Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Description of protein–DNA complexes in terms of electron-density topological features

Andy Becue, Nathalie Meurice, Laurence Leherte and Daniel P. Vercauteren*

Laboratoire de Physico-Chimie Informatique, Facultés Universitaires Notre-Dame de la Paix, Rue de Bruxelles 61, B-5000 Namur, Belgium

Correspondence e-mail: daniel.vercauteren@fundp.ac.be

Received 18 February 2003
Accepted 1 September 2003

This paper presents a computer-modelling approach for the generation of low-resolution representations of two protein–DNA complexes, NF-κB and HIF-1. The representation is obtained by topological analysis of electron-density maps at 2.85 Å crystallographic resolution, which leads to a set containing a limited number of critical points (CP). Analyses of the structure and properties of the CP graphs (density at peak position, volume and ellipticity), as well as comparisons with other reduced representations, are performed in order to assess the usefulness of such representations in molecular-docking applications at medium resolution.

1. Introduction

Proteins play a key role in all living beings, as they are implicated in nearly all biological phenomena and are responsible for a huge number of very different functions. Among the vast protein domain, DNA-recognition proteins are of particular interest because they participate efficiently in cellular differentiation and gene regulation. What makes these processes so efficient is that the biological functions of a protein are determined by its chemical composition, the amino-acid sequence, but also by its three-dimensional spatial structure. A given amino-acid sequence leads to a unique three-dimensional structure, *i.e.* the native conformation of the protein, which contains cavities, protuberances, regions of specific physico-chemical properties *etc.* that are implicated in the recognition process. Hence, studies of the way protein–ligand and protein–DNA partners bind to each other logically require knowledge of the three-dimensional structure and interaction parameters of each partner implicated in the molecular-complementarity processes.

DNA is now considered as a possible target in biotechnology for the development of new and efficient therapeutic agents, for example antitumour compounds (Das & Jayaram, 1998; Yang *et al.*, 1998). Hence, most current studies concentrate on the way ligands or proteins can selectively bind to the major or minor groove of DNA at particular locations with respect to the base-pair sequence in order to regulate DNA transcription, the ultimate goal of this work being the development of new drugs that bind to the DNA target in a specific manner.

In spite of the rapid increase in computer performance, theoretical prediction of the recognition mechanism of nucleotide sequences by proteins (or ligands) still constitutes a very complex problem because of the highly complicated potential energy surface that must be explored, with many local energy minima. Search strategies for the global minimum of a native protein-folding conformation or of the structure of

a host/guest macromolecular complex must therefore be very efficient in order to retrieve this minimum among the large number of possible solutions. Moreover, the study of large molecular assemblies using atomistic descriptions and force fields is often hampered by the huge number of atoms to consider and the large computational resources needed to adequately sample the relevant degrees of freedom. Obviously, quantum-mechanical calculations should be the most reliable way to describe precisely all interactions between all atoms of the molecular or supramolecular systems, but such calculations are only applicable to small- or medium-size molecules (a few hundred atoms at most) and become totally unrealistic for systems as large as protein–DNA complexes. An alternative to atomistic descriptions could therefore consist of using simplified or reduced representations of the molecular partners. Reduced models have indeed already been proposed, for example in three-dimensional structural similarity studies of biopharmacological ligands to allow fast superimposition of multiple molecules, *i.e.* more than pairwise superimpositions (Meurice *et al.*, 1998), in molecular-complementarity studies to allow fast exploration of the intermolecular three-dimensional space between small-size partners (Altuvia *et al.*, 1995; Leherte *et al.*, 1995) and, even more logically, for huge systems with a large number of degrees of freedom as in protein and/or DNA complexes (Leherte & Allen, 1994).

With respect to protein folding, Levitt and coworkers (Levitt & Warshel, 1975; Levitt, 1976) were the first to show that reduced representations, *i.e.* C^{α} for the backbone and a virtual atom at the side-chain centroid, could capture the characteristics of the overall folds. Subsequently, Wallqvist & Ullner (1994) and Hassinen & Peräkylä (2001) proposed protein models in which the backbone is also represented by the C^{α} atom only and the side chains by one or more points according to their structure and properties, *e.g.* their hydrophobicity. Liwo *et al.* (1997) used a two-site representation for the backbone, *i.e.* one point located at the C^{α} position and the other at the peptidic bond (between two C^{α}), and a single-site representation for each side chain. Maggiora *et al.* (2001), who worked on the three-dimensional structural similarity of proteins, proposed an original representation based on adaptable spherical Gaussian functions located on the individual atoms of the proteins. This representation consisted of a very interesting flexible description of the underlying fold geometry of the protein components that can be adjusted by changing the 'width' of each Gaussian. More 'atom-like' descriptions were obtained by reducing the width, while fuzzier representations were generated by increasing the width. Other protein models were particularly specific to docking studies. They were generally based on a rigid-body approximation and on geometric criteria, assuming that the docked proteins were structurally similar to the undocked partners. For example, Ausiello *et al.* (1997) used a representation based on discrete slices of each partner coupled to three modules evaluating steric bumps, electrostatic clashes and shape complementarity. Palma *et al.* (2000) described each protein as a three-dimensional matrix composed of small cubic cells of 1 Å size, with the possible introduction of eventual softness into the algorithm. Gardiner *et al.* (2001) based their strategy on the representation of proteins in terms of dot surfaces calculated using the Connolly method, which extracts the critical points of the Connolly surface, *i.e.* knobs and holes. These points were expected to match in the case of a successful docking, *i.e.* the knobs of one partner superposed on the holes of the other.

For nucleic acids, fewer reduced representations are available in the literature. Tan & Harvey (1989) proposed a pseudo-atomic model of DNA in which each base pair is described by three points: the first point represents the intersection of the helix axis in the base-pair plane, the second the position of the sugar-phosphate backbone and the last point lies in the major groove. These three points describe a plane which corresponds to a base-pair plane, similar to that obtained by an atomic representation. Von Kitzing & Schmitt (1995) simply chose to divide the nucleic acids, DNA or RNA, into the three natural substructures: phosphate, ribose and bases (A, T, G, C, U).

Many of the above macromolecular representations were thus based on similar principles, *i.e.* virtual points determined from the geometry of the subunits only, *e.g.* amino acid, nucleotide *etc.* However, better descriptions than representations based only on the geometry can be proposed. For example, it is well known in drug design that various ligands of different three-dimensional structures may present similar affinity and/or activity for a common receptor; this means that they must present structural and physico-chemical similarities at a lower level of detail and, as a consequence, that it is not necessary to work at atomic resolution. In this way, the electron density $\rho(\mathbf{r})$ of molecules constitutes an important property to consider because it is directly related to their physico-chemical behaviour. The $\rho(\mathbf{r})$ distribution reflects the anisotropy of the chemical functions or subunits of a molecule. Considering this, we propose a method based on the use of the $\rho(\mathbf{r})$ distributions of the molecules calculated at medium crystallographic resolution and their reduction in terms of critical points (CPs). The aim is to obtain reduced representations of macromolecules, *i.e.* protein and DNA, that could be used in docking studies and which present advantages compared with the existing models in terms of efficiency and rapidity. Additionally, in order to adequately represent the interactions implicated in the recognition phenomena, the reduced representations of the molecular partners could be completed by other physico-chemical properties, *e.g.* steric hindrance, hydrophobicity, charge and/or hydrogen-bonding capabilities.

Our method has been applied to two test protein–DNA complexes, NF-κB and HIF-1, which were chosen for their scientific interest and their structural properties, *i.e.* they were medium-size complexes and, in the case of NF-κB, cover both the dimerization and DNA-recognition loops. A very brief biological and structural description of the two systems is presented in the next section. It is followed by a presentation of our methodology, *i.e.* the calculation of the electron-density maps and the generation of the critical points. Finally, results

are discussed in order to validate our strategy compared with other reduced-representation methods that can be found in the literature.

2. Materials

NF-κB (nuclear factor κB) and HIF-1 (hypoxia-induced factor 1), two transcription factors, are both currently and widely studied for their implication in various diseases involved in inflammatory processes or in cell reactions towards aggression.

The butterfly-shaped structure of NF-κB (Fig. 1) was determined by two different research groups (Ghosh *et al.*, 1995; Müller *et al.*, 1995). In its active form, NF-κB is a dimer that is mostly composed of two DNA-binding subunits P50 and RelA (P65). The coordinates of both DNA-binding subunits P50 and P65, composed of 312 and 273 amino acids, respectively, and of the 5′-TGGGGACTTTCC-3′ DNA oligomer bound to NF-κB are stored in PDB file 1vkx (Chen *et al.*, 1998).

HIF-1 is a heterodimeric complex composed of two protein subunits, HIF-1α and ARNT (aryl hydrocarbon receptor nuclear translocator), which consist of 826 and 789 amino acids, respectively. They both possess two successive domains: bHLH (basic helix–loop–helix) and PAS (PER–ARNT–SIM) at their amino-terminal extremity (Semenza *et al.*, 1997). The coordinates of the bHLH motifs of both DNA-binding proteins ARNT and HIF-1α, as well as the 5′-GCCC-TACGTGCTGC-3′ DNA oligomer bound to HIF-1, are stored in PDB file 1d7g (Michel *et al.*, 2000); their three-dimensional representation is presented in Fig. 2. The amino-terminal halves of HIF-1α and ARNT are sufficient for dimerization and binding to DNA (Semenza, 2001). The domains considered here are composed of 59 amino acids for HIF-1α instead of 826 and 59 amino acids for ARNT instead of 789. It is important to notice that the terms 'HIF-1α' and 'ARNT' used further in this article actually refer to 'the bHLH domain of HIF-1' and 'the bHLH domain of ARNT', respectively.

