Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Hierarchical analysis of promolecular full electron-density distributions: description of protein structure fragments

Laurence Leherte

Laboratoire de Physico-Chimie Informatique, Facultés Universitaires Notre-dame de la Paix, Rue de Bruxelles 61, B5000 Namur, Belgium

Correspondence e-mail: laurence.leherte@fundp.ac.be

Received 20 February 2004
Accepted 5 May 2004

A theoretical method is applied to describe protein structures in terms of hierarchically related substructures. The approach is based on the location of local maxima (peaks) in promolecular electron-density (ED) distributions established at various smoothing levels. Promolecular ED distributions are generated using either the Promolecular Atom Shell Approximation (PASA) representation or the ED are calculated using the Cromer–Mann coefficients. The analysis of the decomposition patterns of a protein structure in its native and hypothetical extended conformations showed that the amino-acid residues have a similar decomposition pattern regardless of their position in the protein sequence, the protein conformation and the influence of the crystal packing. A link is proposed with model-building tools in protein structure determination from diffraction data.

1. Introduction

Fragments of molecular structures (groups of atoms, templates) are useful in the field of molecular recognition/matching (Lemmen *et al.*, 1998; Gillet *et al.*, 2003; Jain, 2003; Krä*mer et al.*, 2003), in building molecular structures or determining their properties from transferable structural fragments (Walker & Mezey, 1994; Mezey, 1996; Matta & Bader, 2002), in X-ray diffraction data interpretation *etc*. Of the methods proposed in the literature to fit fragments into a three-dimensional electron-density (ED) distribution, some allow the location of molecular structural features below atomic resolution. More particularly, in the field of protein crystallography, the concept of fragments is of great importance, *e.g.* in the building of molecular models from diffraction-derived data such as ED distributions. The first application of structural fragments in macromolecular model building was proposed by Jones & Thirup (1986) and implemented in the program *FRODO* (Jones, 1978, 1985). The authors showed that retinol-binding protein could be reconstructed from substructures of three other unrelated proteins.

In a crystal structure determination of a biological macromolecule, the interpretation of an ED map in terms of a macromolecular model is an iterative procedure with several refinement/reconstruction steps (Lamzin & Perrakis, 2000). Software which is popular among protein crystallographers involves the use of databases of fragments, *i.e. TURBO-FRODO*, which is a derivative of *FRODO* (Jones, 1978, 1985), *O* (Jones *et al.*, 1991; Jones & Kjeldgaard, 1997), *XtalView* (McRee, 1999) or *MAIN* (Turk, 1992). The '*warpNtrace*' method (Perrakis *et al.*, 1999), part of the *ARP*/*wARP* suite (Morris *et al.*, 2003), brings a high degree of automation to the process of building and refining a crystallographic macromolecular model. It is based on an iterative procedure in

which a 2.5 Å (or higher) resolution map is interpreted in terms of a hybrid model, a combination of auto-built protein fragments and individual density centres.

In the literature, tracing a protein skeleton is proposed to be achievable using different approaches such as core tracing as presented in Swanson (1994) and references therein. In the most recent papers, the skeleton is traced by searching for ridges in the ED obtained at a resolution value of 2.5 Å, as implemented in the program *MAID* (Levitt, 2001), or at 3.5 Å, as in *QUANTA* (Oldfield, 2002). Levitt (2001) used the so-called skeletonized bones as trial points for the initial positioning of a helix or sheet, while Oldfield (2002) applied a pattern-recognition algorithm for α-helix and β-strand motifs based on a principal component analysis of the inertia tensor matrix at selected skeleton points. Previously, Fortier *et al.* (1997) and Leherte *et al.* (1997) defined geometrical templates to search for secondary-structure motifs in 3 Å resolution ED maps. These templates were built from the coordinates of $C^{\alpha}-C{=}O-N$ centres of mass (COM), which were found to be very close to the local ED maxima obtained using a critical point (*i.e.* points where the gradient of the ED is zero) analysis approach (Johnson, 1977) in 3 Å resolution maps.

Protein fragments/templates are often defined as short sequences of amino-acid residues. For example, Kleywegt & Jones (1997) used five-residue-long poly-Ala sequences in their work on feature detection in macromolecular ED maps at 2.7–3 Å resolution. Pavelcik *et al.* (2002) used rigid arbitrary single-residue or two-residue-long segments and poly-Ala fragments of up to 18 atoms. These fragments were represented in terms of spherical harmonic Bessel expansions of their ED at crystallographic resolution values of around 3 Å.

The positioning of fragments in a map often consists of a six-dimensional search (three translations and three rotations) as in the program *ESSENS* (Kleywegt & Jones, 1997), while in Cowtan's method the three-dimensional translational search is replaced by a stage of three Fourier transformations as implemented in *FFFEAR* (Cowtan, 2001). In this latter program, skeleton fragments are defined as nine-residue-long poly-Ala (or poly-Gly) sequences that were identified by cluster analysis of a large subset of the Protein Data Bank (PDB; Berman *et al.*, 2000). Such fragments are representative of helix, strand, turn and helix-end secondary-structure features. Additional five- and ten-residue-long fragments were also theoretically generated from information extracted from the Ramachandran plot. Terwilliger (2003*a*) described an automated procedure implemented in the program *RESOLVE* for macromolecular model building of polypeptide backbones. Fragment libraries of helices and β-strands obtained from refined protein structures are positioned at previously determined potential locations using templates which are averaged over ED distributions at a crystallographic resolution of 3 Å. These templates are six amino-acids long and four amino-acids long for helices and strands, respectively. The four fragment libraries that are considered were built from selected PDB coordinates: 17 α-helices made of 6–24 residues, 17 β-strands made of 4–9 residues, a library composed of three-amino-acid-long skeletons including C^{β} coordinates and a fourth library composed of three-residue-long segments in which the first residue is represented by its $C^{\alpha}-C{=}O$ atoms. The approximate locations and orientations of helices and β-strands are identified using the templates and the FFT-based convolution method proposed by Cowtan (2001). Once a main chain is built, the side-chain assignment is performed. For example, Terwilliger (2003*b*) used templates built from averaged amino-acid side-chain densities in 574 refined PDB structures to calculate probability measures by a Bayesian approach.

Based on different identification methods, the approach proposed by Ioerger & Sacchettini (2002) involves a procedure called *CAPRA* that predicts coordinates of C^{α} atoms in ED maps and connects them. *CAPRA* is based on pattern-recognition techniques and a neural network algorithm to predict which potential atoms in an 2.4–2.8 Å ED trace are closest to true C^{α} atoms. *CAPRA* is part of a more general system called *TEXTAL* (Ioerger *et al.*, 1999; Gopal *et al.*, 2003) in which the identification of structural features in an unsolved map using rotation-invariant descriptors is achieved by case-based reasoning, *i.e.* based on previously solved structures. Such rotation-invariant features are computed from limited regions of ED distributions and are, for example, the mean ED value, the moment of inertia, the eigenvalues and their ratio, distances from centre-of-mass *etc.* Other references to works based on geometrical reconstruction of protein backbones from C^{α} coordinates can be found in Iwata *et al.* (2002).

The definition of molecular fragments can be achieved by partitioning ED distributions into separate regions of space through surfaces. The Atom-In-Molecule (AIM) approach developed by Bader (2001) is known to assign a basin to an atom through the definition of zero-flux surfaces, which are defined by

$$\nabla\rho(\mathbf{r}) = 0. \tag{1}$$

Walker and Mezey have developed the so-called additive fuzzy ED fragmentation method which is the basis of the Molecular ED Lego Assembler (MEDLA) method (Walker & Mezey, 1994; Mezey, 1996) for the computation of *ab initio* quality ED for large molecules. Within the conventional Hartree–Fock/self-consistent field/linear combination of atomic orbitals *ab initio* scheme, an ED distribution is defined as

$$\rho(\mathbf{r}) = \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}\varphi_i(\mathbf{r})\varphi_j(\mathbf{r}), \tag{2}$$

where $\varphi_i(\mathbf{r})$ is an atomic orbital and P_{ij} is an element of the $n \times n$-dimensional density matrix $\mathbf{P}$. Similarly, the ED of a fuzzy ED fragment specified as a subset k of nuclei is given by

$$\rho^k(\mathbf{r}) = \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}^k\varphi_i(\mathbf{r})\varphi_j(\mathbf{r}), \tag{3}$$

where P_{ij}^k is a function of P_{ij}. On this basis, Walker and Mezey reconstructed protein structures using a set of 21 different fragments (Walker & Mezey, 1994) such as carbonyl, amine or ammonium, carboxylate groups *etc.*

Table 1
PASA 3-21G $w_{a,i}$ (no units) and $\zeta_{a,i}$ (bohr^{-2}) coefficients for selected atoms.

See http://iqc.udg.es/cat/similarity/ASA/funcset.html for more precision (Amat & Carbó-Dorca, 1997).

