Solution structure of the cellular factor BAF responsible for protecting retroviral DNA from autointegration

Mengli Cai[1], Ying Huang[2], Ronglan Zheng[2], Shui-Qing Wei[2], Rodolfo Ghirlando[2], Myung Soo Lee[2], Robert Craigie[2], Angela M. Gronenborn[1] and G. Marius Clore[1]

The solution structure of the human barrier-to-autointegration factor, BAF, a 21,000 M_r dimer, has been solved by NMR, including extensive use of dipolar couplings which provide *a priori* long range structural information. BAF is a highly evolutionarily conserved DNA binding protein that is responsible for inhibiting autointegration of retroviral DNA, thereby promoting integration of retroviral DNA into the host chromosome. BAF is largely helical, and each subunit is composed of five helices. The dimer is elongated in shape and the dimer interface comprises principally hydrophobic contacts supplemented by a single salt bridge. Despite the absence of any sequence similarity to any other known protein family, the topology of helices 3–5 is similar to that of a number of DNA binding proteins, with helices 4 and 5 constituting a helix-turn-helix motif. A model for the interaction of BAF with DNA that is consistent with structural and mutagenesis data is proposed.

Integration of a DNA copy of the viral genome into host DNA is an essential step in the replication cycle of retroviruses[1]. The viral DNA, made by reverse transcription within the cytoplasm of an infected cell, forms part of a large nucleoprotein complex that is derived from the core of the infecting virion. Although the protein components of these preintegration complexes have not been well defined, they include the viral integrase protein that carries out the key enzymatic steps of viral DNA integration. Preintegration complexes, isolated from cells after infection with Moloney murine leukemia virus (MLV)[2] or human immunodeficiency virus type 1 (HIV-1)[3,4], efficiently integrate their DNA into an exogenous target DNA *in vitro*. A striking feature of this reaction is the strong preference for intermolecular integration into the cellular target DNA and avoidance of intramolecular integration into the viral DNA. Such autointegration *in vivo* would destroy the viral DNA before it is able to integrate into the host genome and, consequently, would abort the viral replication cycle. MLV preintegration complexes that have been incubated at high ionic strength lose the protection against autointegration and almost exclusively use their own DNA as the target for integration[5]. This loss of protection against autointegration is due to liberation of a cellular factor from the complexes. After separating salt-stripped complexes from free proteins, the protection against autointegration can be reconstituted by incubating the complexes with an extract of uninfected NIH 3T3 cells[5].

The cellular factor responsible for preventing autointegration, known as barrier-to-autointegration factor or BAF, has recently been identified using this reconstitution reaction as an assay for purification[6]. BAF also restores the activity of salt-stripped HIV preintegration complexes (H. Chen & A. Engelman, pers. comm.). BAF is an 89-residue cellular protein (Fig. 1) that does not exhibit any significant amino acid sequence similarity to any known protein, although a search of sequence databases reveals that several species express a transcript that can encode a protein

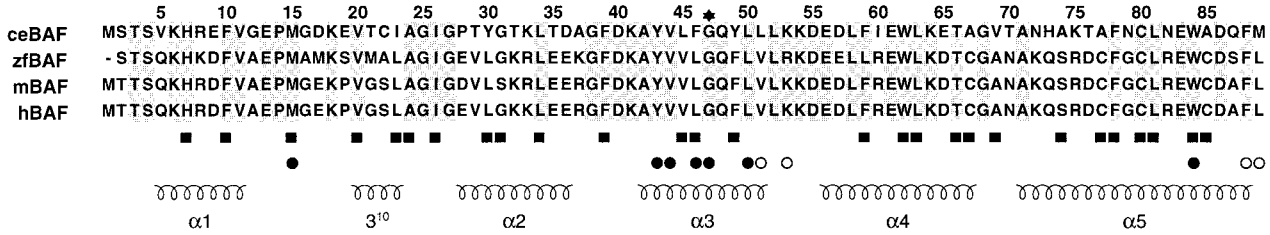

Fig. 1 Sequence alignment of human, mouse, zebrafish and *C. elegans* BAF, together with a summary of the secondary structure. Residues in the monomer that have a surface accessibility of ≤ 20% of that in an isolated Gly-X-Gly extended tripeptide (where X is any amino acid) are indicated by the filled-in-squares; residues that are buried upon dimerization and have surface accessibilities of ≤ 20% in the dimer are indicated by filled-in-circles; residues that have a surface accessibility of 20–50% in the dimer and whose surface accessibility is reduced by a factor of 2 or more relative to that in the monomer are indicated by open circles. When both a filled-in square and circle are shown, the surface accessibility of that particular residue is not only less than 20% in the monomer, but is reduced by a factor of 2 or more in the dimer. The asterisk above Gly 47 indicates the approximate location of the C_2 axis of the dimer. Only human and mouse BAF proteins have been expressed and purified. The sequence of zebrafish BAF is derived from an EST (Genbank accession number AA495322). The *C. elegans* BAF sequence is derived from an EST and genomic clone (Genbank accession numbers C41921 and B0464 respectively). Abbreviations: hBAF, mBAF, zfBAF and ceBAF, human, mouse, zebrafish and *C. elegans* BAF respectively.

[1]Laboratories of Chemical Physics and [2]Molecular Biology, Building 5, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, USA.

Correspondence should be addressed to G.M.C. *email: clore@speck.niddk.nih. gov* or A.M.G. *email: gronenborn@vger.niddk.nih.gov*

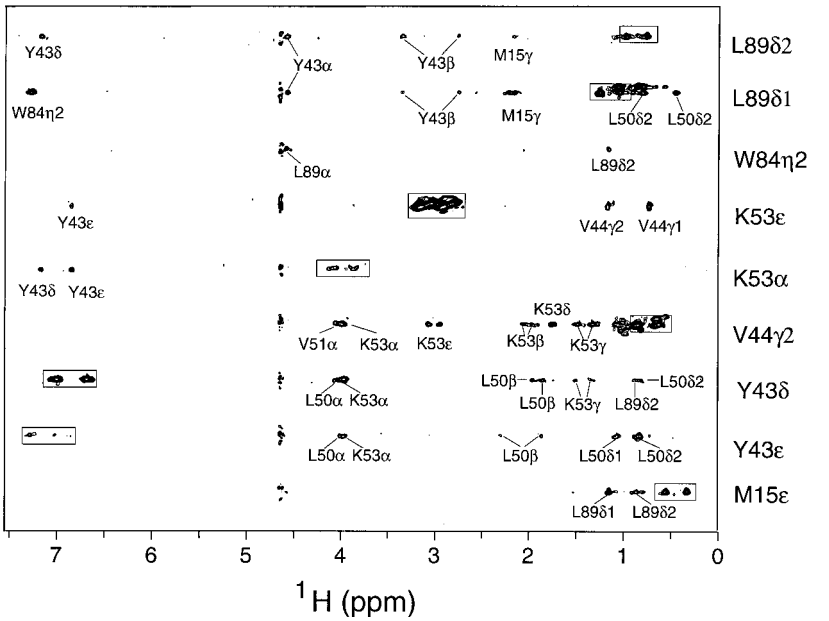

Fig. 2 Examples of strips taken from the 3D ^{13}C-separated/^{12}C-filtered NOE spectrum recorded on a 1:1 mixture of $^{13}C/^{14}N$ and $^{12}C/^{15}N$ labeled BAF in D_2O, illustrating intersubunit NOEs from protons attached to ^{13}C (in the indirect dimension) to protons attached to ^{12}C (in the acquisition dimension). The peaks that are boxed-in are residual diagonal peaks arising from ^{13}C-attached protons.