3. Methodology

As noted in §1, our method is based on the calculation of the electron-density (ED) functions for each partner of the macromolecular systems in question and their simplification in terms of graphs of critical points (CPs) by topological analysis.

The ED functions were generated by simulating an X-ray diffraction experiment at a specific crystallographic diffraction resolution R. As we know the atomic positions, the nature of each atom of the molecular partners and the crystal cell dimensions, it is possible to calculate an adequate set of structure factors $F(\mathbf{h})$ using the program *XTAL* (Hall *et al.*, 2000). The corresponding three-dimensional electron-density map (EDM) is then obtained by Fourier transform of $F(\mathbf{h})$.

As we wish to develop an original docking strategy, the EDMs of each partner of the complexes formed by NF-κB and HIF-1 have been generated separately. The atomic structure of each partner has been centred in a $P1$ symmetry cell with an extra 10 Å on each side to avoid border effects. Grid spacings

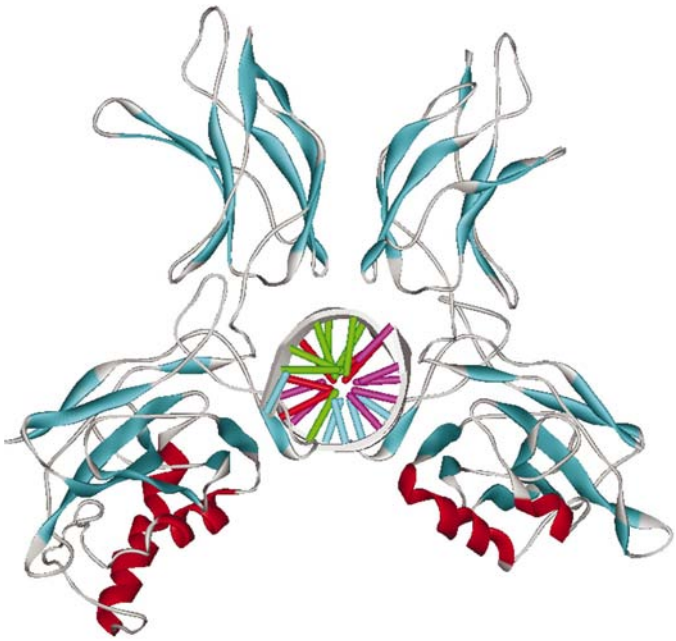

Figure 1
Ribbon representation (top view) of the NF-κB P50–P65 heterodimer, bound to the DNA oligomer. P50 is on the left, P65 on the right and the DNA oligomer in the middle. Visualization was performed using *WebLab Viewer* (http://www.accelrys.com/viewer/index.html).

Figure 2
Ribbon representation (top view) of the HIF-1 ARNT–HIF-1α heterodimer bound to the DNA oligomer. ARNT is at the top left, HIF-1α at the top right and the DNA oligomer at the bottom. Visualization by *WebLab Viewer*.

of 0.5 and 0.8 Å were selected for the EDM calculations of HIF-1 and NF-κB, respectively. One important step was the choice of an adequate diffraction resolution value. At atomic resolution, *i.e.* 1.0 Å, the ED function $\rho(\mathbf{r})$ is concentrated on the atomic positions and topological analysis would lead to a classic 'atom-like' representation; no data reduction would be observed. On the other hand, too low a resolution, *e.g.* 5.0 Å, would lead to excessive reduction: the ED function would spread all over the molecule as a quasi-spherical shape and important structural information about the chemical functions could be lost. At about 3.0 Å, $\rho(\mathbf{r})$ is indicative of the chemical groups and functions (Leherte & Allen, 1994; Leherte *et al.*, 1996) and constitutes a good resolution value. We thus considered different resolution values around 3.0 Å to find a crystallographic resolution value which best catches the chemical features of the macromolecular partners.

The topological analysis of the ED in terms of gradients and Laplacians of $\rho(\mathbf{r})$ was developed by Bader in his 'Atoms in Molecules' (AIM) theory (Bader, 1995). Such an analysis of the $\rho(\mathbf{r})$ distribution allows its CPs to be obtained. By definition, CPs are the points where the gradient $\nabla\rho(\mathbf{r})$ vanishes along each of the three main spatial directions. The Hessian matrix ($\mathbf{H}$) of the continuous three-dimensional $\rho(\mathbf{r})$ function,

$$\mathbf{H}(\mathbf{r}) = \begin{pmatrix} \partial^2\rho/\partial x^2 & \partial^2\rho/\partial x\partial y & \partial^2\rho/\partial x\partial z \\ \partial^2\rho/\partial y\partial x & \partial^2\rho/\partial y^2 & \partial^2\rho/\partial y\partial z \\ \partial^2\rho/\partial z\partial x & \partial^2\rho/\partial z\partial y & \partial^2\rho/\partial z^2 \end{pmatrix} \quad (1)$$

is a real and symmetric matrix that can be diagonalized by finding a rotation of the original coordinate system which aligns the new coordinate axes with the three principal axes of the CP,

$$\mathbf{H}'(\mathbf{r}) = \begin{pmatrix} \partial^2\rho/\partial x'^2 & 0 & 0 \\ 0 & \partial^2\rho/\partial y'^2 & 0 \\ 0 & 0 & \partial^2\rho/\partial z'^2 \end{pmatrix}. \quad (2)$$

The three diagonal elements of $\mathbf{H}'(\mathbf{r})$, or eigenvalues of $\mathbf{H}$, correspond to the local curvature of $\rho(\mathbf{r})$ with respect to a principal axis.

The Laplacian $\nabla^2\rho(\mathbf{r})$ of the electron density, obtained as the trace of $\mathbf{H}'$, has a clear physical meaning as its sign indicates whether the ED is locally depleted [$\nabla^2\rho(\mathbf{r}) > 0$] or concentrated [$\nabla^2\rho(\mathbf{r}) < 0$]. The ellipticity of the charge density provides a measure of the extent to which the charge is preferentially accumulated in a given plane. It can be calculated as

$$\text{ellipticity} = \ln(|\text{ev}_{\max}|/|\text{ev}_{\min}|), \quad (3)$$

where $\text{ev}_{\min}$ and $\text{ev}_{\max}$ are the smallest and highest eigenvalues of $\mathbf{H}(\mathbf{r})$, respectively. Finally, CPs can be identified and differentiated knowing the rank of $\mathbf{H}'$, which is the number of non-zero eigenvalues, and the signature S, *i.e.* the algebraic sum of their signs. When the rank of $\mathbf{H}'$ is 3 (three non-zero eigenvalues), four cases are possible: (i) $S = -3$, which corresponds to a local maximum or peak (PK); $\rho(\mathbf{r})$ then adopts a maximum value along the three principal directions x', y', z'; (ii) $S = -1$, which corresponds to a saddle point or pass (PS); (iii) $S = +1$, which corresponds to a second kind of saddle point or pale (PL); (iv) $S = +3$ for a local minimum or pit (PT). Bader, who worked at atomic resolution, deduced that the PKs and PSs are in general associated with atoms and chemical bonds, respectively (Bader, 1995). PLs are located inside rings and PTs can be found in cages.

Several years ago, Johnson (1977) developed a topological analysis approach dedicated to the representation of the EDMs of protein structures in terms of their CPs and their linkage. This approach was implemented in the computer program *ORCRIT* (Johnson, 1977) as part of the Crysalis Project (Terry, 1983). The purpose of the project was to build an expert system for the automated interpretation of experimental protein EDMs in the 2.0–2.5 Å resolution range. In the same spirit, in 1993 Fortier and coworkers initiated the Molecular Scene Analysis project (Fortier *et al.*, 1993) aimed at reconstructing and interpreting crystal and molecular structures in terms of structural motifs using the same topological approach for the segmentation of medium-resolution (3.0 Å) maps of proteins (Leherte *et al.*, 1994).

The first major problem to be solved in our work was to determine how local information associated with each CP can lead to a three-dimensional shape description containing sufficient information to represent the actual spatial distribution of the ED. As we wish to refer to groups of atoms, our attention focused on the PKs only. We note that at the CP locations the main three curvatures of the $\rho(\mathbf{r})$ function, *i.e.* the eigenvalues of the Hessian matrix constructed from the second derivatives, constitute local information that can be transferred to the space surrounding the CP of interest. Hence, it is possible to evaluate (or reconstruct) a three-dimensional function in the close neighbourhood of each CP. To a first degree of approximation, each PK can thus be considered as the centre of expansion of a Gaussian function,

$$\rho(\mathbf{r}) = \rho(0)\exp[\alpha\mathbf{r}^{\mathbf{T}}\mathbf{H}\mathbf{r}/\rho(0)], \quad (4)$$

where α has been set equal to 2.0 to fit the *ORCRIT* results (Leherte & Vercauteren, 1997).

A volume can further be associated with each PK by integrating the exponential term of the Gaussian function over the space within the frame of an ellipsoid characterized by three main axes r_x, r_y and r_z

$$V = \frac{\pi^{3/2}\rho_{(0)}^{3/2}}{2^{3/2}|\text{ev}_1|^{1/2}|\text{ev}_2|^{1/2}|\text{ev}_3|^{1/2}} = \frac{4\pi}{3}r_x r_y r_z, \quad (5)$$

where ev_1, ev_2 and ev_3 are the three eigenvalues along the three main axes x', y', z' of the system. This mathematical operation provides a method of representing the shape anisotropy of each CP and this can be included as a steric contribution to the potential interaction expression in complementarity studies (Leherte & Allen, 1994). The three-dimensional representations of the ellipsoids obtained were generated using a hardcore interaction between the PKs and a probe. Each macromolecule is centred into a three-dimensional grid with a grid spacing of 0.25 Å. When a grid point falls outside V the potential is set to zero; otherwise it is set to an arbitrary value of 99. The isopotential surfaces drawn

Table 1
Percentage values of the significant chemical groups, *i.e.* backbones, phosphates and sugars, which are represented by one peak (PK) for the four proteins studied, *i.e.* P50, P65, HIF-1α and ARNT, and the two DNA oligomers, respectively, obtained by critical point analysis of their electron-density maps at resolutions of 2.7, 2.8, 2.85, 2.9, 3.0 and 3.2 Å.