	C	N	O	S
w_1	0.507950	0.482876	0.487933	0.432988
ζ_1	0.206511	0.272475	0.347803	0.222102
w_2	0.188827	0.258506	0.286441	0.453579
ζ_2	0.555343	0.733120	0.988549	4.10644
w_3	0.176511	0.148777	0.125730	0.0517120
ζ_3	7.35320	10.0967	13.0164	46.0626
w_4	0.112597	0.0980039	0.0893239	0.0549096
ζ_4	22.0876	30.4880	39.3558	142.191
w_5	0.0141154	0.0118374	0.0105727	0.00681092
ζ_5	107.346	149.788	195.531	711.403

Popelier *et al.* (2003) have extended the AIM approach to the analysis of the Laplacian of the valence charge density. Critical points are computed from molecular wavefunctions and then connected to build the so-called two-dimensional L-graph. From a set of 31 small organic molecules, the authors observed that, with basis sets of triple-ζ quality or higher, 16 invariant (transferable) common motifs assigned to functional groups could be extracted.

Periodic Nodal Surfaces (PNS) are especially used to partition space in high and low ED regions. These surfaces are generated by Fourier summations involving only a small set of reflections in reciprocal space. More precisely, reflections are chosen near the reciprocal-space vector $\mathbf{h} = 0$. From the resulting ED distribution, three-dimensional periodic surfaces can be extracted (von Schnering & Nesper, 1991). Applications are found in the field of phase-transition characterization of inorganic crystalline materials through the analysis of the PNS symmetry elements (Leoni & Nesper, 2000, 2003) or to facilitate structural solution of powder diffraction data as proposed by Brenner *et al.* (1997, 2002) for zeolites and tripeptides.

In the present paper, a method to partition protein structures into fragments as a function of the ED smoothing degree is presented. It is based on a decomposition procedure of promolecular ED distributions presented previously by Leherte *et al.* (2003). At any smoothing degree of an ED distribution function, local maxima are located and their corresponding fragment content is determined. Although the smoothing degree is correlated with the overall isotropic temperature factor rather than with the crystallographic resolution, the results of such a protein structure analysis might be seen as a guide to define resolution-dependent protein fragments that can be fitted around high-density centres in ED maps.

In the present work, promolecular representations were calculated using two different models: the Promolecular Atomic Shell Approximation (PASA) representation (Amat & Carbó-Dorca, 1997), which assumes a promolecular ED distribution as a summation over three-dimensional atomic Gaussian-type ED distributions, and a crystallography-based model, which involves the so-called Cromer–Mann coefficients for modelling atomic scattering factors (Cromer & Mann, 1968). The locations of the local maxima found in the ED distributions are also compared with the residue centres defined by Guo *et al.* (1999) in their approach to calculating low-resolution protein ED maps using 'globs' rather than atoms.

In the next section, we describe how to calculate smoothed PASA and crystallography-derived ED distributions as well as PASA-derived scattering factors. The hierarchical merging/clustering algorithm for molecular decomposition is then presented. Results are then detailed for protein structure 4pti, a 58-residue sequence, in its native conformation and in a hypothetical extended geometry.

2. Theoretical background

2.1. Promolecular atomic shell approximation

Promolecular models have often been shown to be a very good approximation level to model electron-density distributions in chemical bond analysis or molecular-similarity applications (for example, Tsirelson, Abramov *et al.*, 1998; Tsirelson, Avilov *et al.*, 1998; Gironés *et al.*, 1998; Mitchell & Spackman, 2000; Gironés *et al.*, 2001; Downs *et al.*, 2002; Bultinck *et al.*, 2003). In the Promolecular Atomic Shell Approximation (PASA) approach, a promolecular ED distribution ρ_M is calculated as a weighted summation over atomic ED distributions ρ_a, which are described in terms of series of squared 1*s* Gaussian functions fitted from atomic basis-set representations (Amat & Carbó-Dorca, 1997),

$$\rho_a(\mathbf{r}-\mathbf{R}_a) = Z_a \sum_{i=1}^{5} w_{a,i}\left[\left(\frac{2\zeta_{a,i}}{\pi}\right)^{3/4}\exp(-\zeta_{a,i}|\mathbf{r}-\mathbf{R}_a|^2)\right]^2, \quad (4)$$

where Z_a and $\mathbf{R}_a$ are the atomic number and the position vector of atom a, respectively. $w_{a,i}$ and $\zeta_{a,i}$ are the fitted parameters as reported in Amat & Carbó-Dorca (1997). ρ_M is then calculated as

$$\rho_M = \sum_a \rho_a, \quad (5)$$

where there is a change of notation with respect to the original work (Amat & Carbó-Dorca, 1997), in which $\rho_M = \sum_a Z_a\rho_a$. As examples, $w_{a,i}$ and $\zeta_{a,i}$ coefficients fitted from the atomic 3-21G basis set for selected atom types are reported in Table 1.

In the present approach, to generate smoothed three-dimensional functions, an ED map is a deformed version of ρ_M that is directly expressed as the solution of the diffusion equation according to the formalism presented by Kostrowicki *et al.* (1991),

$$\rho_{a,t}(\mathbf{r}-\mathbf{R}_a) = Z_a \sum_{i=1}^{5} \alpha_{a,i}(1+4\beta_{a,i}t)^{-3/2}\exp\left(\frac{-\beta_{a,i}|\mathbf{r}-\mathbf{R}_a|^2}{1+4\beta_{a,i}t}\right), \quad (6)$$

where

$$\beta_{a,i} = 2\zeta_{a,i},$$
$$\alpha_{a,i} = w_{a,i}\left(\frac{\beta_{a,i}}{\pi}\right)^{3/2},$$
$$\sum_{i=1}^{5} w_{a,i} = 1. \qquad (7)$$

In this context, the smoothing parameter t is seen as the product of a diffusion coefficient with time. On a similar basis, Duncan & Olson (1993) also generated smoothed molecular surfaces by convolving the density function with a three-dimensional Gaussian function of appropriate variance.

In the next two subsections, atomic scattering factors are first derived from the PASA description of an ED distribution. The resulting mathematical relationships are then used to generate an analytical expression of a promolecular ED distribution function based on the well known Cromer–Mann coefficients.

2.2. Atomic scattering factors

The intensity of X-rays diffracted by a crystalline structure is proportional to the modulus of their corresponding structure factor $F(\mathbf{h})$,

$$F(\mathbf{h}) = \sum_{j=1}^{\text{nat}} f_j \exp\left[-B_j\left(\frac{\sin\theta}{\lambda}\right)^2\right]\exp(2\pi i\mathbf{h}\cdot\mathbf{r}_j), \qquad (8)$$

where f_j and B_j are the atomic scattering factor and the isotropic temperature factor of atom j, respectively, 2θ is the angle between the diffracted and the primary beams of wavelength λ and $\mathbf{h}$ is a reciprocal-space vector. Within the crystallographic approach, the ED distribution function $\rho(\mathbf{r})$ is calculated as the Fourier Transform (FT) of $F(\mathbf{h})$,

$$\rho(\mathbf{r}) = \frac{1}{V}\sum_{\mathbf{h}=-\infty}^{+\infty} F(\mathbf{h})\exp(-2\pi i\mathbf{h}\cdot\mathbf{r}). \qquad (9)$$

In practice, the number of known structure factors occurring in (9) is not infinite and varies with resolution. For a set of reflections characterized by $\theta_{\max}$, the maximum value of θ, the resolution $d_{\min}$ is defined by Bragg's law as

$$\frac{\sin\theta_{\max}}{\lambda} = \frac{1}{2d_{\min}}. \qquad (10)$$

In the crystallographic package *XTAL* (Hall *et al.*, 2002), the atomic scattering factors $f_{\text{CM}}(s)$ are fitted by the so-called Cromer–Mann expression which involves nine coefficients $\{a\}$, $\{b\}$ and c,

$$f_{\text{CM}}(s) = \sum_{i=1}^{4} a_i \exp(-b_i s^2) + c, \qquad (11)$$

with $s = \sin\theta/\lambda$ and $f(0) = Z_a$. Selected Cromer–Mann coefficients, as implemented in the *XTAL* package (Hall *et al.*, 2002), are presented in Table 2.

For a given atom type, the relation between $f(\mathbf{h})$ and the atomic ED distribution function is (Ladd & Palmer, 2003),

$$f(\mathbf{h}) = \int \rho(\mathbf{r})\exp(2\pi i\mathbf{h}\cdot\mathbf{r})\,\mathrm{d}\mathbf{r} \qquad (12)$$

or, for spherical symmetry,

$$f(s) = 4\pi\int_0^{\infty}\rho(r)\frac{\sin(4\pi sr)}{4\pi sr}r^2\,\mathrm{d}r. \qquad (13)$$

The application of (13) to the analytical expression of ρ_{PASA} (6) allows the establishment of a mathematical expression for the corresponding atomic scattering factor of atom a,

$$f_{\text{PASA}}(s) = Z_a\sum_{i=1}^{5} w_{a,i}\exp\left[-\frac{2\pi^2(1+8\zeta t)s^2}{\zeta_{a,i}}\right]. \qquad (14)$$

A comparison of the scattering-factor values calculated with (14) using $t = 0.0$ bohr^2 (coefficients given in Table 1) with the Cromer–Mann expression (equation 11, coefficients given in Table 2) for a given atom type shows extremely slight differences between the two kinds of atomic scattering factors f.

2.3. Promolecular electron-density distribution based on Cromer–Mann coefficients

In the present subsection, the inverse approach is applied in order to generate ED distributions ρ_{CM} from the Cromer–Mann atomic scattering factors.

