Journal of Microscopy, Vol. 228, Pt 2 2007, pp. 174–184
Received 24 January 2007; accepted 16 May 2007

Resolving the geometry of biomolecules imaged by cryo electron tomography

L. BONGINI*, D. FANELLI†, S. SVENSSON‡, M. GEDDA‡, F. PIAZZA# & U. SKOGLUND¶
**Dipartimento di Fisica, Università di Firenze, Florence, Italy*
†Theoretical Physics, School of Physics and Astronomy, University of Manchester, Manchester, U.K.
‡Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, Uppsala, Sweden
#Laboratoire de Biophysique Statistique ITP/SB, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
¶Cell and Molecular Biology Department, Karolinska Institutet, Stockholm, Sweden

Key words. Brownian dynamics, fuzzysets, immunoglobulin G, principle component analysis, protein data bank, volume images.

Summary

In this paper, we describe two methods for computerized analysis of cryo electron tomography reconstructions of biomolecules. Both methods aim at quantifying the degree of structural flexibility of macromolecules and eventually resolving the inner dynamics through automatized protocols. The first method performs a Brownian dynamics evolution of a simplified molecular model into a fictitious force field generated by the tomograms. This procedure enables us to dock the simplified model into the experimental profiles. The second uses a fuzzy framework to delineate the subparts of the proteins and subsequently determine their interdomain relations. The two methods are discussed and their complementarities highlighted with reference to the case of the immonoglobulin antibody. Both artificial maps, constructed from immunoglobulin G entries in the Protein Data Bank and real tomograms are analyzed. Robustness issues and agreement with previously reported measurements are discussed.

1. General background

Nowadays it is becoming increasingly evident that dynamics play a crucial role in the biological functions of macromolecules. Proteins exhibit a variable degree of structural flexibility that is related to their intrinsic ability of functioning as a molecular machine. Collective motions of domains are in fact believed to greatly enhance proteins' ability to bind other molecules. Flexible units might act as dynamical gates that govern the accessibility of specific sites and indirectly control the cascade of reactions triggered by a binding event. A satisfactory comprehension of protein functional dynamics is therefore a great challenge, at the frontier of biology and physics. In this perspective, a crucial requirement is to combine single molecule techniques allowing for individual particle imaging and automatic (computerized) tools to extract the relevant structural information from the obtained images.

Recent developments of single molecule detection techniques have made it possible to unravel the structural variability of macromolecular assemblies through specific experiments. Among others, electron tomography (ET) is often seen to be most promising for studying large macromolecular complexes within their cellular context (Sali *et al.*, 2003), but can also be successfully used to access the structure of individual biomolecules in solution. Hence, by analyzing a large gallery of reconstructed profiles, we can qualitatively inspect the inherent, large-scale, flexibility of the various molecular domains involved. To quantify the degree of variability as displayed by the available tomograms, it is important to develop automatic procedures for image analysis. Throughout the years, many methods for this purpose have been presented, mostly focusing on proteins in cellular context. Just recently, one issue of *Journal of Structural Biology* was even completely devoted to *Software tools for macromolecular microscopy* (Carragher *et al.*, 2007), a topic that includes also automatic image analysis methods. One commonly used strategy amounts to combine the relatively low-resolution ET data with high-resolution structures of proteins as determined by X-ray crystallography (or by single particle ET), i.e. docking of high- and low-resolution density maps. The docking is performed interactively or, alternatively,

Correspondence to: Stina Svensson. Tel: +46 18 4713465; fax: +46 18 553447; e-mail: stina@cb.uu.se

following more automatic procedures, see the survey by Wriggers and Chacó (2001). Several program packages for handling ET data have been developed, which includes docking possibilities, e.g. Situs by Wriggers *et al.* (1999). In general, it is indeed an interesting approach to virtually 'enhance' the ET resolution through a docking analysis. In practise, however, this method relies on the assumption that detailed maps of individual subunits are accessible through X-ray crystallography (or by single particle ET), which is not always the case. More importantly, current docking procedures assume rigid molecular structures, and do not accommodate for the degree of structural flexibility exhibited by several macromolecular specimens.

