Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Domain identification by iterative analysis of error-scaled difference distance matrices

Thomas R. Schneider

FIRC Institute of Molecular Oncology, Via Adamello 16, 20139 Milan, Italy, and European Institute of Oncology, Via Ripamonti 435, 20141 Milan, Italy

Correspondence e-mail: schneider@ifom-firc.it

Received 19 March 2004
Accepted 21 September 2004

Iterative interpretation of error-scaled difference distance matrices is suggested as a means of dividing a protein into structural domains on the basis of conformational differences between different models. Two conformers of Src kinase {PDB codes 1fmk [Xu *et al.* (1997). *Nature (London)*, **385**, 595–602] and 2src [Xu *et al.* (1999). *Mol. Cell*, **3**, 629–638]} in the inactive state with and without a substrate analogue bound are analysed in order to demonstrate the approach. SH3, SH2 and the N- and C-terminal lobes of the kinase domain are detected as structural modules that move with respect to each other. Notably, a relative movement between the SH3 and SH2 domains is detected although both structures of Src kinase are in the 'assembled' state. Detailed analysis shows that Arg318, a residue topologically located in the N-terminal lobe of the kinase domain, structurally belongs to the C-terminal lobe. The movement of this residue together with the C-terminal lobe upon substrate binding leads to the loss of a salt bridge between Arg318 and Asp117, a residue in the SH3 domain, providing an explanation for the increased mobility of the SH3 domain.

1. Introduction

As the crystal structures of biological macromolecules can be determined more and more rapidly, concepts and tools for the interpretation of structural models become ever more important. In particular, the comparison of structures to detect meaningful similarities and significant differences is an important source of information. To derive statistically sound results from structural data, we are developing a framework for structure comparison in which the coordinate uncertainties are explicitly taken into account in all steps.

A possible approach to rationalize differences between different conformers of a molecule is to divide the molecule into domains of preserved geometry and ascribe a part of the conformational differences to relative movements of the domains (Wriggers & Schulten, 1997). Different definitions of such domains and various methods for identifying these domains have been put forward. For example, Nichols *et al.* (1995) define a rigid domain as a set of atoms selected such that all differences between distances between any two corresponding atoms in different conformers are smaller than a user-defined threshold ε measured in ångströms. Technically, the analysis is realized by analysing thresholded difference distance matrices. Another approach, the adaptive selection procedure (Wriggers & Schulten, 1997), employs repeated least-squares superpositions of lists of atoms that are updated depending on whether or not individual atoms fulfil a tolerance criterion after superposition. The actual tolerance criterion used depends on the structures being compared, but

doi:10.1107/S0907444904023492

can be determined by calibrating the coordinate errors against a theoretical curve.

Using a heuristic formula for coordinate uncertainties suggested by Cruickshank (1999), we have recently suggested a formalism in which the largest conformationally invariant region of a protein can be identified on the basis of the condition that all interatomic distances within the region are identical within error (Schneider, 2000). In this approach, the criterion for two distances being identical is that their difference must be less than 2σ (where σ is the uncertainty in the measurement of the difference). The σ values are determined individually for every single difference in interatomic distances and thus allow a consistent treatment of all atoms despite the fact that their coordinate uncertainties can vary substantially between structures and between different atoms in the same structure. Changes in interatomic distances normalized to their uncertainties can be conveniently represented in error-scaled difference distance matrices (Schneider, 2000) and an algorithm to find automatically the largest conformationally invariant region with respect to a set of conformers by analysing these matrices has been described (Schneider, 2002).

A prominent family of proteins in which the relative movements of domains play a decisive role for the protein function are the Src kinases (for a review, see Huse & Kuriyan, 2002). Src kinases consist of two pairs of domains in which the kinase part of the molecule harbours the active site between its N-terminal and C-terminal domains and the SH2/SH3 tandem of domains is involved in the regulation of the activity and the binding of substrate molecules. The structure of Src kinase in the inactive conformation has been determined with and without substrate analogues bound to the active site (Xu *et al.*, 1997, 1999). The mobility of the domains with respect to each other has been studied in a molecular-dynamics simulation on Src kinase itself (Young *et al.*, 2001) and by experimental investigations using both NMR spectroscopy and X-ray crystallography on a close homologue of Src kinase, Fyn kinase (Arold *et al.*, 2001; Ulmer *et al.*, 2002).

Here, we suggest the use of the previously described method for the automatic analysis of error-scaled difference distance matrices in an iterative fashion to divide a protein molecule into structural domains. To demonstrate its validity, we apply the proposed approach to the analysis of two conformers of Src kinase originating from crystal structures and evaluate the results.