The analytical expression of f_{CM} given in (11) was modified into a sum over four Gaussian terms only,

$$f_{\text{CM}'}(s) = \sum_{i=1}^{4} a'_{a,i}\exp(-b'_{a,i}s^2). \qquad (15)$$

This step was achieved using the program *ODRPACK* (Boggs *et al.*, 1992), which was configured so as to find a solution using an ordinary least-squares procedure with user-supplied derivatives. The resulting coefficients, $w'_{a,i}$ and $\zeta'_{a,i}$, were obtained by fitting 300 $f_{\text{CM}}(s)$ values regularly distributed in the range $0 \le \sin\theta/\lambda \le 1.5$ Å^{-1}. These coefficients are reported in Table 3 together with the values of the quadratic deviation r.m.s. that provides information about the quality of the fits for atom types C, N, O and S. Thus,

$$\rho_{\text{CM}'}(r) = \sum_{i=1}^{4} a'_{a,i}\left(\frac{4\pi}{b'_{a,i}}\right)^{3/2}\exp\left(-\frac{4\pi^2 r^2}{b'_{a,i}}\right), \qquad (16)$$

where, by analogy with (4),

$$a'_{a,i} = Z_a w'_{a,i},$$
$$b'_{a,i} = \frac{2\pi^2}{\zeta'_{a,i}}. \qquad (17)$$

A comparison of the ED distributions calculated using the PASA approximation (4) and the Cromer–Mann-derived expression (16) shows that extremely slight differences in the scattering factors f lead, however, to large differences in the ED distributions (Fig. 1). Both PASA and CM′ ED models will thus be used in the present work.

The smoothing of $\rho_{\text{CM}'}$ can be performed by setting

$$b'_{a,i} = \frac{2\pi^2(1+8\zeta'_{a,i}t)^2}{\zeta'_{a,i}}. \qquad (18)$$

At this point, it is interesting to determine the link between t, the smoothing factor, and s, the crystallographic resolution. A combination of (8) and (15) considering (18) shows that t is

Table 2
Cromer–Mann scattering-factor coefficients, a (e^-), b ($Å^{-2}$) and c (e^-), as implemented in the program *XTAL* (Hall *et al.*, 2002).

	a_1	b_1	a_2	b_2	a_3	b_3	a_4	b_4	c
C	2.31	20.8439	1.02	10.2075	1.5886	0.5687	0.865	51.6512	0.2156
N	12.2126	0.0057	3.1322	9.8933	2.0125	28.9975	1.1663	0.5826	−11.529
O	3.0485	13.2771	2.2868	5.7011	1.5463	0.3239	0.867	32.9089	0.2508
S	6.9053	1.4679	5.2034	22.2151	1.4379	0.2536	1.5863	56.172	0.8669

Table 3
Modified PASA-type $w'_{a,i}$ (no units) and $\zeta'_{a,i}$ ($bohr^{-2}$) coefficients and corresponding r.m.s. as obtained using the program *ODRPACK* (Boggs *et al.*, 1992) to fit the modified Cromer–Mann parameters onto the original expression as defined in *XTAL* (Hall *et al.*, 2002).

	C	N	O	S
w'_1	0.2393	0.287387	0.132836	0.100089
ζ'_1	0.12966	0.190588	0.179536	0.0987856
w'_2	0.44569	0.447482	0.415172	0.324466
ζ'_2	0.36378	0.558535	0.461571	0.249377
w'_3	0.04602	0.126283	0.232106	0.440556
ζ'_3	2.739	8.54393	1.08496	3.80834
w'_4	0.26854	0.138076	0.219764	0.134873
ζ'_4	13.722	29.7433	21.8289	47.8362
R.m.s.	1.5412×10^{-3}	6.75425×10^{-6}	5.86318×10^{-4}	1.34458×10^{-4}

actually correlated with B, the overall isotropic temperature factor,

$$\left[\sum_{i=1}^{4} a'_{a,i} \exp(-b'_{a,i} s^2)\right] \exp(-Bs^2) = \left\{\sum_{i=1}^{4} a'_{a,i} \exp[-b'_{a,i}(1 + 8\zeta'_{a,i} t)s^2]\right\}. \quad (19)$$

Since $B = 8\pi^2 u$, where u is the mean-square atomic displacement, it follows that $u = 2t$. This equality can be verified by calculating ED distributions using the program *XTAL* (Hall *et al.*, 2002) at a given crystallographic resolution value using, either $f_{CM'}$ calculated at $t = 0$ and $u \neq 0$ or at $t = u/2$ and $u = 0$ (in practice, the very small value $u = 0.001$ $Å^2$ was selected). ED results are indeed identical.

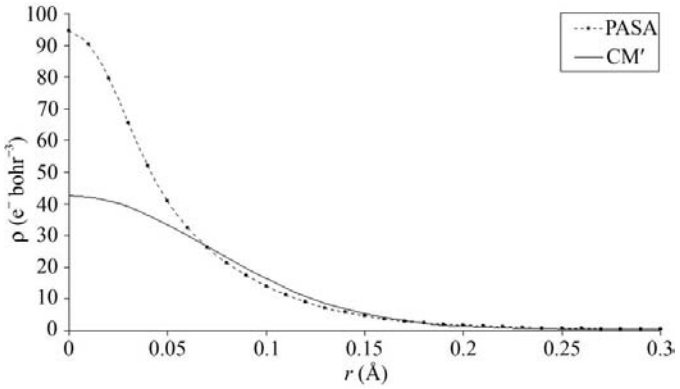

Figure 1
Atomic promolecular ED distributions calculated using the PASA representation (4) and the modified Cromer-Mann atomic scattering factors (16).

2.4. Hierarchical merging/clustering algorithm

In order to follow the pattern of local maxima (peaks) in a molecular ED distribution as a function of the smoothing parameter t, the algorithm described by Leung *et al.* (2000) was implemented. This algorithm was originally established to cluster data by modelling the blurring effect of lateral retinal interconnections based on scale-space theory. In scale-space theory, it is considered that an image $f(\mathbf{r})$ is embedded into a continuous family $F(\mathbf{r}, t)$ of gradually smoother versions of it. The original image corresponds to the scale $t = 0$ and increasing the scale simplifies the image. The work of Leung and coworkers was adapted to decompose a molecular structure from its ED distribution function $F(\mathbf{r}, t)$ $[\rho(\mathbf{r}, t)]$ (Leherte *et al.*, 2003). The various steps of the resulting merging/clustering algorithm are as follows.

(i) At scale $t = 0$, each atom of a molecular structure is considered as a local maximum (peak) of the promolecular ED distribution function. They are consequently considered as the starting points of the merging procedure described below.

(ii) As t increases from 0.0 to a given maximal value, each peak moves continuously along a gradient path to reach a location in the three-dimensional space where $\nabla\rho = 0$. From a practical point of view, this consists of following the trajectory of the peaks obtained at t on the ED distribution surface calculated at $t + \Delta t$,

$$\mathbf{r}_{\text{peak}}(t) = \mathbf{r}_{\text{peak}}(t - \Delta t) + \frac{\Delta}{\rho_{\text{peak}}(t - \Delta t)} \nabla\rho_{\text{peak}}(t). \quad (20)$$

The trajectory search is stopped when $\nabla\rho_{\text{peak}}(t)$ is lower or equal to a limit value, grad_{lim}. Once all peak locations are found, close peaks are merged if their interdistance is lower than the initial value of $\Delta^{1/2}$. The procedure is repeated for each selected value of t.

If the initial Δ value is too small to allow convergence towards a local maximum of the ED within the given number of iterations, its value is doubled (a scaling factor that is arbitrarily selected) and the procedure is repeated until final convergence.

(iii) The procedure can be carried out until the whole set of maxima becomes one single point. This would be the ultimate stopping criteria of the merging procedure. In this case, the final point that would be obtained will correspond to the whole molecular structure.

This approach is inverse of that presented by Glick and coworkers (Glick, Grant *et al.*, 2002; Glick, Robinson *et al.*, 2002), who developed a method for ligand-binding site identification on a protein through the multiscale concept. In their work, a hierarchy of models generated using a k-mean clustering algorithm for the ligand under consideration is established starting from the lowest resolution representation of the

ligand, *i.e.* one single point located at the mean position of the ligand atoms.

The results obtained using the present algorithm can be interpreted in terms of dendrograms. Visual results were generated using the web version of the program *Phylodendron* (Gilbert, 1996). Input data were written in the adequate format using *DENDRO* (Dury, 2002), a home-made program implemented using Delphi, an object-oriented programming language that allows the representation and processing of data in terms of classes of objects.

