BIOLOGICAL
CRYSTALLOGRAPHY

ISSN: 1399-0047
journals.iucr.org/d

Protein secondary-structure description with a coarse-grained model

Gerald R. Kneller and Konrad Hinsen

Acta Cryst. (2015). D**71**, 1411–1422

BIOLOGICAL CRYSTALLOGRAPHY

ISSN 1399-0047

Protein secondary-structure description with a coarse-grained model

Gerald R. Kneller[a,b,c]* and Konrad Hinsen[a,b]

[a]Centre de Biophysique Moléculaire, CNRS, Rue Charles Sadron, 45071 Orléans, France, [b]Synchrotron SOLEIL, L'Orme de Merisiers, BP48, 91192 Gif-sur-Yvette, France, and [c]Université d'Orléans, Chateau de la Source, Avenue du Parc Floral, 45067 Orléans, France. *Correspondence e-mail: gerald.kneller@cnrs-orleans.fr

Received 1 October 2014
Accepted 10 April 2015

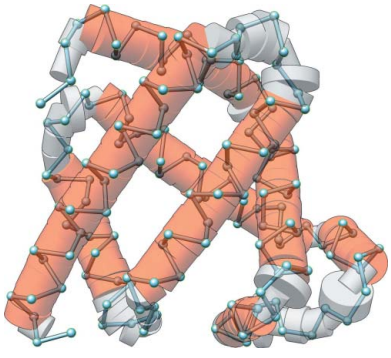
Keywords: secondary-structure description; coarse-grained model; *ScrewFit*.

Supporting information: this article has supporting information at journals.iucr.org/d

A coarse-grained geometrical model for protein secondary-structure description and analysis is presented which uses only the positions of the C^{α} atoms. A space curve connecting these positions by piecewise polynomial interpolation is constructed and the folding of the protein backbone is described by a succession of screw motions linking the Frenet frames at consecutive C^{α} positions. Using the ASTRAL subset of the SCOPe database of protein structures, thresholds are derived for the screw parameters of secondary-structure elements and demonstrate that the latter can be reliably assigned on the basis of a C^{α} model. For this purpose, a comparative study with the widely used *DSSP* (*Define Secondary Structure of Proteins*) algorithm was performed and it was shown that the parameter distribution corresponding to the ensemble of all pure C^{α} structures in the RCSB Protein Data Bank matches that of the ASTRAL database. It is expected that this approach will be useful in the development of structure-refinement techniques for low-resolution data.

1. Introduction

Protein secondary-structure elements (PSSEs) are the basic building blocks of proteins and their form and arrangement is of fundamental importance for protein folding and function. They were first predicted by Pauling and Corey on the basis of hydrogen bonding (Pauling & Corey, 1951; Pauling *et al.*, 1951) and were later confirmed by X-ray diffraction experiments. The localization of PSSEs in protein structure databases is one of the most basic tasks in bioinformatics and various methods have been developed for this purpose. We mention here *DSSP* (*Define Secondary Structure of Proteins*; Kabsch & Sander, 1983) and *STRIDE* (*STRuctural IDEntification*; Frishman & Argos, 1995), which assign PSSEs on the basis of geometrical, energetic and statistical criteria and which are the most widely used approaches. This results in contiguous domains along the amino-acid sequence of the protein, which are labelled 'α-helix', 'β-strand' *etc.* There is no precise and universally accepted definition for PSSEs, and therefore each method produces slightly different results. The geometrical variability of these PSSEs, which depends on the global protein fold, is not explicitly considered by these approaches. In order to account for structural variability owing to protein flexibility, an extension of the *DSSP* method has been proposed which uses a continuous assignment of PSSEs on the basis of *DSSP* analyses with different thresholds for the hydrogen-bond geometry (Andersen *et al.*, 2002). The more recently published *ScrewFit* method (Kneller & Calligari, 2006; Calligari & Kneller, 2012) allows by construction both assignment and geometrical description of PSSEs. It describes the geometry of the whole protein backbone by a succession of screw motions



linking successive C—O—N groups in the peptide bonds, from which PSSEs can be assigned on the basis of statistically established thresholds for the local helix parameters. The latter have been derived by screening the ASTRAL database (Chandonia *et al.*, 2004), which provides representative protein structure sets containing essentially one secondary-structure motif. The *ScrewFit* description is intuitive and bears some resemblance to the *P-Curve* approach proposed by Sklenar *et al.* (1989), in the sense that both methods lead to a sequence of local helix axes, the ensemble of which defines an overall axis of the protein under consideration. *ScrewFit*, however, uses a minimal set of parameters and was originally developed to pinpoint changes in protein structure owing to external stress.

The experimental basis for the automated assignment of PSSEs in proteins is X-ray crystallography, which yields information about the positions of the heavy atoms in a protein. Although the number of resolved protein structures has increased almost exponentially during the last two decades, the fraction of proteins for which the atomic structure is known is still very small. Among the approximately 100 000 protein structures in the RCSB Protein Data Bank (Kirchmair *et al.*, 2008), there are about 600 structures for which only the C^{α} positions on the protein backbone are given. Such structures cannot be analyzed with the widely used *DSSP* method, and the description of the global protein fold as well as the assignment of PSSEs require methods which use only the C^{α} positions. To our knowledge, Levitt and coworkers were the first to publish a method of secondary-structure assignment on the basis of the C^{α} positions (Levitt & Greer, 1977), and different approaches for this purpose have been published since then (Dupuis *et al.*, 2004; Labesse *et al.*, 1997; Park *et al.*, 2011). Like *DSSP* and *STRIDE*, these methods aim at assigning PSSEs on a true/false basis, and the underlying models for this decision are not exploited or not exploitable for a more detailed description of protein folds.

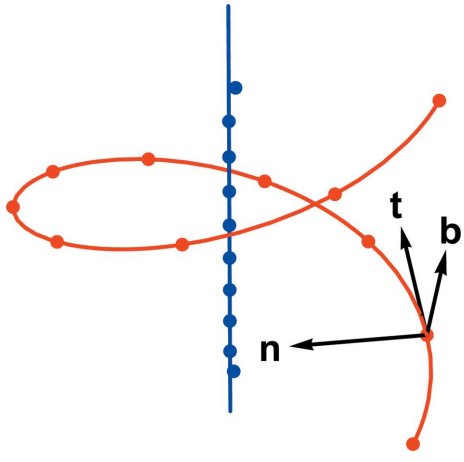


Figure 1
Frenet frame $\{\mathbf{t}, \mathbf{n}, \mathbf{b}\}$ at one point of the helicoidal curve defined in (33) (red solid line). Setting $R = 1$ and $h = 0.3$, the latter is shown for one turn, together with $N = 11$ equidistantly spaced sampling points (red points). The blue line is the helix axis and the blue points correspond to the rotation centres $\mathbf{s}_j^{(c)}$ ($j = 1, \ldots, N-1$). The figure was produced with the *Mathematica* software (Wolfram Research, 2014).

A recent tendency in structural biology is the exploitation of low-resolution images, often from electron microscopy (Grimes *et al.*, 1999; Marabini *et al.*, 2013). Such data sets do not permit structure refinement with all-atom models, but require coarse-grained models for interpretation. Since the vast majority of coarse-grained models for proteins that have been proposed use the C^{α} positions among their variables (Tozzini, 2005), a description of secondary structure based on these positions is likely to become more important in the future. Moreover, such descriptions can be integrated into the structure-refinement method itself, using the regularity of PSSEs as model constraints in much the same way as known constraints on the chemical bond structure are exploited in structure refinements with all-atom models.

The motivation of this paper was to develop an extension of the *ScrewFit* method which works only with the C^{α} positions, maintaining the capability of *ScrewFit* (i) to describe the global fold of a protein by a minimalistic model, (ii) to assign PSSEs and (iii) to characterize variations in PSSEs. In the context of low-resolution modelling, we expect this approach to be most useful as part of a structure-refinement procedure, rather than as an *a posteriori* analysis of a structure refined by other means. Our method is described in §2 and illustrations are presented and discussed in §3. These illustrations have the main goal of showing that our description of protein secondary structure is reasonable. A short résumé with an outlook concludes the paper.

2. A coarse-grained model for the fold of a protein

2.1. C^{α} space curve and Frenet frames

We consider the ensemble of the C^{α} positions, $\{\mathbf{R}_1, \ldots, \mathbf{R}_N\}$, as a discrete representation of a space curve, $\mathbf{r}(\lambda) = \sum_{k=1}^{3} r_k(\lambda)\mathbf{e}^{(k)}$, where $\lambda \in [\lambda_a, \lambda_b]$ and $\mathbf{e}^{(k)}$ ($k = x, y, z$) are the basis vectors of a space-fixed Euclidean coordinate system. Imposing that

$$\mathbf{r}(\lambda_j) = \mathbf{R}_j, \quad j = 1, \ldots, N, \tag{1}$$

at equidistantly sampled values of λ,

$$\lambda_j = \lambda_a + (j-1)\Delta\lambda, \quad \Delta\lambda = (\lambda_b - \lambda_a)/N, \tag{2}$$

we define a continuous space curve by a piecewise polynomial interpolation of the C^{α} positions. The values for λ_a and λ_b are arbitrary and one may in particular choose $\lambda_a = 0$ and $\lambda_b = N$, such that $\Delta\lambda = 1$. At each C^{α} position, we construct the local Frenet basis (Fig. 1) from the interpolated space curve,

$$\mathbf{t}(\lambda) = \frac{\dot{\mathbf{r}}(\lambda)}{|\dot{\mathbf{r}}(\lambda)|}, \tag{3}$$

$$\mathbf{n}(\lambda) = \frac{\dot{\mathbf{t}}(\lambda)}{|\dot{\mathbf{t}}(\lambda)|}, \tag{4}$$

$$\mathbf{b}(\lambda) = \mathbf{t}(\lambda) \wedge \mathbf{n}(\lambda), \tag{5}$$

where $\mathbf{t}$, $\mathbf{n}$ and $\mathbf{b}$ are the tangent vector, the normal vector and the binormal vector to the curve, respectively. The dot denotes a derivative with respect to λ. Interpolating the space curve

