doi:10.1006/jmbi.2000.3854 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 300, 563–574

JMB

Solution Structural Studies and Low-resolution Model of the *Schizosaccharomyces pombe* sap1 Protein

Michael Bada[1], Dirk Walther[2], Benoît Arcangioli[3], Sebastian Doniach[1] and Marc Delarue[4]*

[1]*Department of Applied Physics and SSRL, Stanford University Stanford, CA 94305, USA*

[2]*Incyte Pharmaceuticals, Palo Alto, CA 94306, USA*

[3]*Unité des Virus Oncogènes Institut Pasteur, 25 rue du Dr Roux, 75015 Paris, France*

[4]*Unité de Biochimie Structurale, Institut Pasteur 25 rue du Dr Roux 75015 Paris, France*

**Corresponding author*

Sap1 is a DNA-binding protein involved in controlling the mating type switch in fission yeast *Schizosaccharomyces pombe*. In the absence of any significant sequence similarity with any structurally known protein, a variety of biophysical techniques has been used to probe the solution low-resolution structure of the sap1 protein. First, sap1 is demonstrated to be an unusually elongated dimer in solution by measuring the translational diffusion coefficient with two independent techniques: dynamic light-scattering and ultracentrifugation. Second, sequence analysis revealed the existence of a long coiled-coil region, which is responsible for dimerization. The length of the predicted coiled-coil matches estimates drawn from the hydrodynamic experimental behaviour of the molecule. In addition, the same measurements done on a shorter construct with a coiled-coil region shortened by roughly one-half confirmed the localization of the long coiled-coil region. A crude T-shape model incorporating all these information was built. Third, small-angle X-ray scattering (SAXS) of the free molecule provided additional evidence for the model. In particular, the $P(r)$ curve strikingly demonstrates the existence of long intramolecular distances. Using a novel 3D reconstruction algorithm, a low resolution 3D model of the protein has been independently constructed that matches the SAXS experimental data. It also fits the translation diffusion coefficients measurements and agrees with the first T-shaped model. This low-resolution model has clearly biologically relevant new functional implications, suggesting that sap1 is a bifunctional protein, with the two active sites being separated by as much as 120 Å; a tetrapeptide repeated four times at the C terminus of the molecule is postulated to be of utmost functional importance.

Keywords: DNA-protein interactions; low-resolution modelling; ultracentrifugation analysis; dynamic light-scattering; small angle X-ray scattering

Introduction

The fission yeast *Schizosaccharomyces pombe* switches its mating type mitotically, producing a cell population of both mating types called P (for plus) and M (for minus). The switching process occurs by a gene conversion event, from one of the two donor loci, *mat2*P and *mat3*M, to the acceptor locus *mat1*. The mating type switching is controlled by a chromosomal imprinting event that marks one strand of the *mat1* locus (Klar, 1987; Arcangioli, 1998; Dalgaard & Klar, 1999). The protein binds DNA as a dimer at about 140 bp away from *mat1* (Arcangioli & Klar, 1991), at a specific site called the SAS1 element (switch activating site). The cloning of the gene *sap1* allowed to demonstrate that this gene is essential to cell growth independently of mating type switching (Arcangioli *et al.*, 1994); it is believed that sap1 is essential for chromosomal DNA organization (B.A., unpublished results).

The binding of sap1 to its DNA target is well documented and has been the subject of several biochemical studies (Ghazvini *et al.*, 1995). Its most

Abbreviations used: SAS, switch activating site; DLS, dynamic light-scattering; *M*, molecular mass; SAXS, small-angle X-ray scattering; PDB, Protein Data Bank.

E-mail address of the corresponding author: delarue@pasteur.fr

0022-2836/00/030563–12 $35.00/0

favourable DNA-binding site is a direct repeat of five nucleotides separated by 5 bp pairs. Biochemical studies have localized the DNA binding domain in the N terminus of the protein, flanked by two distinct dimerization domains. Moreover, it appears that sap1 bends the DNA when it binds to its specific recognition site (M. Ghazvini & B.A., unpublished results).

To date, sequence databases searches revealed no sequence similar to sap1 using state-of-the-art software (e.g. BLAST2P, FASTA in Swissprot and Sptrembl sequence databases). Even though the structural characterization of specific DNA-protein interaction is growing for transcription factors (Pabo & Sauer, 1992), methyltransferases and endonucleases, there is still a need to increase our knowledge of new protein topologies interacting with specific sequences of DNA. In particular, most of the current body of knowledge of proteins interacting with tandem repeats concerns zinc finger motifs (Rastinejad *et al.*, 1995), homeodomains (Li *et al.*, 1995), or the Rel-NFκb family (Muller *et al.* 1995), which are absent from sap1.

We have performed a series of biophysical measurements to further characterize the architecture and structural organization of this small (254 amino acid residues) protein in solution. In particular, in the course of crystallization trials, the dynamic light-scattering (DLS) technique was used to assess the monodispersity of the solution (Ferré d'Amaré & Burley, 1994); it revealed a highly asymmetric molecule. We have used a number of other biophysical methods to confirm this result and derive a low-resolution model of the protein. Finally, we make use of this molecular description of the protein and its very peculiar shape to derive a specific and testable functional hypothesis.

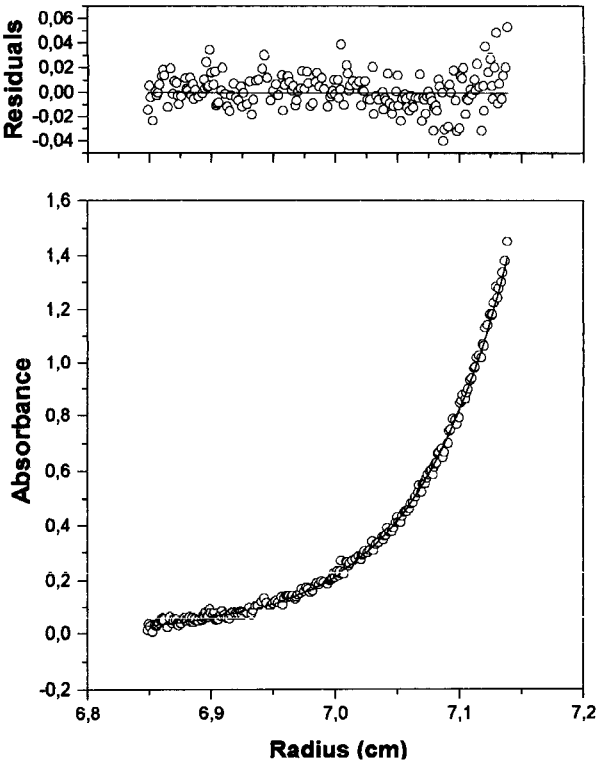

Figure 1. Sedimentation equilibrium of sap1 1-10 construct at 18,000 rpm and 20 °C. The top panel represents the goodness of fit of the experimental data with a model of a single species in solution of mass 46 kDa. UV absorption was read at 290 nm after 24 hours of equilibration.

Results

sap1 is monodisperse in solution and forms a dimer of 46 kDa

The ultracentrifugation measurement clearly shows that sap1 is a dimer in solution. Indeed, the fit is excellent with only one species in solution (Figure 1). Assuming a specific volume of 0.724 g/cm^3, we got a molecular mass of 46-50 kDa in two separate independent experiments. The range of the concentration spanned in the cell is 0.1-5 mg/ml; the expected molecular mass of the dimer is 47 kDa (203 residues, six His and four additional residues due to the cloning itself made the 1-10 construct 213 residues long, with an estimated molecular mass of 23.5 kDa for each monomer, see Figure 6).