3. Applications

The hierarchical merging/clustering algorithm that is described in the theoretical section does not require any calculation of the ED maps. It is based solely on the knowledge of the analytical expression of the promolecular ED function and its first derivative. The algorithm was applied to the structure of protein 4pti as retrieved from the PDB (Berman *et al.*, 2000). This protein consists of 58 amino-acid residues and crystallizes in space group $P2_12_12_1$, with unit-cell parameters a = 43.1, b = 22.9, c = 48.6 Å, α = 90.0, β = 90.0, γ = 90.0°. The primary structure is Arg1-Pro2-Asp3-Phe4-Cys5-Leu6-Glu7-Pro8-Pro9-Tyr10-Thr11-Gly12-Pro13-Cys14-Lys15-Ala16-Arg17-Ile18-Ile19-Arg20-Tyr21-Phe22-Tyr23-Asn24-Ala25-Lys26-Ala27-Gly28-Leu29-Cys30-Gln31-Thr32-Phe33-Val34-Tyr35-Gly36-Gly37-Cys38-Arg39-Ala40-Lys41-Arg42-Asn43-Asn44-Phe45-Lys46-Ser47-Ala48-Glu49-Asp50-Cys51-Met52-Arg53-Thr54-Cys55-Gly56-Gly57-Ala58. H atoms were not added to the structure. The decomposition of the protein structure was achieved at t values ranging from 0 to 2.5 bohr^2, *i.e.* B = 0–110.6 Å^2, with a step of 0.05 bohr^2. The initial value $\Delta_{\rm init}$ was set equal to 0.0001 bohr^2 and grad$_{\rm lim}$ to 0.0001 e$^-$ bohr^{-4}. The value of Δ was doubled if convergence was not observed after 2000 iterations. Five different decomposition calculations were performed: (i) a PASA-based partitioning of the 4pti structure using space-group and periodicity information; (ii) a PASA-based partitioning of an isolated 4pti structure in its native conformation; (iii) a CM′-based partitioning of an isolated 4pti structure in its native conformation; (iv) a PASA-based partitioning of an isolated hypothetical extended geometry of the 4pti sequence; (v) a CM′-based partitioning of an isolated hypothetical extended geometry of the 4pti sequence. The calculations provided, at each value of t, the local maxima (peaks) of the ED and the corresponding fragment contents. They took 67 h 49 min, 9 h 51 min, 8 h 11 min, 4 h 10 min and 3 h 36 min, respectively, on a Pentium III 1 GHz processor. The hypothetical extended geometry of the initial 4pti amino-acid sequence was generated in order to check the influence of the protein conformation on the decomposition results. This single strand was obtained using the program *SwissPDBViewer* (Guex & Peitsch, 1997).

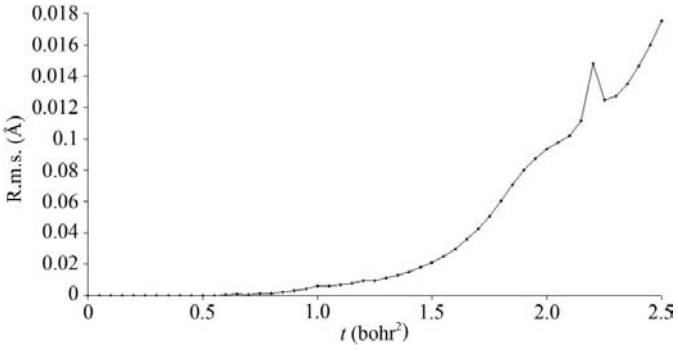

Figure 2
R.m.s. calculated over distances between corresponding peaks observed in the two PASA ED distributions of protein structure 4pti in its native conformation, calculated with and without crystal-packing information.

3.1. Influence of the crystal packing

The first two calculations, (i) and (ii), provided almost identical results in terms of the number and locations of the ED local maxima (Fig. 2). Fig. 2 shows values of an r.m.s. function computed over all distances between corresponding peaks in the two calculations at each smoothing value t. It is observed that as t increases, the peak locations are separated by larger distances. This is expected since at very high resolution values, local maxima are essentially located on atoms, while at higher smoothing degrees peaks are representative of larger regions of space and are thus more susceptible to being affected by the crystal packing. However, even at t = 2.5 bohr^2, the r.m.s. stays below 0.02 Å. A discontinuity is observed at t = 2.2 bohr^2, with r.m.s. = 0.015 Å which arises from a local maximum that is separated by a distance of 0.093 Å between the crystalline (i) and the isolated (ii) versions. In the isolated case, this peak corresponds to a fragment composed of seven atoms, $(C-O-CB-OG1-OG2)_{Thr32}(N-CA)_{Phe33}$, while in the crystalline version it contains only four atoms: $(C-O-CB)_{Thr32}(N)_{Phe33}$. On the whole, the number of local maxima is identical, with or without crystal-packing consideration, and both decompositions lead to identical fragments at each t value. The r.m.s. calculated over the values of ρ at the peak locations also evolves progressively from 0 to the very small value of 2.69×10^{-4} e$^-$ bohr^{-3} at t = 2.5 bohr^2. Owing to such small differences between the results obtained using the crystalline packing information and the isolated molecule, only these last conditions were considered further in the present work. All other calculations were thus carried out without the time-consuming consideration of the crystalline periodic information, especially for the artificially built strand used in calculations (iv) and (v).

3.2. Analysis of the isolated native conformation

In this subsection, the t-dependent hierarchical description of the ED peaks is detailed and the corresponding fragment contents are compared, when possible, with literature data at specific crystallographic resolution values. A complete representation of the decomposition results is too large to be shown here. However, part of the dendrogram obtained from calculation (ii) is shown in Fig. 3 for the first eight amino-acid residues (*i.e.* 67 atoms). It is seen that the merging first occurs between the C and O atoms of the backbone carbonyl groups

at $t = 0.45$ bohr^2. Between $t = 0.50$ and 0.55 bohr^2 the merging of C and O atoms of Glu and Asp side chains occurs. The C^{β}—S fragment of Cys5 (fragment IX in Fig. 4) appears at $t = 0.65$ bohr^2, as well as that of Cys30. The chain atoms of Cys51 and Cys14 are merged at $t = 0.60$ and 0.70, respectively. Thus, below this last value of t, the fragments that were obtained are similar in size (the atomic content is one to three non-H atoms) to most of the fragments used by Walker and Mezey in their MEDLA application to protein structures (Walker & Mezey, 1994). At the value of $t = 0.9$ bohr^2, all Pro five-membered rings are fully merged (except for Pro2, which is clustered with Arg1) and the hierarchical structure of these three merging patterns observed up to that point is identical. It can also be seen in Fig. 4 that the value $t = 1.5$ bohr^2 allows the partitioning of the protein structure into parts located either on the backbone or on the side chains of the protein. This is actually valid for the entire protein sequence. At this smoothing value, 43% of the backbone fragments correspond to the merging of the —(C═O)—N—C_{α} backbone atoms and 29% to the merging of —(C═O)—N—C_{α}—C^{β} atoms. Exceptions are listed: Pro shows a different behaviour in that some backbone and side-chain atoms are merged; the Pro backbone fragments can be composed of the carbonyl atoms only as in Pro8 (fragment XV in Figs. 3 and 4); the first two residues, Arg1 and Pro2, have backbone and ring atoms, respectively, that are merged into one single point at $t = 1.25$ (fragment I in Figs. 3 and 4); the side chains of Ile18 and Ile19 are clustered into a single fragment together with some backbone atoms of both residues; the Lys backbone fragments are composed of —(C═O)—N—C^{α}—C^{β}—C^{γ} atoms.

At $t = 1.5$ bohr^2, most of the side chains are represented by one maximum located at an average distance of 0.80 Å from the side-chain COMs. Exceptions are Gly and Ala residues which do not present any side-chain peaks, Val with only one peak for the whole residue, the two Ile which are merged into a single fragment, Pro and Ser (Table 4). The single peak of Val is formed at $t = 1.1$ bohr^2. The fragments observed for each type of amino-acid residue are described in Table 4 together with the average distance with

