Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

ScrewFit: combining localization and description of protein secondary structure

Paolo A. Calligari[a] and Gerald R. Kneller[b,c]*

[a]Department of Chemistry, Ecole Normale Supérieure, 24 Rue Lhomond, 75005 Paris, France, [b]Centre de Biophysique Moléculaire, CNRS UPR 4301, Rue Charles Sadron, 45071 Orléans, France, and [c]Synchrotron Soleil, L'Orme de Merisiers, BP 48, 91192 Gif-sur-Yvette, France

Correspondence e-mail: gerald.kneller@cnrs-orleans.fr

Received 17 April 2012
Accepted 12 September 2012

A new application of the *ScrewFit* algorithm [Kneller & Calligari (2006), *Acta Cryst.* D**62**, 302–311] is presented which adds the detection of protein secondary-structure elements to their detailed geometrical description in terms of a curve with intrinsic torsion. The extension is based on confidence and persistence criteria for the *ScrewFit* parameters which are established by analyzing the structural fluctuations of standard motifs in the SCOP fold classes. The agreement with the widely used *DSSP* method is comparable with the general consensus among other methods in the literature. This combination of secondary-structure detection and analysis is illustrated for the enzyme adenylate kinase.

1. Introduction

The localization and geometrical description of protein secondary-structure elements is one of the standard tasks in structural biology. In the last few decades, a variety of methods have been developed for this purpose, which handle either the localization (Kabsch & Sander, 1983; Richards & Kundrot, 1988; Frishman & Argos, 1995; Taylor, 2001) or the geometrical description (Barlow & Thornton, 1988; Sklenar *et al.*, 1989; Thomas, 1994; Hanson *et al.*, 2011). The *ScrewFit* algorithm that we published more recently in this journal (Kneller & Calligari, 2006) belongs *a priori* to the second group and is particularly suited to localizing changes in protein structure. It describes the winding of the protein main chain through local screw motions, relating the C—O—N atoms in successive peptide bonds. An application to structural biology has been published in Calligari *et al.* (2009), in which the method was used to quantify the impact of ligand binding on the neuraminidase enzyme from different influenza viruses.

The purpose of this communication is to demonstrate that *ScrewFit* can easily be extended to allow both localization and geometrical description of protein secondary-structure elements. The method is based on Chasles' theorem, which states that any rigid-body motion can be described by a screw motion, *i.e.* by a roto-translation where the axes of rotation and translation are parallel. The corresponding active coordinate transformation $\mathbf{r}\rightarrow\mathbf{r}'$ for the Cartesian coordinates of a position vector $\mathbf{r}$ is given by

$$\mathbf{r}' = \mathbf{R}_x + \mathbf{D}(\mathbf{n},\varphi)\cdot(\mathbf{r}-\mathbf{R}_x) + \alpha\mathbf{n}, \qquad (1)$$

where $\mathbf{D}(\mathbf{n},\varphi)$ is a rotation matrix which is parametrized by a unit vector $\mathbf{n}$ in the direction of the rotation axis and a rotation angle φ. The column vector $\mathbf{R}_x$ contains the coordinates of the reference point for the rotation, which is located on the axis of the screw motion, and α is a real parameter describing the translation along the screw axis. The operation (1) is applied to map the positions of the {C, O, N} atoms in a given peptide bond i to those in peptide bond $i + 1$, considering the {C—O—N} groups as rigid bodies. The parameters of the roto-translation (1) are constructed by a quaternion-based rigid-body fit $\{\mathrm{C—O—N}\}(i)\rightarrow\{\mathrm{C—O—N}\}(i+1)$, which yields four quaternion parameters $\{q_0, q_1, q_2, q_3\}$ and the translation vector $\mathbf{t} = \mathbf{R}_{C,i+1} - \mathbf{R}_{C,i}$. Here, $\mathbf{R}_{C,i}$ and $\mathbf{R}_{C,i+1}$ denote the positions of the C atom in peptide planes {C—O—N}(i) and {C—O—N}(i + 1), respectively. The quaternion parameters obey the normalization condition $q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1$ and the positions $\mathbf{R}_{C,i}$ and $\mathbf{R}_{C,i+1}$ are chosen to be the respective centres for the superposition fit. The unit vector $\mathbf{n}$ and the rotation angle φ can be computed from the quaternion parameters. Defining $\mathbf{t}_{\|} = \mathbf{n}\cdot\mathbf{t}$ and $\mathbf{t}_{\perp} = \mathbf{t} - \mathbf{t}_{\|}$ to be the components of the translation vector $\mathbf{t}$ parallel and orthogonal to the rotation axis, respectively, the radius ρ of the screw motion is given by

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

and the translation along the screw axis is $\alpha = |\mathbf{t}_{\|}|$.