around each C^α position with a second-order polynomial involving the respective left and right neighbours, we obtain

$$\dot{\mathbf{r}}(\lambda_j) = \frac{\mathbf{R}_{j+1} - \mathbf{R}_{j-1}}{2\Delta\lambda}, \tag{6}$$

$$\ddot{\mathbf{r}}(\lambda_j) = \frac{\mathbf{R}_{j+1} - 2\mathbf{R}_j + \mathbf{R}_{j-1}}{\Delta\lambda^2} \tag{7}$$

for $j = 2, \ldots, N-1$. At the end points of the chain one can only use forward and backward differences, respectively, and a second-order interpolation of the C^α space would lead to identical $\{\mathbf{t}, \mathbf{n}\}$ planes at the first and last two C^α positions, which is not compatible with a helicoidal curve. In this case we resort to third-order interpolation, such that

$$\dot{\mathbf{r}}(\lambda_1) = \frac{-11\mathbf{R}_1 + 18\mathbf{R}_2 - 9\mathbf{R}_3 + 2\mathbf{R}_4}{6\Delta\lambda}, \tag{8}$$

$$\ddot{\mathbf{r}}(\lambda_1) = \frac{2\mathbf{R}_1 - 5\mathbf{R}_2 + 4\mathbf{R}_3 - \mathbf{R}_4}{\Delta\lambda^2}, \tag{9}$$

$$\dot{\mathbf{r}}(\lambda_N) = \frac{-2\mathbf{R}_{N-3} + 9\mathbf{R}_{N-2} - 18\mathbf{R}_{N-1} + 11\mathbf{R}_N}{6\Delta\lambda}, \tag{10}$$

$$\ddot{\mathbf{r}}(\lambda_N) = \frac{-\mathbf{R}_{N-3} + 4\mathbf{R}_{N-2} - 5\mathbf{R}_{N-1} + 2\mathbf{R}_N}{\Delta\lambda^2}. \tag{11}$$

We note here that the Frenet frames constructed at the C^α positions 2–N are identical to the so-called 'discrete Frenet frames' introduced in Hu *et al.* (2011).

2.2. Relating Frenet frames by screw motions

Having constructed the Frenet frames, the next step consists of constructing the screw motions which link consecutive frames along the protein main chain. For this purpose, the basis vectors $\{\mathbf{t}(\lambda_j), \mathbf{n}(\lambda_j), \mathbf{b}(\lambda_j)\} \equiv \{\mathbf{t}_j, \mathbf{n}_j, \mathbf{b}_j\}$ must be referred to their respective anchor points $\mathbf{R}_j$. Defining

$$\boldsymbol{\varepsilon}_j^{(1)} = \mathbf{t}_j, \quad \boldsymbol{\varepsilon}_j^{(2)} = \mathbf{n}_j, \quad \boldsymbol{\varepsilon}_j^{(3)} = \mathbf{b}_j, \tag{12}$$

the 'tips' of the Frenet basis vectors are located at

$$\mathbf{x}_j^{(k)} = \mathbf{R}_j + \boldsymbol{\varepsilon}_j^{(k)} \quad (k = 1, 2, 3), \tag{13}$$

and the mathematical problem consists of finding the screw parameters for the mappings $\{\mathbf{x}_j^{(k)}\} \rightarrow \{\mathbf{x}_{j+1}^{(k)}\}$ for $j = 1, \ldots, N-1$.

2.2.1. Screw motions. In general, a rigid-body displacement $\mathbf{x} \rightarrow \mathbf{y}$ can be expressed in the form

$$\mathbf{y} = \mathbf{x}^{(c)} + \mathbf{D} \cdot [\mathbf{x} - \mathbf{x}^{(c)}] + \mathbf{t}, \tag{14}$$

where $\mathbf{x}^{(c)}$ is the centre of rotation, $\mathbf{D}$ is a rotation matrix and $\mathbf{t}$ is a translation vector. By construction,

$$\mathbf{t} = \mathbf{y}^{(c)} - \mathbf{x}^{(c)}. \tag{15}$$

The elements of the rotation matrix can be expressed in terms of three independent real parameters. One possible choice is to use the rotation angle φ and the unit vector $\mathbf{n}$ pointing in the direction of the rotation axis. For this parametrization, $\mathbf{D}$ has the form (Altmann, 1986)

$$\mathbf{D}(\mathbf{n}, \varphi) = \cos\varphi\mathbf{1} + (1 - \cos\varphi)\mathbf{P} + \sin\varphi\mathbf{N}, \tag{16}$$

where $\mathbf{P} = (n_i n_j)$ $(i, j = 1, 2, 3)$ is the projector on $\mathbf{n}$ and $\mathbf{N}$ is a skew-symmetric 3×3 matrix which is defined by the relation $\mathbf{N} \cdot \mathbf{v} = \mathbf{n} \wedge \mathbf{v}$ for an arbitrary vector $\mathbf{v}$. The elements of $\mathbf{N}$ are $N_{ij} = -\sum_k \varepsilon_{ijk} n_k$, where ε_{ijk} $(i, j, k = 1, 2, 3)$ are the components of the totally antisymmetric Levi–Civita tensor. We recall that $\varepsilon_{ijk} = \pm 1$ for an even and odd permutation of 123, respectively, and $\varepsilon_{ijk} = 0$ otherwise. The parameters of the rigid-body displacement (14) depend on the choice of the rotation centre, $\mathbf{x}^c$, and there is a special choice, $\mathbf{x}^c = \mathbf{s}$, for which the translation vector $\mathbf{t}$ points in the direction of the rotation axis $\mathbf{n}$, such that $\mathbf{t} \cdot \mathbf{n} > 0$. This is known as Chasles' theorem (Chasles, 1830) and the corresponding rigid-body displacement describes a screw motion,

$$\mathbf{y} = \mathbf{s} + \mathbf{D}(\mathbf{n}, \varphi) \cdot (\mathbf{x} - \mathbf{s}) + \alpha\mathbf{n}. \tag{17}$$

Using that $\mathbf{D}(\mathbf{n}, \varphi) \cdot \mathbf{n} = \mathbf{n}$, one can easily show that α is the projection of the translation vector on the rotation axis,

$$\alpha = \mathbf{t} \cdot \mathbf{n}. \tag{18}$$

The position $\mathbf{s}$ is not uniquely defined, but stands for all points on the screw axis. Defining $\mathbf{s}^c$ to be the point for which the distance $|\mathbf{s} - \mathrm{x}^c|$ is a minimum, the screw axis is defined through

$$\mathbf{s} = \mathbf{s}^{(c)} + \mu\mathbf{n}, \quad -\infty < \mu < +\infty, \tag{19}$$

where

$$\mathbf{s}^{(c)} = \mathbf{x}^{(c)} + \frac{1}{2}[\mathbf{t}^\perp + \cos(\varphi/2)\mathbf{n} \wedge \mathbf{t}], \tag{20}$$

and $\mathbf{t}^\perp = \mathbf{t} - (\mathbf{n} \cdot \mathbf{t})\mathbf{n}$ is the component of $\mathbf{t}$ which is perpendicular to the rotation axis. We note that $[\mathbf{s}^{(c)} - \mathbf{x}^{(c)}] \cdot \mathbf{t} = 0$. The radius of the screw motion is defined through $\rho = |\mathbf{x}^{(c)} - \mathbf{s}^{(c)}|$ and it follows from (20) that

$$\rho = \frac{|\mathbf{t}^\perp|}{2}[1 + \cot(\varphi/2)^2]^{1/2}. \tag{21}$$

2.2.2. Determining the screw parameters. Assuming that the Frenet frames at the C^α positions have been constructed, the fold of a protein is defined by the sequence of screw motions $\mathbf{x}_j^{(k)} \rightarrow \mathbf{x}_{j+1}^{(k)}$, where

$$\mathbf{x}_{j+1}^{(k)} = \mathbf{s}_j^{(c)} + \mathbf{D}(\mathbf{n}_j, \varphi_j) \cdot [\mathbf{x}_j^{(k)} - \mathbf{s}_j^{(c)}] + \alpha_j\mathbf{n}_j, \tag{22}$$

for $j = 1, \ldots, n-1$ and $k = 1, 2, 3$. The corresponding parameters are computed as follows.

(i) Determine the translation vectors

$$\mathbf{t}_j = \mathbf{R}_{j+1} - \mathbf{R}_j. \tag{23}$$

(ii) Perform a rotational least-squares fit (Kneller, 1991) $\{\boldsymbol{\varepsilon}_j^{(k)}\} \rightarrow \{\boldsymbol{\varepsilon}_{j+1}^{(k)}\}$ by minimizing the target function

$$m(Q_j) = \sum_{k=1}^{3} \left| \boldsymbol{\varepsilon}_{j+1}^{(k)} - \mathbf{D}(Q_j) \cdot \boldsymbol{\varepsilon}_j^{(k)} \right|^2 \tag{24}$$

with respect to four quaternion parameters, $\mathrm{Q} = \{q_0, q_1, q_2, q_3\}$, which parametrize the rotation matrix according to

$$\mathbf{D}(Q) = \begin{bmatrix} q_0^2+q_1^2-q_2^2-q_3^2 & 2(q_1q_2-q_0q_3) & 2(q_0q_2+q_1q_3) \\ 2(q_1q_2+q_0q_3) & q_0^2-q_1^2+q_2^2-q_3^2 & -2(q_0q_1-q_2q_3) \\ -2(q_0q_2-q_1q_3) & 2(q_0q_1+q_2q_3) & q_0^2-q_1^2-q_2^2+q_3^2 \end{bmatrix}. \tag{25}$$

The quaternion parameters are normalized such that $q_0^2, q_1^2, q_2^2, q_3^2 = 1$, which leaves three free parameters describing the rotation. We note here only that the minimization of (24) leads to an eigenvector problem for the optimal quaternion, which can be efficiently solved by standard linear algebra routines, and that the corresponding eigenvalue is the squared superposition error (Kneller, 1991). The latter is zero for superposition of Frenet frames, since two orthonormal and equally oriented vector sets can be perfectly superposed. It is also worthwhile noting that the upper limit in the sum in (24) can be changed from 3 to 2, since two linearly independent vectors with the same origin, here $\mathbf{t}_j$ and $\mathbf{n}_j$, suffice to define a rigid body.