with a high degree of sequence identity. BAF is a DNA binding protein that bridges together segments of double stranded DNA. A model in which BAF compacts the viral DNA can simply account for its activity in blocking autointegration; in particular, viral DNA bridged by BAF would be les10 • octobers accessible as a target for integration[6]. Since BAF is not present in virions, the preintegration complex must acquire it from the cytoplasm of the infected cell. Given the important role that BAF plays in retroviral integration, we have solved the three-dimensional solution structure of human BAF.

Multimeric state of BAF

Sedimentation equilibrium experiments with BAF in phosphate buffer pH 6.5 containing 0.2 M NaCl yielded a measured molecular mass of 21,180 ± 450 g mol^{-1}. This compares with a calculated mass of 20,702 g mol^{-1} for a dimer of BAF with the additional N-terminal GSH residues from the thrombin cleavage site. Thus BAF exists as a monodisperse dimer population under these conditions. Size-exclusion chromatography on a Superdex 200 column also shows that BAF is dimeric under these conditions (data not shown). The ^{15}N $T_1/T_{1\rho}$ ratios at 500 MHz and 40 °C range from 7.4–10.1 with a most probable value of 8.0, resulting in an effective correlation time of ~10.4 ns, a diffusion anisotropy of ~1.34, and a ratio of the y to x components of the diffusion tensor of ~1.1[7]. These data are also fully consistent with a dimer that is slightly elongated in shape.

Structure determination

The solution structure of human BAF was solved by multidimensional heteronuclear NMR spectroscopy[8–10], making use of uniformly labeled ^{15}N and $^{15}N/^{13}C$-labeled protein, as well as heterodimers comprising a 1:1 mixture of uniformly labeled $^{13}C/^{14}N$ and $^{15}N/^{12}C$ subunits to identify intersubunit NOEs. An example of the quality of the NMR data is shown in Fig. 2 which depicts strips from the 3D ^{13}C-separated/^{12}C-filtered NOE spectrum recorded on a mixed-labeled heterodimer. This experiment specifically detects intersubunit NOEs from ^{13}C-attached protons (in the indirect dimension) to ^{12}C-attached protons (in the acquisition dimension). The structure was solved on the basis of 1,646 experimental NMR restraints per monomer. These include 48 intersubunit NOEs and 250 residual dipolar couplings per monomer. The latter directly provide information on the orientation of various interatomic vectors (N-H_N, Cα-C', N-C'and H_N-C') relative to the molecular alignment tensor, and consequently provides unique information that defines long range order *a priori*, including the relative orientation of the two subunits in the dimer[11]. Since the dimer is two-fold symmetric, the z axis of the alignment tensor lies along the long axis of the dimer, and the x axis of the alignment tensor coincides with the C_2 symmetry axis. For the purposes of defining the relative orientations of the subunits, the $^1D_{NH}$ dipolar couplings are sufficient. The other dipolar couplings principally serve to better define the backbone torsion angles. A summary of the structural statistics is provided in Table 1, and a superposition of the final ensemble of 40 simulated annealing structures is shown in Fig. 3*a*.

Description of the monomer

Each monomer of BAF is approximately globular in shape with a radius of gyration of ~12.4 Å, and composed of five helices (residues 5–10, 28–36, 42–51, 56–67 and 71–88) and a small 3_{10}-helical turn (residues 20–23) (Figs 3*b*, 4*a*,*b* and 5*b*). One face of the molecule is formed by helices 1 and 2, the helical turn and the N-terminus of helix 3, while the other face is formed by helices 4 and 5 and the C-terminus of helix 3 (Fig. 4*a*). Helix 3 thus serves to bridge the two faces. Helices 1 and 5 are approximately parallel to each other (interhelical angle ~ 21°), helix 2 is oriented at an angle of ~75°, ~90° and ~70° to helices 3, 4 and 5 respectively; helix 3 is oriented at ~140° and ~90° to helices 4 and 5 respectively; and the angle between helices 4 and 5 is ~120°. The core, formed by the five helices is tightly packed and composed principally of hydrophobic (aliphatic and aromatic) residues (Figs 1, 3*b*). In addition to all the helical backbone hydrogen bonds, there is a single long range hydrogen bond between the backbone amide of Ala 42 and the backbone carbonyl of Lys 18 which, together with a salt bridge between Lys 41 and Glu 17, serves to position the loop connecting helix 1 to the helical turn in close proximity to the loop connecting helices 2 and 3. All four cysteine residues are located in the protein interior. Cys 67 is protected from solvent by the long side chain of Lys 33 which reduces its solvent accessible surface area to ~20% of that in an extended Gly-Cys-Gly tripeptide segment, and is packed against Leu 30 and Leu 63. The other three cysteines are completely buried in the hydrophobic core with accessible surface areas less than 5% of that in an extended Gly-Cys-Gly tripeptide: Cys 77 is packed against Ile 26 and Leu 63; Cys 80 is packed against His 7, Phe 10 and Ala 24; and finally Cys 85 is packed against Leu 46, Phe 49, Leu 50, Trp 84 and Leu 89. The sulfhydryl groups of the four cysteines are all involved in hydro-