2. Outline of the method

The standard difference distance matrix (DD matrix) between atoms of a molecule in two conformations contains elements Δ_{ij}^{ab} that correspond to the difference in distance between atoms i and j in two conformations, a and b, respectively (Nishikawa *et al.*, 1972),

$$\Delta_{ij}^{ab} = d_{ij}^{a} - d_{ij}^{b}. \tag{1}$$

If Δ_{ij}^{ab} is positive, the interatomic vector between i and j in conformation a is contracted with respect to that in b. Conversely, negative elements of the difference distance matrix indicate an expansion of the interatomic vector.

As differences between interatomic distances are small numbers calculated as differences between large numbers, the elements of DD matrices should be related to their experimental uncertainties. If the uncertainty of the coordinates of atoms i and j in models a and b, σ_i^a, σ_j^a, σ_i^b and σ_j^b, are known, an estimate for the uncertainty for the change in interatomic distance Δ_{ij}^{ab} can be derived by error propagation,

$$\sigma(\Delta_{ij}^{ab}) = \left[(\sigma_i^a)^2 + (\sigma_j^a)^2 + (\sigma_i^b)^2 + (\sigma_i^b)^2\right]^{1/2}. \tag{2}$$

This error estimate can be used to normalize the difference distance matrix, so that the error-scaled difference distance matrix (EDD matrix) with elements E_{ij}^{ab} can be calculated (Schneider, 2000),

$$E_{ij}^{ab} = \Delta_{ij}^{ab}/\sigma(\Delta_{ij}^{ab}). \tag{3}$$

In small-molecule crystallography, standard uncertainties (s.u.s) for atomic positions are routinely determined *via* the inversion of the normal matrix of the refinement. For macromolecular structure determinations this is (apart from some exceptional cases) not feasible and approximations have to be made. Cruickshank has suggested an heuristic formula that allows the calculation of the coordinate error $\sigma(x, B_{\rm avg})$ for an atom with the average value B in a macromolecular structure (Cruickshank, 1999),

$$\sigma(x, B_{\rm avg}) = (N_i/N_{\rm obs})^{1/2} C^{-1/3} R_{\rm free} d_{\rm min}, \tag{4}$$

where N_i is the number of fully occupied sites, $N_{\rm obs}$ is the number of unique diffraction data, C and $d_{\rm min}$ are the completeness and the limiting resolution of the diffraction data used in refinement and $R_{\rm free}$ is the free R value for the final model. It should be noted that $\sigma(x, B_{\rm avg})$ can be calculated without having the diffraction data available. This is an important technical advantage as experimental data are often not deposited in the respective databases (Jiang *et al.*, 1999). Assuming a linear relationship between the coordinate uncertainty of an atom and its B value, B_i, an error estimate for the coordinate error $\sigma_{x,i}$ of an individual atom i can be obtained *via* (Schneider, 2000),

$$\sigma_{x,i} = [\sigma(x, B_{\rm avg})/B_{\rm avg}]B_i, \tag{5}$$

and inserted (as one possible approximation to the true coordinate uncertainties) into (2).

After construction of an error-scaled difference distance matrix, a conformationally invariant region can be identified by collecting a set of atoms for which all respective matrix elements are smaller than a given threshold (typically 2σ; Schneider, 2002). The problem of finding such a set of atoms is mathematically equivalent to the maximum clique problem (finding the largest set of nodes that are all connected by edges) in graph theory. The exact solution of this problem is known to be NP-hard, *i.e.* the computational resources required to solve the problem grow exponentially with its size

Table 1
Calculation of coordinate error estimates.

Values for the number of fully occupied non-H-atom sites, N_i, the number of observables, $N_{\rm obs}$, the completeness of the diffraction data, cpl, the free R value, $R_{\rm free}$, and the maximum resolution, $d_{\rm min}$, were taken from the headers of the respective PDB files with the following exceptions. For 1fmk, the number of refined atoms was set to the count of non-H atoms in the PDB file and the number of reflections used in refinement and the completeness of the data were set to the values stated in Table 1 of Xu *et al.* (1997). For 2src, the number of reflections used in refinement was set to the number stated in Table 2 of Xu *et al.* (1999). The number of C$^\alpha$ atoms (#CA) and the mean, the standard deviation and the minimum and maximum values for the coordinate error, σ_i, are shown in the right-hand half of the table.