The value given in parentheses is the total number of amino-acid residues, phosphate and sugar rings contained in the atomic structure (first column) or represented by at least one PK in the PK graphs (other columns).

	2.7 Å	2.8 Å	2.85 Å	2.9 Å	3.0 Å	3.2 Å
Amino acid						
Gly (45)	97.6 (42)	97.6 (42)	97.6 (41)	97.6 (42)	97.6 (41)	94.3 (40)
Ala (40)	97.4 (39)	94.9 (39)	97.4 (38)	100.0 (39)	97.5 (40)	100.0 (40)
Val (52)	92.3 (52)	96.1 (52)	94.1 (51)	92.3 (52)	94.1 (51)	85.4 (51)
Leu (57)	84.2 (57)	89.4 (57)	93.0 (57)	89.5 (57)	91.2 (57)	91.2 (57)
Ile (35)	94.3 (35)	88.2 (34)	91.4 (35)	91.4 (35)	91.4 (35)	88.6 (35)
Pro (45)	95.4 (44)	91.1 (45)	86.7 (45)	77.3 (44)	75.0 (44)	62.2 (44)
Ser (46)	95.6 (46)	95.6 (46)	93.5 (46)	89.1 (46)	92.7 (41)	93.0 (43)
Thr (37)	91.9 (37)	97.2 (36)	91.7 (36)	75.0 (36)	77.8 (36)	76.5 (34)
Asp (39)	84.6 (39)	87.2 (39)	87.2 (39)	84.6 (39)	76.9 (39)	79.5 (39)
Glu (49)	87.8 (49)	77.5 (49)	87.8 (49)	89.8 (49)	91.8 (49)	89.8 (49)
Asn (23)	91.3 (23)	91.3 (23)	87.0 (23)	91.3 (23)	87.0 (23)	69.6 (23)
Gln (30)	72.4 (29)	86.2 (29)	89.7 (29)	93.1 (29)	93.1 (29)	93.1 (29)
Lys (44)	84.1 (44)	84.1 (44)	90.7 (43)	90.9 (44)	86.4 (44)	86.4 (44)
Arg (60)	81.7 (60)	91.7 (60)	91.7 (60)	91.7 (60)	88.3 (60)	86.7 (60)
His (21)	95.2 (21)	85.7 (21)	85.7 (21)	76.2 (21)	85.7 (21)	90.5 (21)
Cys (16)	93.8 (16)	100.0 (16)	100.0 (16)	81.2 (16)	62.5 (16)	56.2 (16)
Met (13)	76.9 (13)	84.6 (13)	100.0 (13)	100.0 (13)	92.3 (13)	84.6 (13)
Phe (25)	96.0 (25)	100.0 (25)	88.0 (25)	95.8 (24)	92.0 (25)	88.0 (25)
Trp (2)	100.0 (2)	50.0 (2)	100.0 (2)	100.0 (2)	100.0 (2)	100.0 (2)
Tyr (24)	79.2 (24)	79.2 (24)	91.7 (24)	79.2 (24)	83.3 (24)	79.2 (24)
Total (703)	89.6 (697)	88.4 (696)	92.2 (693)	89.3 (695)	87.8 (690)	84.7 (689)
DNA unit						
PO_4 (48)	100.0 (48)	100.0 (48)	100.0 (48)	100.0 (48)	100.0 (48)	100.0 (48)
Sugar (52)	98.1 (52)	98.1 (52)	100.0 (52)	100.0 (52)	100.0 (52)	98.1 (52)
Total (100)	99.1 (100)	99.1 (100)	100.0 (100)	100.0 (100)	100.0 (100)	99.1 (100)

later in the paper using *Data Explorer* (IBM; http://www.research.ibm.com/dx) correspond to a potential value of 99.

4. Results

4.1. Graphs of critical points for NF-κB and HIF-1

After having generated the three-dimensional EDMs at different crystallographic resolution values ranging between 2.7 and 3.2 Å, the CPs were extracted by application of *ORCRIT*. A lower limit cutoff value of 1.0 e Å^{-3} was used. It prevents *ORCRIT* from generating too many insignificant CPs from the lowest density regions, *i.e.* $\rho(\mathbf{r})$ fluctuations arising from Fourier Transform truncation errors.

In order to associate each PK with an amino acid or a nucleotide, we needed to compare the graphs of PKs with the molecular structures of the macromolecules. A basic simplified model of each partner was built. In this model, each amino acid was reduced to two centroids, 'backbone' and 'side chain', located at the centre of mass (COM) of the N—C$^{\alpha}$—(C═O) subunit and of the amino-acid side chain, respectively. The DNA nucleotides were partitioned into three centroids, 'phosphate', 'sugar' and 'base', located at the COMs of the PO_4 group, the ribose ring and the nucleic base, respectively. Each PK of the graphs was associated with a specific subunit by comparing its position with the position of the closest subunit centroid.

Within the 2.7–3.2 Å resolution range, the 2.85 Å resolution value was selected as the best working resolution. To make this choice, we were particularly attentive to the structural regularity in the CP graph representation, *i.e.* one PK for each (N—C$^{\alpha}$—C═O) group in the case of amino-acid residues, as well as one PK per phosphate and one PK per sugar ring for the DNA-backbone graphs. The percentage values of each significant chemical group which is represented by one PK only are given in Table 1 for the studied resolution values. The total number of amino-acid residues, phosphates and sugar rings contained in the atomic structure (first column) or represented by PKs, *i.e.* associated with at least one PK, in the PK graphs (other columns) is indicated in parentheses. For example, only 41 out of 45 Gly residues are detected in the 2.85 Å resolution map. Graphs of PKs obtained for a protein partner, ARNT of HIF-1, and a DNA oligomer are illustrated in Figs. 3 and 4, respectively.

At the best chosen resolution (2.85 Å), the backbone of nearly all amino-acid residues, *i.e.* 92.2%, is represented by only one PK. It constitutes the highest percentage among the other resolution results (Table 1). We can also observe that for each resolution value some amino acids are not associated with any PK, *i.e.* 6, 7, 10, 8, 13 and 23 (giving a total of 703) at 2.7, 2.8, 2.85, 2.9, 3.0 and 3.2 Å, respectively. At 2.85 Å, ten amino acids are not represented; these are mainly small amino-acid residues, *i.e.* Gly, Ala and Val. This resolution, however, remains the best resolution if we look at the percentage of '0-PK' and '2-PK' per backbone. Indeed, increasing the resolution value results in a higher percentage of amino-acid backbones that are no longer represented ('0-PK'), for example at 3.2 Å; inversely, when decreasing the resolution value, *e.g.* to 2.7 Å, more residues are represented by two PKs per backbone. The 2.85 Å resolution value thus suits very well the regularity condition for protein-backbone representations, *i.e.* one PK per backbone. At this crystallographic resolution, side chains are generally represented by zero to two PKs as a function of the size and type of functional group of the amino acid (Table 2). Small amino acids, *i.e.* Gly and Ala, logically have no side-chain PK. Medium-sized amino acids possess one PK and large amino acids, *i.e.* Lys, Arg, Trp and Tyr, are represented by one or more PKs. For example, 11.8% of all valine side chains are represented by no PK and 88.2% by one PK. This is a consequence of the different environment of each amino-acid

side chain making the electron density adopt a more or less globular shape.

Table 2
Total number of amino acids in the protein sequences, total number of associated amino acids in the graphs of peaks and percentages of amino acids with zero, one and two peaks per backbone and per side chain for the reduced representations of the four studied proteins, *i.e.* P50, P65, HIF-1α and ARNT, obtained by critical point analysis of the electron-density maps at a resolution of 2.85 Å.