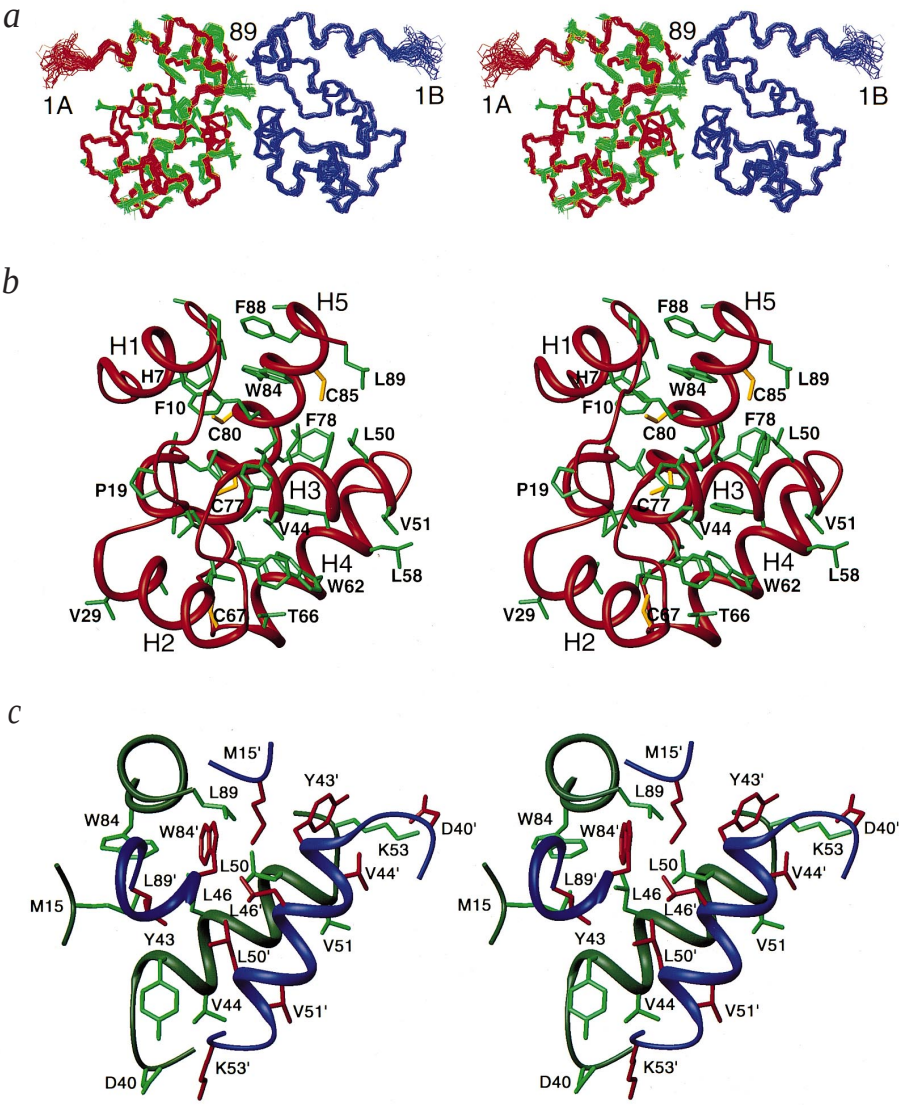

Fig. 3 Stereoviews of BAF. ***a***, Superposition of the final ensemble of 40 simulated annealing structures of the BAF dimer. The backbone of one subunit is shown in red and of the other blue; the side chains of only one subunit are shown in green. ***b***, The BAF monomer illustrating the packing of the hydrophobic core. The backbone is shown as a red tube, side chains are in green and the four reduced cysteines in yellow. ***c***, The dimer interface. The backbone (displayed as a tube) and side chains of one subunit are shown in dark and light green respectively, and of the other in blue and red respectively. Residue labels for the two subunits are distinguished by the absence (green subunit) or presence (blue subunit) of a prime after the residue number.

gen bonding interactions: the Sγ atoms of Cys 67 and Cys 80 accept hydrogen bonds from the backbone amide of Ala 69 and Ala 24 respectively; while the SγH atoms of Cys 77 and Cys 85 donate hydrogen bonds to the backbone carbonyls of Ala 24 and Leu 81 respectively. Thus, the interactions involving the sulfhydryls of Cys 77 and Cys 80 may play a role in stabilizing the interactions between the helical turn (residues 20–23) and helix 5. The involvement of the four cysteine side chains in stabilizing the protein core is consistent with mutational data. Replacing any of the four cysteine residues with an alanine reduces the melting temperature by 3–5 °C relative to that of the wild type (61.85 °C). Interestingly Cys 67, Cys 77 and Cys 85 are substituted by Ala in *Caenorhabditis elegans* BAF (Fig. 1). This is a conservative substitution since Ala is similar in size and hydrophobic character to Cys.

There are two other polar residues that are either completely buried (Ser 74) or partially buried (His 7) in the protein core. The hydroxyl group of Ser 74 is hydrogen bonded to the backbone carbonyl of Asn70. In *C. elegans* BAF, Ser 74 is replaced by Ala which is of similar size but more hydrophobic in character. His 7 is located at the interface of helices 1 and 5, and packed against the methyl groups of Val 11 and Ala 24. At pH 6.5, the imidazole group of His 7 is in the neutral Nε2-H tautomer (with ^{15}N chemical shifts for the Nε2 and Nδ1 atoms of 175 and 225 p.p.m. respectively) and the Hε2 proton donates a hydrogen bond to the carboxylate of Glu 83. Both His 7 and Glu 83 are conserved in human, mouse, zebrafish and *C. elegans* BAF.

The overall sequence identity between human and mouse BAF, human and zebrafish BAF, and human and *C. elegans* BAF is ~97%, ~85% and ~58% respectively (Fig. 1), and all the buried residues within the monomer are either completely conserved in the case of human, mouse and zebrafish BAF or substituted conservatively in the case of *C. elegans* BAF (for example, Leu→Phe or Tyr, Cys→Ala, Ser→Ala). Thus, one can safely conclude that the overall architecture of the BAF monomer is preserved across the four species.

Dimeric structure of BAF

BAF forms a stable dimer in solution, approximately cylindrical in shape, with dimensions of about 51 × 33 × 30 Å. The interface is formed by helix 3, the C-terminus of helix 5 and the loop connecting helix 1 to the helical turn (Figs 3*c*, 4*a*,*b*), and some of the intersubunit NOEs defining the interface are illustrated in Fig. 2. Approximately 540 Å^2 per subunit are buried upon dimerization. Helix 3 of one subunit is oriented antiparallel (at an angle of ~150°) to helix 3 of the other subunit, and the C_2 axis is located approximately between Gly 47 of one subunit and Gly 47' of the other. A summary of the residues that are buried upon dimerization is provided in Fig. 1, and a stereoview showing details of the dimer interface is shown in Fig. 3*c*. With the exception of a salt bridge between the side chains of Lys 53 and Asp 40', the intersubunit contacts are all hydrophobic. Specifically, the aliphatic portion of the Lys 53 side chain packs against Tyr 43' and Val 44'; Leu 50 interacts with Tyr 43', Leu 46', Gly 47' and Leu 50'; Val 51 interacts with Phe 39', Val 44', Gly 47'; Gly 47 interacts with Gly 47', Leu 50' and Val 51'; Lys 54 interacts with Tyr 43'; and Leu 89 interacts with Pro 14', Met 15', Leu 46' and Trp 84'.