The DLS experiments also indicated only one species in solution, because the polydispersity of the solution was very low. Indeed, the width of the distribution of the molecular dimensions of the particles undergoing Brownian motion was estimated to be less than 15% of their mean value. The fit of the autocorrelation function decay was excellent (Figure 2) using the theoretical curve allowing for only one species in solution (equation (2)).

The translational diffusion coefficient is incompatible with a model of a spherical molecule

However, the estimated molecular mass (M) derived from the DLS experiments departed very much from the expected value. We took this as an indication of a non-spherical shape of the molecule. Indeed, the molecular mass estimated by DLS is reliable only if the molecule is a sphere, that is to say if one can write $f = 6\pi\eta R$ (see equation (3)).

We measured the translation diffusion coefficient at different protein concentrations for the sap1 1-10 construct and then extrapolated it to zero concentration. The measured D_{trans} is 5.2 10^{-7} cm^2/s (Figure 3). The expected D_{trans} for a sphere of mass 47 kDa is about 7.3 10^{-7} cm^2/s.

We also measured the D_{trans} of a shorter construct of sap1, namely 1-9, where 28 residues have been deleted from the C terminus of the molecule

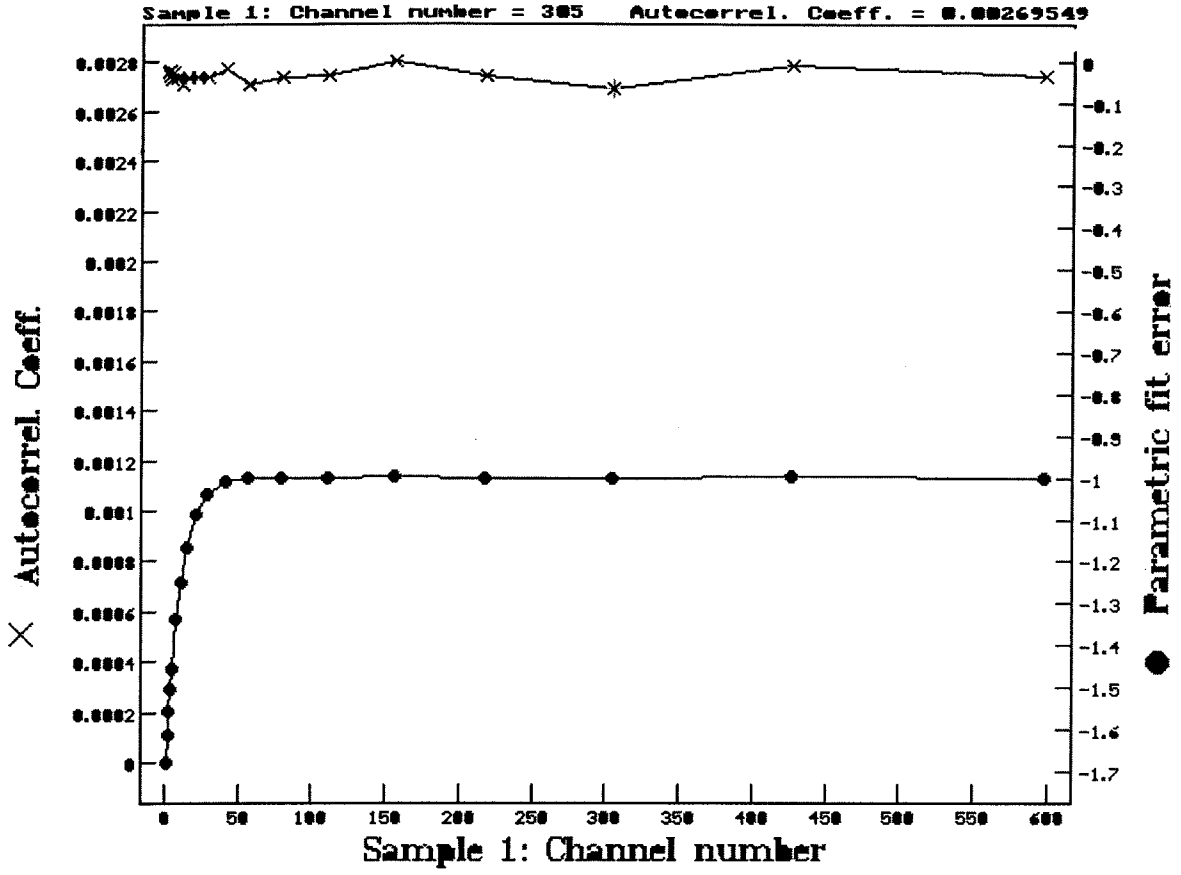

Figure 2. Dynamic light-scattering experiment of the 1-10 construct at 25 °C. The top part of the Figure represents the goodness of fit with equation (3) of the text (scale on the left), while the decay of the autocorrelation function, whose time constant gives the translation diffusion coefficient, is at the bottom of the Figure, with an inverted representation (scale on the right). The concentration was 2.5 mg/ml in 400 mM NaCl, 100 mM Hepes (pH 7.0).

(see Figure 6). The D_{trans} at zero concentration is measured to be 6.9 10^{-7} cm^2/s (Figure 3), while the expected D_{trans} for a sphere of mass 41 kDa is 7.7 10^{-7} cm^2/s.

A spherical model of the molecule would imply a ratio:

$$D_{trans}(1\text{-}9)/D_{trans}(1\text{-}10) = (M(1\text{-}10)/M(1\text{-}9))^{1/3} = 1.05$$

since D_{trans} is proportional to the inverse of the radius of the molecule. Experimentally, we measured a ratio for the two constructions of 1.33. This indicates that the C-terminal part of the molecule is deeply involved in its non-globularity.

The D_{trans} measurements of both 1-10 and 1-9 were confirmed by another technique, namely ultracentrifugation, which gave similar numbers, albeit with a larger experimental error in the data points. Nevertheless, by using a linear fit of all the experimental points, we could estimate the extrapolation to zero concentration as 5.09(±0.12) 10^{-7} cm^2/s for 1-10 sap1 and 6.8(±0.17) 10^{-7} cm^2/s for the 1-9 sap1 construct (Figure 3), in good agreement with the DLS.

For the sap1 1-10 construct, we can also calculate f_{exp}/f_{sphere} to be 1.40, where f is the friction coefficient related to the diffusion coefficient by the usual Einstein relationship (equation (3)). Assuming the 47 kDa dimer molecule to have an ellipsoidal shape, we then can estimate the molecule to be very elongated with an axial ratio a/b of 5-10, because $f_{ellipsoid}$ values have been tabulated for different ratios a/b of a prolate ellipsoid (Cantor & Schimmel, 1980). Moreover, assuming a specific volume of 0.724 cm^3/g, the dimensions of the molecule were estimated to be roughly $2a = 110$ Å and $2b = 18$ Å, which seems rather odd for a protein. It is worth mentioning that the molecule shape is not necessarily best represented by a prolate ellipsoid and could assume various forms; for instance a T, or a dumbbell shape could also be tried. Nevertheless, this approximate calculation sets the stage (and orders of magnitude) for subsequent reasoning.