In this paper, we shall present two complementary methods, which will constitute an alternative approach to ET data analysis, and eventually provide an additional tool to shed light on the issue of proteins' dynamics. It is worth emphasizing that an ensemble made of structural parameters, each set describing a different spatial arrangement of a given molecule within the collection, will constitute a unique input to develop simplified, coarse-grained, models of the macromolecules under study. By further elaborating this structural information, we can aim at resolving the relevant physical interactions directly from the experimental measurements, as it is has been outlined by Bongini *et al.* (2004).

The first method discussed in the following allows to extract geometrical features from observed cryo electron tomography (Cryo-ET) profiles, once a simplified model of the geometry of the molecule is put forward. The method uses a Brownian dynamics algorithm to evolve the resulting mechanical model in a fictitious force field generated by the raw data density. In the end of this procedure, the model is dynamically adjusted to the experimental structure and consequently the intrinsic geometry resolved. This amounts to quantitatively characterize the mutual orientation of the units that are assumed to form the coarse-grained model through direct estimates of their linear dimensions and relative angles.

In general, it is however difficult to *a priori* propose a realistic description of a macromolecular assembly in terms of a collection of massive entities, by simple visual inspection. Indeed, such a process should be assisted and complemented by dedicated strategies that are able to decompose the 3D profiles in individual building blocks. Traditionally, from an image analysis point of view, the position of distinct subparts is recovered after having delineated the object of interest, in this case a macromolecule, from the surrounding background, i.e. by first applying a *crisp* segmentation algorithm. However, for Cryo-ET data, this approach is not the ideal one. Nevertheless it is still used, e.g. in Volkmann (2002) and Baker *et al.* (2006), where subunit identification is to some extent based on an initial segmentation into macromolecule and background. There are two reasons why we suggest a different strategy. Firstly, the imaged macromolecules are usually so small that, despite the high resolution, they will consist only of a limited number of image elements. Feature extraction based on a crisp segmentation then is nonrobust, as small errors in delineation of the actual borders can be dramatically magnified in the subsequent analysis. Secondly, the transmission electron microscopy (TEM) images used to reconstruct the tomogram are recorded in a low-dose setting to prevent radiation damages which results in low-contrast Cryo-ET images, i.e. there is actually no distinct border between object and background visible in the image. Conversely, to take full advantage of the information contained into the grey-level distribution of the imaged macromolecules, we can resort to other, more appropriate, strategies. In this specific case, we use a fuzzy segmentation instead of a crisp one, and base our analysis on the so called fuzzy object. In the fuzzy object, the grey-level in each element is proportional to the degree of membership the element has to the object. A method that exploits the automatic decomposition of fuzzy objects is proposed here and is shown to provide a robust decomposition of the experimental volumes. Once the building blocks are identified, all structural parameters relative to their mutual orientation can be extracted. All these are sensible information, of paramount importance when aiming at developing a realistic model of the macromolecule.

In this paper, we provide an account of the above two approaches. The methods are described separately in Section 4 and 5, respectively. We illustrate the methods by focusing on the case of the Immunoglobulin G (IgG) antibodies. In particular, we consider density maps constructed from the Protein Data Bank (PDB) models, Berman *et al.* (2000), as well as available Cryo-ET data sets. The reasons for choosing the IgG antibody as a case study are essentially twofold. First of all, IgG are well-characterized macromolecules from a structural point of view and the complete X-ray crystallographic map, deposited in PDB, will serve as an input to generate a full set of phantom data. More importantly, we have access to a collection of Cryo-ET reconstructions of the IgG, already published in Sandin *et al.* (2004), which will be used to test the performance of our analysis tools versus real data.

As we shall discuss in Sections 4 and 5, the two approaches are indeed highly complementary and hold promise to result in novel postprocessing tools for quantitatively measuring the degree of spatial variability as seen from Cryo-ET experiments. This issue is further addressed in Section 6.