With only three exceptions comprising conservative substitutions (Leu46→Phe and Leu89→Met in *C. elegans* BAF and

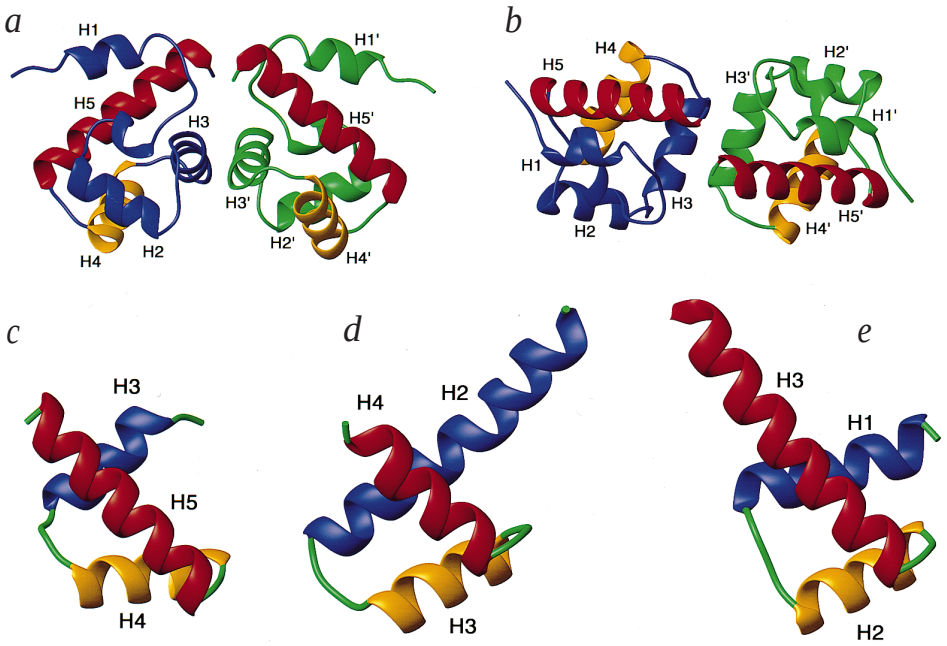

Fig. 4 ***a***,***b***, Ribbon diagrams showing two orthogonal views of the BAF dimer. ***c***–***e***, Comparison of the topology of helices 3–5 of BAF (*c*) with helices 2-4 of the Iβ subdomain of MuA transposase[13] (*d*) and helices 1–3 of the Drosophila paired homeodomain[14] (*e*). The first and second (DNA recognition) helices of the helix-turn-helix motif are shown in yellow and red respectively; other helices are shown in blue for one subunit and green for the other.

Lys53→Arg in zebrafish BAF), the residues at the dimer interface are conserved in human, mouse, zebrafish and *C. elegans*. Thus, one can conclude that the dimeric structure of BAF is also evolutionarily conserved.

Similarities to other proteins

Despite the absence of any sequence similarity, the topology of helices 3–5 is very similar to that found in a number of helix-turn-helix DNA binding proteins[12] (Fig. 4*c*–*e*). The closest structural resemblance is with helices 2– 4 of the Iβ subdomain of phage Mu transposase[13] (2.0 Å for 40 Cα atoms; Fig. 3*d*) and helices 1–3 of the Drosophila paired homeodomain[14] (2.1 Å for 42 Cα atoms; Fig. 3*e*). Helices 3 and 4 of the Iβ subdomain and helices 2 and 3 of the homeodomain constitute the helix-turn-helix motif (HTH). The HTH is characterized by an interhelical angle of ~120° and a turn ranging from 3–6 residues; DNA recognition is principally achieved by the second helix of the HTH. Thus, by analogy, helices 4 and 5 of BAF constitute a helix-turn helix motif. The actual turn between helices 4 and 5 of BAF is three residues in length, with a Gly at position 1 of the turn, typical of many HTH sequences, and has a conformation that is very similar to to that seen in the paired homeodomain.

A model for DNA binding by a BAF monomer

In the presence of DNA, BAF forms a large oligomeric nucleoprotein complex (R.Z. and R.C., unpublished data). Given this complexity, we have restricted ourselves to proposing a possible model for the interaction of a BAF monomer with DNA. This is illustrated in Fig. 5. The N-terminal end of helix 5 (the recognition helix of the HTH motif) readily fits into the major groove, permitting interactions with the DNA involving Asn 70, Gln 73, Lys 72 and Lys 75 (Fig. 5*a*). In this orientation, Gln 5 and Lys 6 of helix 1 are positioned along the phosphate backbone, and Arg 60 of helix 4 (the first helix of the HTH motif) can provide supplementary electrostatic interactions with the DNA (Fig. 5*a*). These seven residues are conserved in human, mouse and zebrafish BAF, and present a large positively charged patch on the surface of the protein (Fig. 5*b*). In the case of *C. elegans* BAF, Gln 5 is substituted by Val, Arg 60 by Ile, Asn 70 by Thr, Lys 72 by Asn, Gln 73 by His and Arg 75 by Lys. Thus, four out of the six substitutions are conservative in nature, although the extent of positive charge on the potential DNA binding surface will be reduced in *C. elegans* BAF relative to that of the other three species of BAF.

Correlation with biochemical data

To test the predicted location of the DNA binding surface, a series of single site directed mutants were made in which positively charged Lys or Arg residues were substituted by Glu. Mutation of Lys 6, Lys 33, Arg 60, Lys 64, Lys 72 and Arg 75, all of which map to the postulated DNA binding surface (Fig. 5*b*), abolished DNA binding. Mutation of Lys 54 also abolished DNA binding, possibly due to disruption of a salt bridge between Lys 54 and Asp 86 affecting the relative orientation of helices 4 and 5 which comprise the HTH motif. Mutation of Lys 53 also results in a significant reduction in DNA binding, presumably due to destabilization of the dimer as a result of disruption of the inter-subunit salt bridge between Lys 53 and Asp 40'. Other mutations that decrease DNA binding include Arg 8 and Lys 32, both of which are located at the edges of the postulated DNA binding surface. Mutation of Arg 37, Lys 41 and Arg 82, on the other hand, all of which are distant from the postulated DNA binding surface, have no effect on DNA binding.