Another parameter that can be extracted from the rotational superposition fit is the angular distance between the peptide planes $\{\mathrm{C{-}O{-}N}\}(i)$ and $\{\mathrm{C{-}O{-}N}\}(i+1)$. Defining $d(\mathcal{X}, \mathcal{X}') = \sum_{\alpha} w_{\alpha}(\mathbf{x}_{\alpha} - \mathbf{x}'_{\alpha})^2$ (α = C, O, N) to be the weighted Euclidean distance between the two vector sets $\mathcal{X} = \{\mathbf{R}_{\mathrm{O},i} - \mathbf{R}_{\mathrm{C},i}, \mathbf{R}_{\mathrm{N},i} - \mathbf{R}_{\mathrm{C},i}\}$ and $\mathcal{X}' = \{\mathbf{R}_{\mathrm{O},i+1} - \mathbf{R}_{\mathrm{C},i+1}, \mathbf{R}_{\mathrm{N},i+1} - \mathbf{R}_{\mathrm{C},i+1}\}$, the orientational distance is defined by

$$\Delta = \left[\frac{d(\mathcal{X}, \mathcal{X}')}{d(\mathcal{X}, \mathcal{X}')_{\max}}\right]^{1/2}, \qquad (3)$$

where $d(\mathcal{X}, \mathcal{X}')_{\max}$ is the maximum Euclidean distance. We note here that $d(\mathcal{X}, \mathcal{X}')_{\max}$ equals the maximum eigenvalue in the quaternion-based rotational superposition problem $\{\mathrm{C{-}O{-}N}\}(i) \rightarrow \{\mathrm{C{-}O{-}N}\}(i+1)$, which can be formulated as an eigenvector problem for the optimal quaternion. As described in Kneller (1991) and Kneller & Calligari (2006), the resulting eigenvalues correspond to the squares of the respective fit errors and therefore $d(\mathcal{X}, \mathcal{X}')_{\max} = \lambda_{\max}$, where $\lambda_{\max}$ is the largest eigenvalue. *ScrewFit* is implemented in a Python open-source code freely available at http://dirac.cnrs-orleans.fr/plone/software/screwfit/screwfit/.

2. Secondary-structure assignments

Secondary-structure motifs are generally defined with respect to the regular winding of the main chain in model polypeptides, which is associated with specific hydrogen-bond patterns. However, significant deviations from the ideal conformations of these motifs are found in experimentally determined protein structures. This structural variety can be used to establish confidence intervals for the *ScrewFit* parameters which are associated with a given structural motif. For this purpose, we analyzed 1027 α-helices and 1336 β-strands from the SCOP+ASTRAL database (Chandonia *et al.*, 2004), which contains the coordinates of secondary-structure elements for each domain