PDB	N_i	$N_{\rm obs}$	cpl (%)	$R_{\rm free}$ (%)	$d_{\rm min}$ (Å)	DPI (Å)	#CA	$\langle\sigma_i\rangle$ (Å)	$\sigma(\sigma_i)$ (Å)	Min(σ_i) (Å)	Max(σ_i) (Å)
1fmk	4006	48728	66.1	26.4	1.5	0.163	438	0.130	0.048	0.042	0.274
2src	3915	55186	93.2	28.1	1.8	0.138	450	0.126	0.065	0.050	0.442

(Nichols *et al.*, 1995; Lesk, 1995). We have recently described a genetic algorithm that finds acceptable approximate solutions to the problem using moderate computing resources (Schneider, 2002).

3. Calculations

3.1. Error estimates for the models

Two crystal structures of Src kinase were selected for analysis: PDB code 1fmk (Xu *et al.*, 1997) and PDB code 2src (Xu *et al.*, 1999). In both structures, Src kinase is phosphorylated at Tyr527 and present in the 'assembled' (Xu *et al.*, 1999) conformation. 2src was cocrystallized with the non-hydrolysable ATP analogue AMP-PNP. The structures were determined from different crystals forms, both in the orthorhombic space group $P2_12_12_1$, with unit-cell parameters a = 51.8, b = 87.4, c = 101.3 Å (1fmk) and a = 50.6, b = 73.0, c = 172.7 Å (2src).

To estimate the coordinate errors in the two models, the numbers necessary for evaluation of (4) were initially taken directly from the PDB files. However, it turned out that some of the fields in the PDB file did not contain the standard entries; for example, for both structures the field for the number of reflections used in refinement contained the number of collected diffraction intensities and not the number of unique reflections. The results summarized in Table 1 are based on numbers taken from the respective publications. As the completeness of the data for 1fmk was only 66.1%, the values obtained from application of the standard error model [equations (4) and (5)] were scaled by a factor of 1.2 [corresponding to assuming a high-resolution limit of 1.8 instead of 1.5 Å; see (4)] to make the coordinate-error estimate more realistic.

3.2. Iterative analysis of the error-scaled difference distance matrix

The error-scaled difference distance matrix between the C$^\alpha$ atoms of 1fmk and 2src (Fig. 1) was analysed in terms of conformationally invariant regions by iterative application of the genetic algorithm described by Schneider (2002). The first round of the iteration was run on an EDD matrix corresponding to all 436 atoms present in both models and all atoms identified as conformationally invariant were flagged. Before the second iteration, the flagged atoms were deleted from the stored lists of atoms. A new EDD matrix corresponding to the reduced list of atoms was then calculated and analysed,