Since the overall BAF–DNA complex comprises a large nucleoprotein assembly, it seems possible that BAF may associate with other proteins in the cell and be part of a multiprotein–DNA complex. In this regard, several oligomerization modules involved in DNA compaction have recently been elucidated structurally: in particular, the nucleosome core particle which comprises an octamer of H2A–H2B and H3–H4 dimers[15], and the Drosophila TAF42–TAF62 dimers for which a histone-like octamer TAF complex within the transcription factor TFIID has been suggested[16]. All the proteins involved are helical and make use of helices as dimerization elements to form homo- and heterodimers. The structural data and chemi-

cal mapping experiments on the histone core complex show that the DNA is wrapped around the multimeric protein complex, thereby allowing for maximal compaction of the DNA. Clearly this type of homo- and heterodimerization may be possible for BAF, as well as for structurally related proteins. If a related type of arrangement would occur with BAF and DNA, this would readily explain why the retroviral DNA is protected from autointegration.

Concluding remarks

Here we have presented the three-dimensional structure of BAF, a dimeric HTH DNA binding protein, with no significant sequence similarity to any previously identified protein of known function. BAF is a highly evolutionarily conserved cellular protein (Fig. 1) which is abundantly expressed in numerous tissues (M.S.L. and R.C., unpublished data). Although its physiological role in the cell remains to be elucidated, its presence is clearly vital to the cell since inhibitory RNA experiments in *C. elegans* have shown that the absence of BAF results in arrest of embryonic development (M.S.L., R.C. and M. Krause, unpublished data). Retroviruses have hijacked the properties of BAF to inhibit autointegration of viral DNA and thus facilitate integration of viral DNA into the host chromosome. Further studies on the oligomeric BAF–DNA complex are required to answer questions concerning the mechanism by which BAF exerts its role in retroviral integration, as well as to provide clues with respect to its cellular function.

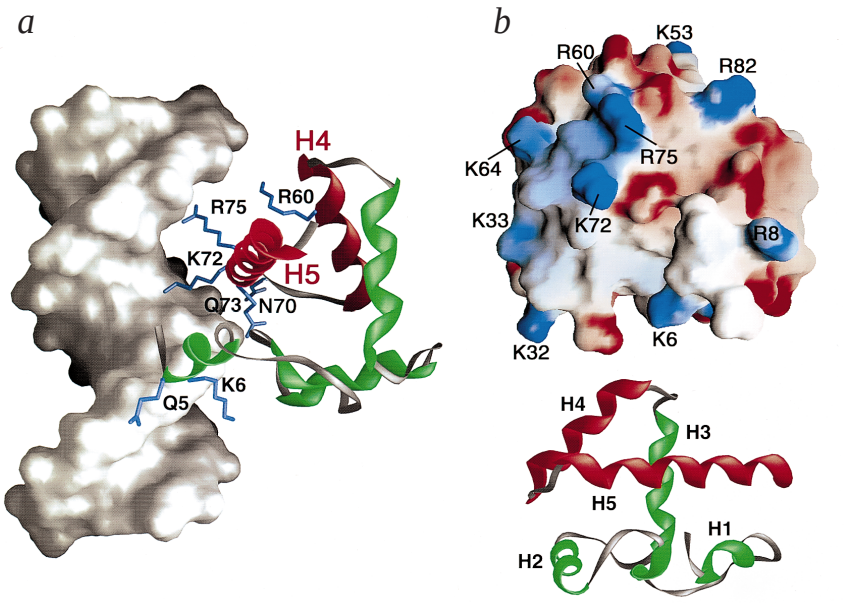

Fig. 5 Model for the binding of a BAF monomer to DNA. ***a***, BAF docked onto DNA. The DNA is shown as a molecular surface, the backbone of BAF as a ribbon diagram with the HTH motif in red, and potential side chains contacting the DNA in blue. The N-terminal end of helix 5 is located in the major groove of the DNA. ***b***, Electrostatic potential mapped onto a molecular surface of a BAF monomer (top) together with a ribbon diagram of BAF in the same orientation. The postulated DNA binding surface is positively charged.

Methods

Expression and purification. The human BAF coding sequence (GenBank database accession number AF070447) was amplified from a liver cDNA library (Clonetech) by PCR and cloned into the BamHI and NdeI sites of pET15b as described for the murine clone[6] and expressed in *E. coli* BL21(DE3) grown in minimal medium. Uniform ^{15}N and ^{13}C labeling was carried out using $^{15}NH_4Cl$ and $^{13}C_6$-glucose as the sole nitrogen and carbon sources, respectively, using the protocol described[17]. Cells were grown at 37 °C, induced with isopropyl β-D-thiogalactopyranoside, harvested 3 h after induction by centrifugation, and lysed by incubation with 0.01% w/v lysozyme in 'lysis' buffer (25 mM HEPES pH 7.5, 150 mM NaCl) for 20 min at 0 °C, followed by sonication. After centrifugation, BAF was found exclusively in the pellet fraction. BAF was solubilized in buffer (25 mM HEPES pH 7.5, 150 mM NaCl, 5 mM imidazole) containing 6 M guanidinium chloride, and the solubilized protein was applied to a Ni^{2+} affinity column and, after thorough washing with the solubilizing buffer, eluted with an imidazole gradient (20 mM–0.7 M). The fractions containing BAF were dialyzed overnight at 4 °C in buffer containing 100 mM potassium phosphate pH 6.5, 5 mM EDTA and 200 mM NaCl. The histidine tag was removed by cleavage with thrombin (10 units mg^{-1} of BAF for 2 h at room temperature), and the thrombin was subsequently removed by adsorption to a column of benzamidine sepharose. The eluted protein was incubated with 100 mM DTT for 2 h at 40 °C to ensure full reduction of any disulfide bonds, and further purified by reverse phase HPLC. The HPLC purified protein still contains a small proportion of a higher molecular weight species that migrates as a series of bands close to 60,000 M_r in SDS-PAGE. This species, which is inactive in DNA binding assays, was removed by gel-filtration on a Superdex 200 column. Purified BAF was characterized by electrospray ionization mass spectroscopy and had a molecular mass of 10,383.

Site directed mutagenesis. Positively charged residues in human BAF were individually changed to Glu, and Cys residues to Ala using the QuikChange Site-Directed Mutagenesis Kit (Stratagene). Mutant proteins were expressed in pET15b and purified by Ni-affinity chromatography using a Ni-NTA spin kit (Qiagen) under denaturing conditions. After removal of the His-tag by thrombin cleavage, the mutant proteins were further purified on a Superdex 75 gel filtration column. The DNA binding activity of the mutant BAF proteins was measured by a gel-shift assay (R.Z. and R.C., unpublished data).

Analytical ultracentrifugation. Sedimentation equilibrium experiments were conducted at 20 °C at three different rotor speeds 14,000, 18,000 and 22,000 r.p.m. on an Optima XL-A analytical ultracentrifuge. Equilibrium was achieved within 26 h. The data were analyzed in terms of a single ideal solute to obtain buoyant molecular mass, M(1-νρ) using the Optima XL-A data analysis software (Beckman).