The small-angle X-ray scattering (SAXS) spectrum indicates very long (>120 Å) intramolecular vectors

The Guinier plots of the SAXS spectra allowed calculation of the radius of gyration (R_g) after extrapolation to zero concentration. They are 42.0

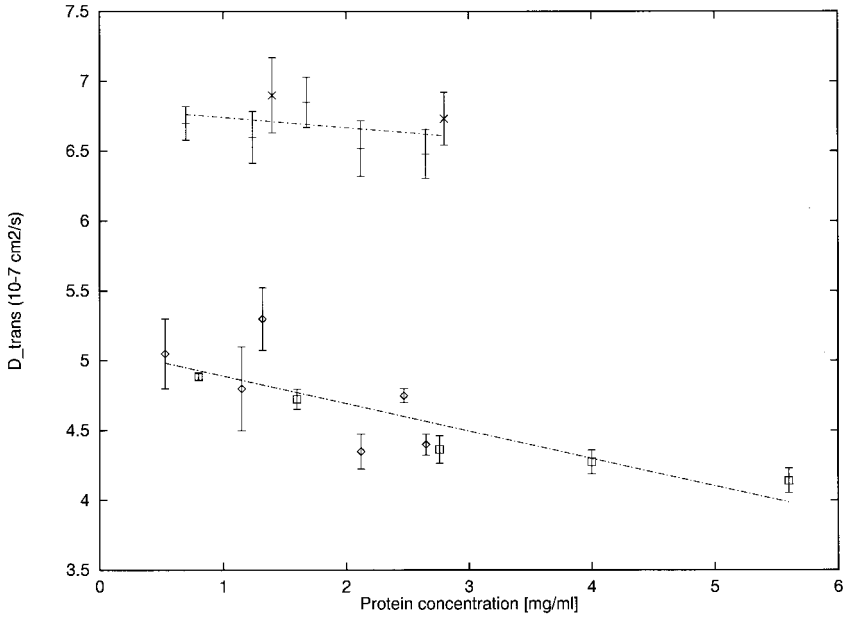

Figure 3. Translational diffusion coefficient for sap1 1-10 and 1-9 constructs. Both the measurements done with dynamic light-scattering (DLS) technique (crosses and diamonds) and the ones with centrifugation (hyphens and squares) are represented. They were made at different protein concentrations (in mg/ml) and the result is plotted in units of 10^{-7} cm^2/s. The error bars of the DLS points correspond to the highest and lowest values measured in a set of six to eight independant experiments. The error bars in the centrifugation points represent the highest and lowest values calculated on one set of time-points measured on the same sample, but with alternative fits of the the $z(t)$ curves, i.e. by excluding some of the ''outlier'' experimental points or not. Linear regression fits to all experimental data points are also represented for both constructs.

and 31.5 Å for the 1-10 and 1-9 constructs, respectively (see Table 1). This is also incompatible with modeling the protein as a sphere, since we expect roughly 25 and 22 Å for molecules of these masses, respectively. The difference between the 1-10 and 1-9 experimentally observed radius of gyration is especially striking, since they differ only by the removal of 28 amino acid residues, i.e. 13 % of the mass of the protein.

The SAXS spectra also allowed us to calculate the $P(r)$ curves, shown in Figure 4 for both the sap1 1-10 and 1-9 constructs. This led to an evaluation of the maximum distance between any couple of points in the molecule, D_{max}, at the rightmost part of the $P(r)$ curve, which could be estimated to be 157(±5) Å for 1-10 and 110(±5) Å for 1-9 sap1. This is compatible with the DLS and ultracentrifugation measurements of the diffusion coefficients, which pointed to a very elongated molecule. Assuming that the difference in distance between the two constructs is covered by an α-helix of pitch 1.54 Å, we get a peptide of length

Table 1. Calculated and measured R_g and D_{trans} values for the 1-10 and 1-9 constructs of sap1

	R_g^{exp} (Å)	R_g^{calc} (Å)	D_{trans}^{exp} (cm^2/s)	D_{trans}^{calc} (two models) (cm^2/s)
sap1: 1-10	42.0	40.2	$5.1(\pm0.12) \times 10^{-7}$	$6.3\text{-}4.5 \times 10^{-7}$
sap1: 1-9	31.5	26.7	$6.8(\pm0.7) \times 10^{-7}$	$8.4\text{-}6.9 \times 10^{-7}$
sap1: 1-10 +31 bp DNA	46.0	53.0		

The R_g values are measured by the Guinier method with SAXS data and extrapolated to zero concentrations (Cantor & Schimmel, 1980). The D_{trans} coefficients were measured both by DLS and by centrifugation, and the extrapolation to zero concentration was obtained by linear regression (Press *et al.*, 1992). The calculated values were obtained by first transforming atomic coordinates into an assembly of spheres (program AtoB, Byron, 1997) and then processed through the HYDRO program (Garcia de la Torre *et al.*, 1994). For the R_g values, only the calculated values of the T-shaped model described in the text are presented, since the second model, obtained by fitting directly the SAXS data, automatically gives the right experimental R_g value (Walther *et al.*, 2000). For the D_{trans} coefficients, two values are given: the first value is the one of the T-shaped model, for which it can be seen that there is less matter in the model than in the experiment, whereas the second value comes from the 3D reconstruction from SAXS data. The measured $D_{trans}(1\text{-}9)/D_{trans}(1\text{-}10)$ ratio is 1.33, while the expected one, if the molecules were spheres, would be $D_{trans}(1\text{-}9)/D_{trans(1\text{-}10)} = R_2/R_1 = (M_2/M_1)^{1/3} = 1.05$, and this points to a very elongated model for the molecule.

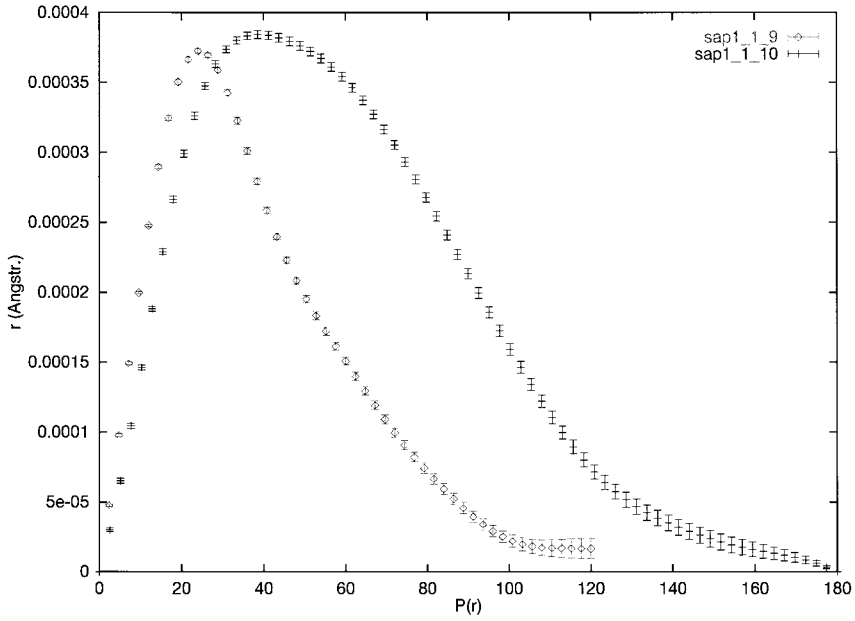

Figure 4. $P(r)$ curve for both the 1-10 and 1-9 constructs. The curves were calculated from the experimental $I(s)$ curve with the program GNOM (Semenyuk & Svergun, 1991). It clearly shows that sap1 1-10 contain intramolecular vectors as long as 160 Å , whereas sap1 1-9 does not extend much further away than 110 Å (see the text). The protein concentration was 5 mg/ml, in 400 mM NaCl at pH 7.0.