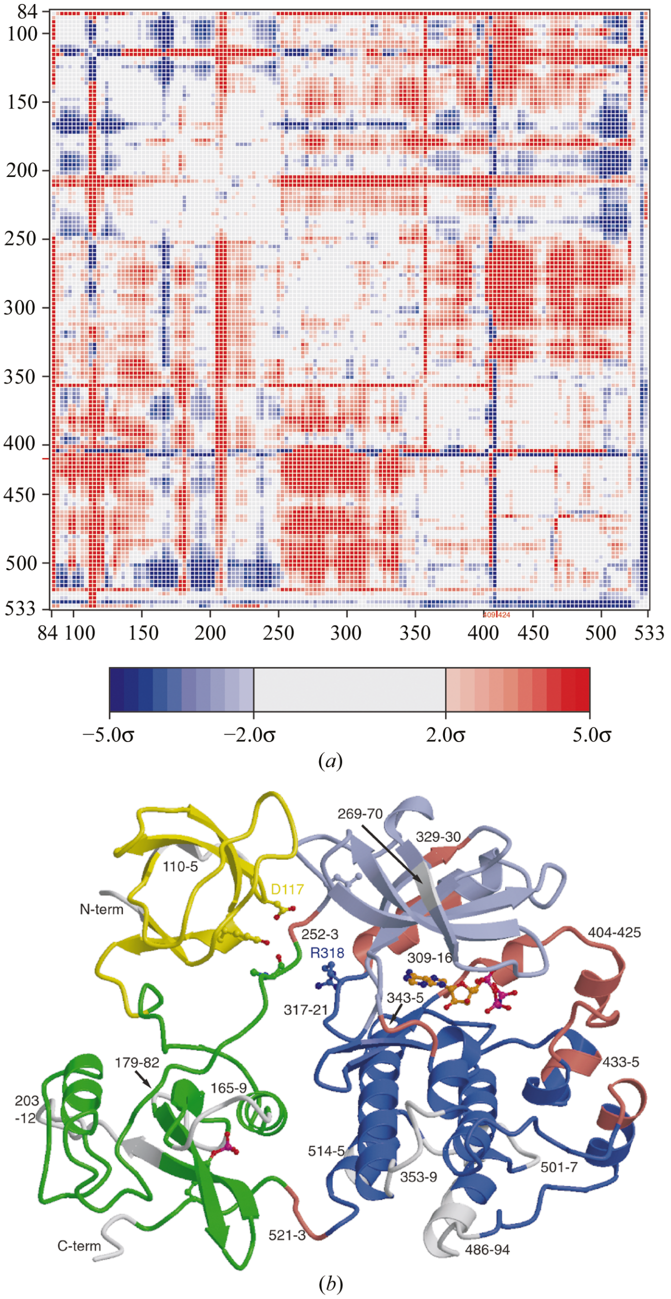

Figure 1
(*a*) Error-scaled difference distance matrix comparing the C$^\alpha$-atom positions of 1fmk and 2src; (*b*) schematic view of Src kinase, with conformationally invariant regions indicated by different colours. The substrate analogue is shown with C atoms in orange. Selected side chains are depicted with C atoms in the colour of the respective domain. Yellow: SH3 domain, Tyr136, Asp117. Green: SH2 domain, Pro250, Tyr527. Light blue: N-terminal lobe of the kinase domain, Lys255. Dark blue: C-terminal lobe of the kinase domain, Arg318. Flexible parts are shown in grey and red; grey indicates that the respective residues are involved in a crystal contact in at least one of the two crystal forms.

followed by flagging and deletion of the atoms found to be conformationally invariant. This scheme of reducing and analysing the atom lists was repeated until, in round five, only 21 of the remaining 93 atoms were found in the next conformationally invariant region.