			Backbone			Side chain		
Amino acid	No. atomic amino acids	No. reduced amino acids	0-PK	1-PK	2-PK	0-PK	1-PK	2-PK
Gly	45	41	0.0	97.6	2.4	100.0	0.0	0.0
Ala	40	38	0.0	97.4	2.6	94.7	5.3	0.0
Val	52	51	5.9	94.1	0.0	11.8	88.2	0.0
Leu	57	57	1.7	93.0	5.3	1.7	98.3	0.0
Ile	35	35	8.6	91.4	0.0	5.7	85.7	8.6
Pro	45	45	13.3	86.7	0.0	20.0	80.0	0.0
Ser	46	46	4.3	93.5	2.2	15.2	84.8	0.0
Thr	37	36	8.3	91.7	0.0	5.6	94.4	0.0
Asp	39	39	10.3	87.2	2.5	0.0	100.0	0.0
Glu	49	49	8.1	87.8	4.1	2.0	98.0	0.0
Asn	23	23	8.7	87.0	4.3	0.0	100.0	0.0
Gln	30	29	3.4	89.7	6.9	0.0	100.0	0.0
Lys	44	43	7.0	90.7	2.3	4.7	55.8	39.5
Arg	60	60	1.7	91.7	6.6	0.0	45.0	55.0
His	21	21	4.8	85.7	9.5	0.0	100.0	0.0
Cys	16	16	0.0	100.0	0.0	6.2	93.8	0.0
Met	13	13	0.0	100.0	0.0	0.0	100.0	0.0
Phe	25	25	8.0	88.0	4.0	0.0	84.0	16.0
Trp	2	2	0.0	100.0	0.0	0.0	0.0	100.0
Tyr	24	24	8.3	91.7	0.0	4.1	29.2	66.7

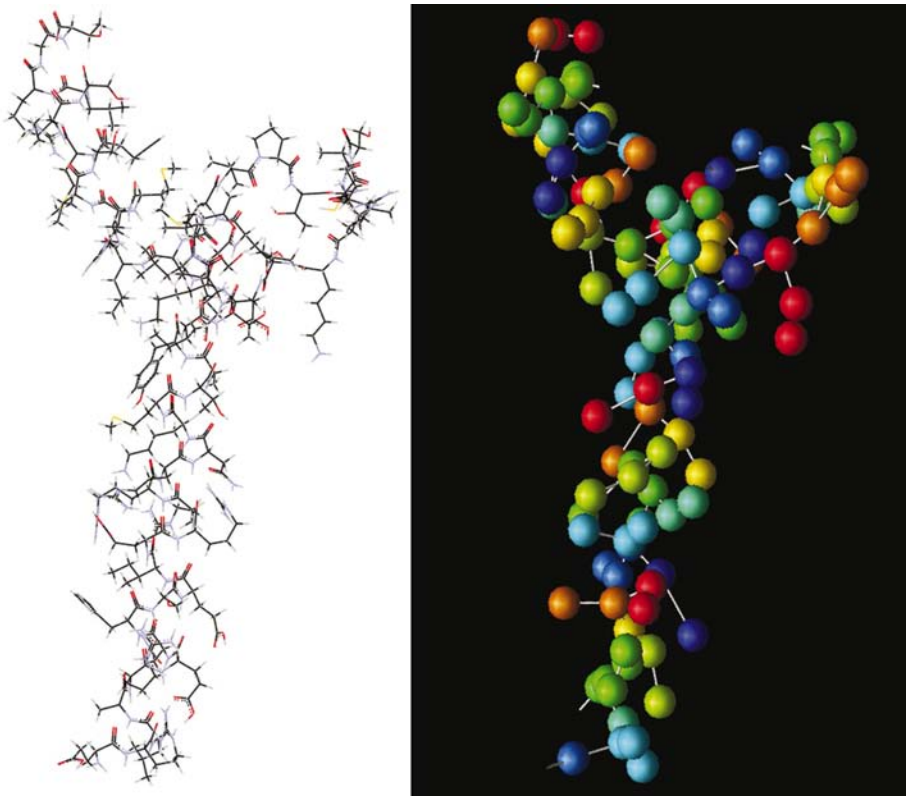

Figure 3
Comparison between the atomic representation (left) and the graph of peaks (right) for ARNT obtained at a resolution of 2.85 Å by critical point analysis of the electron-density maps. Colour code: a group of PKs of the same colour are associated with the same amino acid. Visualization of the atomic and reduced representations was performed using *WebLab Viewer* and *Data Explorer* (IBM; http://www.research.ibm.com/dx), respectively.

Table 3 reports information collected for the DNA graphs of PKs at 2.85 Å resolution. The regularity condition is very well fulfilled: 100.0% of sugar rings and 100.0% of the phosphate groups are represented by one PK only. The base pairs are constituted by one to three PKs: purine bases (A and G), which are two-ring structures, are represented by two or more PKs, while pyrimidine bases (T and C), which are single-ring structures, are represented by one or two PKs (Fig. 4). The PKs are spread around 1.4 Å of the COM of the base (Fig. 5*a*) as they are generally centred on the amino groups contained in the base rings. The single PKs associated with sugar rings and phosphates are very well localized, as seen in Fig. 5(*a*), in which we represent the distance between the PK position and the centroid of the subunits. Most of the sugar-ring and phosphate PKs are indeed located at distances of about 0.3 and 1.1 Å from their corresponding COM, respectively. The narrow distribution in Fig. 5(*a*) indicates the good regularity in the representation of these subunits in terms of $\rho(\mathbf{r})$ PKs. With our method, DNA is thus clearly represented by a double reduced strand constituted of alternating phosphate and sugar-ring PKs. The main structural information of the helix is thus preserved.

The topological analyses of P50 and P65, composed of 312 and 273 amino-acids, led to 570 and 511 PKs, respectively; those of HIF-1α and ARNT, both composed of 59 amino acids, led to 126 and 127 PKs, respectively. For the DNA oligomer of NF-κB, composed of 12 nucleotides, we obtained 98 PKs; for HIF-1, with 14 nucleotides, we obtained 111 PKs. The main structural information for each macromolecular partner, *i.e.* the number of atoms, of amino acids or base pairs, and the corresponding number of PKs at 2.85 Å resolution, is presented in Table 4.

By comparison with the reduced representations constructed on the atomic positions presented in §1, our method based on the electron density $\rho(\mathbf{r})$ constitutes a good starting point to model the anisotropy of the subunits, which can be useful when considering steric interactions between the mole-

Table 3
Composition of the molecular partners in terms of atoms, amino acids or base pairs, phosphates and sugar rings for the four studied proteins, *i.e.* P50, P65, HIF-1α and ARNT, and the two DNA oligomers, respectively, as well as the number of peaks obtained by critical point analysis of the electron-density maps at a resolution of 2.85 Å.

	Total No.	1-PK	2-PK	3-PK
PO_4	48	100.0	0.0	0.0
Sugar	52	100.0	0.0	0.0
A	9	0.0	100.0	0.0
T	9	33.3	66.7	0.0
G	17	0.0	35.3	64.7
C	17	29.4	70.6	0.0

Table 4
Total number and percentages of DNA subunits, *i.e.* phosphate, sugar rings and nucleic bases, with 0, 1, 2 and 3 peaks per subunit for the reduced representations of the two studied DNA oligomers, *i.e.* NF-κB and HIF-1, obtained by critical point analysis of the electron-density maps at a resolution of 2.85 Å.

(*a*) Protein.

	No. of atoms	No. of amino acids	No. of peaks
P50 (NF-κB)	2454	312	570
P65 (NF-κB)	3176	273	511
HIF-1α (HIF-1)	1011	59	126
ARNT (HIF-1)	987	59	127

(*b*) DNA oligomers.

DNA strings	No. of atoms	No. of base pairs (bp)/ phosphates (p)/sugars (s)	No. of peaks
DNA string (NF-κB)	486	12 bp/22 p/24 s	98
DNA string (HIF-1)	884	14 bp/26 p/28 s	111

cular partners. Indeed, solid representations of the graphs of PKs of ARNT and the HIF-1 DNA oligomer superimposed on the atomic structures (Figs. 6 and 7) show that in proteins, and especially in DNA strands, subunits are characterized by very distinct shapes. In Fig. 6, we can observe that the PK volumes cover the major side-chain structures, *e.g.* a PK associated with an amino acid containing an aromatic ring will be flat, oriented in the plane of the ring and located at its centre; an amino acid containing an amino-terminal side chain will be associated with a PK located at the extremity of the chain. The DNA representation in terms of ellipsoids is more discriminating when looking at the subunits: the phosphate PKs are small and spherical, while the sugar-ring PKs are spread all over the ring structure. The ellipsoids associated with the base subunits are flat and aligned with the base pairs. These visual representations reflect the ability of the graphs of PKs to represent the major structural features. Table 5 shows that phosphate groups which present few differences between the three eigenvalues have a low ellipticity value, explaining the spherical shape; whereas sugar and bases show higher ellipticity values owing to the flat PK shapes. In proteins, we can observe that there is no major difference between backbone and side-chain PKs, according to their eigenvalues and ellipticity values. A more precise study of the shape anisotropy is presented in the next section in order to correlate this visual information with other quantitative descriptors.

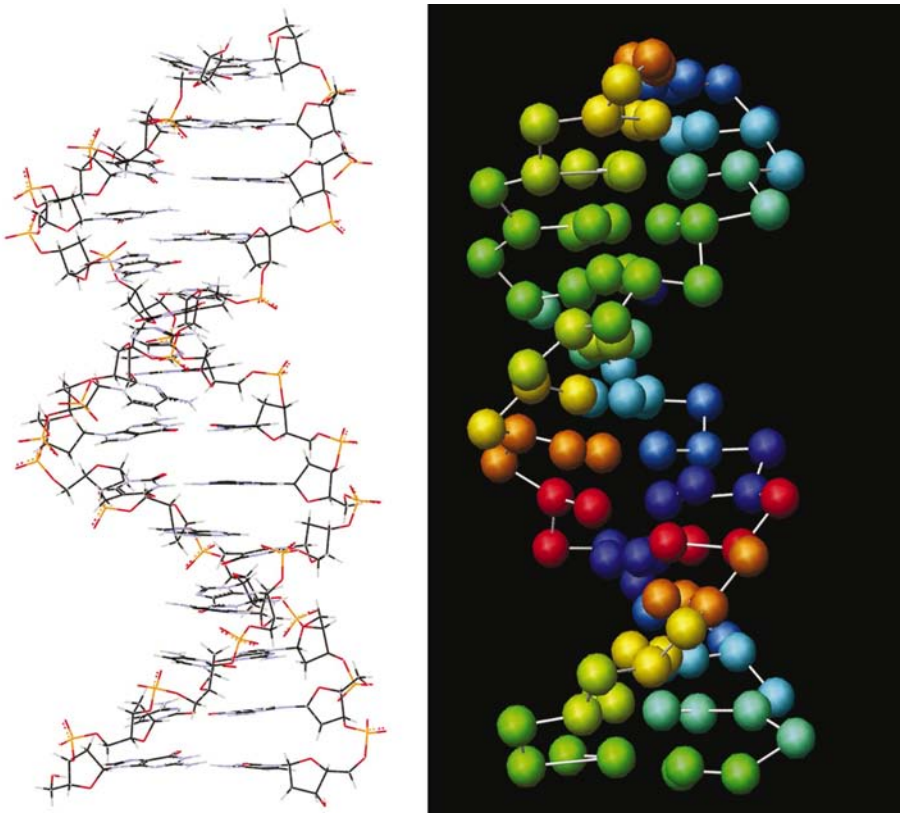

Figure 4
Comparison between the atomic representation (left) and the graph of peaks (right) of HIF-1 DNA oligomer obtained at a resolution of 2.85 Å by critical point analysis of the electron-density maps. Colour code: a group of PKs of the same colour are associated with the same nucleotide. Visualization of the atomic and reduced representations was performed using *WebLab Viewer* and *Data Explorer*, respectively.