(iii) Extract $\mathbf{n}_j$ and φ_j from the quaternion parameters Q_j. This can easily be achieved by expoiting the relations

$$\left.\begin{aligned} q_0 &= \cos(\varphi/2) \\ q_1 &= \sin(\varphi/2)n_x \\ q_2 &= \sin(\varphi/2)n_y \\ q_3 &= \sin(\varphi/2)n_z \end{aligned}\right\}. \tag{26}$$

Here and in the following the index j is dropped. Several cases have to be considered. If $(q_1^2, q_2^2, q_3^2)^{1/2} > \varepsilon$, where ε depends on the machine precision of the computer being used, we compute a 'tentative rotation axis'

$$\mathbf{n}_t = \frac{1}{(q_1^2+q_2^2+q_3^2)^{1/2}} \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}. \tag{27}$$

Then we check whether $\mathbf{t}\cdot\mathbf{n}_t \geq 0$. If this is the case, we set

$$\mathbf{n} = \mathbf{n}_t, \tag{28}$$

$$\varphi = 2\arccos(q_0). \tag{29}$$

In the case that $\mathbf{t}\cdot\mathbf{n}_t < 0$, we set

$$\mathbf{n} = -\mathbf{n}_t, \tag{30}$$

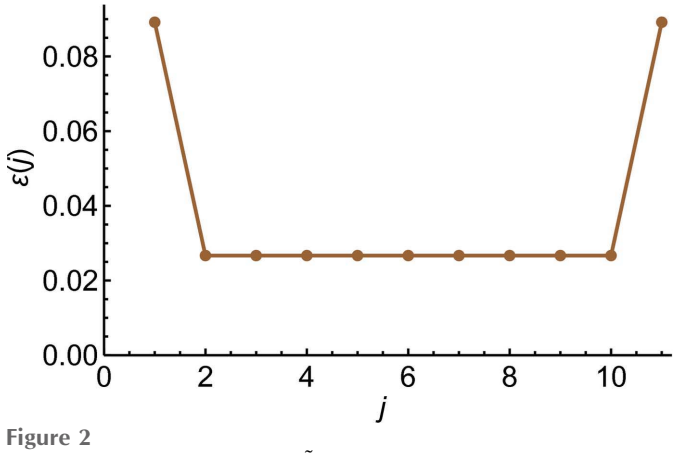


Figure 2
Overlap error (35) for the bases $\tilde{\mathbf{F}}(\lambda_j)$ and $\mathbf{F}(\lambda_j)$ at the red points in Fig. 8.

$$\varphi = 2\arccos(-q_0). \tag{31}$$

This corresponds to replacing $Q \rightarrow -Q$ before evaluating $\mathbf{n}$ and φ according to (28) and (29). Such a replacement is possible since the elements of $\mathbf{D}(Q)$ are homogeneous functions of order two in the quaternion parameters, such that $\mathbf{D}(Q) = -\mathbf{D}(Q)$.

For the sake of completeness, we finally mention the case that $(q_1^2, q_2^2, q_3^2)^{1/2} \leq \varepsilon$, which corresponds to a pure translation and cannot occur in our application to protein backbones. In this case, one would set $\varphi = 0$ and $\mathbf{n} = \mathbf{t}/|\mathbf{t}|$.

(iv) Using the parameters $\{\mathbf{n}_j, \varphi_j\}$ and defining the positions $\mathbf{R}_j$ to be the rotation centres, $\mathbf{x}^{(c)} = \mathbf{R}_j$, compute for $j = 1, \ldots, N-1$ the positions $\mathbf{s}_j^{(c)}$ on the local screw axes according to relation (20) and the local helix radii according to relation (21).

2.2.3. Regularity of PSSEs. To quantify the regularity of PSSEs, we introduce the distance measure

$$\delta(j) = \left|\mathbf{s}_j^{(c)} + \mathbf{t}_j^{\parallel} - \mathbf{s}_{j+1}^{(c)}\right|, j = 1, \ldots, N-2, \tag{32}$$

where $\mathbf{t}_j^{\parallel} = (\mathbf{n}_j \cdot \mathbf{t}_j)\mathbf{n}_j$. For an ideal PSSE, where all consecutive Frenet frames are related by the same screw motion, $\delta(j)$ is strictly zero. This measure of non-ideality deviates from the 'straightness' parameter in the *ScrewFit* algorithm (Kneller & Calligari, 2006), which is defined as $\sigma_j = \boldsymbol{\mu}_{j+1}\cdot\boldsymbol{\mu}_j/(|\boldsymbol{\mu}_{j+1}||\boldsymbol{\mu}_j|)$, with $\boldsymbol{\mu}_j = \mathbf{s}_{j+1}^{(c)} - \mathbf{s}_j^{(c)}$, and which defines the ideality of PSSEs through the cosine of the angle between subsequent local screw axes.

2.3. Numerical test

To test the numerical construction of Frenet frames, we consider a perfect helicoidal curve and compare the exact Frenet frames with the corresponding numerical approximations. The parametric representation of the curve is

$$\mathbf{r}(\lambda) = \rho\cos(\lambda)\mathbf{e}^{(x)} + \rho\sin(\lambda)\mathbf{e}^{(y)} + h\lambda\mathbf{e}^{(z)}, \tag{33}$$

where $\rho > 0$ is the radius of the helix and its pitch is $p = h/2\pi$. Fig. 4 shows the form of the curve (33) for one complete turn (red line), setting $R = 1$ and $h = 0.3$ in arbitrary length units. Defining the matrix $\mathbf{F}(\lambda) = [\mathbf{t}(\lambda), \mathbf{n}(\lambda), \mathbf{b}(\lambda)]$, it follows from (33) that

$$\mathbf{F}(\lambda) = \begin{bmatrix} -\dfrac{R\sin(\lambda)}{(h^2+R^2)^{1/2}} & -\cos(\lambda) & \dfrac{h\sin(\lambda)}{(h^2+R^2)^{1/2}} \\ \dfrac{R\cos(\lambda)}{(h^2+R^2)^{1/2}} & -\sin(\lambda) & -\dfrac{h\cos(\lambda)}{(h^2+R^2)^{1/2}} \\ \dfrac{h}{(h^2+R^2)^{1/2}} & 0 & \dfrac{R}{(h^2+R^2)^{1/2}} \end{bmatrix}. \tag{34}$$

Using the method described in §2.1, we construct numerical approximations $\tilde{\mathbf{F}}(\lambda_j)$ of the Frenet bases (34) at $N = 11$ equidistant sampling points, $\mathbf{R}_j$, which are shown as red dots in Fig. 4. From these Frenet bases we construct the axis points $\mathbf{s}_j^{(c)}$ (blue dots), which are shown together with the exact screw axis (blue line). For the first and the last axis point one notices a visible offset from the latter. We quantify the error of the numerically computed Frenet bases, $\tilde{\mathbf{F}}(\lambda_j)$, as

$$\varepsilon(j) = \{\mathrm{tr}[\mathbf{\Delta}(j)^T \cdot \mathbf{\Delta}(j)]\}^{1/2}, \quad (35)$$

where

$$\mathbf{\Delta}(j) = \tilde{\mathbf{F}}(\lambda_j)^T \cdot \mathbf{F}(\lambda_j) - \mathbf{1}. \quad (36)$$

For a perfect overlap of $\tilde{\mathbf{F}}(\lambda_j)$ and $\mathbf{F}(\lambda_j)$ one should have $\tilde{\mathbf{F}}(\lambda_j)^T \cdot \mathbf{F}(\lambda_j) = \mathbf{1}$, such that $\varepsilon(j) = 0$. We note that $\varepsilon(j)$ is the Frobenius norm (Golub & van Loan, 1996) of $\mathbf{\Delta}(j)$. Fig. 2 shows $\varepsilon(j)$ corresponding to the Frenet basis in Fig. 4 and confirms the slight offset of the first and last axis points from the ideal screw axis.

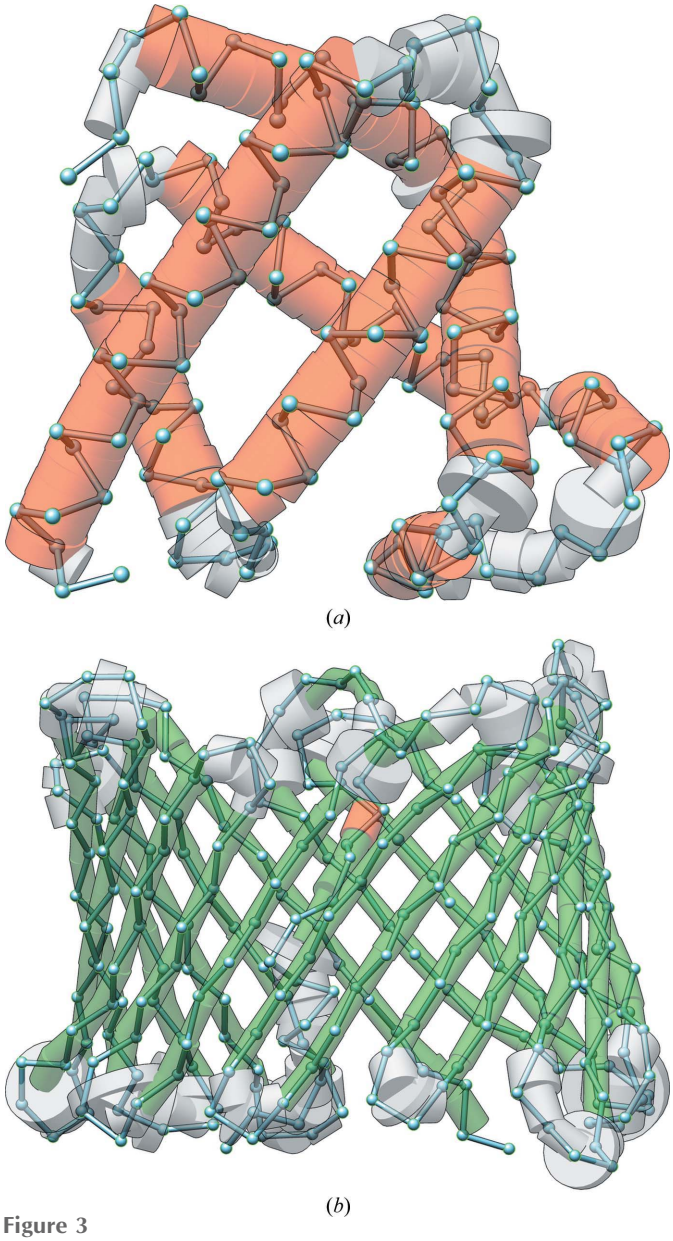


Figure 3
(*a*) C^α trace (cyan) of myoglobin (PDB entry 1a6g, essentially α-helices) and the tube representation computed from the *ScrewFrame* parameters. The details are described in the text. (*b*) The corresponding figure for human VDAC-1 (PDB entry 2k4t, essentially β-strands). The colouring scheme for the tubes is green for β-strands and red for α-helices.