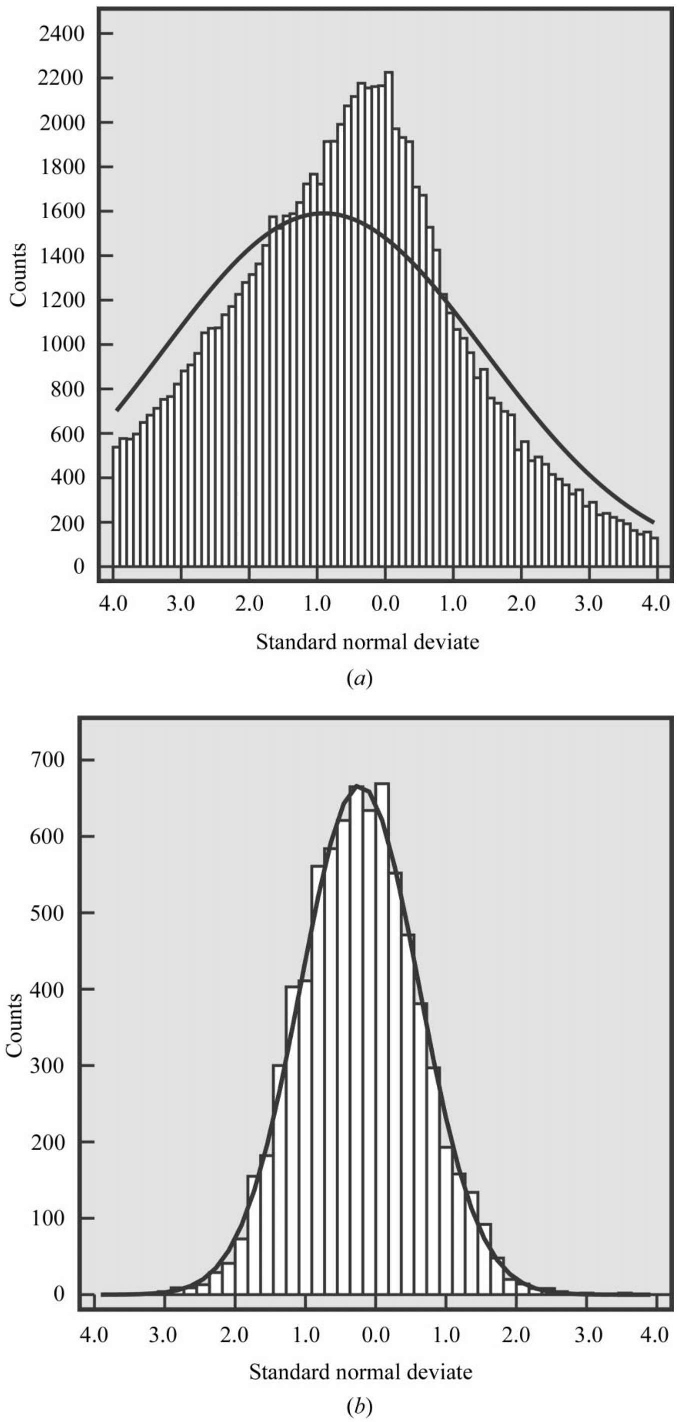

Figure 2
Histograms for the error-scaled difference distance matrix elements for all distances between (*a*) all 436 C$^\alpha$ atoms present in 1fmk and 2src [$\langle E\rangle = -0.91$, $\sigma(E_i) = 2.38$, $\chi^2 = 60.2$] and (*b*) 125 C$^\alpha$ atoms attributed to the first conformationally invariant region. The curves represent Gaussians with the respective means and standard deviations. For statistics regarding the histograms for the other domains, see Table 2.

Table 2
Results of the iterative analysis of 436 C$^\alpha$ atoms present in both 1fmk and 2src.

$n_{\rm it}$ is the number of the iteration, '0' stands for the atoms remaining after iteration 4. #ci is the number of atoms found to be conformationally invariant. The residue numbers for these atoms are given; residues involved in a crystal contact in at least one of the structures are shown in bold. R.m.s.d. denotes the r.m.s.d. obtained for a superposition of the C$^\alpha$ atoms of the respective region. #*E*, $\langle E\rangle$ and $\sigma(E)$ are the count, the mean and the standard deviation for error-scaled difference distance matrix elements. $\chi^2 = 1/n_{\rm bin}\sum_{i=1}^{n_{\rm bin}}(O_i - E_i)^2/E_i$, where $n_{\rm bin}$ is the number of bins in the histogram and O_i and E_i are the observed counts and the number of counts expected from a normal distribution for a bin *i*, respectively. The number of atoms in the second column adds up to 450, as atoms that are only present in one conformer (here 2src) are automatically counted as flexible.

$n_{\rm it}$	#ci	Residue Nos.	#*E*	$\langle E\rangle$	$\sigma(E)$	χ^2	R.m.s.d. (Å)	Remarks
1	125	317–321, 346–352, 360–403, 426–432, 436–466, 472–485, 495–500, 508–513, 516–520	7750	−0.23	0.84	2.2	0.24	CKIN
2	95	144–164, 170–178, 183–202, 213–251, 524–529	4465	−0.01	0.66	0.9	0.29	SH2
3	72	254–268, 271–308, 322–328, 331–342	2556	−0.06	0.84	1.0	0.34	NKIN
4	51	87–109, 116–143	1275	0.07	0.73	2.4	0.22	SH3
0	107	84–86, 110–115, **165–169**, **179–182**, **203–212**, 252–253, **269–270**, 309–316, 329–330, 343–345, **353–359**, 404–425, 433–435, **486–494**, **501–507**, **514–515**, 521–523, 530–533	n/a	n/a	n/a	n/a	n/a	Flexible

For the individual iterations, the genetic algorithm was run with standard parameters as described by Schneider (2002); in particular the lower and upper thresholds for the elements of the error-scaled difference distance matrix were set to $\varepsilon_{\rm l} = 2\sigma$ and $\varepsilon_{\rm h} = 5\sigma$. The rather complex error-scaled difference distance matrix (Fig. 1) was readily interpreted in four iterations taking 6 s on a 3.2 GHz Xeon-based PC running Suse-Linux 9.0.

Least-squares superpositions of coordinates were carried out with *LSQKAB* from the *CCP*4 suite of programs (Kabsch, 1976; Collaborative Computational Project, Number 4, 1994). The results are summarized in Table 2. *Xfit* (McRee, 1999) was used for structure analysis. For the relative movements of the domains, the following hinge atoms were identified by visual inspection: Gln144 CA for SH3 *versus* SH2, Gly344 CA for NKIN *versus* CKIN and Trp188 CA for NKIN *versus* SH3, where NKIN and CKIN are the N- and C-terminal lobes of the kinase domain

4. Results and discussion

4.1. Identification of domains

The analysis identified four conformationally invariant regions that correspond to the four canonical modules of Src kinase (Table 2 and Fig. 1*b*): the SH3- and SH2-domains and the N- and C-terminal lobes of the kinase domain. The conformationally invariant region found in the fifth iteration contained only 21 atoms and was not continuous in space, indicating that the noise level had been reached.