4.2. Analysis of the peak descriptors

The first studied PK descriptor is $\rho(0)$, *i.e.* the ED value at the PK position. Fig. 8 illustrates the occurrence of $\rho(0)$ values expressed as a percentage of the total number of PKs associated with the proteins (Fig. 8*a*), the particular amino acids Cys and Met (Fig. 8*b*) and the DNA oligomers (Fig. 8*c*). $\rho(0)$ of the protein PKs adopts a regular distribution centred at 1.9 $e^- Å^{-3}$ (Fig. 8*a*), whereas DNA presents a double-peak distribution centred at 2.2 and 3.4 $e^- Å^{-3}$ (Fig. 8*c*). The 2.2 $e^- Å^{-3}$ peaks are associated with the sugar PKs and bases, while the

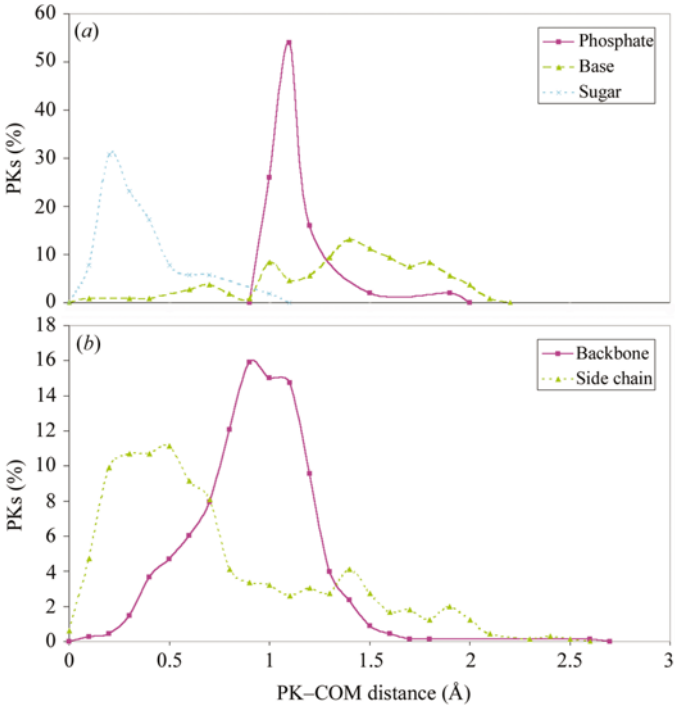

Figure 5
Interdistance distribution between the reduced representations (PK) obtained by critical point analysis of the electron-density maps at resolution of 2.85 Å and the atomic reduced points used for the association of peaks (*a*) with the nucleic subunits 'phosphate', 'sugar ring' and 'nucleic base' and (*b*) the protein subunits 'backbone' and 'side chain'.

Table 5
Eigenvalues and ellepticity average values with standard deviation for the two protein subunits, *i.e.* backbone and side chain, and the three DNA subunits, *i.e.* phosphate, sugar ring and bases, of the graphs of peaks obtained by critical point analysis of the electron-density maps at a resolution of 2.85 Å.

Subunit	ev1	ev2	ev3	Ellipticity
Backbone	−0.58 ± 0.10	−0.34 ± 0.08	−0.18 ± 0.07	−1.22 ± 0.42
Side chain	−0.58 ± 0.13	−0.35 ± 0.10	−0.20 ± 0.09	−1.15 ± 0.46
Phosphate	−0.78 ± 0.12	−0.62 ± 0.07	−0.55 ± 0.09	−0.37 ± 0.27
Sugar ring	−0.55 ± 0.07	−0.33 ± 0.07	−0.18 ± 0.05	−1.14 ± 0.28
Base	−0.87 ± 0.15	−0.32 ± 0.08	−0.17 ± 0.07	−1.70 ± 0.51

3.4 e^{-} Å^{-3} peaks are associated with the phosphate PKs. This observation will be correlated with the volume distribution studied later in this paper. The narrow and regular distribution of the sugar rings, as for the phosphate groups, simply arises from the regularity of the ED behaviour for these subunits, which present no major electronic differences along the DNA chains. Phosphate groups have higher ED peaks compared with sugar rings constituted of C, N and O. The nucleic base distribution presents a second peak at 2.6 e^{-} Å^{-3}. These last PKs are associated with the A and G bases, the purines composed of two rings. A detailed analysis of the protein $\rho(0)$ distribution (Fig. 8*a*) shows that the backbone PK distribution, located at 2.0 e^{-} Å^{-3}, is narrower than the side-chain PK distribution, which is spread between 1.2 and 3.2 e^{-} Å^{-3} and is centred at 1.9 e^{-} Å^{-3}. This difference in the $\rho(0)$ distribution

Figure 6
Ellipsoid representation of the graph of peaks of ARNT using a hardcore potential between peaks and a zero radius probe (isopotential surface = 99). Visualization was performed using *Data Explorer* (left view).

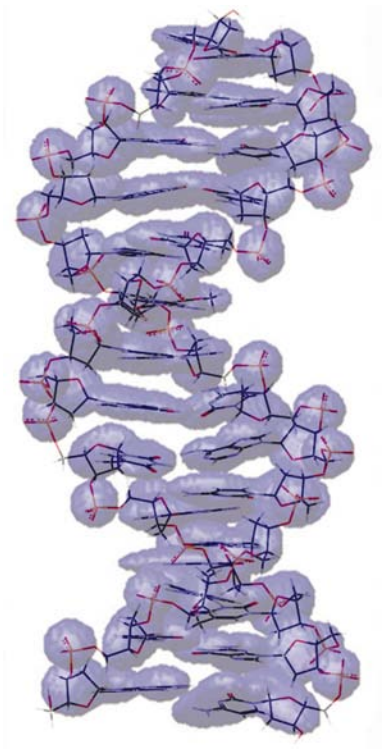

Figure 7
Ellipsoid representation of the graph of peaks of HIF1 DNA oligomer using a hardcore potential between peaks and a zero radius probe (isopotential surface = 99). Visualization was performed using *Data Explorer* (diagonal view).

is explained by the fact that the backbone N—C^{α}—C(=O) structure is similar for all amino acids. The various structures of the amino-acid residues explain the broader distribution for the side-chain PKs. The side-chain distribution shows that the higher values of $\rho(0)$ arise from the Cys and Met residues, which contain an S atom. By comparing the $\rho(0)$ values of these residues for the backbone and side-chain PKs (Fig. 8*b*), we can see that the higher values of $\rho(0)$ arise from the side-chain contribution only.

We have also considered the ellipticity as a descriptor for the peaks. It is related to the ratio between the highest and smallest eigenvalues as defined in §3 (equation 3). Analysis of the ellipticity-value distributions for the protein graphs (Fig. 9*a*) shows that they adopt a similar shape for both backbone and side-chain PKs. The distribution is broad, ranging from 0.2 to 3.5 and centred at 1.2. The phosphate PKs are associated with a smaller ellipticity value, *i.e.* 0.4, than the sugar rings and nucleic groups (Fig. 9*b*). Sugar rings present a medium-value distribution, *i.e.* 1.1, while bases are characterized by a broad range centred at 1.7. The narrow and low ellipticity-value distribution of the phosphate PKs reflects the spherical and well centred distribution of the electrons around this subunit. This descriptor is thus useful for discriminating between the phosphate, sugar ring and nucleic base groups, but does not allow differentiation of the protein subunit backbones and side chains.

Studies of the volume distributions (equation 5) show that the values are quite similar for proteins and DNA subunits, as illustrated in Figs. 10(*a*) and 10(*b*), respectively. Broad distributions centred around 30.0 Å^3 are obtained for the backbone and side-chain PKs, as well as for base and sugar PKs. In relation to the analysis made for the high $\rho(0)$ values of the DNA phosphate groups and the two sulfur-containing residues, Cys and Met, we have observed that these subunits present a lower volume value, *i.e.* 23.7 and 28.0 Å^3 (not presented in Fig. 10*a*), respectively, compared with other groups. This reflects the more confined distribution of the ED around the PO_4 groups and S atoms. The other subunits present a broader distribution of values.