Sample preparation. The purified protein was concentrated and dialyzed overnight under argon in buffer containing 100 mM potassium phosphate pH 6.5, 200 mM NaCl and 5 mM DTT. 50 mM deuterated DTT was then added, and the sample extensively degassed in an NMR tube by repeatedly switching between a weak vacuum and argon purging, prior to flame sealing the NMR tube. Four NMR samples were prepared: 1.5 mM (monomer concentration) ^{15}N labeled and $^{15}N/^{13}C$ labeled BAF samples in 95% H_2O/5% D_2O; and heterodimeric BAF (~3 mM monomer concentration) containing a 1:1 mixture of ^{13}C labeled and ^{15}N labeled protein in 95% H_2O/5% D_2O and 99.996% D_2O. For the mixed heterodimers, ^{13}C and ^{15}N labeled protein was mixed after gel filtration purification, denatured with 6 M guanidinium chloride, and subsequently refolded by dialysis against buffer containing 100 mM potassium phosphate pH 6.5, 200 mM NaCl and 5 mM DTT.

Table 1 Structural statistics[1]

	<SA>	$(\overline{SA})r$
R.m.s. deviations from experimental distance restraints (Å)[2]		
All (864)	0.022 ± 0.001	0.022
Intramolecular		
Interresidue sequential (|i - j| = 1) (288)	0.015 ± 0.002	0.013
Interresidue medium range (1 < |i - j| ≤ 5) (267)	0.026 ± 0.002	0.026
Interresidue long range (1 < |i - j| ≤ 5) (190)	0.023 ± 0.003	0.019
Intraresidue (7)	0.000 ± 0.000	0.000
H-bonds (64)	0.017 ± 0.003	0.012
Intermolecular (48)	0.036 ± 0.012	0.047
R.m.s. deviations from exptl dihedral restraints (°) (257)[2]	0.161 ± 0.066	0.128
R.m.s. deviations from $^3J_{HN\alpha}$ coupling constants (Hz) (66)[2]	0.97 ± 0.02	0.94
R.m.s. deviations from secondary ^{13}C shifts (p.p.m.)		
$^{13}C\alpha$ (87)	1.06 ± 0.02	1.05
$^{13}C\beta$ (78)	1.19 ± 0.02	1.20
R.m.s. deviations from methyl ^1H shifts (p.p.m.) (44)	0.12 ± 0.01	0.12
R.m.s. deviations from residual dipolar couplings (Hz)[3]		
$^1D_{NH}$ (Hz) (76)	0.42 ± 0.01	0.42
$^1D_{C\alpha C'}$ (Hz) (76)	1.34 ± 0.01	1.32
$^1D_{NC'}$ (Hz) (55)	0.48 ± 0.001	0.48
$^2D_{HNC'}$ (Hz) (52)	1.22 ± 0.03	1.20
Deviations from idealized covalent geometry		
bonds (Å) (1,419)	0.003 ± 0.000	0.005
angles (°) (2,566)	0.432 ± 0.012	0.543
impropers (°) (742)	0.489 ± 0.034	0.552
Measures of structure quality		
E_{LJ} (kcal mol^{-1})[4]	-883 ± 13	-861
PROCHECK[5]		
Residues in most favorable region of Ramachandran plot	91.4 ± 1.5	89.5
No. of bad contacts per 100 residues	3.5 ± 1.3	2.8
Coordinate precision (Å)[6]		
backbone (N, Cα, C', O)	0.34 ± 0.08	
All non-hydrogen atoms	0.71 ± 0.06	

[1]The notation of the NMR structures is as follows: <SA> are the final 40 simulated annealing structures; $\overline{SA}$ is the mean structure obtained by averaging the coordinates of the individual SA structures best-fitted to each other (residues 3–89 of both subunits); $(\overline{SA})r$ is the restrained regularized mean structure obatined by restrained regularization of the mean structure $\overline{SA}$. The number of terms per monomer for the various restraints is given in parentheses.
[2]None of the structures exhibited interproton distance violations greater than 0.5 Å, dihedral angle violations greater than 3°, or $^3J_{HN\alpha}$ coupling constant violations greater than 2 Hz. The torsion angle restraints consist of 84 ϕ, 78 ψ, 60 χ_1, 29 χ_2 and 6 χ_3 angles. Protein backbone hydrogen bonding restraints (two per hydrogen bond, $r_{NH\text{-}O}$ = 1.5–2.8 Å, $r_{N\text{-}O}$ = 2.4–3.5 Å) were introduced during the final stages of refinement according to standard criteria.
[3]The expected r.m.s. deviations for a set of random vectors, given by $\{2D_a^2[4 + 3R^2]/5\}^{1/2}$, based on a value of -14.9 Hz for D_a^{NH} and a rhombicity R of 0.17, are 19.05, 3.56, 2.11 and 6.27 Hz for the $^1D_{NH}$, $^1D_{C\alpha C'}$, $^1D_{NC'}$ and $^1D_{HNC'}$ dipolar couplings respectively.
[4]The Lennard-Jones van der Waals energy was calculated with the CHARMM PARAM19/20 parameters[36] and is not included in the target function for simulated annealing or restrained minimization.
[5]The overall quality of the structure was assessed using the program PROCHECK[37]. There were no ϕ/ψ angles in the disallowed region of the Ramachandran plot. The dihedral angle G-factors for ϕ/ψ, χ_1/χ_2, χ_1, χ_3/χ_4 are 0.39 ± 0.04, 0.75 ± 0.06, 0.16 ± 0.13 and 0.19 ± 0.12 respectively.
[6]Defined as average r.m.s.difference (residues 3–89 of both subunits) between the final 40 simulated annealing structures and the mean coordinates.