The r.m.s.d. values for the least-squares superposition of the two copies of the respective parts of the molecule range between 0.22 and 0.34 Å (Table 2). These values are substantially lower than the r.m.s.d. obtained for the superposition of all 436 atoms (1.14 Å) and compare well with the estimated coordinate errors of the two structures.

To further validate the choice of conformationally invariant regions, the statistics for the elements of the EDD matrices were inspected. If the only differences between two structural models are of random nature and the uncertainty estimates are correct, the distribution of error-scaled difference distance matrix elements is expected to be Gaussian with a mean of zero, a standard deviation of one and a χ^2 (for the definition of χ^2 see Table 2) for the comparison between the expected and the observed distribution of approximately 1.0.

As expected, the distribution of EDD-matrix elements between the entire molecules of Src kinase is not Gaussian; the mean and the standard deviation for 94 830 EDD-matrix elements are −0.91 and 2.38, respectively, and the distribution does not have a Gaussian shape (Fig. 2*a*; $\chi^2 = 60.2$). The non-Gaussian behaviour of the distribution indicates that some of the differences observed originate from systematic displacements of atoms on their own or in groups.

For the regions selected as conformationally invariant, the distributions are much closer to Gaussian behaviour (Table 2 and Fig. 2*a*). For example, the 7750 elements corresponding to the error-scaled difference distance matrix comparing the conformationally invariant part of the C-terminal lobe of the kinase domain have a mean value of −0.23 with a standard deviation of 0.84. These values in combination with a χ^2 of 2.2 indicate that both the error model and the choice of atoms for the conformationally invariant region are reasonable.

4.2. Description of domains

Beginning at the N-terminus of the molecule, the first conformationally invariant region identified corresponds to the SH3 domain extending from Thr84 to Ile143. Apart from the three N-terminal residues and a part of the n-Src loop (Ile110–Glu115; Xu *et al.*, 1997), the polypeptide backbone of this part of the molecule is identical within the errors between the two structures (Fig. 1). Both flexible regions are involved in crystal contacts. For the n-Src loop, a high mobility has also been observed in solution studies of single SH3 domains of Fyn (Morton *et al.*, 1996) and the human MIA protein (Stoll *et al.*, 2003) and in crystallographic and solution studies of the SH2/SH3 tandem domain from Fyn kinase (Arold *et al.*, 2001; Ulmer *et al.*, 2002).

The second domain, the SH2 domain, begins at residue Gln144, immediately after the end of the SH3 domain and extends to residue Gln251. This domain comprises several flexible parts that are involved in crystal contacts (Ala165–Arg169, Thr179–Gly182, Lys203–Phe212). The location of the boundary between the SH3 and SH2 domains between Ile143 and Gln144 is consistent with the finding in the recent study on the SH2/SH3 domain pair from Fyn (Arold *et al.*, 2001) that the equivalent residue of Ile143 in Fyn (Ile144) is highly mobile, suggesting that a hinge between the SH2 and SH3 domains could be located in this region. The C-terminal part of this domain contains residues that are normally not included in the SH2 domain itself but are considered as the SH2 linker connecting the SH2 domain and the N-terminal lobe of the kinase domain. This linker contains Pro250, which is known to interact with the peptide-binding region of the SH3 domain. A hydrogen bond with a constant length of 2.6 Å between Pro250 O and Tyr136 OH is in fact present in both structures. However, as the analysis of the error-scaled difference distance matrices is limited to C^α atoms, the relative movement of Pro250 CA and Tyr136 CA [$C^\alpha - C^\alpha$ distance = 8.2 (1fmk) *versus* 7.6 Å (2src)] is detected as significant. This observation indicates that while the backbone of the SH3 domain is moving with respect to the SH2 linker, side chains (in this case the side chain of Tyr136) involved in binding can compensate for such a motion in order to maintain key interactions.

The C-terminus of the molecule (Glu524–Pro529) is attached to the SH2 domain *via* the phosphorylated Tyr527. Although this region is distant from the SH2 domain in terms of sequence, the automatic algorithm has nevertheless added it to the structural module containing the SH2 domain.

