article is part of a Special Issue on Coiled-Coil, Fibrous & Repeat Proteins

Optimal data-driven parameterization of coiled coils☆

Dmytro Guzenko, Sergei V. Strelkov*

Department of Pharmaceutical and Pharmacological Sciences, KU Leuven, Leuven 3000, Belgium

ARTICLE INFO

Keywords:
Coiled coils
Protein structure

ABSTRACT

α-Helical coiled coils (CCs) represent an important, highly regular protein folding motif. To date, many thousands of CC structures have been determined experimentally. Their geometry is usually modelled by theoretical equations introduced by F. Crick that involve a predefined set of parameters. Here we have addressed the problem of efficient CC parameterization from scratch by performing a statistical evaluation of all available CC structures. The procedure is based on the principal component analysis and yields a minimal set of independent parameters that provide for the reconstruction of the complete CC structure at a required precision. The approach is successfully validated on a set of canonical parallel CC dimers. Its applications include all cases where an efficient sampling of the CC geometry is important, such as for solving the phase problem in crystallography.

1. Introduction

α-Helical coiled coils (CCs) representing a superhelical arrangement of two or more α-helices are an important protein structure motif (Truebestein and Leonard, 2016). As much as 8% of all human genome proteins are estimated to include CC motifs (Rose et al., 2005). The basis of the CC formation is the favourable interaction of α-helices that show a periodic pattern (most typically, a heptad) in the hydrophobicity of the residues along their sequence (Lupas and Gruber, 2005). Early on, an analytical representation of a CC structure through a set of simple equations was proposed (Crick, 1953). This representation assumed a regular and symmetric arrangement of the α-helices forming the CC. Its parameters included the radius and pitch of the superhelix, the phase of the first residue (often expressed by the so-called Crick angle, also known as the interface angle (Wood et al., 2017)), and the radius, pitch and rise per residue of the α-helix (Strelkov and Burkhard, 2002). In the past, two main approaches to derive the set of Crick parameters from a given experimental structure were proposed. The first approach was based on a best fit between a theoretical structure parametrized through Crick equations and the experimentally determined CC (Offer et al., 2002; Grigoryan and DeGrado, 2011). The second approach involved a direct measurement of the local Crick parameters along the structure using the program Twister (Strelkov and Burkhard, 2002). While the Crick parameterization is a mathematical abstraction, for naturally occurring CCs the corresponding parameter space is quite restricted with some parameters clearly correlated (Grigoryan and DeGrado, 2011). At the same time, additional parameters beyond the Crick representation, such as the axial offset of individual α-helices forming the CC for example, are sometimes introduced to better describe the geometry of natural CCs (Wood et al., 2014).

In this manuscript, we reconsider the problem of optimal CC parameterization from scratch *i.e.*, not assuming any predefined analytical representation. Our goal is to obtain the most accurate description that would use a minimal number of parameters. To this end, we analyze a large set of experimentally determined CC structures and infer the relevant structural parameters in a statistically rigorous way. This is done through the principal component (PC) analysis of all such structures, whereby the orthogonal PCs are derived without the need to explicitly define them. The given CC structure is then described by a set of PC scores corresponding to the amplitudes of deviations from the mean CC structure. Since the PCs can be ranked by their 'importance' (*i.e.*, the explained variance ratio in the dataset), a certain number of top-scoring parameters can be included towards achieving the required modelling accuracy with respect to the experimental structure. While we will focus the description of our algorithm on parallel CC dimers, other CC structures are treated in the same principle way. The present code implementation can handle parallel and antiparallel dimeric CC structures, as well as trimers.

2. Sets of experimental α-helical and CC fragments

As a starting point, all parallel dimeric CC fragments of 15 residues per chain were extracted from the CC+ database (Testa et al., 2009) and truncated to polyalanine. Following the procedures described in

☆ This Special Issue is edited by Andrei N. Lupas and Derek N. Woolfson, and is published as a companion to the meeting on Coiled-Coil, Fibrous & Repeat Proteins at Alpbach, Austria, September 3–8, 2017.

* Corresponding author.
E-mail address: sergei.strelkov@kuleuven.be (S.V. Strelkov).

https://doi.org/10.1016/j.jsb.2018.07.001
Received 10 January 2018; Received in revised form 30 May 2018; Accepted 1 July 2018

Guzenko et al. (2017), a non-redundant set of 14175 CC fragments was obtained. Here non-redundancy was defined in terms of structural dissimilarity, so that the root mean square deviation (RMSD) between any two fragments upon their optimal superposition was at least 0.2 Å. A non-redundant set of 15760 15-residue α-helices with a minimal pairwise RMSD of 0.1 Å was constructed in a similar way. All RMSD values for 15-residue α-helices and CC fragments given here and below are for all-atom polyalanine models.

3. PC analysis and parameterization of an α-helix

Multiple structural alignment of the non-redundant set of 15-residue α-helices was obtained using iterative superposition method (Gil and Guallar, 2013). Thereafter the aligned coordinates were subjected to a PC analysis. This resulted in a decomposition of each α-helix into a linear combination of the PC vectors:

$$H_i = \overline{H} + \sum_{j=1}^{N} a_i jC_j, \quad (1)$$

where $\overline{H}$ is the mean α-helix, C_j is the j-th component and a_j is its score.

Each PC represents a mode of variation of the α-helical geometry, with the explained variance in the dataset proportional to the eigenvalues of the coordinate covariance matrix (Fig. 1a). The resulting three dominant modes including two perpendicular 'bend' modes (Fig. 1a) and one 'twist' (Fig. 1c) mode and their variance ratios are consistent with those previously reported (Emberly et al., 2003). The encoding system consisting of the mean helix $\overline{H}$ and scores of the first three components allows to model individual α-helices from the non-redundant set with a mean RMSD of 0.23Å (Fig. 1d).