Table 1
Screw radii (nm) for standard model structures generated with *Chimera* (Pettersen *et al.*, 2004). Since *ScrewFit* uses the C atoms in the peptide planes as reference points for the (pure) rotations, whereas *ScrewFrame* uses the C^α atoms, the radii determined by *ScrewFit* are systematically smaller than those obtained from *ScrewFrame*.

	α-Helix	Parallel β-strand	Antiparallel β-strand	3_{10}-Helix	π-Helix
ScrewFit	0.165	0.061	0.051	0.122	0.165
ScrewFrame	0.227	0.098	0.080	0.187	0.227

3. Applications

We will now consider four applications of the coarse-grained protein model described above, which will be referred to as *ScrewFrame* in the following.

(i) The first application concerns the construction of a tube model for two proteins whose secondary-structure elements are, respectively, essentially α-helices and β-strands.

(ii) In the second application, we explore the stability of the most important *ScrewFrame* parameters, ρ and δ, under perturbations of the protein structure.

(iii) The third application is a comparative study of *ScrewFrame* and *DSSP* for secondary-structure assignment. We provide a comparison with the *de facto* standard *DSSP* as a proof of the validity of our approach.

(iv) The fourth application is devoted to a secondary-structure analysis of all protein structures in the RSCB Protein Data Bank for which only the positions of the C^α atoms are known.

3.1. Tube representation of a protein

As a first application, we present *ScrewFrame* analyses of myoglobin, an oxygen-binding globular protein in muscular tissues which contains essentially α-helices (PDB entry 1a6g; Vojtechovsky *et al.*, 1999), and of the integral human membrane protein VDAC-1, in which the predominant PSSEs are β-strands (PDB entry 2k4t; Hiller *et al.*, 2008). Figs. 3(*a*) and 3(*b*) show tube models of the respective proteins which have been constructed from the *ScrewFrame* parameters. The tube is a succession of cylinders whose radii are defined by the *ScrewFrame* parameter ρ (see equation 21), which describes the radius of the screw motion linking consecutive Frenet frames. The axis of each cylinder is the local screw axis and its height is the distance between two consecutive screw-motion centres $\mathbf{s}_j^{(c)}$ and $\mathbf{s}_{j+1}^{(c)}$ (see equation 20) on that axis. By definition, the C^α atoms are on the surface of the tube. As in the original *ScrewFit* algorithm, the screw radius allows the discrimination of different types of PSSEs (see Table 1) and the tube is coloured red to indicate α-helices and green for β-strands. The protein main axis is the concatenation of all local screw axes and it plays the same role as the 'overall protein axis' in the *P-Curve* algorithm (Sklenar *et al.*, 1989), although its construction is different. We provide the tube models for these two proteins as Supporting Information in the form of BILD files for the molecular-visualization program *Chimera* (Pettersen *et al.*, 2004).

Figs. 4(*a*) and 4(*b*) display the parameter ρ for myoglobin and VDAC-1, respectively, as a function of the residue index (blue line). For comparison we also show the α-helices found by *DSSP*, which are indicated by the vertical stripes in dark grey. The horizontal stripes in light grey indicate the tolerance interval for the ρ parameter for α-helices and β-strands, respectively, the definition of which will be described in the following section. Fig. 5 shows the corresponding regularity measure δ (see equation 32) for myoglobin (Fig. 5*a*) and VDAC-1 (Fig. 5*b*), which plays an important role in the attribution of secondary-structure elements that will be discussed in the following section.

3.2. Perturbation analysis

In view of the application of our approach to low-resolution data, it is interesting to explore the influence of perturbations in the C^{α} positions on the resulting *ScrewFrame* parameters. We limit ourselves to the parameters ρ and δ, which are the most important parameters for characterizing the secondary structure of a protein. As examples, we consider the PDB structures 1a6g for myoglobin and 2k4t for the VDAC-1 protein, which were treated in the previous section. The Cartesian coordinates of all C^{α} atoms are shifted by random numbers which are drawn from a normal distribution with zero mean and a prescribed width ε. For the latter we chose ten values, increasing from 0.01 to 0.1 nm in steps of 0.01 nm, and for each value of ε we generate 1000 random configurations. Fig. 6 displays the results of the perturbation analysis for myoglobin (Fig. 6*a*) and VDAC-1 (Fig. 6*b*). In both cases we show

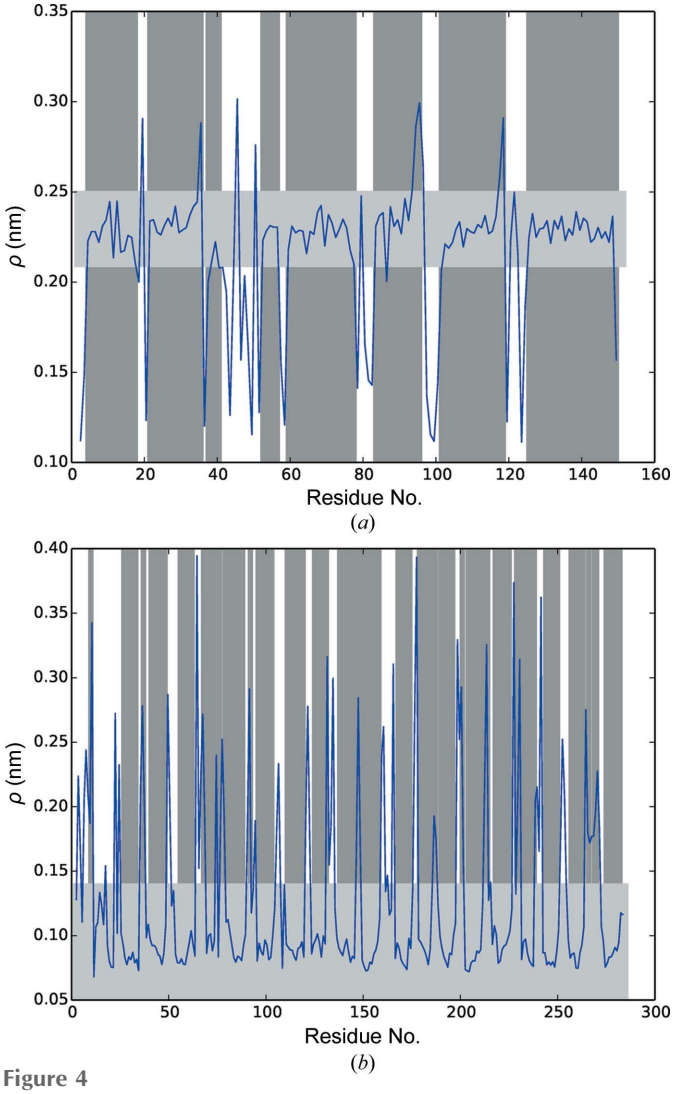


Figure 4
(*a*) The *ScrewFrame* parameter ρ (blue line) for myoglobin (PDB entry 1a6g) as a function of the residue number. The vertical dark grey stripes indicate the α-helices found by *DSSP* and the horizontal light grey stripe indicates the tolerance for the ρ parameter in the case of an α-helix (see equation 37). (*b*) The same representation for human VDAC-1 (PDB entry 2k4t), where the horizontal grey stripe indicates the tolerance for the ρ parameter in the case of a β-strand (see equation 40).

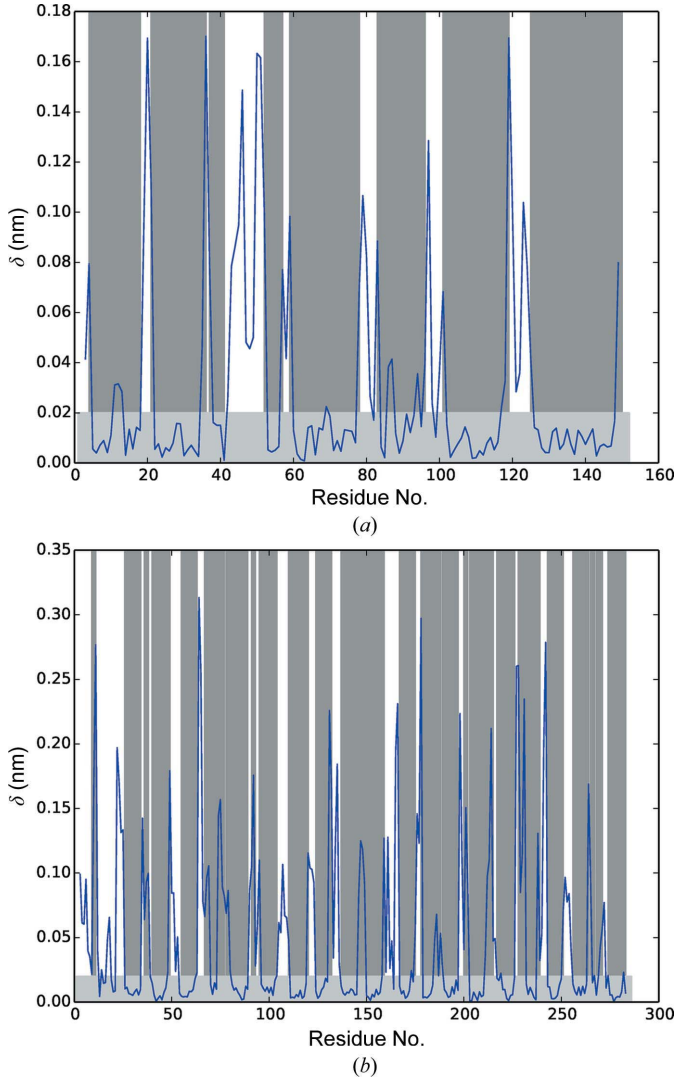


Figure 5
(*a*) The *ScrewFrame* parameter δ (blue line) for myoglobin (PDB entry 1a6g) as a function of the residue number. The vertical dark grey stripes indicate the α-helices found by *DSSP* and the horizontal light grey stripe indicates the tolerance for the δ parameter in the case of an α-helix (see equation 37). (*b*) The same representation for human VDAC-1 (PDB entry 2k4t), where the horizontal grey stripe indicates the tolerance for the δ parameter in the case of a β-strand (see equation 40).