30 ± 5 residues, i.e. close to the difference in mass between the two molecules (28 residues).

The plot of log $QI(Q)$ as a function of Q^2 (see Boehm *et al.*, 1999) clearly showed a two-regime decreasing curve; the low-resolution regime gives a cross-sectional radius of gyration of $R_{XS\text{-}1} = 22.1$ Å. This was converted into the corresponding length of the underlying cylinder $h = 137$ Å (Hjelm, 1985), which is compatible with the D_{max} measurents and a T-shaped model (see below).

Taken together, these data support the notion that sap1 contains long coiled-coil helices. This is not totally unexpected, since long coiled-coil helices are actually quite common in DNA-binding proteins (for a review, see Lupas, 1996).

The R_g of sap1 alone or complexed with its DNA target argues for a T-shaped molecule

The SAXS spectrum of a short oligonucleotide (31 bp long) complexed to the sap1 protein was measured in solution. The almost unchanged measured R_g (see Table 1) led us to restrict our tentative low-resolution modelling of sap1 to T-shaped molecules, where the upper part of the T would be the N-terminal DNA-binding domain, and the bar of the T would be the coiled-coil dimerization domain. The 1:1 stoichiometry of the complex was carefully adjusted by first measuring as accurately as possible the concentrations of both protein and DNA stock solutions (before mixing) using UV absorption spectroscopy, collecting UV spectra at several concentrations and checking the linearity of the measurement; however, we cannot exclude a 5% error in the actual concentration of the protein, since it is based on a calculated value of the extinction coefficient of the protein, which is assumed in this method to depend solely on its amino acid composition.

Prediction of very long coiled-coil regions and sequence analysis

The program COILS (Lupas *et al.*, 1990) was used on the sap1 sequence and indeed strongly predicted a long coiled-coil region in the C-terminal part of the 1-10 and 1-9 constructs (Figure 5). This region spans residues 145-210 of sap1, i.e. about $65 \times 1.54 = 100$ Å. In fact, removing the last 28 amino acid residues of 1-10 to obtain 1-9 is equivalent to reduce the length of the coiled-coil region by a factor of about 2 (Figure 6). This can be used to assess the validity of low-resolution models based on the 1-10 hydrodynamic behaviour alone (see below). The resulting molecule is still a dimer in solution (data not shown).

There is another region of potentially helical character between positions 110 and 140, of which we will make use below. Also, there is some tendency towards helical conformation in the N-terminal end of the molecule, which has been implicated in previous studies as another potential dimerization region, allowing cooperative DNA-binding of pairs of dimers (Ghazvini *et al.*, 1995).

We looked in the sequence for potential hinge regions between the different structural elements and we noted the following peculiar stretches of sequences that are rich in amino acid types known to be involved in flexible regions (i.e. P, A, G, S and T), residues 16-26 ASxSPSSSPA, where x stands for a non-remarkable residue type. This is located just after the ''alternative dimerization domain'' and before the DNA-binding domain itself, of unknown topology.

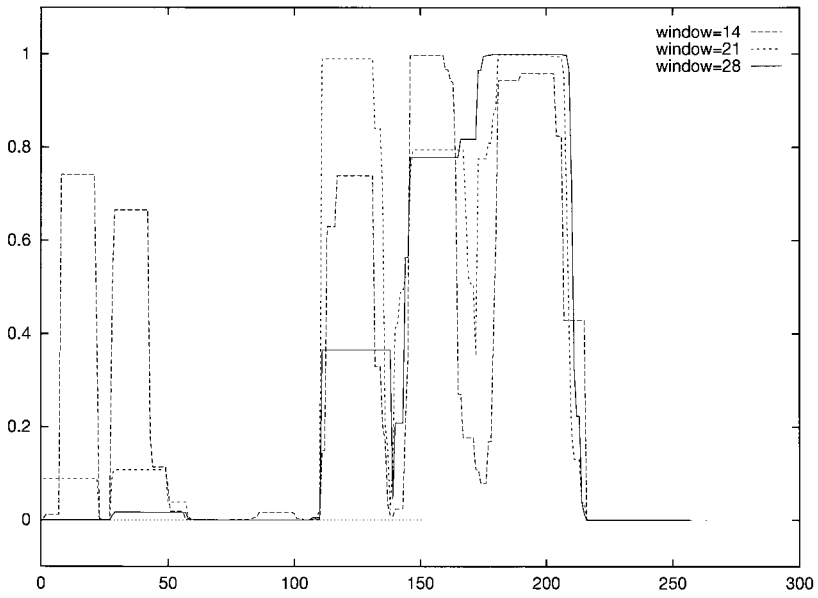

Figure 5. Result of the COILED COILS prediction method of Lupas *et al.* (1990) on the sap1 sequence; the calculation was done using different window sizes; even the most stringent window reveals a high probability of a coiled-coil structure in at least the 148-210 region.

We also found the two remarkable sequences PGxxTTP at positions 98-107, just before the first long helix predicted by the COILS algorithm, which presumably signals the end of the DNA binding domain, and residues 212-218 PPSAxAP, just after the strongly predicted coiled-coil region.

We have summarized this information in Figure 6, which conveys what we believe is the structural organization of the sap1 protein. In addition, we have indicated the remarkable stretch of residues CDC just before the beginning of the coiled-coil region, because these two cysteine residues have been implicated in the dimerization mode of the protein (B.A., unpublished results). In particular, it could be shown that these two cysteine residues may lead to inactive molecules at high concentrations of the protein, as tested through their *in vitro* DNA-binding activity, probably through illegitimate intramolecular disulfide bridges. Mutating these cysteine residues into alanine solved this problem.

Finally, there is a very peculiar sequence feature at the C terminus of the protein, between residues 220 and 235; namely, the almost perfect repetition (four times) of a tetrapeptide MG(T/A)N. We looked in the sequence database for the tetrapeptide MGTN. It was found only once as a repetition of several (in this case three) tandem repeats in the calponin family, which is involved in the formation of actin networks.

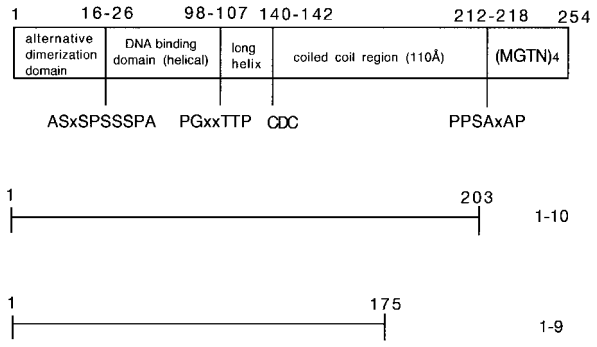

Figure 6. A representation of the structural organization of sap1; the C-terminal tetrapeptide repeat is indicated, as well as long stretches of (P,A,G,S,T) sequences, which are often involved in domain separations; in addition, the CDC sequence, implicated in the dimerization of the protein (see the text) is indicated; residue numbers are provided on top. The two constructs 1-9 and 1-10 are indicated, with respect to the original full protein.