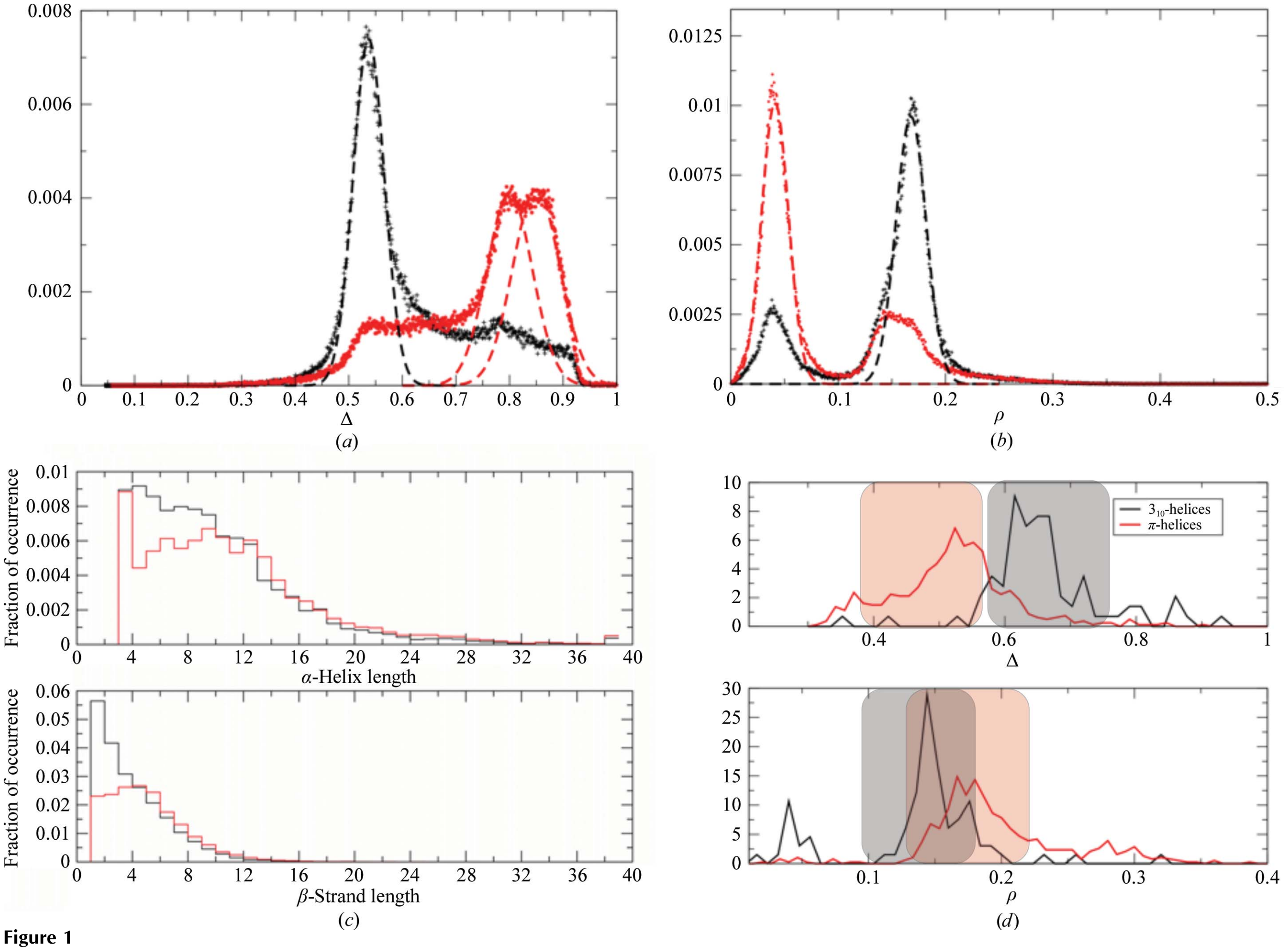

Figure 1

(*a*, *b*) Normalized distribution (solid lines) from data sets $\mathcal{A}$ (black) and $\mathcal{B}$ (red) and fitted Gaussian functions (dashed lines) for Δ (*a*) and ρ (*b*). The corresponding correlation coefficients are between 0.90 and 0.96. The two peaks found in $\mathcal{B}$ for the parameter Δ were assigned to β-strand and to extended conformation using the values obtained on model peptides from Kneller & Calligari (2006) as a reference. (*c*) Length distribution for α-helices and β-strands found by *ScrewFit* (black histograms) and *DSSP* (red histograms). (*d*) Distribution of the parameters Δ and ρ for two distinct data sets of protein structures containing only π-helices and 3_{10}-helices (black and red, respectively), which have been constructed by combining the *DSSP* algorithm with visual inspection. The vertical stripes indicate the corresponding confidence ranges defined in the text for these motifs.

Table 1
ScrewFit parameters for different structural motifs from a model polypeptide (Kneller & Calligari, 2006) and from screening of the data sets presented in this work.