The region corresponding to the N-terminal lobe of the catalytic domain is connected to the SH2 linker *via* two flexible residues (Thr252 and Gln253) and starts with Gly254, which is immediately followed by Leu255, whose side chain is anchored in a deep hydrophobic pocket of the N-terminal lobe. In the N-terminal lobe, one pair of residues, Leu269 and Glu270, is involved in a crystal contact. The difference in conformations found for a second pair of residues, Val329 and Ser330, may be caused by compensating shifts owing to an inaccuracy in the modelling of the neighbouring residue Val328: although the environment of the side chain of this residue is identical in both structures, it has been modelled in two different rotamer conformations with a difference in χ_1 of ~65° in 1fmk and 2src. Part of the polypeptide that topologically belongs to the N-terminal lobe of the kinase domain (Leu317–Lys321) moves together with the C-terminal lobe of the kinase domain. This part is coupled to the C-terminal lobe by hydrogen bonds (atoms involved and distances in 1fmk/2src: Leu317 N···Tyr276 OH = 2.9/3.0 Å, Lys321 N···Val502 O = 2.9/2.9 Å, Lys231 NZ···Gln369 OE1 = 2.9/2.9 Å). The relative motion of this loop region with respect to the N-terminal lobe is facilitated by the C-terminal part of the α C helix, which shows some flexibility (Gln309–Lys316).

A linker containing a glycine (Lys343, Gly344, Leu345) provides the connection to the C-terminal lobe of the kinase domain. In this domain, a number of residues are affected by different crystal contacts (Glu353–Arg359, Glu486–Leu494). The activation loop becomes ordered upon binding of the substrate analogue (Asp404–Pro425), which has knock-on effects on the adjacent regions Ala433–Leu435 and Val467–Val471. The C-terminal domain connects to the C-terminus of the molecule that is attached to the SH2 domain by a short flexible linker consisting of residues Thr521–Thr523.

4.3. Relative motion of the domains

Between 1fmk and 2src, all four domains of Src kinase move with respect to one another. Superposition based on the different conformationally invariant regions allows us to quantify the respective movements and to understand the effects that these movements have on individual atoms.

As observed previously (Xu *et al.*, 1999), when a substrate analogue is bound the two lobes tilt against each other (here by ~5°) to accommodate the ligand in a cleft between them.

Interactions between the SH3 domain and the N-terminal lobe are mostly mediated by residues located in the SH2 linker. On the SH3 side of the linker, hydrogen bonds are found between Lys249 NZ and Asp91 O, Pro250 O and Tyr136 O, Thr252 OG1 and Arg95 NH1, and Gln253 O and Trp118 NE1. On the kinase-domain side of the linker, the side chains of Leu255 and Gln241 interact with the N-terminal lobe of the kinase domain. Interestingly, the majority of these interactions stay intact despite a relative movement of ~3° (assuming a hinge approximately at Tyr118 CA) of the SH3 domain relative to the N-terminal lobe of the kinase domain. For such a small movement, the side chains can compensate for differences in relative backbone positions.

In 1fmk, an important interaction between the kinase domain and the SH3 domain is the salt bridge formed between the guanidino moiety of Arg318 and the carboxylate group of Asp117. Upon binding of the substrate analogue, the backbone of Arg318 moves together with the C-terminal half of the kinase domain and the distance between the C^α atoms of Arg318 and Asp117 increases from 12.1 to 13.0 Å. This movement leads to a loss of this salt bridge (the distances between the carboxylate O atoms and the guanidino N atoms increase from 2.7 and 3.1 Å to more than 9 Å; Fig. 1*b*). The breaking of this salt bridge may endow the SH3 domain with more conformational freedom, allowing it to alter its position with respect to the SH2 domain by ~4° (around a hinge in Ile144 CA). Such flexibility in the relative positioning of the two regulatory domains of Src kinase has been found to be possible in crystallographic and solution studies of the SH2/SH3 pair alone (Arold *et al.*, 2001). However, for the assembled state of the kinase, molecular-dynamics simulations have predicted a strong dynamic coupling between the two domains (Young *et al.*, 2001) that is mostly mediated by the linker region around residue 144. The detection of a relative motion of the SH2/SH3 domain between two assembled forms of Src kinase may indicate that in the assembled form interactions of the regulatory domains with the kinase domain (in particular the salt bridge between Arg318 and Asp117) are necessary to provide a scaffold for stabilizing the conformation of the SH2/SH3 domain pair. An intact linker between the SH2 and SH3 domains would thus be a necessary but not sufficient condition for a strong dynamic coupling between the two domains.

5. Conclusion and perspectives

Four conformationally invariant regions, corresponding to the four canonical domains of Src kinase, have been identified by automatic analysis of the error-scaled difference distance matrix between two conformers. The detailed analysis of the conformationally invariant and the flexible regions shows that the proposed method can be used for delineating structural domains in proteins. In particular, the method is able to detect parts of a structural domain that are distant in sequence from the bulk of a domain. Examples in the present study are the C-terminal part of the molecule containing the phosphorylated tyrosine, which is included in the group of atoms containing the SH2 domain, and the loop containing Arg318, which has been found to be dynamically coupled to the C-terminal lobe of the kinase domain, although in terms of sequence it is located in the N-terminal half of the kinase domain.