Discussion

Low resolution modelling

We could model the entire coiled-coil region from the structure of collagen, as taken from the PDB; the side-chains were modelled using the most common rotamers in the graphics program O (Jones *et al.*, 1991). Then the side-chain rotamers were scanned and optimized so as to avoid van der Waals clashes (Koehl & Delarue, 1994). The N-terminal DNA-binding domain was very crudely modelled as the lambda repressor, since it has approximately the right number of amino acid residues. The exact model is irrelevant, since we are aiming at a low-resolution model and since the coordinates will subsequently be transformed into an assembly of spheres anyway (see below), i.e. transformed into a model of resolution around 10 Å only. The respective orientation of the coiled-coil and the DNA-binding regions were chosen to be orthogonal, mostly because of the R_g results of the sap1-DNA complexes. The additional long

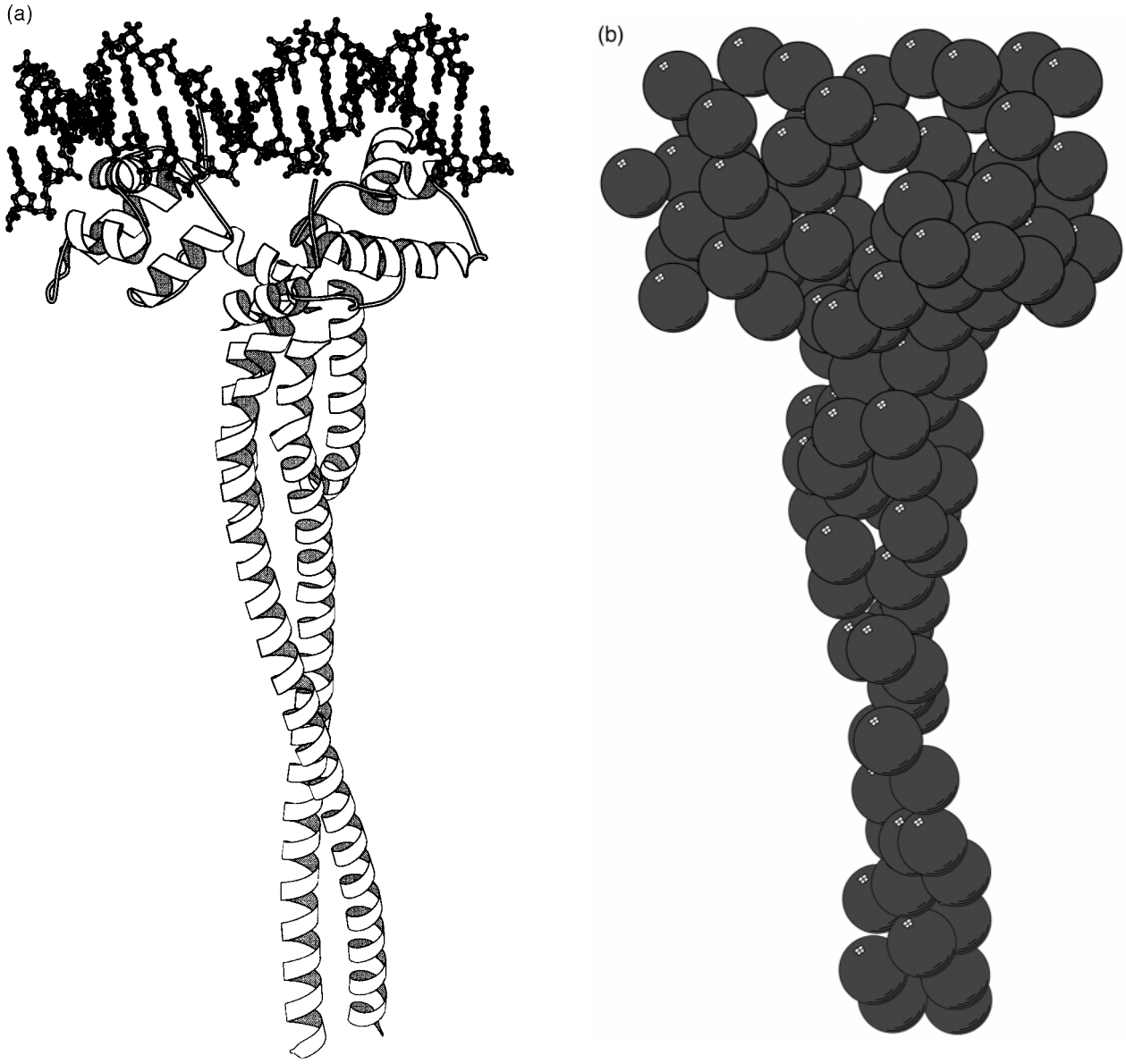

Figure 7. (a) Possible T-shaped model of the sap1 1-10 construct, complexed with a DNA 18 bp oligonucleotide (MOLSCRIPT); the model is made of a coiled-coil part, taken from collagen and mutated to the sap1 sequence with O (Jones *et al*., 1991); its length was adjusted using coiled-coil sequence predictions, and D_{trans} and R_g measurements reflecting the hydrodynamic behaviour of the molecule; there is also a DNA-binding domain (lambda repressor), and an additional helix, built in such a way as to form a helix bundle with the coiled-coil first and the N terminus third; the T-shape of the model is supported by the fact that the experimental R_g value does not vary very much for the free and DNA-bound sap1 molecule. (b) Same as (a), but converted to an assembly of spheres by the AtoB program written by Byron (1997). In this case, the size of the spheres is constant. This model could in turn be plugged into the HYDRO program written by Garcia de la Torre *et al*. (1994) to calculate its hydrodynamic parameters (i.e. R_g and D_{trans}).

helix was modelled so as to form an α-helix bundle (four helices in the dimer) at the bottom of the coiled-coil (see Figure 7(a)). The 1-9 construct was simply modelled from the 1-10 model by removing the 28 most C-terminal residues. We call this model the T-shaped model.

We could calculate the hydrodynamic properties of this model using the program HYDRO, after proper transformation of the *x*,*y*,*z* coordinates of the atoms into an assembly of spheres of either variable of constant radius using a Fortran program developed by O. Byron (Figure 7(b)). The beads radii are adjusted so that their added volume is equal to the estimated volume of the protein, which is deduced from the sequence directly, using tabulated specific volume of amino acid constituents. Its value for the 1-10 construct is 23,939 $Å^3$, with a specific volume of 0.724 cm^3/g.

The HYDRO program gives the radius of gyration as well as the translation diffusion coefficient.

The results, presented in Table 1, are in reasonable agreement with the experimental data, for both the sap1 1-10 and 1-9 constructs.

This model may not be the only possible one, but it has the virtue of being able to account for all the experimental observations.

The above model also fits quite nicely to the model produced independently by a recently developed algorithm (Walther *et al*., 2000), which works without any *a priori* hypothesis on the protein shape and is based on the goodness of fit of the model with the experimental SAXS data. In Figure 8, we show a superposition of ten runs, together with the coiled-coil model described above. In Figure 9 we compare the reconstruction of a 3D model of the truncated molecule sap1 1-9 with that for the full sap1 1-10. As may be seen, the truncated model is roughly half the size of the full model, consistent with the effects of removal of the 28 C-terminal residues being confined to the coiled-coil region of the molecule.

The average of the ten simulations of the program SAXS3D (Walther *et al*., 2000) gives a set of points representing the SAXS data. For each simulation, they sit on an hexagonal lattice, but this is no longer the case after averaging several simulations; however, each point could be assigned to an adjustable radius, in very much the same way as described above for atomic coordinates. In this way, we can also calculate translation diffusion coefficients for both sap1 1-10 and 1-9 constructs. The results are indicated in Table 1 and they are in good agreement with the experimental measurements.