Motif	$\rho_{\rm ideal}$ (nm)	$\overline{\rho} \pm \varepsilon_{\rm p}$ (nm)	$\Delta_{\rm ideal}$	$\overline{\Delta} \pm \varepsilon_{\Delta}$
α-Helix	0.171	0.168 ± 0.055	0.582	0.537 ± 0.091
3_{10}-Helix	0.146	0.146 ± 0.055	0.670	0.670 ± 0.091
π-Helix	0.178	0.178 ± 0.055	0.471	0.471 ± 0.091
β-Strand	0.055	0.041 ± 0.040	0.875	0.850 ± 0.129
Extended	0.037	0.041 ± 0.040	0.754	0.800 ± 0.114

classified according to the SCOP fold classes (Murzin *et al.*, 1995). The motifs are taken from proteins with less than 40% identity in the amino-acid sequence. In the following, we refer to the coordinate subsets for α-helices and β-strands as $\mathcal{A}$ and $\mathcal{B}$, respectively. The corresponding *ScrewFit* parameters scatter substantially, reflecting the conformational variability of the structural motifs in the respective data sets. Their distributions nevertheless display well defined peaks (Figs. 1*a* and 1*b*) which can be clearly separated for subsets $\mathcal{A}$ and $\mathcal{B}$. Fitting these peaks by Gaussian functions,

$$g(x) = \frac{1}{(2\pi)^{1/2}\sigma}\exp\left[-\frac{(x-\overline{x})^2}{2\sigma^2}\right],$$

we obtained estimations of the confidence intervals for the *ScrewFit* parameters Δ and ρ from the respective width parameters σ_Δ and σ_ρ. These parameters and the respective mean values $\overline{\Delta}$ and $\overline{\rho}$ are listed in Table 1 together with the *ideal* values for model polypeptides in Kneller & Calligari (2006). The confidence interval of each parameter was set to twice the corresponding standard deviation,