All four domains move with respect to each other when a substrate analogue binds to the active site. Notably, the relative motions include a change between the SH3 and SH2 domain, indicating that in the inactive form the SH2/SH3 tandem is not as rigid as indicated by molecular-dynamics simulations. However, it should be noted that the substrate-bound form of Src kinase studied here is not fully assembled, as an important interaction between the kinase half and the regulatory half is lost as a result of the relative movement of the SH3 domain and the C-terminal half of the kinase domain.

The flexible regions within and between the conformationally invariant parts of the molecule are either involved in crystal contacts or their flexibility is related to the function of the protein. The exception is a pair of residues in the N-terminal lobe of the kinase domain, for which no obvious reason for different conformations could be found. A possible source for the differences observed is a possible inaccuracy in the modelling of a neighbouring residue in one of the two structures. With the help of the experimental diffraction data, this hypothesis could be checked easily.

From a statistical point of view, the choice of atoms forming conformationally invariant regions can be validated by inspecting the distribution of error-scaled difference distance matrix elements for these atoms: for an accurate error model and a correct choice of atoms, these elements should be normally distributed. For the four domains found, the corresponding distributions are not perfectly Gaussian, but given the crude assumptions made for estimating the coordinate uncertainties, the agreement of the observed with the expected distribution is acceptable. In the future, the expected distribution of the error-scaled difference distances for pairs of very closely related structures (*e.g.* two models refined against two data sets from the same crystal) could be used as a target in the empirical optimization of error models for coordinate uncertainties. In fact, the adjustment of error models until the expected statistics for the resulting data are obtained is common practice in the processing of X-ray diffraction data (Borek *et al.*, 2003), and similar approaches could be implemented for structural data. Such and other improved models for the coordinate uncertainties in structural models are necessary to provide a statistical basis for the creation of robust tools for structural bioinformatics.

I am grateful to Barbara Brannetti for comments on the manuscript. A computer program, *ESCET*, implementing the ideas described is available as a Linux executable from the author on request.

References

Arold, S. T., Ulmer, T. S., Mulhern, T. D., Werner, J., Ladbury, J. E., Campbell, I. D. & Noble, M. E. (2001). *J. Biol. Chem.* **276**, 17199–17205.

Borek, D., Minor, W. & Otwinowski, Z. (2003). *Acta Cryst.* D**59**, 2031–2038.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.

Huse, M. & Kuriyan, J. (2002). *Cell*, **109**, 275–282.

Jiang, J., Abola, E. & Sussman, J. L. (1999). *Acta Cryst.* D**55**, 4.

Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.

Lesk, A. M. (1995). *Combinatorial Pattern Matching, Proceedings of the 6th Annual Symposium, CPM 95, Espoo, Finland, July 5–7, 1995*, edited by Z. Galil & E. Ukkonen, pp. 248–260. Berlin: Springer.

McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.

Morton, C. J., Pugh, D. J. R., Brown, E. L. J., Kahmann, J. D., Renzoni, D. A. C. & Campbell, I. D. (1996). *Structure*, **4**, 705–714.

Nichols, W. L., Rose, G. D., Ten Eyck, L. & Zimm, B. H. (1995). *Proteins*, **23**, 38–48.

Nishikawa, K., Ooi, T., Yoshinori, I. & Saito, N. (1972). *J. Phys. Soc. Jpn*, **32**, 1331–1347.

Schneider, T. R. (2000). *Acta Cryst.* D**56**, 714–721.

Schneider, T. R. (2002). *Acta Cryst.* D**58**, 195–208.

Stoll, R., Renner, C., Buettner, R., Voelter, W., Bosserhoff, A.-K. & Holak, T. A. (2003). *Protein Sci.* **12**, 510–519.

Ulmer, T. S., Werner, J. M. & Campbell, I. D. (2002). *Structure*, **10**, 901–911.

Wriggers, W. & Schulten, K. (1997). *Proteins*, **29**, 1–14.

Xu, W., Doshi, A., Lei, M., Eck, M. & Harrison, S. (1999). *Mol. Cell*, **3**, 629–638.

Xu, W., Harrison, S. & Eck, M. (1997). *Nature* (*London*), **385**, 595–602.

Young, M. A., Gonfloni, S., Superti-Furga, G., Roux, B. & Kuriyan, J. (2001). *Cell*, **105**, 115–126.