Figure 8. SAXS 3D reconstruction of sap1 1-10 construct. The ten reconstructed images are superimposed on the coiled-coil ''atomic'' model (residues 142-212). Small spheres denote scatterers used for the reconstruction (grey scale denotes depth). The coiled-coil (light grey) is represented on a finer scale.

Functional implications

Based on a modular architecture of the protein, it may be speculated that sap1 is a bifunctional molecule: the N-terminal domain of sap1 is a DNA-binding protein, which, when fixed to its specific sequence target, sends a ''flag'', in the form of a tetrapeptide repeated four times (eight times for the dimer), about 120 Å away from the DNA-binding site. Sap1 may then be thought of as a mechanical device in charge of physically coupling two well-separated molecules: one of them is the DNA, the other may possibly be part of the nucleo-skeleton. This last suggestion is based on the fact that the repeated tetrapeptide found at the C-terminal part of sap1 is found also in calponins, a family of proteins involved in binding networks of actin. Failure to transmit this signal properly would make the cell-cycle go astray. Therefore, the structural architecture of the sap1 protein suggests a natural mechanical explanation for a transmission of signal between the DNA molecule and its nuclear organization on one hand and the molecules involved in the nucleo-skeleton superstructure on the other hand. This points to a role of sap1 at the chromosome segregation step and may be related to the imprinting event involved in mating type switching.

Experiments to assess the role of the most C-terminal part of sap1 *in vivo* have been undertaken, as well as attempts to characterize the phenotype of cells where this part of the protein is overexpressed.

We believe that the general architecture of sap1, coupling two distinct functional domains through a dimerization domain that separates them by a long (120 Å) coiled-coil arm, is quite general and will be found in several other nucleic acid-binding proteins. One recent example of this is the rotavirus protein NSP3, whose function is to bring the viral mRNA into close contact with the host cell translation initiation machinery (Piron *et al*., 1999) through exactly the same general architecture.

Materials and Methods

Protein expression and purification

The sap1 protein was purified as an His-Tag construction in a phage T7 polymerase-inducible *Escherichia coli* strain and purified using a Pharmacia Ni-column. The protein was eluted from the column by 0.45 M imidazole buffer (pH 7), after washing the column extensively with

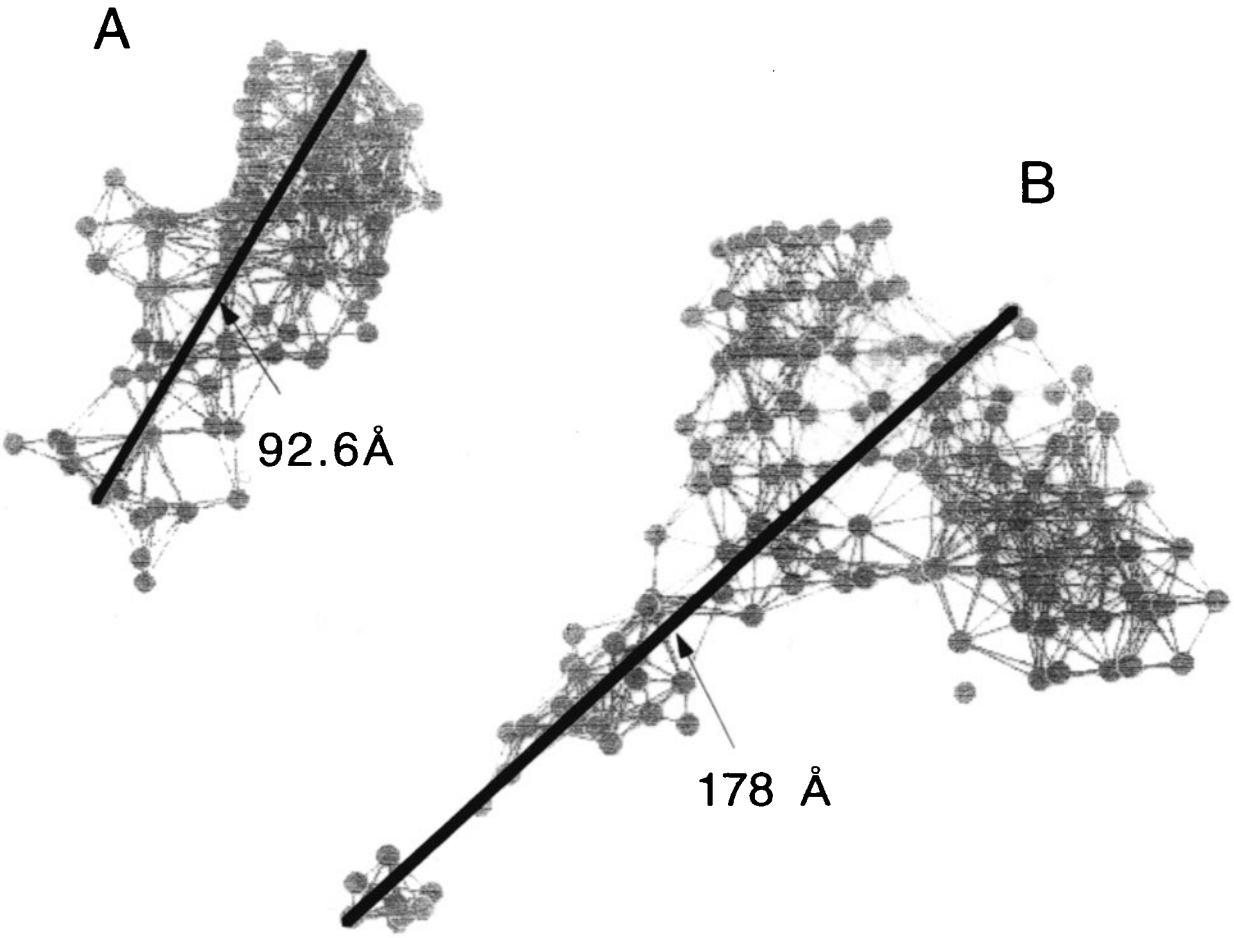

Figure 9. Comparison of SAXS 3D reconstruction of (a) the truncated version of sap1 1-9 with (b) that of the full construct 1-10. Note that (a) and (b) are not on the same scale.

10 mM imidazole buffer. The protein was estimated to be more than 99 % pure by gel electrophoresis and used without further purification. It was dialyzed extensively in a 10 mM Hepes (pH 7), 400 mM NaCl buffer before concentrating it (Amicon, Millipore Corp., Bedford, MA, USA).

Two constructs were studied during this work (see Figure 6): the 1-10 construct, which is 203 amino acid residues longs (i.e. the 51 C-terminal residues have been deleted; they are not important for DNA binding), and the 1-9 construct in which another 28 residues have been deleted from the C terminus region (175 amino acid residues in total). Due to the cloning strategy and the His-tag at the N terminus, the actual length of the 1-10 construct is 213 residues.

Light-scattering experiments

The protein solution polydispersity was analysed in a DynaPro 801 apparatus (Protein Solutions, Inc., Charlottesville, VA). The samples were passed through 0.100 μm Micropore filters prior to injection into the measurement cell. The temperature was read at the time of the experiment from an internal sensor and was not regulated; it varied from 20 °C to 30 °C. The temperature-dependency of the diffusion coefficient was corrected using the formula:

$$D_1/D_2 = T_1/T_2\eta(T_2)/\eta(T_1) \quad (1)$$

and the temperature-dependency of the viscosity was taken from the Handbook of Chemistry and Physics, CRC Press, 77th edit. (1996-1997).