2. Cryo-electron tomography: the case of the antibody

Tomography means imaging by sections or sectioning. In the case of Cryo-ET, an electron microscope is used to capture projection images of the biological specimen. In order to reduce the radiation damage, the specimen is instantaneously frozen at the liquid nitrogen temperature (about -180°C), before inspection. Quick-freezing causes the water to form vitreous ice around the proteins, preserving their hydrated structure

and immobilizing them in the states they lastly occupied. The biological specimen is then imaged in a TEM by irradiating it from different angles. The full 3D structure is recovered by backprojecting the set of 2D images. Standard reconstruction tools can be complemented by the Constrained Maximum Entropy Tomography (COMET), a refinement procedure, which enhances the signal-to-noise ratio in an iterative manner, thus allowing more details to be included in the tomograms Skoglund *et al.* 1996).

In the data collection strategy for individual 3D reconstruction, evenly distributed projections covering the sphere in all directions are ideally used. In practise, this is difficult to realize. The simplified strategy normally used, is to record projections onto a stationary CCD device while rotating the specimen in even discrete steps around a single axis orthogonal to the electron beam. What can be usually varied is exposure times over the projections to compensate for specimen thickness variation after tilting. This is called a single axis tilt series (SATS) and is the data collection strategy which has been used to collect the data for all the specimen discussed in our paper.

With single, or double, axis data collection schemes, a large portion of data is not accessible. In the SATS case, the projection angles are commonly ranging between ± 60 degrees, which in turn implies that a large ensemble of possible directions remain unexplored. This is the so called 'missing wedge' phenomenon, also referred to as 'missing valley' because the inaccessible lines fill a V-shaped region in Fourier space. When adopting a double tilt acquisition scheme, the amount of missing data is reduced but not gone.

The missing valley of data is clearly manifested in the 3D volume as a tendency for the observed specimen density to disappear along the direction of the electron beam. This mechanism affects the resolution of the reconstructed tomogram, which is therefore sensitive to the direction (anisotropy). Conversely, in 3D reconstruction originating from, e.g. single particle ET, all projection directions are equally sampled due to the random orientation of the molecules in the specimen and, consequently, the missing data problem is less crucial.

As an additional source of difficulties, one should mention the 'dose problem' which arises due to specimen damage induced by electron-specimen interaction. The extent of the damage is proportional to the amount of energy deposited on the specimen and this in turn limits the number of images that can be collected before altering the structural properties of the sample under investigation. As a consequence, TEM images are to be recorded in a low-dose setting: for an ice–embedded biological sample the total dose cannot exceed 2000–5000 e^-/nm^2 depending on the acceleration voltage of the microscope, a limit that prevents severe damage to occur. In a typical ET experiment, the total dose has to be partitioned among different projections and therefore each micrograph appears to be extremely noisy (low signal-to-noise ratio). We shall return on these issues in Sections 4 and 5, when discussing the importance of representing the tomograms in a fuzzy setting.

In a recent paper by Sandin *et al.* (2004), Cryo-ET experiments on the monoclonal murine antibody IgG2a were described. Antibodies are crucial constituents of our immunological defence system. They bind to foreign agents and target them, for instance, for destruction. The IgG antibody is the most abundant antibody in blood and has a molecular weight of about 150 kDa. It is composed of three subunits, two fragment antigen binding arms (abbreviated as 'Fab' arms) and a stem ('Fc'). The connections are provided by a flexible hinge that allows for a significant relative mobility of the two arms, as demonstrated by 2D electron microscopy analysis (Roux, 1999). The methods presented in this paper (Section 4 and 5) are illustrated using data from the experiment reported by Sandin *et al.* (2004). We refer to that article for details.

3. From PDB entry to volume image

As previously anticipated, the forthcoming analysis is carried out on antibodies imaged through Cryo-ET experiments. In addition, to fully validate the proposed strategies we have constructed a set of artificial profiles from the antibodies maps deposited in the PDB (Berman *et al.*, 2000).

The only crystallographic structure of an intact IgG2 antibody, corresponds to murine IgG2A mAb 231, as reported in Harris *et al.* (1997) (PDB code 1IGT). In the following study, we also consider intact murine IgG1 mAb 61.1.3, as reported in Ollmann Saphire *et al.* (2002) (PDB code 1IGY).