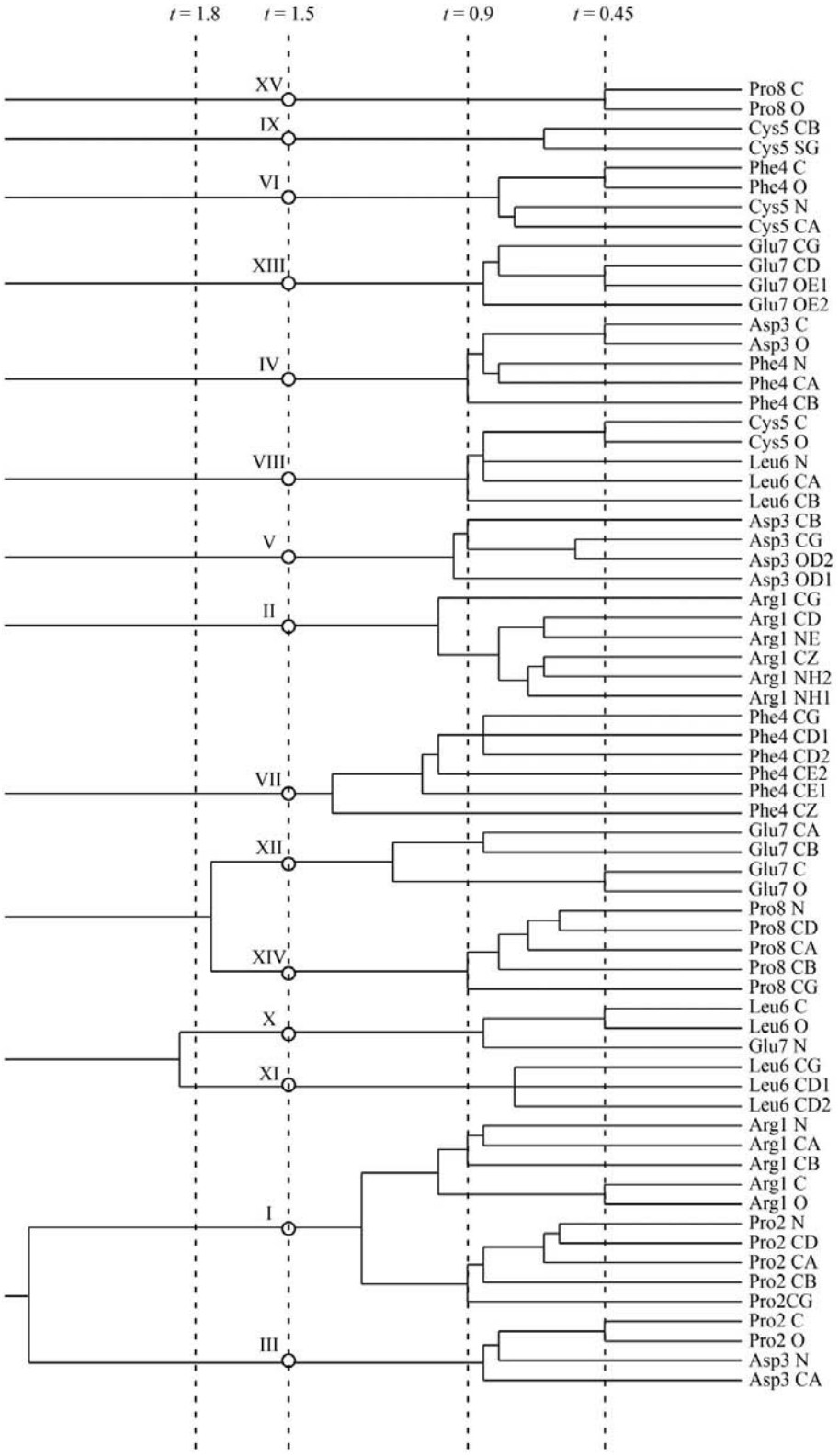

Figure 3
Dendrogram depicting the results of the hierarchical merging/clustering algorithm applied to the PASA ED distribution of protein structure 4pti in its native conformation. Only the first eight residues are shown in this figure. Results at various values of t (0.45, 0.9, 1.5 and 1.8 bohr^2) are emphasized using vertical lines.

respect to the corresponding side-chain COM. On the whole, this corresponds to the decomposition observed in various critical point analysis of protein structures carried out at resolution values between 2.85 (Becue *et al.*, 2003) and 3 Å (Leherte *et al.*, 1997) and to the globbic description level proposed by Guo *et al.* (1999) in which backbones and side-chain globs are located on C^{α}(C═O)N and side-chain COMs, respectively. In the cases of the four Tyr side chains, a single fragment is formed at t = 1.65 and 1.70 bohr^2 for Tyr10 and Tyr23 and for Tyr21 and Tyr35, respectively. It might thus be thought that t = 1.7 bohr^2 is a more appropriate value for a one-to-one correspondence between ED peaks and residue COMs; however, at t values larger than 1.5 bohr^2 side-chain groups and backbone fragments begin to be merged. For example, at t = 1.65 bohr^2 the side-chain and carbonyl groups of each Thr and the N—C^{α} group of their next residue are merged. Also, at t = 1.8 and 2.0 bohr^2, the two disulfide bridges are formed, *i.e.* side-chain groups are merged in pairs. A display of the ED distributions and their local maxima obtained at t = 0.45, 0.9, 1.5 and 2.5 bohr^2 is shown in Fig. 5 for the amino-acid residue sequence 1–8.

A comparison between the local maxima obtained with the original PASA coefficients and the CM′ coefficients is illustrated in Fig. 6 through the values of the square root of the quadratic deviation (r.m.s.) between the locations of corresponding peaks. The figure shows that the number of local maxima and their location are similar, in fact almost identical, from t = 0–0.75 bohr^2 and above t = 1.45 bohr^2, with r.m.s. below 0.01 Å. A rise in the values of r.m.s. between t = 0.75 and 1.45 is a consequence of the different number of local maxima found in the PASA and CM′ ED distributions. Fig. 7 illustrates the variation of the number of peaks as a function of t. It can be seen that the total number of peaks does not differ between both PASA and CM′ promolecular representations. The total number of peaks is relatively constant up to t = 0.4 bohr^2 and for values of t > 1.5 bohr^2, while the total number of points present drastic changes from about t = 0.4 to 1.0 bohr^2 owing to first the formation of C═O clusters and then of large fragments. Up to t = 0.4 bohr^2 (B = 17.7 Å^2), all protein atoms can be detected in the ED distributions. The largest difference δ in the number of peaks between the PASA and CM′ representations occurs at t = 0.75–0.8 bohr^2, where δ = 4. At t = 0.75 bohr^2, all extra CM′ peaks arise owing to the presence of small fragments composed of one to three atoms, *i.e.* $[(N)_{Cys5}]$, $[(CA)_{Cys5}]$, $[(CE1-CE2-CZ)_{Tyr23}]$, $[(OH)_{Tyr23}]$, $[(N)_{Cys51}]$, $[(CA)_{Cys51}]$, $[(N)_{Arg53}]$ and $[(CA)_{Arg53}]$, which are replaced by larger clusters in the PASA ED distribution, *i.e.* $[(N-CA)_{Cys5}]$, $[(CE1-CE2-CZ-OH)_{Tyr23}]$, $[(N-CA)_{Cys51}]$ and $[(N-CA)_{Arg53}]$. At t = 1.5 bohr^2, all fragments but four are identical in both promolecular representations. The four exceptions are fragments $[(C-O)_{Tyr21}(N)_{Phe22}]$, $[(CA-CB-C-O)_{Phe22}(N-CA-CB)_{Tyr23}]$, $[(C-O)_{Gln31}(N)_{Thr32}]$ and $[(CA-C-O)_{Thr32}(N-CA)_{Phe33}]$ in the CM′ representation, which are respectively replaced by $[(C-O)_{Tyr21}(N-CA-CB)_{Phe22}]$, $[(C-O)_{Phe22}(N-CA-CB)_{Tyr23}]$, $[(C-O)_{Gln31}(N\text{-}CA)_{Thr32}]$ and $[(C-O)_{Thr32}(N-CA)_{Phe33}]$ in the PASA representation. It is interesting to note that at t = 1.5 bohr^2, the number of maxima is close to twice the number of amino-acid residues, *i.e.* 103 in both

Table 4
Protein side-chain fragments obtained at t = 1.5 bohr^2 using a hierarchical merging/clustering algorithm applied to the PASA ED distribution of protein structure 4pti in its native conformation without consideration of the crystal packing.

Distances with respect to the side-chain centres of mass are given in Å. Distances corresponding to exceptional cases are given in parentheses.

Amino acid	Fragment content (notation as in PDB)	Frequency	Distance
Arg	CG—CD—NE—CZ—NH1—NH2	5/6	1.17
	CD—NE—CZ—NH1—NH2	1/6	
Asn	CB—CG—OD1—ND2	3/3	0.28
Asp	CB—CG—OD1—OD2	2/2	0.16
Cys	CB—SG	6/6	0.57
Gln	CG—CD—OE1—NE2	1/1	0.65
Glu	CG—CD—OE1—OE2	2/2	0.60
Ile	The two neighbouring Ile are merged	2/2	(3.68)
Leu	CG—CD1—CD2	2/2	0.28
Lys	CD—CE—NZ	4/4	1.38
Met	CG—SD—CE	1/1	0.68
Phe	CB—CG—CD1—CD2—CE1—CE2—CZ	2/4	0.37
	CG—CD1—CD2—CE1—CE2—CZ	2/4	
Pro	N—CA—CB—CG—CD	3/4	(1.94)
	Pro2 merged with Arg1	1/4	
Ser	Side-chain and backbone atoms merged with neighbouring Ala backbone	1/1	(2.62)
Thr	CB—OG1—CG2	3/3	0.67
Tyr	CG—CD1—CD2 and CE1—CE2—CZ—OH (two fragments)	4/4	0.93
Val	There is only one peak for the whole residue	1/1	(3.66)

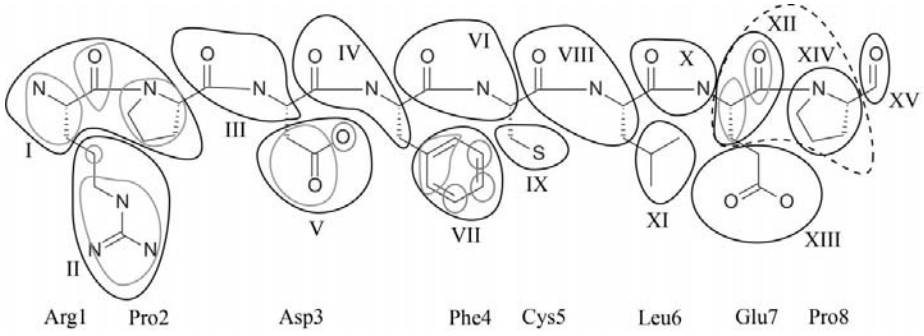

Figure 4
Two-dimensional molecular representation of part of the information contained in Fig. 3. Fragments observed at t = 0.9, 1.5 and 1.8 bohr^2 are shown using grey, black and dashed lines, respectively. Roman numerals indicate the peaks/fragments existing at t = 1.5 bohr^2, as in Fig. 3.