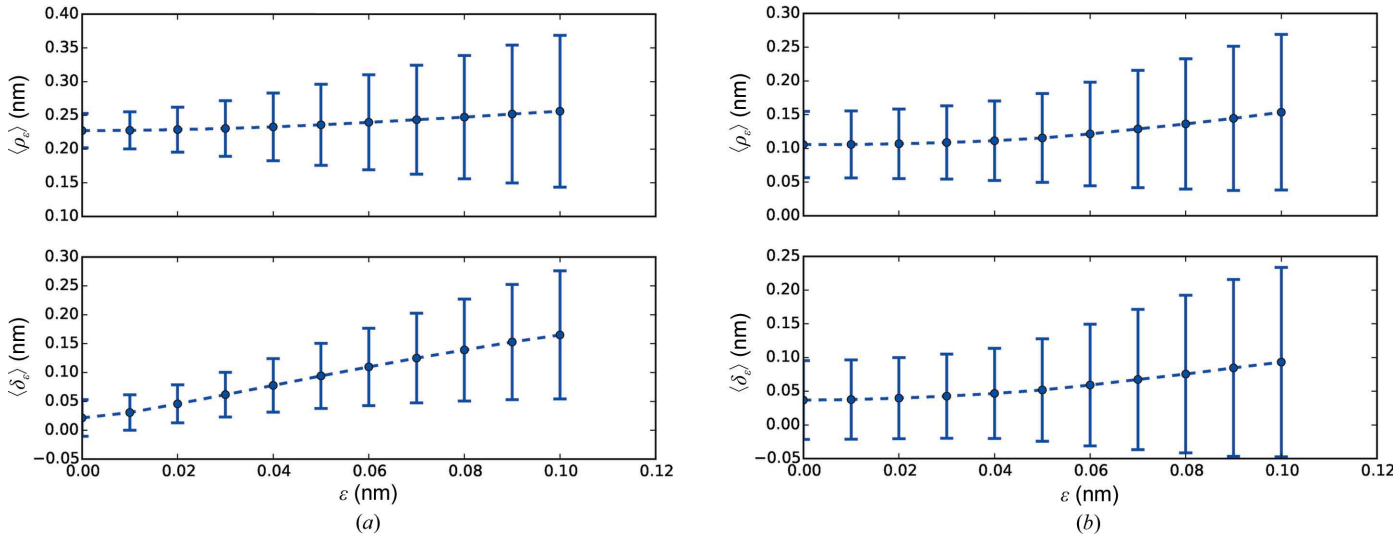


Figure 6
Sensitivity of the *ScrewFrame* parameters to perturbations in the input structure. (*a*) Myoglobin: the upper panel shows the mean value for the helix radius ρ averaged over the residues belonging to α-helices and all random configurations (dashed line), together with the standard deviation (vertical bars). The lower panel shows the corresponding analysis for the regularity parameter δ. (*b*) VADC-1: the upper panel shows the mean value for the helix radius ρ averaged over the residues belonging to β-strands and all random configurations (dashed line), together with the corresponding standard deviation (vertical bars). The lower panel shows the corresponding analysis for the regularity parameter δ.

(i) the mean value for the helix radius ρ averaged over all residues belonging to the respective dominant motif (α-helices in the case of myoglobin and β-strands in the case of VDAC-1) and all random configurations (dashed line), together with the corresponding standard deviation (vertical bars), and

(ii) the corresponding analysis for the regularity parameter δ. The results show that ρ is a robust parameter, which increases slowly with increasing noise, whereas δ reacts much more strongly. It is indeed important for ρ to be robust, because we use it to distinguish between different types of secondary-structure elements. The role of δ is quite different: it is a quality parameter that measures the regularity of the protein fold and is used to distinguish 'good' from 'bad' secondary-structure elements. It is thus to be expected that δ should increase for less well defined input structures. We expect this dependence to become useful in structure-refinement applications, where a restraint on δ can be used to enforce well defined secondary-structure elements.

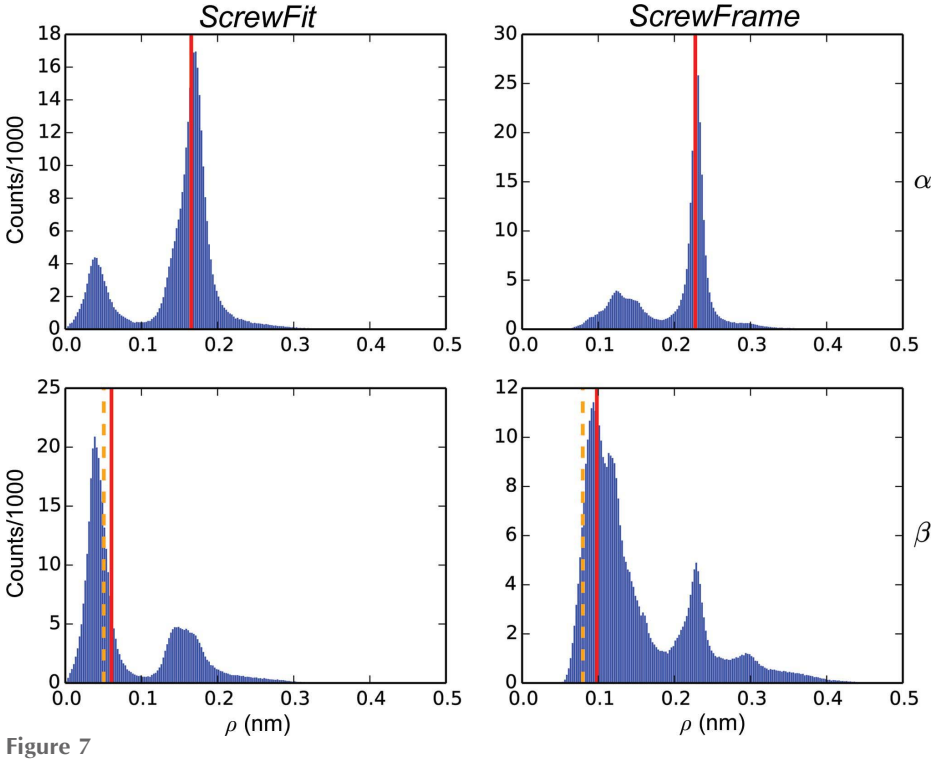


Figure 7
The helix radius ρ for the all-α (top) and all-β (bottom) structures using the *ScrewFit* (left) and *ScrewFrame* (right) methods. Note that the *ScrewFit* radius is based on the C atoms, whereas the *ScrewFrame* radius corresponds to the C^α atoms, which explains the different values. The vertical lines indicate the values for ideal secondary-structure elements. For β-strands there are two ideal values, one for parallel (red) and one for antiparallel (orange, dashed) strands.

3.3. Analysis of the ASTRAL database

In order to compare our C^α-based helicoidal analysis with the original *ScrewFit* method based on peptide planes (Kneller & Calligari, 2006; Calligari & Kneller, 2012), we applied both methods to the 'all-α' and 'all-β' categories of the ASTRAL subset of the SCOPe database (Fox *et al.*, 2013) using the ASTRAL SCOPe 2.04 subset with less than 40% sequence identity. In order to be able to work efficiently with such a large collection of protein structures, we constructed

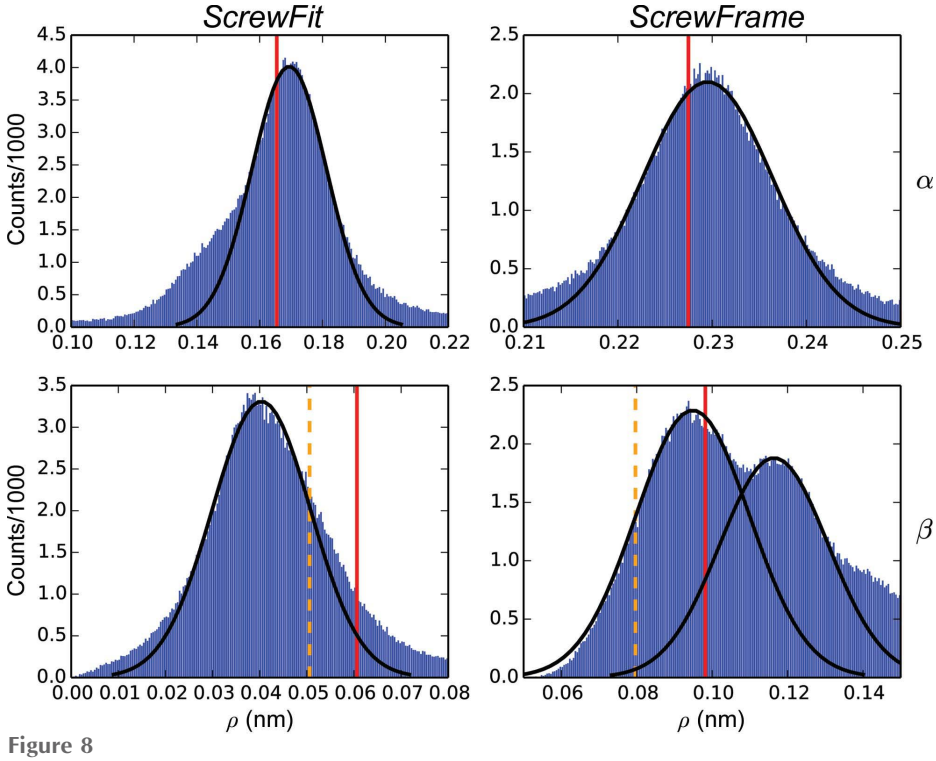


Figure 8
The helix radius ρ around the ideal α value for the all-α structure (top) and around the ideal β value for the the all-β structure (bottom) using the *ScrewFit* (left) and *ScrewFrame* (right) methods. The vertical lines indicate values for ideal secondary-structure elements, as in Fig. 5. The Gaussian distributions fitted to the peaks are drawn in black; their parameters are given in Table 2. The β distribution for *ScrewFrame* can be described well as a superposition of two Gaussian distributions corresponding to parallel and antiparallel strands. The *ScrewFit* method cannot resolve this difference.