From a PDB entry, it is possible to generate a volume image, having a certain *voxel* size (i.e. the size of the *volume picture element*) and resolution, of the corresponding macromolecule. Here, we use a model where a Gauss kernel is placed at each atom position and multiplied by the mass of that atom (Pittet *et al.*, 1999). The total density is then calculated by adding the contributions from Gauss kernels of atoms in the vicinity of the voxel. This results in an image with floating point values, which is linearly stretched and rounded off to an 8-bit integer image. The resolution of the image is 2σ, σ being the variance used to calculate the Gauss kernel. This procedure for creating low-resolution data from PDB entries, mimicking for example Cryo-ET data, has been used in the past, e.g. in the Situs program package for visualization and docking of single molecules (Wriggers & Birmanns, 2001). We assume a voxel size of 5.24 Å, which corresponds to the pixel size of the micrographs recorded in Cryo-ET experiments. Moreover, we construct volumes corresponding to resolution 10, 15, 20 and 25 Å, see Fig. 1.

4. From a coarse-grained representation to the parameters determination

Let us now turn to discussing our first strategy aimed at extracting geometrical features from a Cryo-ET density map.

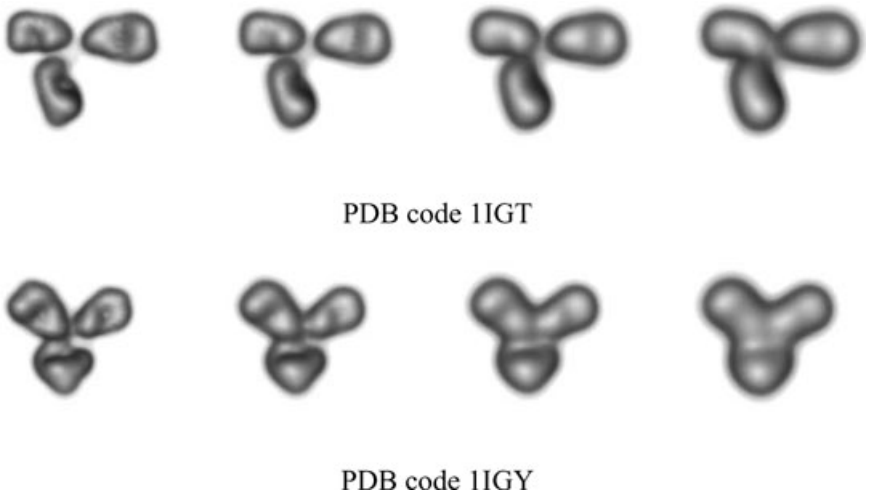

Fig. 1. Volume rendering of IgGs from PDB entries at pixel size 5.24 Å and resolution 10, 15, 20 and 25 Å (from left to right, respectively).

In the spirit of a coarse-grained representation, a set of massive units of appropriate shapes are assumed to represent the protein's rigid domains, the hinged connections between them being modelled with flexible junctions. Such mechanical model is then fitted to the experimental 3D tomograms and the positions of its constituting elements used as reference point to deduce relevant structural information, such as interdomain distances and angles. To shed light onto the technical details of the implementation we shall consider the case of the IgG antibody.

IgG antibodies can be represented by means of three massive spheres jointed together into a common point through as many Hookean springs of given strength. The latter are introduced in order to control the radial extensions of the arms. As mentioned in Section 2, the subunits of an IgG antibody are connected by flexible hinges. The hinge allows not only for angular movement but also radial stretching, an effect which is incorporated in our proposed representation through the springs. A stiff spring will result in a rigid connection, thus preventing excessive stretching to occur. The three spheres can therefore attain different spatial configurations and in principle adjust their relative orientation to match the conformations observed in direct experiments on real antibodies. This simplified model (depicted in Fig. 2) represents a variant of the scheme introduced in Bongini *et al.* (2004).

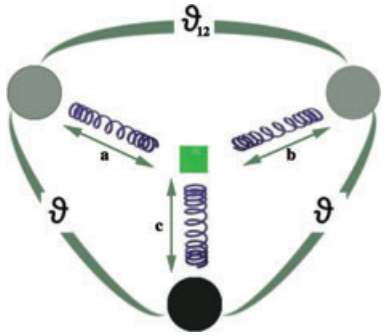

Fig. 2. The schematic model of the IgG antibody and definition of the relevant structural parameters (angles relative to units' mutual orientation and extension of the various arms).