The decay of the autocorrelation function can be described by the following equation:

$$\langle I(t)I(t+\tau)\rangle/\langle I(t)\rangle^2 = 1 + \exp(-D_{\text{trans}}Q^2\tau) \quad (2)$$

where D_{trans} is the translational diffusion coefficient and Q is the scattering vector modulus $Q = 4\pi n\sin(\theta/2)/\lambda$, where θ is the angle of scattering, λ the wavelength and n the refraction index (Harding, 1994).

For a sphere, the Stokes-Einstein equation relating the translation diffusion coefficient to molecular dimensions reads:

$$D_{\text{trans}} = kT/f = kT/6\pi\eta R \quad (3)$$

where R is the radius of the sphere, η is the viscosity and f the friction coefficient.

The autocorrelation function was analysed with the cumulant method, using a single exponential. The goodness of fit was evaluated with a χ^2 test. In equation (2), it can be seen that the baseline should be exactly 1. The measurements presented here are the results of six-eight independent measurements, for which the sum-of-

squares of errors is around 2-3 and the baseline around 1.002-1.001. This method also gives an estimate of the polydispersity of the solute (Murphy, 1997).

The measurements were repeated at different protein concentrations between 1 mg/ml and 6 mg/ml, and then extrapolated to zero concentration. Prior to each series of experimentation, a calibration test was run with a 10 mg/ml solution of hen egg lysozyme.

Ultracentrifugation

The 1-10 and 1-9 constructs of sap1 protein were analyzed using an XL-A ultracentrifuge apparatus (Beckman Instruments, Palo Alto, CA) equipped with UV absorption optics.

The protein solution contained 400 mM NaCl and 100 mM Hepes buffer (pH 7). The protein concentration ranged 1-10 mg/ml in different runs and was estimated from UV absorption, using a molar extinction coefficient at 280 nm $\varepsilon = 13{,}000\ M^{-1}\ cm^{-1}$.

For molecular mass (M) determination, the speed was set at either 18,000 rpm or 16,000 rpm; equilibration was assumed to be attained after 24 hours. Absorbance at a distance r from the rotor axis was read at 290 nm and was fit to the following equation:

$$C_{(r)} = C_0 \exp(-M(1 - v\rho)(r^2 - r_m^2)\omega^2/2RT) \qquad (4)$$

where M is the molecular mass of the molecule, v is the specific volume, ρ the density (1 g/cm^3), R the gas constant, T the temperature and ω the rotor speed (see Stafford, 1997; Hensley, 1996; Hansen *et al.*, 1994).

For the diffusion coefficient determination, the speed was set at 5000 rpm; absorbance measurements were made every ten minutes after injection (single scan mode), for about two hours. The temperature was set to 20 °C. The absorbance profile of the boundary spread function $c(r)$ at different times after injection in the ultracentifugation cell is transformed into a $q(r) = \ln(c(r)/(1 - c(r))$ plot, an almost linear function whose slope serves to build the $z(t)$ function at different times after injection; the $z(t)$ slope, in turn, gives the translation diffusion coefficient. More precisely, the slope of this z-plot is equal to $\pi D_{trans}/4$ and this allows for D_{trans} measurement. This is the method of Muramatsu & Minton (1989), as implemented in the program VELGAMMA of the XL-A software.

The measurement was repeated at five or six different protein concentrations from 1 mg/ml to 5 mg/ml for both 1-10 and 1-9 constructs. Extrapolation to zero concentration was done by a linear regression fit (Press *et al.*, 1992).

Solution X-ray scattering

Experiments were conducted at beamline 4-2 at the SSRL in Stanford. The wavelength was 0.1336 nm (8980 eV) as selected by a pair of Si (111) monochromator crystals. The temperature was set to 20 °C. All protein solutions were centrifuged for about 60 mn at least ten minutes before data collection. Protein concentrations ranged from 1 mg/ml to 8 mg/ml. To avoid radiation-induced protein aggregation, the protein solution was circulated continuously from a reservoir through a 1.3 mm scattering path observation flow-cell with 10 μm thick mica windows. Samples were exposed for about 100 seconds in eight successive frames for each sample; each frame was checked for radiation damage before using it for averaging. Background corrections were performed according to conventional procedures. Typical count rates were about 50,000 counts per second for sample and 30,000 counts per second for buffer. The camera length was calibrated to be 230 cm using a cholesterol myristate sample (see Chen *et al.*, 1996).

The radius of gyration estimates were evaluated according to the Guinier approximation by fitting the log$I(Q)$ *versus* Q^2 curve to a straight line, assuming a slope equal to $-R_g^2/3$ (Cantor & Schimmel, 1980) and then extrapolating to zero concentration. The fitting region was 0.0045 to 0.01 Å^{-1} in Q, the scattering vector modulus. Calculations of the $P(r)$ curves for both 1-10 and 1-9 sap1 constructs were performed according to the indirect transformation method, as implemented in the GNOM program written by Semenyuk & Svergun (1991). The cross-section radius of gyration $R_{XS\text{-}1}$ was calculated by fitting the low-angle region of a log $QI(Q)$ *versus* Q^2 plot to a straight line whose slope is $-R_{XS\text{-}1}^2/2$ (Boehm *et al.*, 1999) and the length of the underlying cylinder model calculated using the formula $h = \sqrt{(12(R_g^2 - R_{XS\text{-}1}^2))}$ (Hjelm, 1985).

Calculated scattering curves were performed by the program CRYSOL (Svergun *et al.*, 1995) and PDB type coordinates. The thickness of the hydration layer was taken to be one layer of water molecules and the excess electron density between this hydration layer and the buffer was taken to be 0.003 e/Å^3.

Sequence analysis

Sequence searches were done using the UWGCG package. The coiled-coils prediction was done using the program COILS, available through the Internet at the IRSEC site at www.isrec.isb-sib.ch/software/COILS_form.html (Lupas *et al.*, 1990).

Hydrodynamics calculations

All hydrodynamic calculations were made with the program HYDRO written by Garcia de la Torre *et al.* (1994). *XYZ* atomic coordinates were reduced and modelled as an assembly of spheres using a Fortran program (Byron, 1997), which was kindly made available to us by the author. Different numbers of spheres with adjustable radii were tried. Best results were obtained with about 100 of them. With more than 150 beads, the HYDRO algorithm becomes prohibitively slow.

The table of $f_{ellipsoid}/f_{sphere}$ friction coefficients for prolate ellipsoids of different axial ratio was taken from Cantor & Schimmel (1980).

Ab initio deconvolution of the SAXS profile

An algorithm for reconstruction of low-resolution 3D-models from SAXS data has been developed and is described elsewhere (Walther *et al.*, 2000). It is based on a ''give'n'take'' procedure of adding and deleting point scatterers on a lattice to produce a best fit to the SAXS data, while constrained by the non-linear physical requirement of positivity and compactness (see also Svergun *et al.* 1996; Chacon *et al.*, 1998).