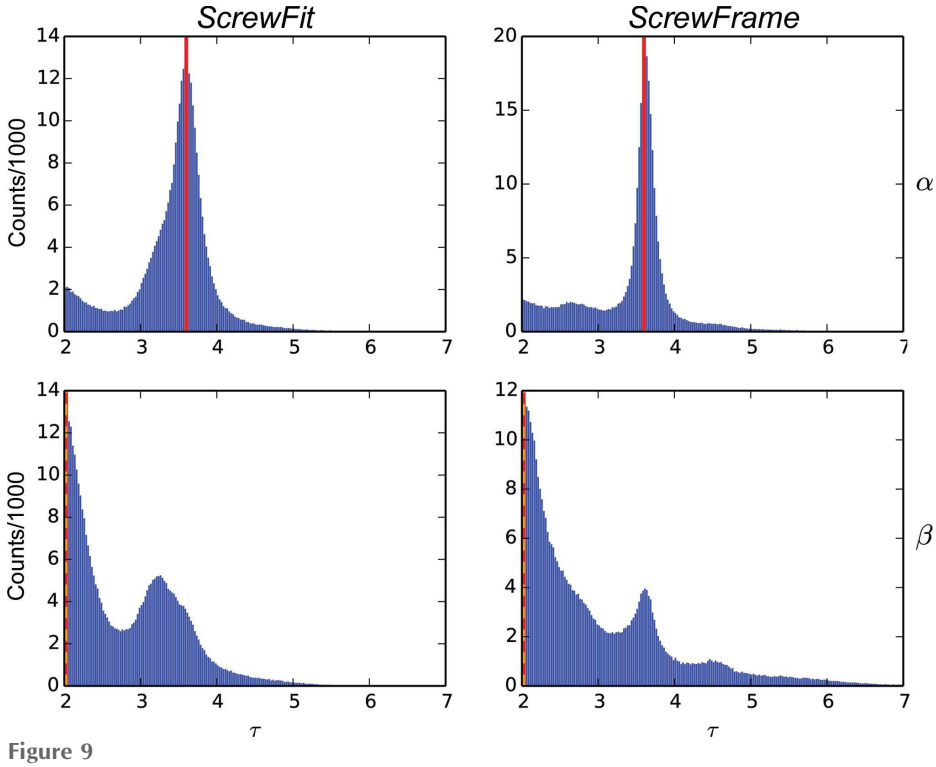


Figure 9
The number of amino-acid residues per full turn, τ, for the all-α (top) and all-β (bottom) structures using the *ScrewFit* (left) and *ScrewFrame* (right) methods. The theoretical minimal value of $\tau = 2$ is very close to the observed value for β-sheets.

an ActivePaper (Hinsen, 2014*a*) containing the structures of the ASTRAL entries in MOSAIC format (Hinsen, 2014*c*). This file is available for download (Hinsen, 2014*b*). In addition to the ASTRAL database of real protein structures, we use ideal secondary-structure elements (α-helix, π-helix, 3_{10}-helix, parallel and antiparallel β-strands) for polyalanine, which were constructed using the program *Chimera* (Pettersen *et al.*, 2004).

We also make comparisons with *DSSP* secondary-structure assignments for this database, using our own implementation of the *DSSP* algorithm which follows the description in the original publication (Kabsch & Sander, 1983) but, like the current version 2 of the *DSSP* software (Hekkelman, 2013), computes an ideal position for the backbone hydrogen positions instead of using experimental values, even if the latter are available.

As a first step, we compute *ScrewFit* and *ScrewFrame* parameters for all structures in the all-α and all-β subsets of the ASTRAL database. In order to avoid inaccuracies introduced by the third-order approximations given by (8)–(11), we do not compute Frenet frames for the first and last residue of each chain. For structures with missing residues, we compute the parameters for each continuous chain segment separately. Since the input structures are dominated by α-helices and β-strands, respectively, we expect the distribution of our parameters to show clear peaks that correspond to these secondary-structure elements.

The most important helix parameter for secondary-structure description is the helix radius ρ, the distribution of which in the ASTRAL database is shown in Fig. 7. The vertical lines show, for comparison, the values for ideal α-helices and β-strands. For the β-strands, the red lines represent parallel strands and the orange dashed lines represent antiparallel strands. A more detailed view is given in Fig. 8, which shows only the region around the dominant peak for each histogram, together with

Table 2
The parameters of the Gaussians fitted to the peaks in the distributions of the *ScrewFrame* parameter ρ (see Fig. 5).

	α-Helix	Parallel β-strand	Antiparallel β-strand
μ_ρ (nm)	0.230	0.116	0.095
σ_ρ (nm)	0.007	0.014	0.015

Gaussian distributions fitted to the peaks. The peaks are rather well described by a Gaussian, and the *ScrewFrame* method even allows the difference between parallel and antiparallel β-strands to be resolved.

Whereas the average ρ value for α-helices is close to the value for an ideal helix, this is not the case at all for β-strands. This can be understood by looking at the distribution of the number of amino acids per full turn, τ, shown in Fig. 9. Since the rotation angle is by definition in the interval $(-\pi \ldots \pi)$, the minimal value of τ is 2. This is also the value that describes an ideal β-strand, which is a flat structure. Any deviation from the ideal β-strand has a larger τ, and because ρ and τ are not independent (the length of the curve arc linking two neighbouring C^α atoms is nearly constant), the deviation in ρ from the ideal value is also asymmetric.

The regularity measure δ, defined in (32), is shown in Fig. 10. It shows that the *ScrewFrame* secondary-structure elements are more regular than those identified by *ScrewFit*, in particular for structures dominated by α-helices. We do not show the distributions of the other parameters defined in the initial *ScrewFit* publication here (Kneller & Calligari, 2006), but they are included in the Supporting Information. We note that the parameter distributions are in general narrower and thus better defined for *ScrewFrame* than for *ScrewFit*. We attribute this fact to fluctuations in the orientations of the peptide planes that have no impact on the C^α geometry.

We use the Gaussian distributions shown in Fig. 8 as the basis for defining secondary-structure elements. We define an α-helix as a sequence of at least four consecutive C^α atoms whose screw transformations satisfy

$$\frac{|\rho - \mu_\rho|}{\sigma_\rho} < 3, \tag{37}$$

$$\delta < 0.02 \text{ nm}, \tag{38}$$

where μ_ρ and σ_ρ are the mean value and standard deviation of the Gaussian distribution for the α peak in Fig. 8. The numerical values of these parameters are shown in Table 2. We define a β-strand as a segment of consecutive C^α atoms whose screw transformations satisfy

$$\min\left(\frac{|\rho - \mu_\rho^{(1)}|}{\sigma_\rho^{(1)}}, \frac{|\rho - \mu_\rho^{(2)}|}{\sigma_\rho^{(2)}}\right) < 1, \tag{39}$$

$$\delta < 0.08 \text{ nm}, \tag{40}$$

where $\mu_\rho^{(1/2)}$ and $\sigma_\rho^{(1/2)}$ are the mean values and standard deviations of the Gaussian distributions for the parallel and antiparallel β peaks in Fig. 8. The numerical parameters in these definitions were chosen to make our definitions match the secondary-structure assignments made by the *DSSP* method.

There is a fundamental difference between our approach and the *DSSP* method for defining β-strands. The *ScrewFrame* approach looks for a regular structure along the peptide chain, whereas the *DSSP* method identifies hydrogen bonds between the strands that make up a β-sheet. *ScrewFrame* thus finds individual strands, which can be paired up to identify sheets in a separate step. A strand must consist of at least three consecutive residues in order to be considered regular; in fact, the regularity measure δ is defined in terms of the difference of two consecutive screw transformations, each of which connects two residues. *DSSP* needs to look at two strands simultaneously in order to identify β structures, but has no minimal length condition and in fact admits β-sheets as

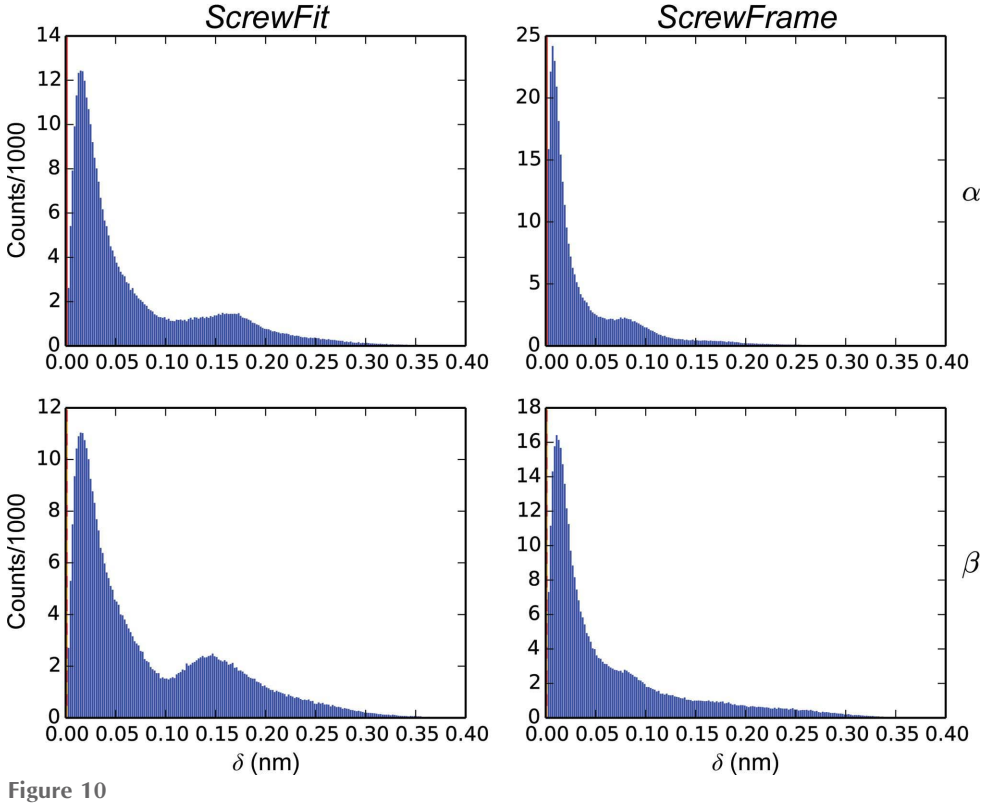


Figure 10
The regularity measure δ defined in (32) for the all-α and all-β subsets of the ASTRAL database (top and bottom, respectively).

small as a single hydrogen-bonded residue pair. For practically relevant β-sheets in real protein structures, these differences are however not important, but they must be understood in order to interpret the following comparison between the two methods.