4.1 Generating a force field from the measured densities

To adjust the coarse grained representation of the molecule into the measured volumes we adopt the following strategy. The Cryo-ET density, hereafter $\rho(\mathbf{r})$, is placed in the middle of a large box and there immobilized, while the schematic bead and spring scaffold is let evolve until it superposes to the true density with the desired accuracy. To accomplish the last step we evolve the model protein according to a Langevin dynamics under the influence of random collisions with a surrounding fictitious solvent (Uhlenbeck & Ornstein, 1930; Reif, 1965). Assuming that no external force field is being imposed, the equation of motion reads:

$$\frac{d^2\mathbf{r}}{dt^2} = -\gamma\frac{d\mathbf{r}}{dt} + \frac{\mathbf{f}(\mathbf{t})}{m}, \qquad (1)$$

where $\mathbf{r}$ labels the particle position and m denotes the mass of the domain. The first contribution on the right hand side is a frictional force that describes the drag on the particle due to the solvent. The magnitude of the drag is related to the friction coefficient γ. The term $\mathbf{f}(\mathbf{t})$ is a fluctuating random force with zero mean that models the impact with the fluid molecules. In order to satisfy a fluctuation-dissipation balance, the mean squared amplitude of the random force and γ must be chosen so that their ratio is fixed and proportional to the temperature of the fictitious solvent.

In the present application we consider each sphere to evolve according to Eq. (1), while respecting the geometrical constraints imposed by the IgG-like topology. In turn, this implies modifying the above equation of motion to account for

an external component of the force, namely $\mathbf{F}_{\text{constr}}$. The latter represents the contribution of the radial springs, that prevents the spheres from leaving the structural assembly and lets them adjust to the experimental density map. Furthermore, the external force should include a repulsive pairwise contribution in order to prevent the spheres from occasionally occupying the same region of space. As the main ingredient of this scheme, we introduce a fictitious force field $\mathbf{F}$, that is generated from the tomographic density $\rho(\mathbf{r})$ as:

$$\mathbf{F}_\rho(\mathbf{r}) = \nabla_{\mathbf{r}}\rho(\mathbf{r}). \tag{2}$$

The latter tends to favour the motion towards the portion of explored volume occupied by the tomogram. In conclusion, Eq. (1) is transformed into:

$$\frac{d^2\mathbf{r}}{dt^2} = -\gamma\frac{d\mathbf{r}}{dt} + \frac{1}{m}[\mathbf{F}_{\text{constr}} + \nabla_{\mathbf{r}}\rho(\mathbf{r}) + \mathbf{f}(\mathbf{t})]. \tag{3}$$

After a transient evolution, the model structure eventually reaches a long-lived steady state and docks into the experimental profile. Due to the repulsive forces introduced in our simulations, each lobe of the Cryo-ET map correctly shows to be visited by one sphere at a time. In order to ensure convergence to sufficiently long-lived states, we found that the temperature of the fictitious solvent should be at least one order of magnitude smaller than the peak density in the tomogram. Likewise, the time step of the numerical integration of Eq. (3) must be chosen at least one order of magnitude smaller than $1/\gamma$ in order to ensure the thermal equilibration of the model with the solvent.

4.2 Results

The procedure of dynamical docking is illustrated in Fig. 3, with reference to the structure 1IGY constructed at 20 Å resolution (see Fig. 1). Different snapshots are displayed relative to subsequent steps in the iterative procedure, thus allowing to visualize the convergence toward the final, steady-state solution. Once the model structure is fitted into the PDB volume, one can calculate the associated angles θ_{12}, θ_{13} and θ_{23}, as defined in Fig. 2.