NMR spectroscopy. NMR experiments were carried out at 40 °C on Bruker DMX500 and DMX600 spectrometers equipped with x,y,z-shielded gradient triple resonance probes. Spectra were processed with the NMRPipe package[18] and analyzed using the programs PIPP and STAPP[19]. ^1H, ^{15}N and ^{13}C resonance assignments were obtained using an array of 3D double and triple resonance heteronuclear correlation experiments[8–10]. $^3J_{HN\alpha}$, $^3J_{\alpha\beta}$, $^3J_{C'C\gamma}$ (aromatic, methyl and methylene), $^3J_{NC\gamma}$ (aromatic, methyl and methylene), $^3J_{C'C'}$ couplings were measured by quantitative J correlation spectroscopy[20]. The tautomeric state of the single histidine was determined by recording a long-range ^1H-^{15}N correlation spectrum[21]. ϕ and ψ backbone torsion angle restraints were derived from $^3J_{HN\alpha}$ and $^3J_{C'C'}$ coupling constants, and the backbone (^{15}N, NH, $^{13}C\alpha$, $^{13}C\beta$, $^{13}C'$ and Hα) secondary chemical shifts using the program TALOS (G. Cornilescu, F. Delaglio & A. Bax, in preparation) which is based on a database of residue triplets correlating ϕ,ψ angles (derived from high resolution X-ray structures) and their corresponding backbone secondary chemical shifts. Side chain torsion angle restraints were derived from NOE/ROE data and three-bond heteronuclear coupling constants. Intramolecular interproton distance restraints were derived from 3D ^{15}N-separated (120 ms mixing time), ^{13}C-separated (35 and 120 ms mixing time), $^{13}C/^{15}N$-separated (120 ms mixing time), $^{15}N/^{15}N$-separated (120 ms mixing time) and $^{13}C/^{13}C$-separated (120 ms mixing time) NOE experiments, and a 3D ^{15}N-separated ROE (35 ms mixing time) experiment. Intermolecular interproton distance restraints were obtained from a 3D ^{13}C-separated/^{12}C-filtered NOE experiment (150 ms mixing time) recorded in D_2O, and from 2D ^{15}N-separated/^{13}C-filtered and ^{13}C-separated/^{15}N-filtered NOE experiments (150 ms mixing time) recorded in H_2O on a heterodimer comprising a 1:1 mixture of $^{13}C/^{14}N$ and $^{12}C/^{15}N$ labeled protein. $^1D_{NH}$, $^1D_{C\alpha C'}$, $^1D_{NC'}$ and $^2D_{HNC'}$ residual dipolar couplings were obtained by taking the difference in the corresponding J splittings measured in oriented (in 6% 3.5:1 DMPC:DHPC at 35°C)[22] and isotropic (in water) BAF. $^1J_{NH}$ and $^1J_{C\alpha C'}$ couplings were obtained from 2D IPAP

{^{15}N,^{1}H}-HSQC[23] and 3D ^{13}Cα-coupled(F_1) HNCO experiments respectively. $^{1}J_{NC'}$ and $^{2}J_{HNC'}$ couplings were obtained from a 2D ^{13}C'-coupled/^{13}Cα-decoupled(F_1) ^{1}H-^{15}N HSQC experiment. The precision of the measured $^{1}D_{NH}$, $^{1}D_{C\alpha C'}$, $^{1}D_{NC'}$ and $^{2}D_{HNC'}$ dipolar couplings was ~0.5-1.0 Hz, ~1.0-1.5 Hz, ~0.5-1.0 Hz and ~1.0-1.5 Hz respectively. The normalization factors (given by $\gamma_N\gamma_H\langle r_{NH}^{-3}\rangle/\gamma_A\gamma_B\langle r_{AB}^{-3}\rangle$ where γ and r represent gyromagnetic ratios and distances respectively) employed for $^{1}D_{C\alpha C'}$, $^{1}D_{NC'}$ and $^{2}D_{HNC'}$ relative to $^{1}D_{NH}$ were 5.36, 9.04 and 3.04 respectively[24]. The magnitude of the axial and rhombic components of the alignment tensor D^{NH} were obtained by examining the distribution of normalized dipolar couplings[25] which yielded values of D_a^{NH} = -14.1 Hz and R = 0.17, where D_a^{NH} is the axial component of the tensor and R is the rhombicity defined as the ratio of the rhombic to axial components of the tensor. Heteronuclear ^{15}N-{^{1}H} NOEs were measured as described[26] and identified only residues 2 and 3 with a ^{15}N-{^{1}H} NOE less than 0.6. ^{15}N T_1 and T_2 values were measured as described[27].

Structure calculations. Approximate interproton distance restraints, derived from multidimensional NOE spectra, were grouped into four distance ranges, 1.8–2.7 Å (1.8–2.9 Å for NOEs involving NH protons), 1.8–3.3 Å (1.8–3.5 Å for NOEs involving NH protons), 1.8–5.0 and 1.8–6.0 Å, corresponding to strong, medium, weak and very weak NOEs respectively[7]. 0.5 Å was added to the upper bound for distances involving methyl groups to account for the higher apparent intensity of the methyl resonances. Distances involving non-stereospecifically assigned methylene protons, methyl groups, and Hδ and Hε protons of Tyr and Phe, were represented as a $(\Sigma r^{-6})^{-1/6}$ sum[28]. The structures were calculated by simulated annealing[29] using the program CNS[30]. The final values for the force constants employed for the various terms in the target function employed for simulated annealing are as follows[31,32]: 1,000 kcal mol^{-1} Å^{-2} for bond lengths, 500 kcal mol^{-1} rad^{-2} for angles and improper torsions (which serve to maintain planarity and chirality), 100 kcal mol^{-1} Å^{-2} for the non-crystallographic symmetry restraint, 4 kcal mol^{-1} Å^{-4} for the quartic van der Waals repulsion term (with the van der Waals radii set to 0.8 times their value used in the CHARMM PARAM19/20 parameters[33]), 30 kcal mol^{-1} Å^{-2} for the experimental distance restraints (interproton distances and hydrogen bonds), 200 kcal mol^{-1} rad^{-2} for the torsion angle restraints, 1 kcal mol^{-1} Hz^{-2} for the $^{3}J_{HN\alpha}$ coupling constant restraints, 0.5 kcal mol^{-1} p.p.m.$^{-2}$ for the secondary ^{13}C chemical shift restraints, 7.5 kcal mol^{-1} p.p.m.$^{-2}$ for the ^{1}H chemical shift restraints, 1.0 kcal mol^{-1} Hz^{-2} for the $^{1}D_{NH}$ dipolar coupling restraints, 0.035, 0.050 and 0.108 kcal mol^{-1} Hz^{-2} for the normalized (relative to $^{1}D_{NH}$), $^{1}D_{C\alpha C}$:NH), $^{1}D_{NC'}$:(NH) and $^{1}D_{HNC'}$:(NH) dipolar coupling restraints respectively[24], and 1.0 for the conformational database potential.

Structure figures were generated using the programs MOLMOL[34] and GRASP[35]. Electrostatic calculations were performed with GRASP[35]. Searching of the Brookhaven protein structure database for structural similarity was carried out using the program DALI[36].

Coordinates. The coordinates of the ensemble of 40 simulated annealing structures, the restrained regularized mean structure, and the complete list of experimental NMR restraints and ^{1}H, ^{15}N and ^{13}C assignments have been deposited in the Brookhaven Protein Data Bank (accession codes 2EZX, 2EZY, 2EZZ and 2EZXMR).

Acknowledgments

We thank D.S. Garrett and F. Delaglio for software support; R. Tschudin for hardware supprot; L. Pannel for mass spectrometry; M. Krause for deriving the *C. elegans* BAF sequence from genomic and EST sequences in the databases; and C. Bewley, D. Garrett, K. Frank, M. Ottiger and N. Tjandra for useful discussions. This work was supported by the AIDS Targeted Antiviral Program of the Office of the Director of the National Institutes of Health (G.M.C., A.M.G. and R.C.)