4. Extended parameterization and initial analysis of the CC structure space

To approach parameterization of a CC structure (of any multiplicity), we first align each of its α-helices to the mean α-helical structure $\overline{H}$. The result of such alignment for each helix is reflected in the scores of the three dominant PCs described above (Fig. 1a). Next, we record the position of the second, third *etc.* helix with respect to the first one. This position is given by the rotation matrix R and the translation vector T (Fig. 2a). The description of a CC structure is then given by:

$$CC_{ext}(N) = \{\overline{H}, A_1, A_2, R_2, T_2, \ldots A_N, R_N, T_N\}, \quad (2)$$

where N is the number of helices in the CC, A_i is the vector of PC scores for the i-th α-helix, R_i and T_i are rotation and translation parameters of the i-th α-helix.

For a CC dimer this process results in an 'extended set' of 18 parameters including three parameters for each of the two α-helices, nine parameters for the rotation matrix and three for the translation vector of the second helix.

Next, we perform a PC analysis on the non-redundant set of 15-residue dimeric CC fragments, each represented by a vector of 'extended' parameters from Eq. (2). The resulting ranking indicates a clear dominance of the first two PCs (Fig. 2b), whereby seven pronounced structural clusters in the space of these two PCs are observed (Fig. 2c). This pattern reflects the fact that the fragment set is dominated by the structures with regular heptad repeats which can start at one of the seven possible heptad positions. The two most prominent clusters correspond to the fragments starting with either *a* or *d* heptad position. This is easily explained by the fact that all fragments recorded in the CC+ database start and end with a core residue.

5. Optimal parameterization of 15-residue CC dimers

As the next step, we focus on a cluster of 15-residue fragments that all have the same heptad register, *e.g.*, those starting at position *a*. Towards this end, all fragments situated in the vicinity of one of the maxima in the space of the first two PCs (Fig. 2c) are selected. Since the initial non-redundant set included overlapping 15-residue fragments, this cluster of fragments alone is structurally representative of all sufficiently long dimeric CCs in the CC+ database.

Results of the PC analysis of this cluster are presented in Fig. 3. The average structure of the cluster is expectedly a canonical left-handed symmetric CC. To model a CC structure with a desired number (k) of parameters, the first k PC vectors are used, and the respective variations are added to the mean parameters vector (Eq. (2)). Technically, the rotation matrix of an individual α-helix restored with this procedure, $\widehat{R}$, may no longer be orthogonal. Therefore we use the nearest orthogonal matrix as in Horn et al. (1988) instead:

$$R = \widehat{R}\,(\widehat{R}^{\mathrm{T}}\widehat{R})^{-\frac{1}{2}} \quad (3)$$

Next, we wanted to explore how efficiently the space of all 15-residue dimeric CC structures can be sampled using a particular number of parameters. To this end, for each structure in the chosen cluster of experimental CCs, we have built the best-matching model upon varying the first k PC parameters. The resulting statistics is presented in Fig. 3c. A zero-PC modelling (which corresponds to simply approximating any structure with the cluster average) results in a median RMSD of 0.5 Å, with a maximal RMSD value observed being 1.2 Å. These relatively low values (obtained for the short 15-residue dimers) reflect the fact that the local CC geometry is quite conserved across all experimental

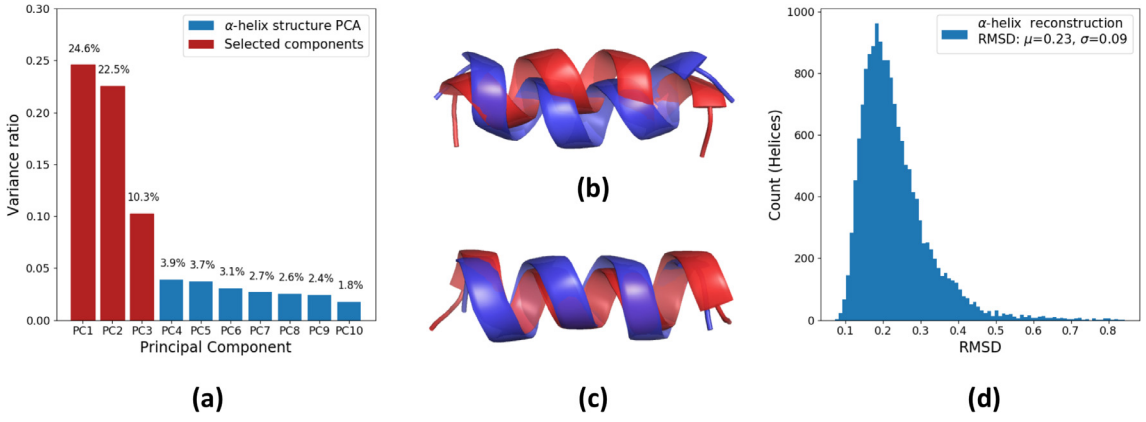

Fig. 1. Principal component analysis of a 15-residue α-helix. (a) Variance ratio per PC of all α-helices in the non-redundant set. (b) The first PC corresponds to the bending of the α-helix in one direction. The second PC represents bending in a perpendicular direction and is not shown. The third PC corresponds to over/underwinding of the α-helix (exaggerated in the figure). (d) Distribution of all-atom RMSD values calculated by superposing model α-helices encoded by the three PC parameters with the original α-helices.