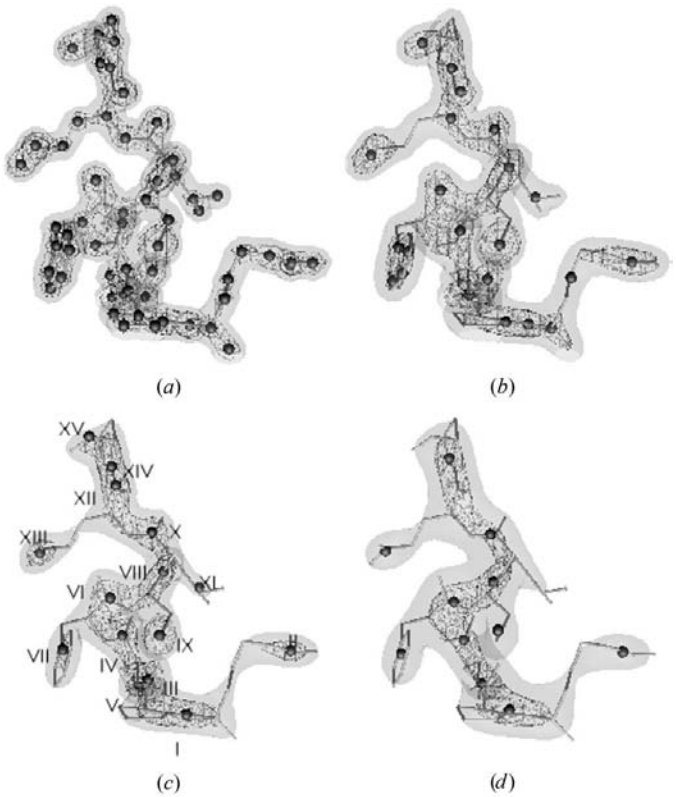

Figure 5
Three-dimensional representations of the first eight amino-acid residue sequence of protein 4pti in its native conformation (sticks) and corresponding local maxima (black spheres) observed in PASA ED distributions smoothed at t = 0.45, 0.9, 1.5 and 2.5 bohr^2. Isodensity contours are as follows. (*a*) t = 0.45 bohr^2: 0.2 (triangulated) and 0.125 e$^-$ bohr^{-3} (solid). (*b*) t = 0.9 bohr^2: 0.15 (triangulated) and 0.1 e$^-$ bohr^{-3} (solid). (*c*) t = 1.5 bohr^2: 0.125 (triangulated) and 0.1 e$^-$ bohr^{-3} (solid). (*f*) t = 2.5 bohr^2: 0.1 (triangulated) and 0.075 e$^-$ bohr^{-3} (solid). Roman numerals indicate the peaks/fragments obtained at t = 1.5 bohr^2, as in Figs. 3 and 4.

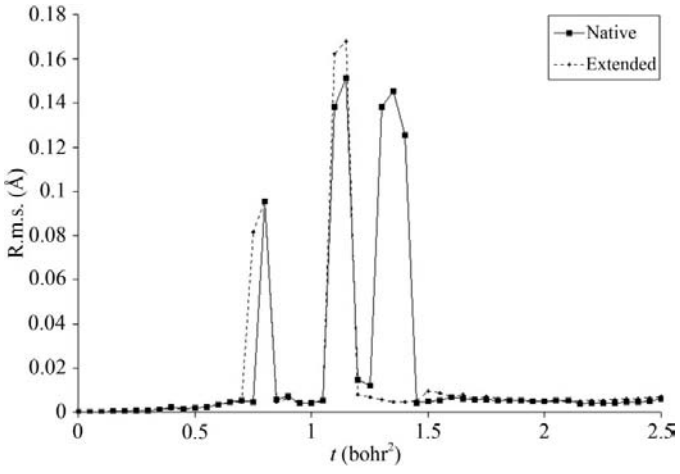

Figure 6
R.m.s. calculated over distances between corresponding peaks observed in the PASA and CM′ ED distributions of protein structure 4pti. Results are reported for the native and hypothetical extended conformations.

the PASA and CM′ approximations. This fits the general usage that there is one maxima per residue backbone and one per side chain, except for the six Ala and six Gly residues. At this smoothing value, it is observed that 61 peaks are close (within 1 Å) to a backbone atom or, more precisely, that 57 peaks are close, at an average distance of 0.197 Å, to C$^\alpha$(C═O)N COMs as defined by Guo *et al.* (1999). The distribution of the average distance d_{av} between peaks and C$^\alpha$(C═O)N COMs as a function of t is shown in Fig. 8, where it is again observed that both promolecular description levels provide extremely similar results. d_{av} is minimum at t = 1.8 (d_{av} = 0.180 Å) rather than 1.5 bohr^2 (d_{av} = 0.197 Å), but 1.8 bohr^2 corresponds to a reduced number of backbone fragments, *i.e.* 55 rather than 57. In fact, backbone fragments are completely formed at t = 0.9 bohr^2, a value from which the number of peaks located at a distance <1 Å from C$^\alpha$(C═O)N COMs is almost constant (Fig. 7).

3.3. Analysis of the isolated hypothetical extended conformation

In this subsection, the decomposition results obtained using the hierarchical merging/clustering algorithm are described

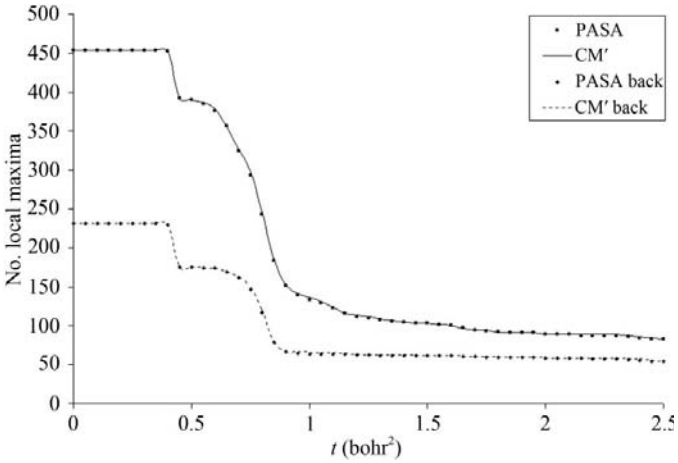

Figure 7
Total number of fragments (peaks) and numbers of peaks close to any backbone atoms within a distance of 1 Å in PASA and CM′ ED distributions of protein 4pti in its native conformation.

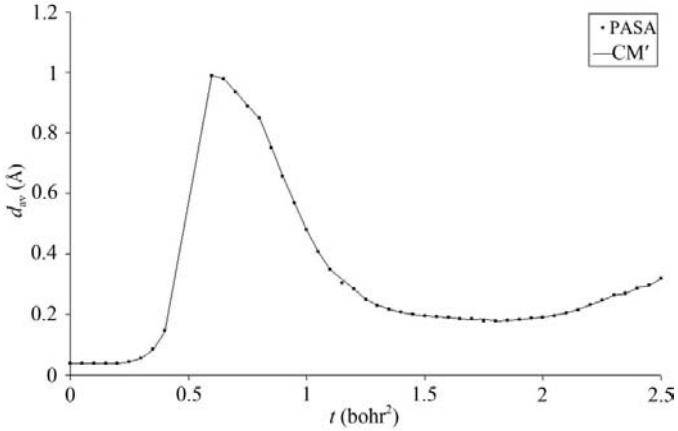

Figure 8
Average distance between ED peaks and C$^\alpha$CON centres of mass of protein 4pti in its native conformation.

for an hypothetical extended conformation of protein 4pti. Comparisons with results for the native geometry are presented in order to elucidate the influence of the protein conformation.

Fig. 9 depicts a three-dimensional representation of the 1–8 amino-acid sequence of 4pti together with the peaks obtained at the PASA description levels at t = 0.45, 0.9, 1.5, and 2.5 bohr^2. At the smoothing value t = 1.5 bohr^2, all but four fragments observed from the decomposition of $\rho_{\rm PASA}$ and $\rho_{\rm CM'}$ are identical and the resulting r.m.s. calculated over distances between corresponding peaks is equal to 0.0095 Å (Fig. 6). The four fragments that differ between both promolecular representations are $[(C-O)_{Arg17}(N-CA)_{Ile18}]$, $[(C-O-CB-CG1-CG2-CD)_{Ile18}(N-CA-CB-CG1-CG2-CD1)_{Ile19}]$, $[(C-O)_{Ala25}(N)_{Lys26}]$ and $[(C-O-CA-CB-CG)_{Lys26}(N-CA-CB)_{Ala27}]$ in the CM′ representations, which are respectively replaced by $[(C-O)_{Arg17}(N-CA-CB-CG1-CG2-CD1)_{Ile18}]$, $[(C-O)_{Ile18}(N-CA-CB-CG1-CG2-CD1)_{Ile19}]$, $[(C-O)_{Ala25}(N-CA-CB)_{Lys26}]$ and $[(C-O-CG)_{Lys26}(N-CA-CB)_{ALA27}]$ in the PASA representation. The largest values of r.m.s. occur at t = 0.8 and 1.15 bohr^2 and correspond to different fragment contents. At t = 0.8 bohr^2, $[N_{Arg20}]_{PASA}$ has its closest corresponding peak $[CA_{Arg20}]_{CM'}$ located at a distance of 1.44 Å, while at t = 1.15 bohr^2, $[CE2_{Phe22}]_{PASA}$ is at a distance of 1.63 Å from $[(CG-CD1-CD2-CE1-CE2-CZ)_{Phe22}]_{CM'}$, *i.e.* no real correspondence