In order to use these graphs as efficient potential reduced representations of biomolecules, it is necessary to add another important descriptor of the PKs: the charge of the subunits at physiological pH. As with the *XTAL* approach $\rho(\mathbf{r})$ distributions are calculated using promolecular models, it is not

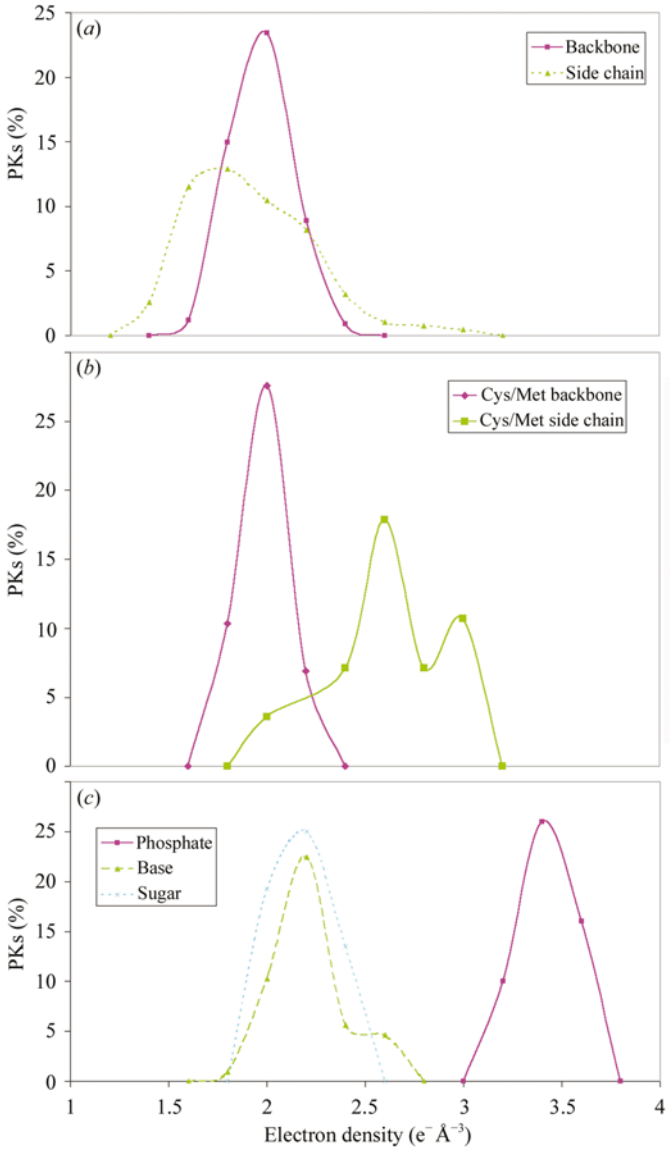

Figure 8
Electron density at peak (PK) position distribution for (*a*) the four studied proteins, *i.e.* P50, P65, HIF-1α and ARNT, (*b*) the amino acids Cys and Met and (*c*) the two DNA oligomers obtained by critical point analysis of the electron-density maps at resolution of 2.85 Å.

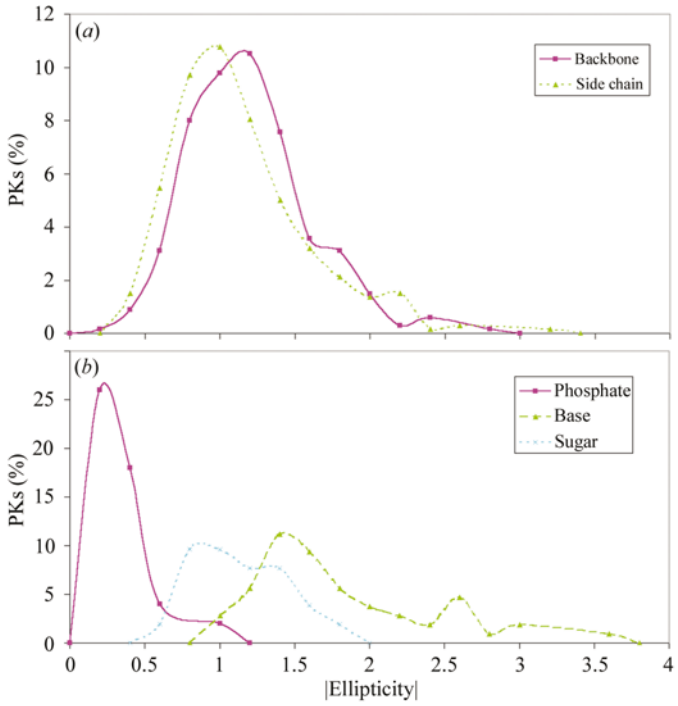

Figure 9
Ellipticity of the peak (PK) distribution for (*a*) the protein subunits 'backbone' and 'side chain' and (*b*) the nucleic acid subunits 'phosphate', 'sugar ring' and 'nucleic base' obtained by critical point analysis of the electron-density maps at resolution of 2.85 Å.

possible to calculate directly the charges associated with the PKs. Indeed, a promolecule basically consists of spherical atoms without interactions. The electronic contribution for each atom is thus equal to its atomic number and is centred on the atomic positions. Therefore, a '−1' charge has been attributed to the 'phosphate', Asp and Glu side-chain PKs and a '+1' charge to Lys, Arg and His side-chain PKs; all other subunits have been considered as neutral groups (Campbell, 1996).

4.3. Discussion

A comparison of our representation, based on the graphs of PKs obtained from topological analysis of the ED calculated at a medium resolution, *i.e.* 2.85 Å, with the methods cited in §1, led to the following observations. PK representations of proteins with one PK per residue at the backbone level are structurally very close to the reduced representation used by Levitt & Warshel (1975), Wallqvist & Ullner (1994), Feig *et al.* (2000) and Hassinen & Peräkylä (2001). At the level of the amino-acid residue side chains, we can compare our results with the representation obtained by Levitt & Warshel (1975) which, in addition to the C^{α} location, included a site on the centroid of the side chain. Moreover, as in models used by Wallqvist & Ullner (1994) and Hassinen & Peräkylä (2001), our representation presents the advantage of discriminating the small amino acids from large amino acids by associating one supplementary side-chain point with the latter. Table 6 presents a comparison between the amino-acid representation in terms of reduced points with the methods cited previously. We can observe that for medium-sized amino acids, *i.e.* Gln and Leu, our method is closer to the Hassinen representation in that it considers only one point for the side chain, instead of

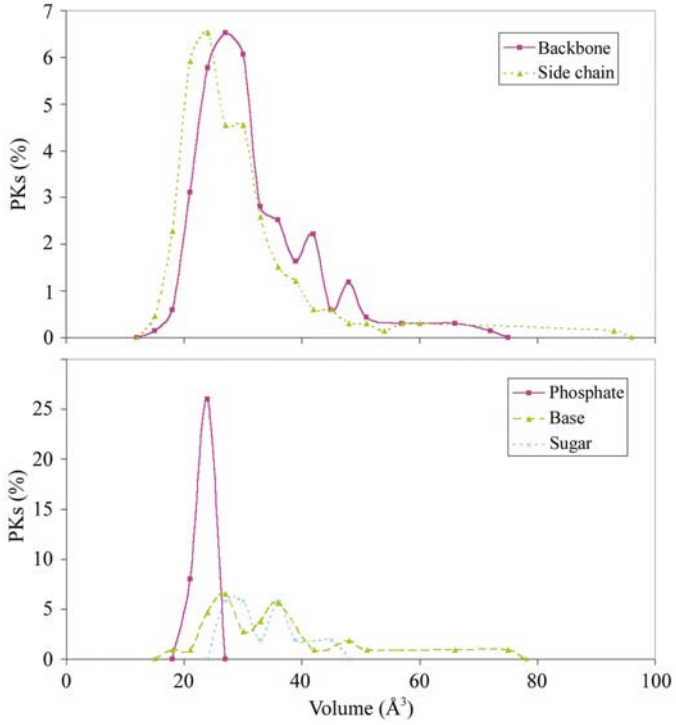

Figure 10
Volume of the peak (PK) ellipsoid distribution for (*a*) the protein subunits 'backbone' and 'side chain' and (*b*) the nucleic subunits 'phosphate', 'sugar ring' and 'nucleic base' obtained by critical point analysis of the electron-density maps at resolution of 2.85 Å.