A one-to-one comparison of secondary-structure elements from two different assignment methods is not of particular interest, because an exact match is the exception rather than the rule. The inherent fuzziness of secondary-structure definitions leads to arbitrary choices and thus inevitable differences. The most frequent deviation between two assignments is at the end points of secondary-structure elements, where a difference of one or two residues is common and acceptable. Another frequent deviation concerns deformed secondary-structure elements, which one method may identify as single elements whereas another method may recognize them as multiple distinct elements.

We therefore chose a statistical comparison to compare the *ScrewFrame* results with those from *DSSP*, which is shown in Fig. 11 for α-helices and Fig. 12 for β-strands. We consider two quantities: (i) the total number of residues of a given structure which are inside a recognized secondary-structure element and (ii) the length of each individual secondary-structure element. We computed the first quantity for both methods and show their joint distribution (Figs. 11*a* and 12*a*). For the vast majority of structures the two residue counts are close to

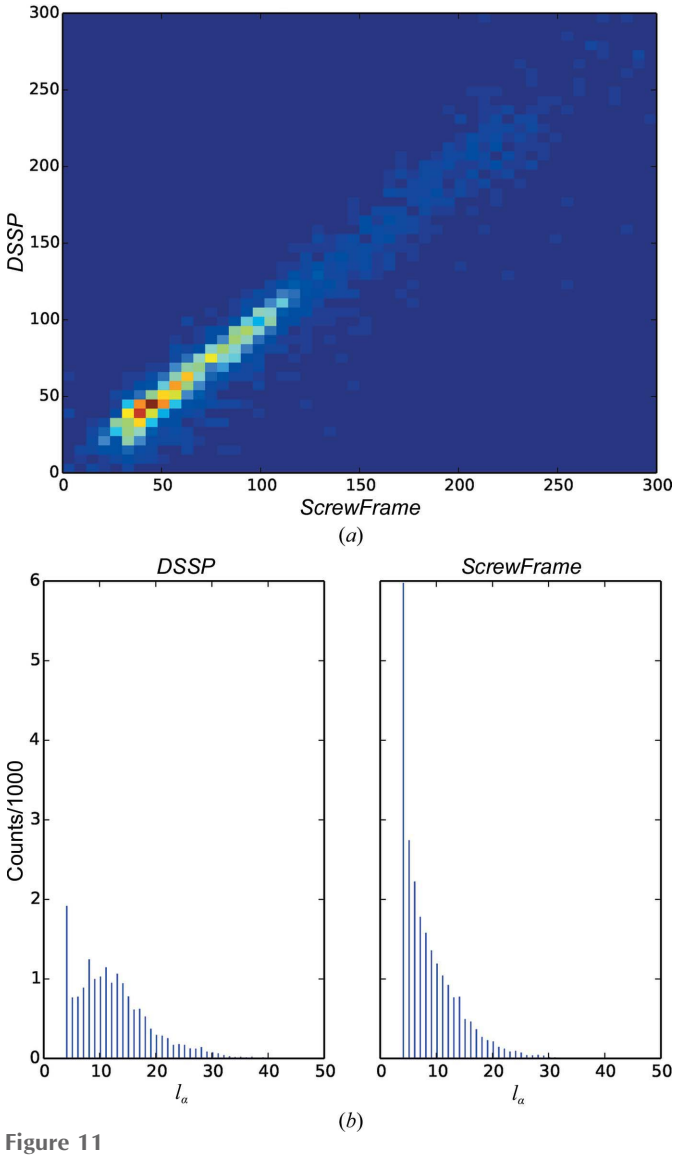


Figure 11
(*a*) A two-dimensional histogram comparing the total number of residues inside α-helices as identified by *ScrewFrame* and *DSSP*. The strong localization of the distribution around the diagonal shows the similarity between these two assignments. (*b*) The distribution of the lengths of identified α-helices; left, *DSSP*; right, *ScrewFrame*. The fatter tail for *DSSP* and the larger number of short helices for *ScrewFrame* are owing to the fact that *ScrewFrame* breaks up strongly deformed helices into several pieces, whereas *DSSP* considers them to be a single helix.

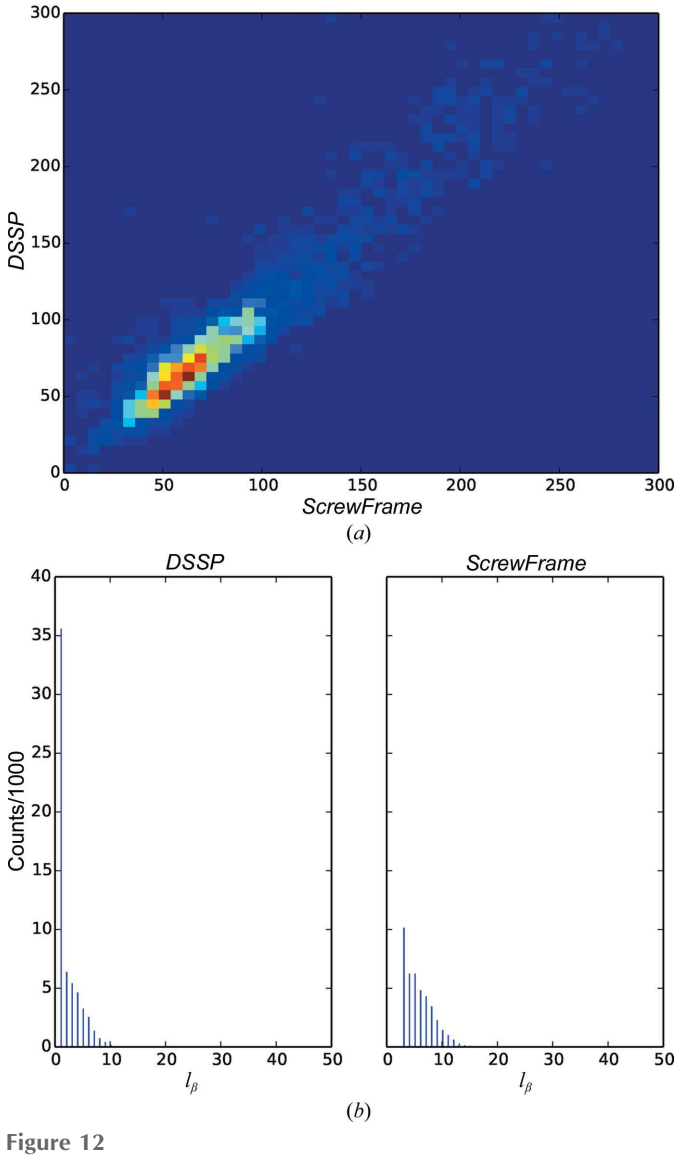


Figure 12
(*a*) A two-dimensional histogram comparing the total number of residues inside β-strands as identified by *ScrewFrame* and *DSSP*. The strong localization of the distribution around the diagonal shows the similarity between these two assignments. (*b*) The distribution of the lengths of identified β-helices; left, *DSSP*; right, *ScrewFrame*. The peak at very short strands in the *DSSP* distribution is absent from the *ScrewFrame* results because *ScrewFrame* needs at least three consecutive residues to recognize a regular structure.

equal, which means that neither method yields systematically more or longer secondary-structure elements than the other. Figs. 11(*b*) and 12(*b*) show the distributions of the lengths of individual secondary-structure elements. For α-helices, *DSSP* has a fatter tail (helices of length 20 or more), whereas *ScrewFrame* identifies a larger number of short helices. The reason for these differences is that *ScrewFrame* tends to split up kinked helices that *DSSP* identifies as single units. For β-strands, we notice that *DSSP* identifies many more very short elements. This is owing to the different definitions: a single β-type hydrogen bond is sufficient to define a β-sheet in *DSSP*, but *ScrewFrame* requires at least three consecutive residues to identify any regular structure.