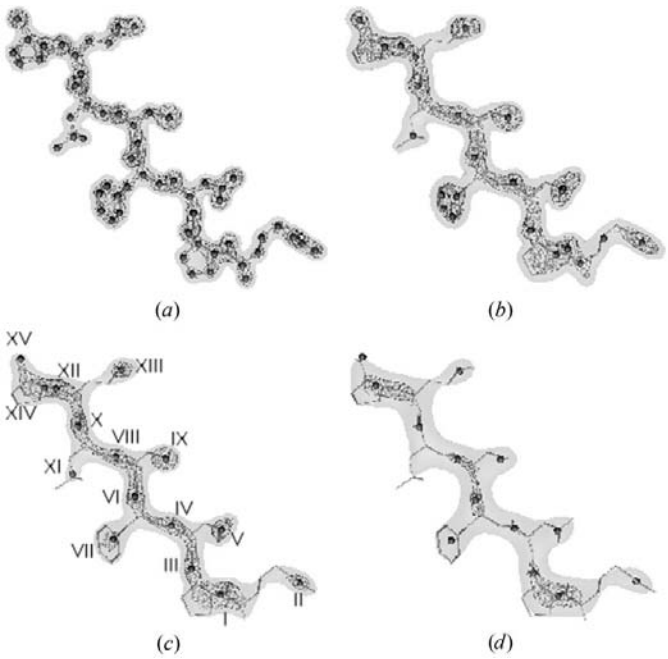

Figure 9
Three-dimensional representations of the first eight amino-acid residue sequence of protein 4pti in its hypothetical extended conformation (sticks) and corresponding peaks (black spheres) observed in PASA ED distributions smoothed at t = 0.45, 0.9, 1.5 and 2.5 bohr^2. Isodensity contours are as follows. (*a*) t = 0.45 bohr^2: 0.2 (triangulated) and 0.125 e$^-$ bohr^{-3} (solid). (*b*) t = 0.9 bohr^2: 0.15 (triangulated) and 0.1 e$^-$ bohr^{-3} (solid). (*c*) t = 1.5 bohr^2: 0.125 (triangulated) and 0.1 e$^-$ bohr^{-3} (solid). (*d*) t = 2.5 bohr^2: 0.1 (triangulated) and 0.075 e$^-$ bohr^{-3} (solid). Roman numerals establish a correspondence with peaks/fragments obtained for the native conformation, as in Fig. 5.

was found. Globally, the values of r.m.s. are lower on a larger range of t values than in the native conformation (Fig. 6), so a better agreement between the two ED models, PASA and CM′, is observed between the local maxima found for the hypothetical extended conformation, owing to the fact that side chains are separated in space and their atoms do not tend to merge together, especially at high values of t.

When comparing the number of peaks obtained for the hypothetical extended conformation and the native geometry, regardless of the promolecular description, PASA or CM′, the main differences are found to occur at t = 0.4 and 0.8 bohr^2. These differences arise from the merging of C═O atoms into fragments which are set at t = 0.4 bohr^2 for the extended conformation rather than at 0.45 bohr^2 for the native conformation. This can be correlated with the average C═O distance, which is slightly lower in the extended conformation (1.229 Å) than in the native conformation (1.254 Å). Thus, very small differences in bond lengths may locally affect the merging pattern. On the whole, in the extended conformation the C═O and $C^\alpha-N$ fragments are merged at t = 0.8 bohr^2, at which point this is not yet the case in the native conformation. At t = 1.5 bohr^2, the main differences between the native and extended conformations arise at the level of the fragment content rather than the fragment number. Except for four observed differences, all fragments are identical; 36% of the backbone fragments correspond to the merging of the $-(C{=}O)-N-C^\alpha$ backbone atoms and 24% to the merging of $-(C{=}O)-N-C^\alpha-C^\beta$ atoms.

The analysis of the decomposition patterns of both native and hypothetical extended geometries shows that the amino-acid residues have a similar decomposition pattern regardless of their position in the protein sequence. The decomposition patterns show that there is no merging before t = 0.4 bohr^2 and backbone fragments appear at about t = 0.9 bohr^2, while unique side-chain motifs are formed around t = 1.5 bohr^2. These three values of t correspond to a mean-square displacement u equal to 0.224, 0.504 and 0.840 Å^2, respectively (1 bohr = 0.529177 Å).

This is consistent with the conclusions raised by Matta & Bader (2002) about the high degree of transferability of geometrical parameters (bond lengths and angles) in amino acids, a prerequisite condition for transferability of their ED distribution. Matta and Bader also observed that although the main-chain group is greatly altered in passing from an isolated amino acid to its residue form owing to the peptidic bonds, there is only little change in the geometric parameters of the remaining main-chain group. Their conclusions were raised from the analysis of non-promolecular ED distributions, which can be *a priori* assumed to be even less transferable than promolecular properties.

In a final step to relate the decomposition results to literature data, a tentative correspondence was established between t and the crystallographic resolution d (rather than u) by visually comparing ED distributions calculated along the axis of a CO_2 molecule using (6) at various values of t and using *XTAL* at various crystallographic resolution values d. Owing to its linear geometry, this small molecule was selected

for simplicity of the visual results. The results are presented in Fig. 10, which shows a similar ED shape at $t = 0.45$ bohr^2 and $d = 2$ Å, $t = 0.8$ bohr^2 and $d = 2.5$ Å, and $t = 1.5$ bohr^2 and $d = 3$ Å. That last correlation is consistent with our fragment contents obtained at $t = 1.5$ bohr^2 and Guo's globbic representation of a protein structure at about 3 Å.

4. Conclusions and perspectives

A method for the hierarchical decomposition of a molecular structure, particularly a protein structure, based on its promolecular ED distributions is presented. The decomposition is achieved by following the trajectories of the atoms in progressively smoothed molecular ED distributions. Various patterns are observed as the smoothing degree t (correlated with B, the overall isotropic temperature factor) increases, with a very interesting situation at $t = 1.5$ bohr^2 ($B = 66.3$ Å^2), where the protein structure is clearly partitioned into backbone and side-chain fragments. At $t = 1.5$ bohr^2, one fragment is observed for each residue backbone, mainly composed of $-(C{=}O)-N-C_\alpha$ or a derivative, and one fragment for each residue side-chain, except for Gly and Ala (no fragment at all) and for Tyr (two fragments). These observations are consistent with several descriptions already proposed in the literature, such as the critical point and the globbic description levels of protein structures, both of which were established at a crystallographic resolution of about 3 Å.

The analysis of the decomposition patterns of both a native and a hypothetical extended geometry showed that the amino-acid residues have a similar decomposition pattern regardless of their position in the protein sequence; also, the decomposition pattern does not vary significantly with the protein conformation (at least up to $t = 2.5$ bohr^2), the promolecular description (Promolecular Atom Shell Approximation or Cromer–Mann) and the influence of the crystal packing.

The calculations presented in this paper were all achieved using ideal Gaussian-type ED distributions without any random noise. A statistical analysis of several factors such as the fragment content, the effect of conformation, crystal packing, solvent molecules, noise *etc.* probably needs to be considered for practical applications to protein crystallography. So far, the results have shown that crystal packing and protein conformation are not crucial in the hierarchical merging patterns that were obtained. Thus, it might be considered that the definition of fragments of a protein structure could be performed without any precise information regarding its three-dimensional geometry. The fragments thus obtained might be used to feed a dictionary of templates for model building with information about their hierarchy as a function of the smoothing degree or resolution. All these practical aspects constitute a further step which will require, in addition to model-building expertise, automation in the analysis of the merging/clustering results.

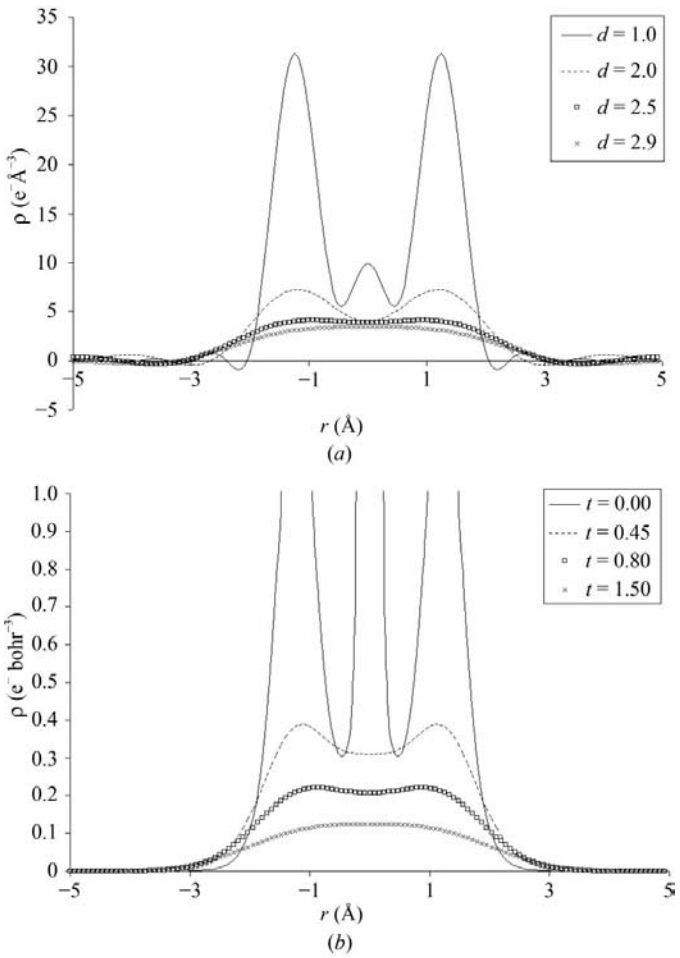

Figure 10
XTAL (*a*) and CM′ (*b*) ED distributions calculated for the CO_2 molecule at various values of d (Å) and t (bohr^2), respectively.