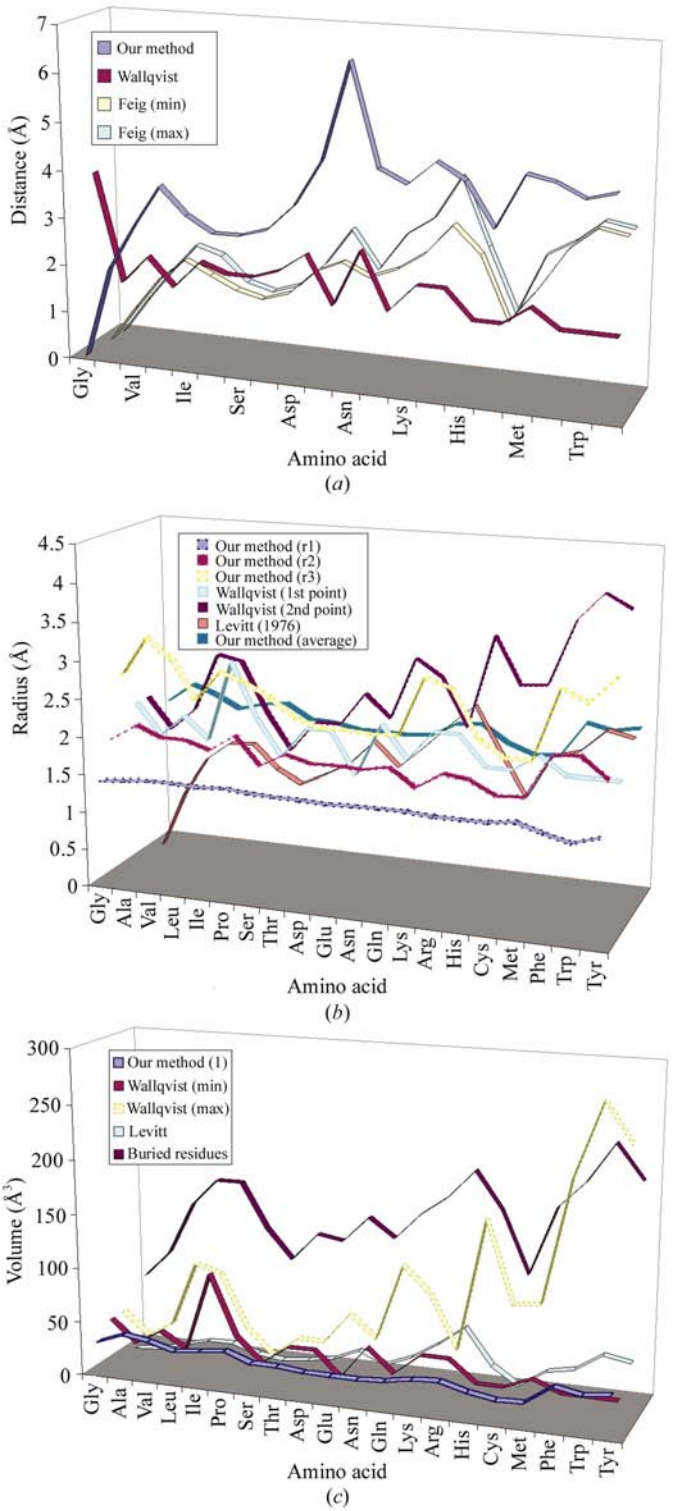

Figure 11
Comparison between our method and those of Wallqvist & Ullner (1994) and Feig *et al.* (2000) for (*a*) the interdistance between the backbone and the side-chain points, (*b*) the radii and (*c*) the volume values of the protein subunit 'side chain'.

Table 6
Comparison between the number of side-chain reduced points per amino acid, obtained by our method and by those of Wallqvist & Ullner (1994), Feig *et al.* (2000) and Hassinen & Peräkylä (2001).

	No. side-chain points			
	Present work	Wallqvist & Ullner (1994)	Feig *et al.* (2000)	Hassinen & Peräkylä (2001)
Gly	0	0	(1)	0
Ala	0	1	1	0
Val	1	1	1	0
Leu	1	2	1	1
Ile	1	2	1	1
Pro	1	1	1	0
Ser	1	1	1	0
Thr	1	1	1	0
Asp	1	1	1	1
Glu	1	2	1	1
Asn	1	1	1	1
Gln	1	2	1	1
Lys	1 or 2	2	1	2
Arg	1 or 2	2	1	2
His	1	2	1	1
Cys	1	2	1	1
Met	1	2	1	1
Phe	1	2	1	1
Trp	2	2	1	2
Tyr	1 or 2	2	1	1

two as in Wallqvist & Ullner (1994). However, for smaller amino acids such as Ser and Val our representation is closer to the Wallqvist representation: it considers one point whereas Hassinen does not associate any side-chain point. Wallqvist & Ullner (1994) based their representation mainly on the hydrophobic/hydrophilic properties of the amino acids in order to decide whether the side chain should be represented by one or two points. In the case of a two-point representation of the side chain, *e.g.* His, Arg *etc.*, the first reduced point is always labelled as hydrophobic, while the second point is either hydrophobic or hydrophilic according to the nature of the residue. This explains why medium-sized amino acids, *e.g.* Cys, Met or Leu, are represented by one PK with our method and by two points using that of Wallqvist. The Feig representation considers only one peak per side chain, which is located at its COM. We note that this method may present some limits for the representation of large amino acids, *i.e.* Trp or Arg. Fig. 11(*a*) compares the distances between the backbone reduced point and its closest side-chain point for the three different methods mentioned above. Feig *et al.* (2000) data represent the average distances between C^{α} and the side-chain COM (min and max). We observe that our values are systematically higher than those of the other methods. This can be explained by the fact that the PK positions do not correspond exactly to the COM of the backbone and side-chain subunits, as shown in Fig. 5(*b*).

In addition, our method presents a non-negligible advantage by considering the anisotropy of the peaks, which allows, for example, discrimination of an aromatic ring from a long amino-terminal side chain. Our approach thus differs mainly by this feature from that of Guo *et al.* (1995, 1999) who also worked with low-resolution electron-density maps in order to model a protein by a continuous chain of 'globs' representing the amide regions of the peptide backbone and the side-chain residues. However, as each 'glob' was treated as a spherical cluster of atoms, this geometrical property was likely to be too approximate to fit the global shape of the underlying chemical groups. Figs. 11(*b*) and 11(*c*) present the radii distributions of the reduced points for each amino acid and the corresponding volume. As we are considering ellipsoids, we have three distributions for each amino acid, *e.g.* small, medium and large radii. It is observed that the medium radius values for each amino-acid ellipsoid are in the range of the radius values reported by Wallqvist & Ullner (1994) for the first interacting point, while the longest ellipsoid radius values are closer to the values reported by the same authors for the second amino-acid reduced point.

Finally, the flexibility of the partners may be an important issue to consider when setting up a docking methodology. In the literature, we note three different ways of considering flexibility in complementarity studies. Firstly, in the cases of protein–ligand complexes, most of the currently developed methods consider a fully flexible ligand and either a rigid protein or a partially flexible protein in the neighbourhood of the active site (Oshiro *et al.*, 1995; Jones *et al.*, 1997; Vieth *et al.*, 1998*a*,*b*). Secondly, another proposed way to consider ligand flexibility is to work with a set of different rigid conformations of the same molecule, as in Meurice (2001). In this last work, the aim was to perform similarity studies between ligands of pharmacological interest using reduced representations in terms of critical points (CPs); a genetic algorithm was specially designed for the automated superimposition of the CP patterns, allowing the simultaneous alignment of several flexible molecules and proposal of a pharmacophore model. Thirdly, in large-molecule studies, *e.g.* in protein–protein docking, most of the existing methods use a rigid-body docking (RBD) approach. In these studies, it is considered that complexed partners ('bound' structure) are only slightly changed compared with the free ('unbound') structure, because most of the conformational changes occur on the surface amino acids (Lo Conte *et al.*, 1999). A 'bound' structure is defined as a co-crystallized molecule which is already deformed and is well suited to an RBD approach; an 'unbound' structure is defined as an independently crystallized partner which could undergo some surface deformations when docked with the other partner. If an RBD method is used with 'unbound' structures, a 'soft' representation of the macromolecular surface provides a way to introduce some fuzziness and thus to implicitly contain the adaptive changes between the 'bound' and 'unbound' partners (Palma *et al.*, 2000). In our case, the topological analysis of the EDMs at medium crystallographic resolution as reduced representations of the amino acids contains some inherent softness in the protein representations and is thus well adapted to an RBD methodology. This approximation can be widely used in automated methods where the molecular flexibility is not expected to affect much of the geometry of the interacting structures; however, it becomes somewhat more difficult when considering nucleic acids. Indeed, when interacting with proteins, it is known that the DNA or RNA polymers may be deformed or

curved by the protein(s). This constitutes a major difficulty when studying protein–DNA complementarity. Giving that very few references have been found in the literature regarding reduced representations of DNA oligomers, it is difficult to immediately know how it may be possible to face the flexibility problem in the case of 'unbound' structures of nucleic acid partners. In the docking method that we are currently developing, we have chosen to work with 'bound' molecules (protein–DNA complexes) as a first step in order to set up our global methodology. After that, it is planned to try the same RBD approach with canonical DNA strings, *i.e.* an 'unbound' DNA string generated by a DNA generator knowing the consensus sequence (Liu & Beveridge, 2001; http://ludwig.chem.wesleyan.edu/dna/) in order to see if the method is sufficiently soft to work with 'unbound' structures of DNA.

5. Conclusion

In this work, we have proposed a method for the generation of reduced representations of macromolecular structures such as protein–DNA complexes. In particular, we have studied two systems of great interest because of their implication in various diseases involved in inflammatory processes or cell reactions: NF-κB and HIF-1. These two complexes both consist of a protein heterodimer, P50/P65 and HIF-1α/ARNT, respectively, bound to a DNA consensus sequence. When considering reduced representations of macromolecules, mainly proteins, most methods presented in the literature are based on a classical 'two-site' model, with a point centred on the C^{α} of the amino acids and the other on C^{β} or on the centre of mass of the residues. These methods are usually based on geometrical considerations only and consequently have the disadvantage of not considering the topology or the physico-chemical properties of the side-chain chemical functions and groups. Our strategy is based on the use of electron-density distributions $\rho(\mathbf{r})$ calculated at a medium crystallographic resolution, *i.e.* 2.85 Å. At this resolution, $\rho(\mathbf{r})$ is concentrated on the chemical groups and functions of the molecular partners and a topological analysis of this three-dimensional property allows the extraction of its critical points (CPs), which are the points where the gradient of $\rho(\mathbf{r})$ is equal to zero. As we focus on simplified representations of the groups of atoms, we only considered peaks (PKs), which are the maxima of $\rho(\mathbf{r})$. Each molecular partner is thus represented by a graph composed of PKs which can be associated with a specific molecular subunit: 'backbone' or 'side chain' for amino acids or 'phosphate', 'sugar ring' or 'nucleic base' for nucleotides. As descriptors for each peak, we evaluated several topological properties, *i.e.* $\rho(\mathbf{r})$, ellipticity, volume and the physico-chemical properties of the subunit attached to each PK, *i.e.* hydrophobicity/hydrophilicity and charge. We showed that the graphs of PKs at 2.85 Å reflect the major structural information of both the protein and the DNA components. For proteins, each amino acid is characterized by at least one backbone PK and up to two side-chain PKs according to its nature. DNA strings are modelled by a succession of 'phosphate' and 'sugar-ring' PKs; the nuclear bases are represented by one to three PKs located on the N- or O-containing functions. With these first observations, we can conclude that our representations based on graphs of PKs constitute a good model to represent the macromolecules with the intrinsic consideration of their electron-density distribution.