Model manipulation, visualization and drawing

All models were visualized with the program O (Jones *et al.*, 1991). Figures were drawn with MOLSCRIPT

(Kraulis, 1991) or a program written by one of us (D.W., program Xlattice, http://www.cmpharm.ucsf.edu/~walther)

Acknowledgments

We thank M.E. Goldberg for help with centrifugation analysis experiments at the Pasteur Institute. We thank Ian Millet and Hiro Tsuruta for help on beamline 4-2 at SSRL, and Keith Hodgson for his general support and encouragement. Work at Stanford was supported, in part, by the U.S. Department of Energy through SSRL/SLAC. This work was also supported by a grant from the Association pour la Recherche sur le Cancer (to B.A.).

References

Arcangioli, B. (1998). A site- and strand-specific DNA break confers asymetric switching potential in fission yeast. *EMBO J.* **17**, 4503-4510.

Arcangioli, B. & Klar, A. J. S. (1991). A novel switch-activating site (SAS1) and its cognate binding factor (sap1) required for efficient mat1 switching in *S. pombe*. *EMBO J.* **10**, 3025-3032.

Arcangioli, B., Copeland, T. D. & Klar, A. J. S. (1994). Sap1, a protein that binds to sequences required for mating type switching, is essential for the viability of *S. pombe*. *Mol. Cell. Biol.* **14**, 2058-2064.

Boehm, M. K., Woof, J. M., Kerr, M. A. & Perkins, S. J. (1999). The Fab and Fg fragments of IgA1 exhibit a different arrangement from that of IgG: a study by X-ray and neutron scattering and homology modelling. *J. Mol. Biol.* **286**, 421-447.

Byron, O. (1997). Construction of hydrodynamic bead models from high-resolution X-ray crystallographic or nuclear magnetic resonance data. *Biophys. J.* **72**, 408-415.

Cantor, C. R. & Schimmel, P. R. (1980). *Biophysical Chemistry*, vol 2, Freeman & Co., San Francisco.

Chacon, P., Moran, F., Diaz, J. F., Pantos, E. & Andreu, J. M. (1998). Low-resolution structures of proteins in solution retrieved from X-ray scattering with a genetic algorithm. *Biophys. J.* **74**, 2760-2775.

Chen, L., Hodgson, K. O. & Doniach, S. (1996). A lysozyme folding intermediate revealed by solution X-ray scattering. *J. Mol. Biol.* **261**, 658-671.

Dalgaard, J. Z. & Klar, A. J. S. (1999). Orientation of DNA replication establishes mating type switching pattern in *S. pombe*. *Nature,* **400**, 181-184.

Ferré-d'Amaré, A. & Burley, S. K. (1994). Use of dynamic light scattering to assess crystallizability of macromolecules and macromolecular assemblies. *Structure,* **2**, 357-359.

Garcia de la Torre, J., Navarro, S., Lopez-Martinez, M. C., Diaz, F. G. & Lopez Cascales, J. J. (1994). HYDRO: a computer program for the prediction of hydrodynamic properties of macromolecules. *Biophys. J.* **67**, 530-531.

Ghazvini, M., Ribes, V. & Arcangioli, B. (1995). The essential DNA-binding protein sap1 of *S. pombe* contains two independent oligomerization interfaces that dictate the relative orientation of the DNA-binding domain. *Mol. Cell Biol.* **15**, 4939-4946.

Hansen, J. C., Lebowitz, J. & Demeler, B. (1994). Analytical ultracentrifugation of complex macromolecular systems. *Biochemistry,* **33**, 13155-13163.

Harding, S. E. (1994). Determination of diffusion coefficients of biological macromolecules by dynamic light scattering. In *Methods in Molecular Biology*, vol 22, Humana Press Inc., Totowa, NJ.

Hensley, P. (1996). Defining the structure and stability of macromolecular assemblies in solution: the re-emergence of analytical ultracentrifugation as a practical tool. *Structure,* **4**, 367-373.

Hjelm, R. P., Jr (1985). The small angle approximation of X-ray and neutron scatter from rigid rods of non-uniform cross section and finite length. *J. Appl. Crystallog.* **18**, 452-460.

Jones, T. A., Zou, J. Y., Cowtan, S. & Kjelgaard, M. (1991). Improved methods for building models in electron density maps and the localization of errors in these models. *Acta Crystallog. sect. A,* **47**, 110-119.

Klar, A. J. S. (1987). Differentiated parental DNA strands confer developmental asymmetry in daughter cells in fission yeast. *Nature,* **326**, 466-470.

Koehl, P. & Delarue, M. (1994). On the use of a self-consistent mean field theory to predict side-chains conformation and estimate their entropies. *J. Mol. Biol.* **239**, 249-275.

Kraulis, P. J. (1991). MOLSCRIPT: a program to generate both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-952.

Li, T., Stark, M. R., Johnson, A. D. & Wolberger, C. (1995). Crystal structure of the MATa1/MATa2 homeodomain bound to DNA. *Science,* **270**, 262-269.

Lupas, A. (1996). Coiled-coils: new structures and new functions. *Trends Biochem. Sci.* **21**, 375-382.

Lupas, A., Van Dyke, M. & Stock, J. (1990). Predicting coiled-coils from protein sequences. *Science,* **252**, 1162-1164.

Muller, C. W., Rey, F. A., Sodeoka, M., Verdine, G. L. & Harrison, S. C. (1995). Structure of the NF-kappa B p50 homodimer bound to DNA. *Nature,* **373**, 278-287.

Muramatsu, B. & Minton, A. P. (1988). An automated method for rapid determination of diffusion coefficients via measurement of boundary spreading. *Anal. Biochem.* **168**, 345-352.

Murphy, R. M. (1997). Static and dynamic light scattering of biological macromolecules: what can we learn? *Curr. Opin. Biotechnol.* **8**, 25-30.

Pabo, C. O. & Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61**, 1053-1095.

Piron, M., Delauney, T., Grosclaude, J. & Poncet, D. (1999). Identification of the RNA-binding, dimerization and EIF4GI-binding domains of the rotavirus non structural protein NSP3. *J. Virol.* **73**, 5411-5421.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1992). *Numerical Recipes: The Art of Scientific Computing*, 2nd edit., Cambridge University Press, Cambridge, UK.

Rastinejad, F., Perlman, T., Evans, R. M. & Sigler, P. B. (1995). Structural determinants of nuclear receptor assembly on DNA direct repeats. *Nature,* **375**, 203-211.

Semenyuk, A. V. & Svergun, D. I. (1991). GNOM: a program package for processing small angle scattering data. *J. Appl. Crystallog.* **24**, 537-540.

Svergun, D. I., Barbereto, C. & Koch, M. H. J. (1995). CRYSOL: a program to evaluate X-ray solution scat-

tering of biological macromolecules from atomic coordinates. *J. Appl. Crystallog.* **28**, 768-773.

Svergun, D. I., Volkov, V. V., Kozin, M. B. & Stuhrmann, H. B. (1996). New developments in direct shape determination from small angle scattering. 2. Uniqueness. *Acta Crystallog. sect. A,* **52**, 419-426.

Stafford, W. F., III (1997). Sedimentation velocity spins a new weave for an old fabric. *Curr. Opin. Biotechnol.* **8**, 14-24.

Walther, D., Cohen, F. E. & Doniach, S. (2000). Reconstruction of low resolution three-dimensional density maps from one-dimensional small angle X-ray solution scattering data for biomolecules. *J. Appl. Crystallog.* **33**, 350-363.

Edited by M. F. Moody

(Received 4 January 2000; received in revised form 11 April 2000; accepted 9 May 2000)