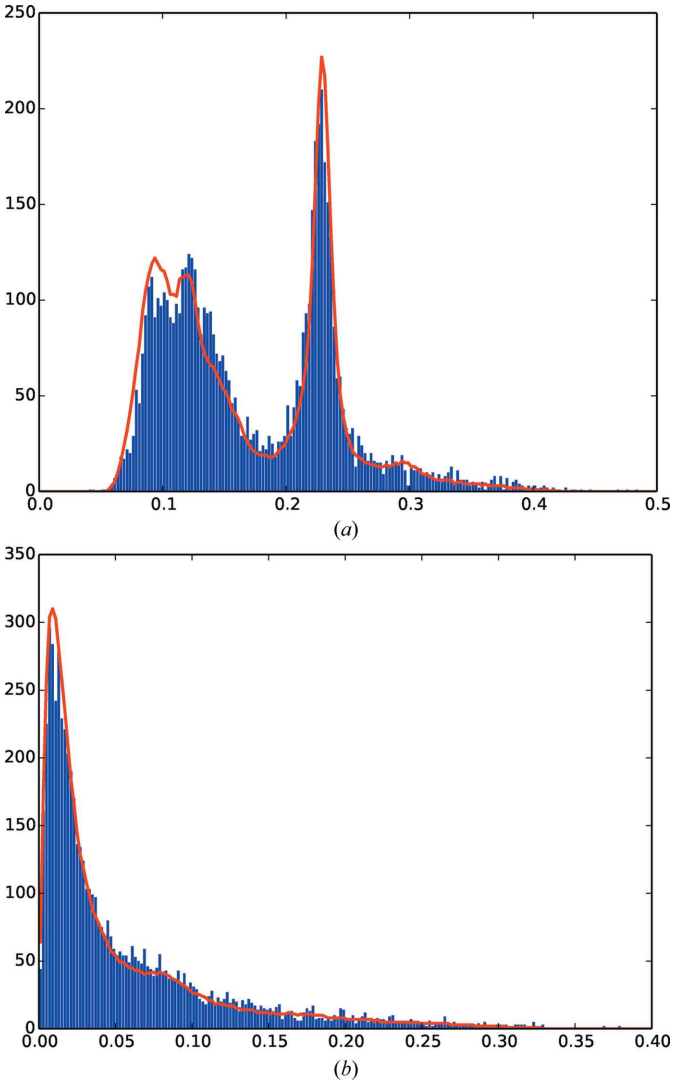


Figure 13
(*a*) Histogram for the ρ parameters corresponding to all pure C^{α} structures in the RCSB Protein Data Bank (blue bars) and the combined histogram for all-α and all-β structures corresponding to the ASTRAL database (red line). Details are given in the text. (*b*) The corresponding figures for the regularity parameter δ. The histograms do not suggest a lower precision for the C^{α} structures compared with all-atom structures.

3.4. Analysis of C^{α} structures in the PDB

The Protein Data Bank contains at this time 595 entries marked as 'CA ATOMS ONLY', which correspond to low-resolution X-ray crystallographic or electron-microscopic data. Secondary-structure assignment methods such as *DSSP*, which are based on an analysis of hydrogen-bond networks, cannot be applied to these entries. The low resolution of the experimental data underlying these structures raises the question whether our approach can still identify secondary structures reliably. The C^{α} positions could be less precise, leading to an increased uncertainty in the *ScrewFrame* parameters that we compute from them.

To investigate this question, we have computed histograms for the *ScrewFrame* parameters for this set of structures in the same way as described above for the ASTRAL database. These histograms are shown in Fig. 13. The red lines show the distributions for the ASTRAL database for comparison. Since the latter are for predominantly α- or β-containing structures, whereas the C^{α}-only PDB entries contain a mixture of all kinds of structures, we must compare with a weighted sum of the histograms of the two ASTRAL categories. The relative weights have been determined empirically: the red lines in the figure correspond to the sum of 0.0067 times the α histogram (upper right plots in Figs. 7 and 10) and 0.01 times the β histogram (lower right plots in Figs. 7 and 10).

The excellent agreement of the histograms suggests that there is no difference in the uncertainty of the *ScrewFrame* parameters between PDB entries for low-resolution data and PDB entries in general. A possible explanation is that there is already no increased uncertainty in the C^{α} positions. Many of the C^{α}-only structures in the PDB have at least partly been obtained by rigid fitting of all-atom protein structures obtained from higher resolution experiments. The relative positions of the C^{α} atoms are therefore no less precise than in an all-atom structure. Unfortunately, the information provided in the PDB entries (and even in the accompanying articles) is not sufficient to identify those parts of any given structure that were constructed with less precise methods, making a more detailed investigation of this question impossible.

4. Conclusions and outlook

We have presented a generalization of the *ScrewFit* method for protein structure assignment and description which uses only the positions of the C^{α} atoms along the protein backbone. As in the *ScrewFit* approach, the global protein fold is described as a succession of screw motions relating consecutive recurrent motifs along the protein backbone, but here the 'motifs' are the tripods (planes) formed by the three (two) orthonormal vectors of the local Frenet bases to the C^{α} space curve. Despite the fact that *ScrewFrame* uses less information than *ScrewFit*, all standard PSSEs are recognized on the basis of thresholds for the local screw radii and a suitably defined regularity measure. *ScrewFrame* even permits parallel and antiparallel β-strands to be distinguished, which the classical

ScrewFit method fails to do. A thorough comparison with the commonly used *DSSP* method in the assignment of PSSEs in the ASTRAL database shows that both methods yield very similar results for the total amount of PSSEs. *ScrewFrame* tends, however, to break long helices into smaller pieces, such that the length distribution of PSSEs is different. Owing to the minimalistic character of the geometrical model for protein folds, the evaluation of the *ScrewFrame* model parameters is very efficient. This allows work with protein structure databases and analysis of simulated molecular-dynamics trajectories of proteins. We have also shown that *ScrewFrame* is robust with respect to perturbations of the input structures. The local helix radius varies only little, whereas the regularity parameter δ increases visibly. This is exactly what is needed in structure refinement of low-resolution data, where thresholds on δ may be used to enforce more or less ideal PSSEs. *ScrewFrame* may also be used a starting point for the development of minimalistic models for protein structure and dynamics, similar to the worm-like chain model (Doi & Edwards, 1986), which has been successfully applied to DNA (Marko & Siggia, 1995). Our method may also be used to analyze dynamic processes such as the folding and unfolding of peptides (Spampinato & Maccari, 2014) and it can describe the fold of intrinsically disordered proteins.

An ActivePaper (Hinsen, 2014*a*) containing all of the software, input data sets and results from this study is available as Supporting Information. The data sets can be inspected with any HDF5-compatible software, *e.g.* the free *HDFView* (The HDF Group, 2013). Running the programs on different input data requires the *ActivePaper* software (Hinsen, 2014*a*).

References

Altmann, S. (1986). *Rotations, Quaternions, and Double Groups.* Oxford: Clarendon Press.

Andersen, C. A., Palmer, A. G., Brunak, S. & Rost, B. (2002). *Structure*, **10**, 175–184.

Calligari, P. A. & Kneller, G. R. (2012). *Acta Cryst.* D**68**, 1690–1693.

Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M. & Brenner, S. E. (2004). *Nucleic Acids Res.* **32**, D189–D192.

Chasles, M. (1830). *Bull. Sci. Math.* **14**, 321–326.

Doi, M. & Edwards, S. (1986). *The Theory of Polymer Dynamics.* Oxford University Press.

Dupuis, F., Sadoc, J.-F. & Mornon, J.-P. (2004). *Proteins*, **55**, 519–528.

Fox, N. K., Brenner, S. E. & Chandonia, J. M. (2013). *Nucleic Acids Res.* **42**, D304–D309.

Frishman, D. & Argos, P. (1995). *Proteins*, **23**, 566–579.

Golub, G. & van Loan, C. (1996). *Matrix Computations.* Baltimore: Johns Hopkins University Press.

Grimes, J. M., Fuller, S. D. & Stuart, D. I. (1999). *Acta Cryst.* D**55**, 1742–1749.

Hekkelman, M. (2013). *DSSP* v.2.2.1. http://swift.cmbi.ru.nl/gv/dssp/.

Hiller, S., Garces, R. G., Malia, T. J., Orekhov, V. Y., Colombini, M. & Wagner, G. (2008). *Science*, **321**, 1206–1210.

Hinsen, K. (2014*a*). *ActivePapers*. http://www.activepapers.org/.

Hinsen, K., (2014*b*). *ASTRAL-SCOPe Subset 2.04 in ActivePapers Format.* doi:10.5281/zenodo.11086.

Hinsen, K. (2014*c*). *J. Chem. Inf. Model.* **54**, 131–137.

Hu, S., Lundgren, M. & Niemi, A. J. (2011). *Phys. Rev. E*, **83**, 061908.

Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.

Kirchmair, J., Markt, P., Distinto, S., Schuster, D., Spitzer, G. M., Liedl, K. R., Langer, T. & Wolber, G. (2008). *J. Med. Chem.* **51**, 7021–7040.

Kneller, G. R. (1991). *Mol. Simul.* **7**, 113–119.

Kneller, G. R. & Calligari, P. (2006). *Acta Cryst.* D**62**, 302–311.

Labesse, G., Colloc'h, N., Pothier, J. & Mornon, J.-P. (1997). *Comput. Appl. Biosci.* **13**, 291–295.

Levitt, M. & Greer, J. (1977). *J. Mol. Biol.* **114**, 181–239.

Marabini, R., Macias, J. R., Vargas, J., Quintana, A., Sorzano, C. O. S. & Carazo, J. M. (2013). *Acta Cryst.* D**69**, 695–700.

Marko, J. F. & Siggia, E. D. (1995). *Macromolecules*, **28**, 8759–8770.

Park, S.-Y., Yoo, M.-J., Shin, J.-M. & Cho, K.-H. (2011). *BMB Rep.* **44**, 118–122.

Pauling, L. & Corey, R. B. (1951). *Proc. Natl Acad. Sci. USA*, **37**, 729–740.

Pauling, L., Corey, R. B. & Branson, H. R. (1951). *Proc. Natl Acad. Sci. USA*, **37**, 205–211.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). *J. Comput. Chem.* **25**, 1605–1612.

Sklenar, H., Etchebest, C. & Lavery, R. (1989). *Proteins*, **6**, 46–60.

Spampinato, G. & Maccari, G. (2014). *J. Chem. Theory Comput.* **10**, 3885–3895.

The HDF Group (2013). *HDFView.* http://www.hdfgroup.org/hdf-java-html/hdfview/.

Tozzini, V. (2005). *Curr. Opin. Struct. Biol.* **15**, 144–150.

Vojtechovsky, J., Chu, K., Berendzen, J., Sweet, R. M. & Schlichting, I. (1999). *Biophys. J.* **77**, 2153–2174.

Wolfram Research (2014). *Mathematica* v.10.0. Champaign: Wolfram Research.