The author thanks Professors Fortier and Glasgow for their continuous interest, Professors Carbó-Dorca and Bultinck for fruitful discussions, Professor D. P. Vercauteren, Director of the Laboratoire de Physico-Chimie Informatique for discussions and a careful reading of the manuscript, L. Piela for discussions on analytical smoothing and L. Dury for the program *DENDRO*. The author also thank the referees for their detailed and useful comments. The FNRS-FRFC, the 'Loterie Nationale' (convention No. 2.4578.02) and the FUNDP are gratefully acknowledged for the use of the Interuniversity Scientific Computing Facility (ISCF) Center.

References

Amat, L. & Carbó-Dorca, R. (1997). *J. Comput. Chem.* **19**, 2023–2039.
Bader, R. F. (2001). *Theor. Chem. Acc.* **105**, 276–283.
Becue, A., Meurice, N., Leherte, L. & Vercauteren, D. P. (2003). *Acta Cryst.* D**59**, 2150–2162.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Boggs, P. T., Byrd, R. H., Rogers, J. E. & Schnabel, R. B. (1992). *ODRPACK* v.2.01. *Software for Weighted Orthogonal Distance Regression.* US Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, USA. http://www.netlib.org/odrpack/.
Brenner, S., McCusker, L. B. & Baerlocher, C. (1997). *J. Appl. Cryst.* **30**, 1167–1172.

Brenner, S., McCusker, L. B. & Baerlocher, C. (2002). *J. Appl. Cryst.* **35**, 243–252.
Bultinck, P., Carbó-Dorca, R. & Van Alsenoy, C. (2003). *J. Chem. Inf. Comput. Sci.* **43**, 1208–1217.
Cowtan, K. (2001). *Acta Cryst.* D**57**, 1435–1444.
Cromer, D. T. & Mann, J. B. (1968). *Acta Cryst.* A**24**, 321–324.
Downs, R. T., Gibbs, G. V., Boisen, M. B. Jr & Rosso, K. M. (2002). *Phys. Chem. Miner.* **29**, 369–385.
Duncan, B. S. & Olson, A. J. (1993). *Biopolymers*, **33**, 231–238.
Dury, L. (2002). *DENDRO*. Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium.
Fortier, S., Chiverton, A., Glasgow, J. I. & Leherte, L. (1997). *Methods Enzymol.* **277**, 131–157.
Gilbert, D. G. (1996). *Phylodendron, for Drawing Phylogenetic Trees*, v.0.8d. Indiana University, USA. http://iubio.bio.indiana.edu/soft/molbio/java/apps/trees/.
Gillet, V. J., Willett, P. & Bradshaw, J. (2003). *J. Chem. Inf. Comput. Sci.* **43**, 338–345.
Gironés, X., Amat, L. & Carbó-Dorca, R. (1998). *J. Mol. Graph. Model.* **16**, 190–196.
Gironés, X., Carbó-Dorca, R. & Mezey, P. G. (2001). *J. Mol. Graph. Model.* **19**, 343–348.
Glick, M., Grant, G. H. & Richards, W. G. (2002). *J. Med. Chem.* **45**, 4639–4646.
Glick, M., Robinson, D. D., Grant, G. H. & Richards, W. G. (2002). *J. Am. Chem. Soc.* **124**, 2337–2344.
Gopal, K., Pai, R., Ioerger, T. R., Romo, T. D. & Sacchettini, J. C. (2003). *Proceedings of the 15th Conference on Innovative Applications of Artificial Intelligence*, pp. 93–100. Menlo Park, CA, USA: AAAI Press.
Guex, N. & Peitsch, M. C. (1997). *Electrophoresis*, **18**, 2714–2723.
Guo, D.-Y., Blessing, R. H., Langs, D. A. & Smith, G. D. (1999). *Acta Cryst.* D**55**, 230–237.
Hall, S., du Boulay, D. & Olthof-Hazekamp, R. (2002). Editors. *The Gnu Xtal System of Crystallographic Software* v.3.7.2. http://xtal.sourceforge.net/.
Ioerger, T. R., Holton, T., Christopher, J. A. & Sacchettini, J. C. (1999). *Proceedings of the 7th Conference on Intelligent Systems for Molecular Biology*, edited by T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.-W. Mewes & R. Zimmer, pp. 130–138. Menlo Park, CA, USA: AAAI Press.
Ioerger, T. R. & Sacchettini, J. C. (2002). *Acta Cryst.* D**58**, 2043–2054.
Iwata, Y., Kasuya, A. & Miyamoto, S. (2002). *J. Mol. Graph. Model.* **21**, 119–128.
Jain, A. N. (2003). *J. Med. Chem.* **46**, 499–511.
Johnson, C. K. (1977). *ORCRIT. The Oak Ridge Critical Point Network Program.* Chemistry Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA.
Jones, T. A. (1978). *J. Appl. Cryst.* **11**, 268–272.
Jones, T. A. (1985). *Methods Enzymol.* **115**, 157–171.
Jones, T. A. & Kjeldgaard, M. (1997). *Methods Enzymol.* **277**, 173–207.
Jones, T. A. & Thirup, S. (1986). *EMBO J.* **5**, 819–822.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Kleywegt, G. J. & Jones, T. A. (1997). *Acta Cryst.* D**53**, 179–185.
Kostrowicki, J., Piela, L., Cherayil, B. J. & Scheraga, H. A. (1991). *J. Phys. Chem.* **95**, 4113–4119.
Krämer, A., Horn, H. W. & Rice, J. E. (2003). *J. Comput.-Aided Mol. Des.* **17**, 13–38.
Ladd, M. & Palmer, R. (2003). *Structure Determination by X-Ray Crystallography*. New York: Kluwer Academic Press/Plenum.
Lamzin, V. S. & Perrakis, A. (2000). *Nature Struct. Biol.* **7**, 978–981.
Leherte, L., Glasgow, J. I., Baxter, K., Steeg, E. & Fortier, S. (1997). *J. Artif. Intell. Res.* **7**, 125–159.
Leherte, L., Dury, L. & Vercauteren, D. P. (2003). *J. Phys. Chem. A*, **107**, 9875–9886.
Lemmen, C., Lengauer, T. & Klebe, G. (1998). *J. Med. Chem.* **41**, 4502–4520.
Leoni, S. & Nesper, R. (2000). *Acta Cryst.* A**56**, 383–393.
Leoni, S. & Nesper, R. (2003). *Solid State Sci.* **5**, 95–107.
Leung, Y., Zhang, J.-S. & Xu, Z.-B. (2000). *IEEE Trans. Pattern Anal. Machine Intell.* **22**, 1396–1410.
Levitt, D. G. (2001). *Acta Cryst.* D**57**, 1013–1019.
McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.
Matta, C. F. & Bader, R. F. W. (2002). *Proteins Struct. Funct. Genet.* **48**, 519–538.
Mezey, P. (1996). *Computational Chemistry – Reviews of Current Trends*, Vol. 1, edited by J. Leszczynski, pp. 109–137. Singapore: World Scientific.
Mitchell, A. S. & Spackman, M. A. (2000). *J. Comput. Chem.* **21**, 933–942.
Morris, R. J., Perrakis, A. & Lamzin, V. S. (2003). *Methods Enzymol.* **374**, 229–244.
Oldfield, T. (2002). *Acta Cryst.* D**58**, 487–493.
Pavelcik, F., Zelinka, J. & Otwinowski, Z. (2002). *Acta Cryst.* D**58**, 275–283.
Perrakis, A., Morris, R. J. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Popelier, P. L. A., Burke, J. & Malcolm, N. O. J. (2003). *Int. J. Quantum Chem.* **92**, 326–336.
Schnering, H. G. von & Nesper, R. (1991). *Z. Phys. B*, **83**, 407–412.
Swanson, S. M. (1994). *Acta Cryst.* D**50**, 695–708.
Terwilliger, T. C. (2003*a*). *Acta Cryst.* D**59**, 38–44.
Terwilliger, T. C. (2003*b*). *Acta Cryst.* D**59**, 45–49.
Tsirelson, V., Abramov, Y., Zavodnik, V., Stash, A., Belokoneva, E., Stahn, J., Pietsch, U. & Feil, D. (1998). *Struct. Chem.* **9**, 249–254.
Tsirelson, V. G., Avilov, A. S., Abramov, Yu. A., Belokoneva, E. L., Kitaneh, R. & Feil, D. (1998). *Acta Cryst.* B**54**, 8–17.
Turk, D. (1992). PhD Thesis, Technische Universität München, Germany.
Walker, P. D. & Mezey, P. G. (1994). *J. Am. Chem. Soc.* **116**, 12022–12032.