In a second step, we analysed the different descriptor values for the graphs of PK of the proteins and DNA strings. We made the following conclusions.

(i) The density at the PK position, $\rho(0)$, is a descriptor that allows differentiation between the phosphate subunits in a DNA graph of PKs and the Cys and Met residues in a protein graph, as they are characterized by a higher value of $\rho(0)$. These observations can be explained by the fact that these three subunits contain atoms with high atomic numbers. Additionally, $\rho(0)$ is implicated in the volume value of each PK and thus constitutes a major descriptor for the discrimination of the PKs.

(ii) Ellipticity also allows discrimination of each molecular subunit, as it presents a specific PK shape anisotropy, particularly for the DNA oligomers: phosphate subunits are spherical and very localized, while sugar rings are represented by a flat ellipsoid oriented in the plane of the ring and base pairs are represented by flat PKs in the direction of the bonding. This value can therefore be employed in an automated procedure of identification of PKs compared with atomic structures, as the anisotropy associated with each PK reflects the local shape of these subunits.

(iii) Electronic charge is an important property as it drives the DNA-recognition and protein-dimerization phenomena. However, it cannot be directly extracted from CP analysis of a promolecular electron-density map, as the electronic contribution for each atom is equal to its atomic number and is centred on the atomic positions. This property has therefore been set to the net charge of the amino-acid side chains.

We have thus shown that our method properly represents the protein and DNA structures and properties and therefore constitutes a basis for the development of an interaction potential for docking purposes. In this sense, an original genetic algorithm (GA) procedure has already been developed in order to manipulate the graphs of PKs needed for the docking. GAs have been chosen as they are well adapted to complex problems with a huge number of possible solutions, as in complementarity studies between macromolecules. Technically, in our implementation each potential solution is represented by a numerical chromosome coding for the position and orientation of each partner and an iterative process allows the evolution and convergence of the system: (i) an initial population of chromosomes is randomly generated, (ii) each solution is evaluated giving them a score, (iii) a 'roulette-wheel' selection is operated, (iv) the biological evolution is simulated by genetic mutations and crossover and (v) an iterative process evaluates the new population and goes back to (i) until a termination condition that ends the process is encountered. The fitness values are directly related to the quality of the solutions and at the end allow the GA to find a

final set of solutions containing the best macromolecular complex configurations. In our case, the GA manages the relative positions of each partner and evaluates an interaction score that is being adapted to the PK representation and the protein–DNA recognition. The score contains three major terms: steric hindrance, which is calculated by a Lennard–Jones potential between the ellipsoid volumes associated with each PK (Leherte & Allen, 1994), a Coulombic electrostatic complementarity calculated according to the charges associated with each amino acid and nucleic subunit, and the evaluation of the amino-acid recognition process by surface calculation and medium distance tables for each amino acid–amino acid and amino acid–nucleic base pair. This last point is presently under development, each contribution of the interaction expression being parameterized with training sets of hundreds of macromolecular complexes.

The authors would like to thank the FUNDP and the Scientific Computing Facility (SCF) centre for computing resources. AB is grateful to Professor M. Raes of the Cellular and Molecular Biology Laboratory of the FUNDP for fruitful discussions and to the 'Fonds National de la Recherche Scientifique' for his PhD Research Fellowship.

References

Altuvia, Y., Schueler, O. & Margalit, H. (1995). *J. Mol. Biol.* **249**, 244–250.
Ausiello, G., Cesareni, G. & Helmer-Citterich, M. (1997). *Proteins Struct. Funct. Genet.* **28**, 556–567.
Bader, R. F. W. (1995). *Atoms in Molecules: A Quantum Theory*, 2nd ed. Oxford: Clarendon Press.
Campbell, N. A. (1996). *Biology*, 4th ed. Menlo Park, CA, USA: Benjamin/Cummings.
Chen, F. E., Huang, D. B., Chen, Y. Q. & Ghosh, G. (1998). *Nature (London)*, **391**, 410–413.
Das, A. & Jayaram, B. (1998). *J. Mol. Liq.* **77**, 157–163.
Feig, M., Rotkiewicz, P., Kolinski, A., Skolnick, J. & Brooks, C. L. (2000). *Proteins Struct. Funct. Genet.* **41**, 86–97.
Fortier, S., Castleden, I., Glasgow, J., Conklin, D., Walmsley, C., Leherte, L. & Allen, F. H. (1993). *Acta Cryst.* D**49**, 168–178.
Gardiner, E. J., Willett, P. & Artymiuk, P. J. (2001). *Proteins Struct. Funct. Genet.* **44**, 44–56.
Ghosh, G., van Duyne, G., Ghosh, S. & Sigler, P. B. (1995). *Nature (London)*, **373**, 303–310.
Guo, D. Y., Blessing, R. H., Langs, D. A. & Smith, G. D. (1999). *Acta Cryst.* D**55**, 230–237.
Guo, D. Y., Smith, G. D., Griffin, J. F. & Langs, A. L. (1995). *Acta Cryst.* A**51**, 945–947.
Hall, S. R., du Boulay, D. J. & Olthof-Hazekamp, R. (2000). Editors. *Xtal3.7 System.* University of Western Australia.
Hassinen, T. & Peräkylä, M. (2001). *J. Comput. Chem.* **22**, 1229–1242.
Johnson, C. K. (1977). *ORCRIT. The Oak Ridge Critical Point Network Program.* Chemistry Division, Oak Ridge National Laboratory, NC, USA.
Jones, G., Willett, P., Glen, R. C., Leach, R. & Taylor, R. (1997). *J. Mol. Biol.* **267**, 727–748.
Kitzing, E. von & Schmitt, E. (1995). *Theochem*, **336**, 245–259.
Leherte, L. & Allen, F. H. (1994). *J. Comput. Aided Mol. Des.* **8**, 257–272.
Leherte, L., Fortier, S., Glasgow, J. & Allen, F. H. (1994). *Acta Cryst.* D**50**, 155–166.
Leherte, L., Latour, T. & Vercauteren, D. P. (1995). *Supramol. Sci.* **2**, 209–217.
Leherte, L., Latour, T. & Vercauteren, D. P. (1996). *J. Comput. Aided Mol. Des.* **10**, 55–66.
Leherte, L. & Vercauteren, D. P. (1997). *J. Mol. Model.* **3**, 156–171.
Levitt, M. (1976). *J. Mol. Biol.* **104**, 59–107.
Levitt, M. & Warshel, A. (1975). *Nature (London)*, **253**, 694–698.
Liu, Y. & Beveridge, D. L. (2001). *J. Biomol. Struct. Dyn.*, **18**, 505–526.
Liwo, A., Oldziej, S., Pincus, M. R., Wawak, R. J., Rackovsky, S. & Scheraga, H. A. (1997). *J. Comput. Chem.* **18**, 849–873.
Lo Conte, L., Chothia, C. & Janin, J. (1999). *J. Mol. Biol.* **285**, 2177–2198.
Maggiora, G. M., Rohrer, D. C. & Mestres, J. (2001). *J. Mol. Graph. Model.* **19**, 168–178.
Meurice, N. (2001). PhD thesis. Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium.
Meurice, N., Leherte, L. & Vercauteren, D. P. (1998). *SAR QSAR Environ. Res.* **8**, 195–232.
Michel, G., Minet, E., Ernest, I., Roland, I., Durant, F., Remacle, J. & Michiels, C. (2000). *J. Biomol. Struct. Dyn.* **18**, 169–179.
Müller, C. W., Rey, F. A., Sodeoka, M., Verdine, G. L. & Harrison, S. C. (1995). *Nature (London)*, **373**, 311–317.
Oshiro, C. M., Kuntz, I. D. & Dixon, J. S. (1995). *J. Comput. Aided Mol. Des.* **9**, 113–130.
Palma, P. N., Krippahl, L., Wampler, J. E. & Moura, J. J. G. (2000). *Proteins Struct. Funct. Genet.* **39**, 372–384.
Semenza, G. L. (2001). *Curr. Opin. Cell Biol.* **13**, 167–171.
Semenza, G. L., Agani, F., Booth, G., Forsythe, J., Iyer, N., Jiang, B. H., Leung, S., Roe, R., Wiener, C. & Yu, A. (1997). *Kidney Int.* **51**, 553–555.
Tan, R. K. & Harvey, S. C. (1989). *J. Mol. Biol.* **205**, 573–591.
Terry, A. (1983). PhD thesis. Stanford University, Stanford, CA, USA.
Vieth, M., Hirst, J. D., Kolinski, A. & Brooks, C. L. III (1998*a*). *J. Comput. Chem.* **19**, 1612–1622.
Vieth, M., Hirst, J. D., Kolinski, A. & Brooks, C. L. III (1998*b*). *J. Comput. Chem.* **19**, 1623–1631.
Wallqvist, A. & Ullner, M. (1994). *Proteins Struct. Funct. Genet.* **18**, 267–280.
Yang, L. J., Feng, X. Z., Lee, I. S. & Bai, C. L. (1998). *J. Mol. Struct.* **444**, 13–20.