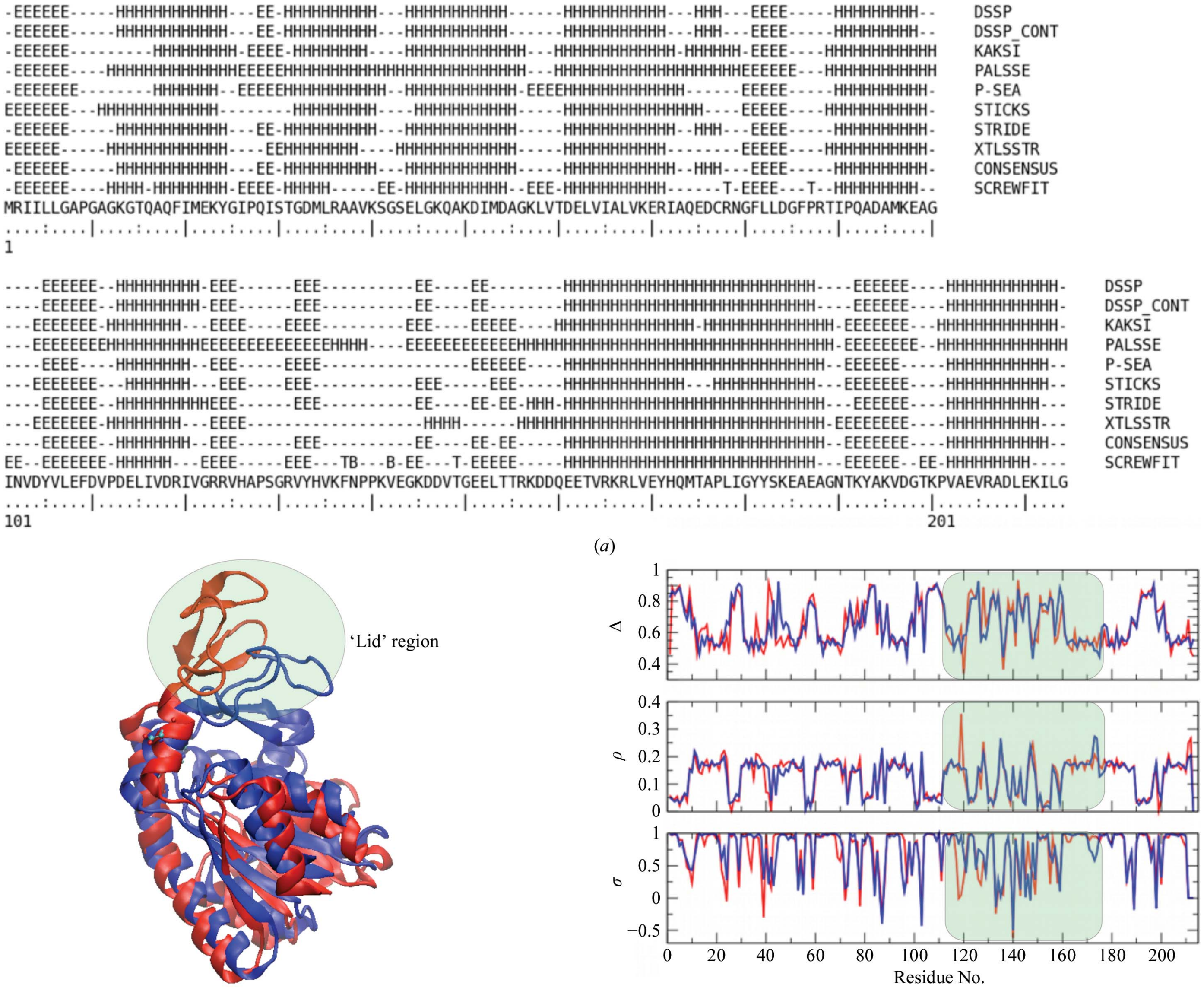

Figure 2
(*a*) Comparison of different secondary-structure assignment algorithms obtained by the 2*Struct* server (Klose *et al.*, 2010) using the crystallographic structure of adenylate kinase (PDB entry 4ake). (*b*) Three-dimensional structure of adenylate kinase in its apo and holo conformations (red and blue, respectively). The upper part of the protein (residues 113–167) defines a 'lid' which opens and closes the binding site. (*c*) *ScrewFit* profiles obtained on apo and holo forms of adenylate kinase from PDB entries 4ake and 1ake. The vertical stripes label the lid region of the protein. The 'open' and 'closed' conformations are well distinguished by the changes in the straightness parameter σ and the local helix radius ρ in the residue ranges 115–120 and 170–175, which locate the hinges of the lid.

$$\varepsilon_k = 2\sigma_k, \qquad (4)$$

where k represents the variable under consideration (ρ or Δ).

To develop an automatic procedure for secondary-structure assignment of a given protein, we adopted the following procedure after having established a profile of the *ScrewFit* parameters as a function of the residue number.

(i) β-Strands and extended configurations are assigned if at least two consecutive residues exhibit values for Δ and ρ in the confidence intervals given in Table 1.

(ii) α-Helices are assigned if Δ and ρ are within the confidence intervals for at least four consecutive residues.

(iii) For 3_{10}-helices and π-helices the mean values of Δ and ρ and the corresponding standard deviations cannot be extracted from a statistical analysis of data set $\mathcal{A}$, since these motifs are rare. They are assigned if $\Delta = \Delta_{\rm ideal} \pm \varepsilon_{\Delta_\alpha}$ and $\rho = \rho_{\rm ideal} \pm \varepsilon_{\rho_\alpha}$ for at least three and five consecutive residues, respectively. Here, $\Delta_{\rm ideal}$ and $\rho_{\rm ideal}$ are the values for the model structure listed in Kneller & Calligari (2006) and $\varepsilon_{\Delta_\alpha}$ and $\varepsilon_{\rho_\alpha}$ are the confidence intervals for α-helices.

The consensus between secondary-structure detection by *ScrewFit* and the well established *DSSP* method (Kabsch & Sander, 1983) may be estimated by the ratio of the number of residues for which both methods give the same assessment for a given motif and the total number of residues assigned by *DSSP*. We performed this comparison on a subset of the PDBSelect25 database (Hobohm & Sander, 1994), which contains 2144 nonredundant chain folds with sequence homology lower than 25% and which should reproduce most of the structural heterogeneity in the whole PDB database. The agreement between *DSSP* and *ScrewFit* on this set was found to be 84% for α-helices and 90% for β-strands. These results are comparable with the general consensus found among different methods for secondary-structure detection (Colloc'h *et al.*, 1993; Dupuis *et al.*, 2004; Martin *et al.*, 2005). The major discrepancies between *ScrewFit* and *DSSP* are found for short motifs (see Fig. 1*c*): *ScrewFit* detects more α-helices of lengths between four and ten residues than *DSSP* and finds a significantly larger number of short β-strands (2–4 residues). These differences probably arise from both the sensitivity of *ScrewFit* to kinks and curvature in the protein backbone and from the known tendency of *DSSP* to overestimate the length of structural motifs in such cases (Cubellis *et al.*, 2005). The reliability of the assignment criteria for the rare 3_{10}-helix and π-helix motifs was specifically verified for two small data sets available in the literature (Pal & Basu, 1999; Fodje & Al-Karadaghi, 2002; see Fig. 1*d*).