Received 7 July, 1998; accepted 25 August, 1998.

1. Katz, R.A. & Skalka, A.M. The retroviral enzymes. *Annu. Rev. Biochem.* **63**, 133–173 (1994).
2. Brown, P.O., Bowerman, B., Varmus, H.E., and Bishop, J.M. Correct integration of retroviral DNA *in vitro*. *Cell* **49**, 347–356 (1987).
3. Ellison, V., Abrams, H., Roe, T., Lifson, J., and Brown, P. Human immunodeficiency virus integration in a cell-free system. *J. Virol.* **64**, 2711–2715 (1990).
4. Farnet, C.M. and Haseltine, W.A. Integration of human immunodeficiency virus type 1 DNA in vitro. *Proc. Natl. Acad. Sci. USA* **87**, 4164–4168 (1990).
5. Lee, M.S. & Craigie, R. Protection of retroviral DNA from autointegration: involvement of a cellular factor. Proc. *Natl. Acad. Sci. USA* **91**, 9823–9827 (1994).
6. Lee, M.S. & Craigie, R. A previously unidentified host protein protects retroviral DNA from autointegration. *Proc. Natl. Acad. Sci. USA* **95**, 1528–1533 (1998).
7. Clore, G.M., Gronenborn, A.M., Szabo, A. & Tjandra, N. Determining the magnitude of the fully asymmetric diffusion tensor from heteronuclear relaxation data in the absence of structural information. *J. Am. Chem. Soc.* **120**, 4889–4890 (1998).
8. Clore, G.M. & Gronenborn, A.M. Structures of larger proteins in solution: three- and four-dimensional heteronuclear NMR spectroscopy. *Science* **252**, 1390–1399 (1991).
9. Clore, G. M. & Gronenborn, A. M. Determining the structures of larger proteins and protein complexes by NMR. *Trends Biotech.* **16**, 22–34 (1998).
10. Bax, A. & Grzesiek, S. Methdological advances in protein NMR *Acc. Chem. Res.* **26**, 131–138 (1993).
11. Tjandra, N., Omichinski, J. G., Gronenborn, A. M., Clore, G. M. & Bax, A. Use of dipolar ^{1}H-^{15}N and ^{1}H-^{13}C couplings in the structure determination of magnetically oriented macromolecules in solution. *Nature Struct. Biol.* **4**, 732–728 (1997).
12. Pabo, C.O. & Sauer, R.T. Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61**, 1053–1095 (1992).
13. Schumacher, S. *et al.* Solution structure of the Mu end DNA-binding Iβ subdomain of phage Mu transposase: modular DNA recognition by two tethered domains. *EMBO J.* **16**, 7532–7541 (1997).
14. Wilson, D.S., Guenther, B., Desplan, C. & Kuriyan, J. High-resolution crystal structure of a paired (PAX) class cooperative homeodomain dimer on DNA. *Cell* **82**, 709–719 (1995).
15. Luger, K., Maeder, A.W., Richmond, R.K., Sargent, D.F. & Richmond, T.J. X-ray structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–259 (1997).
16. Xie, X. *et al.* Structural similarity between TAFS and the hetrotetrameric core of the histone octamer. *Nature* **380**, 316–322 (1996).
17. Cai, M. *et al.* An efficient and cost-effective isotope labeling protocol for proteins expressed in *Escherichia coli*. *J. Biomol. NMR* **11**, 97–102 (1998).
18. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
19. Garrett, D. S., Powers, R., Gronenborn, A. M. & Clore, G. M. A common sense approach to peak picking in two-, three- and four-dimensional spectra using automatic computer analysis of contour diagrams. *J. Magn. Reson.* **95**, 214–220 (1991).
20. Bax, A. *et al.* Measurement of homo- and hetero-nuclear J couplings from quantitative J correlation. *Meth. Enz.* **239**, 79–106 (1994).
21. Pelton, J.G., Torchia, D.A., Meadow, M.D. & Roseman, S. Tautomeric states of the active-site histidines of phosphorylated and unphosphorylated III^{Glc}, a signal transducing protein from *Escherichia coli*, using two-dimensional heteronuclear NMR techniques. *Protein Sci.* **2**, 543–558 (1993).
22. Tjandra, N. & Bax, A. Direct measurement of distances and angles in biomolecules by NMR in dilute liquid crystalline medium. *Science* **278**, 1111–1114 (1997).
23. Ottiger, M., Delaglio, F. & Bax, A. Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. *J. Magn. Reson.* **131**, 173–178 (1998).
24. Bewley, C.A. *et al.* Solution structure of cyanovirin-N, a potent HIV-inactivating protein. *Nature Struct. Biol.* **5**, 571–578 (1998).
25. Clore, G. M., Gronenborn, A. M. & Bax, A. A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information. *J. Magn. Reson.* **133**, 216–221 (1998).
26. Grzesiek, S. & Bax, A. The importance of not saturating H_2O in protein NMR: application to sensitivity enhancement and NOE measurements. *J. Am. Chem. Soc.* **115**, 12593–12594 (1993).
27. Tjandra, N., Wingfield, P.T., Stahl, S.J. & Bax, A. Anisotropic rotational diffusion of perdeuterated HIV protease from ^{15}N NMR relaxation measurements at two magnetic fields. *J. Biomol. NMR.* **8**, 273–284 (1996).
28. Nilges, M. A calculational strategy for the structure determination of symmetric dimers by ^{1}H-NMR. *Proteins Struct. Funct. Genet.* **17**, 297–309 (1993).
29. Nilges, M., Gronenborn, AM., Brünger, A. T. & Clore G. M. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints: application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Prot. Engng.* **2**, 27–38 (1988).
30. Brünger, A.T. *et al.* Crystallography and NMR system (CNS): a new software suite for macromolecular structure determination. *Acta Crystallogr.* D **in the press** (1998).
31. Clore, G.M. & Gronenborn, A.M. New methods of structure refinement for macromolecular structure determination by NMR. *Proc. Natl. Acad. Sci. USA* **95**, 5891–5898 (1998).
32. Clore, G. M., Gronenborn, A. M. & Tjandra, N. Direct structure refinement against residual dipolar couplings in the presence of rhombicity of unknown magnitude. *J. Magn. Reson.* **131**, 159–162 (1998).
33. Brooks, B.R. *et al.* CHARMM: a program for macromolecular energy minimization and dynamics calculations. *J. Comput. Chem.* **4**, 187–217 (1993).
34. Koradi, R., Billeter, M. & Wuthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14** 51-5, 29–32 (1996).
35. Nichols, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11**, 281–296 (1991).
36. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).
37. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).