3. Combining secondary-structure assessment and description

Fig. 2(*a*) shows the assignment of protein secondary-structure elements obtained using *ScrewFit* for the enzyme adenylate kinase from *Escherichia coli* (PDB entry 4ake; Muller *et al.*, 1996), together with the assignments obtained using eight other methods. The consensus between all methods can be read off from the figure and the agreement between *ScrewFit*, *P-SEA* (Labesse *et al.*, 1997) and *STICKS* (Taylor, 2001) is particularly pronounced. This result is not unexpected, as these three methods use similar geometrical concepts to quantify the protein backbone winding. In addition to the detection of secondary-structure elements, *ScrewFit* can be used for what it was originally designed for: a detailed geometrical *description* of protein secondary-structure elements. This point is illustrated in Fig. 2(*c*), which displays the *ScrewFit* parameters obtained from crystallographic structures of adenylate kinase in its apo form and complexed with the inhibitor AP5A (PDB entry 1ake; Muller & Schulz, 1992; see Fig. 2*b*). The evolution of the parameters Δ and ρ along the backbone together with the straightness parameter (Kneller & Calligari, 2006) clearly quantify the structural differences in the detected secondary structure between the holoprotein and the apoprotein. More information is given in the figure caption.

4. Conclusion

We have presented a new application of the *ScrewFit* algorithm which extends its functionality from a geometrical description of protein backbone conformations to the detection of secondary-structure elements. The latter is achieved by using confidence intervals of the *ScrewFit* parameters, which are established by analyzing the natural variability of the standard secondary-structure motifs. The example of the enzyme adenylate kinase illustrates the combination of secondary-structure detection and description. The latter shows that the essential structural changes are exactly in the hinge region of the lid domain of the protein, which opens and closes its active site.

Paolo Calligari acknowledges financial support from the Agence Nationale de la Recherche, contract no. ANR-2010-COSI-001-01.

References

Barlow, D. & Thornton, J. (1988). *J. Mol. Biol.* **201**, 601–619.

Calligari, P., Kneller, G., Giansanti, A., Ascenzi, P., Porrello, A. & Bocedi, A. (2009). *Biophys. Chem.* **141**, 117–123.

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

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B. & Mornon, J. (1993). *Protein Eng.* **6**, 377–382.

Cubellis, M. V., Cailliez, F. & Lovell, S. C. (2005). *BMC Bioinformatics*, **6**, Suppl. 4, S8.

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

Fodje, M. N. & Al-Karadaghi, S. (2002). *Protein Eng.* **15**, 353–358.

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

Hanson, R. M., Kohler, D. & Braun, S. G. (2011). *Proteins*, **79**, 2172–2180.

Hobohm, U. & Sander, C. (1994). *Protein Sci.* **3**, 522–524.

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

Klose, D. P., Wallace, B. A. & Janes, R. W. (2010). *Bioinformatics*, **26**, 2624–2625.

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.

Martin, J., Letellier, G., Marin, A., Taly, J., Brevern, A. & Gibrat, J. (2005). *BMC Struct. Biol.* **5**, 1–17.

Muller, C. W., Schlauderer, G. J., Reinstein, J. & Schulz, G. E. (1996). *Structure*, **4**, 147–156.

Muller, C. W. & Schulz, G. E. (1992). *J. Mol. Biol.* **224**, 159–177.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.

Pal, L. & Basu, G. (1999). *Protein Eng. Des. Sel.* **12**, 811–814.

Richards, F. & Kundrot, C. (1988). *Proteins*, **3**, 71–84.

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

Taylor, W. (2001). *J. Mol. Biol.* **310**, 1135–1150.

Thomas, D. J. (1994). *J. Mol. Graph.* **12**, 146–152.