We further focused on 1IGT and 1IGY low-passed to different resolutions, as discussed in the preceding Section 3. Our mechanical model is dynamically evolved for each of the considered cases and eventually fitted to the maps. The final estimates of the angles are reported in Table 1. Such values are approximately similar in the interesting range of 15–20 Å, corresponding to the resolution estimated for a successful Cryo-ET experiment. Deviations are instead observed at 10 Å and 25 Å. In the former case, the hole located in the middle of the Fc stem and Fab arms becomes distinctly visible. The potential associated with each density lobe is hence bimodal and the spheres can be trapped in proximity of either available state. Conversely, at a lower resolution the map is too smoothed, thus inducing local distortions into the fitted model. The final extension of the arms (corresponding to the stretching of the connecting springs) also shows a robust convergence. The values corresponding to the final steady state are reported in Table 1.

Table 1. Angles and arms extensions (in Å) for the PDB data constructed at different resolutions (pixel size 5.24 Å), after dynamical docking of the simplified model depicted in Fig. 2. The relaxed length of the spring is set to $r_0 = 70$ Å.

Res	θ_{12}	θ_{13}	θ_{23}	a	b	c
		PDB id 1IGT				
10 Å	136°	106°	108°	69 Å	65 Å	72 Å
15 Å	135°	108°	109°	69 Å	66 Å	71 Å
20 Å	131°	107°	115°	67 Å	68 Å	71 Å
25 Å	120°	109°	116°	67 Å	68 Å	71 Å
		PDB id 1IGY				
10 Å	110°	116°	120°	68 Å	64 Å	69 Å
15 Å	111°	111°	128°	66 Å	67 Å	70 Å
20 Å	110°	112°	125°	66 Å	66 Å	66 Å
25 Å	112°	113°	125°	67 Å	67 Å	69 Å

Finally, we applied the method to real Cryo-ET data. The model is shown to correctly interpret the observed profile, as confirmed by inspection of Fig. 4. As previously discussed, 3D tomograms are obtained from noisy 2D micrographs (low-dose condition) and display anisotropic resolution, the latter being associated with unphysical stretching of the reconstructed profiles.

When adjusting the mechanical model to a selected Cryo-ET map, the algorithm may then converge to a global

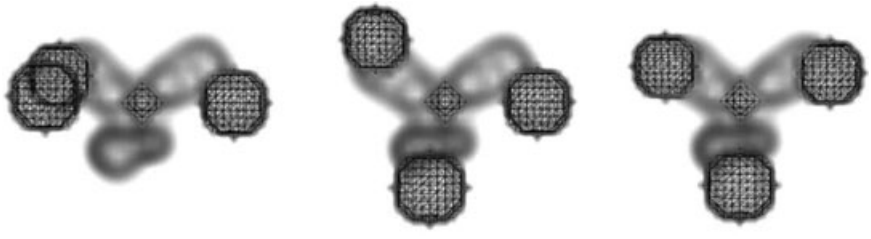

Fig. 3. Dynamical docking of the coarse-grained model into the PDB distribution for 1IGY low-passed to 20 Å. During the iteration steps, the model is evolved according to the Langevin dynamics scheme discussed above and is eventually shown to adapt to the analyzed density distribution.

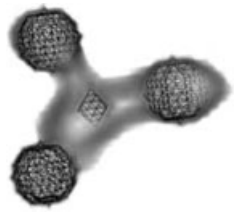

Fig. 4. The dynamical model of IgG fitted to a real tomogram as reconstructed from Cryo-ET. The relative angles are calculated to $\theta_{12} = 109, \theta_{13} = 106$ and $\theta_{23} = 119$. The extensions of arms are $a = b = c \simeq 65$ Å.

minimum which is slightly affected by the aforementioned intrinsic distortions. It must nonetheless be stressed that the reconstructed molecules are randomly oriented in the biological specimen. Hence the combined effect of different and independent distortions will not bias the distribution of the measured structural parameters towards any specific direction but it will simply result in an artificial enhancement of the observed flexibility. On the one hand, by averaging over different replica of the molecule, the most probable conformation is hence correctly displayed. On the other, by reconstructing the distribution of the measured structural parameters, one in general obtain a blurred version of the sought profile, resulting from the convolution with an *a priori* unknown distortion probability function. The same kind of reasoning certainly applies to the shot noise, which can be assumed to be evenly distributed over the 2D image, and hence equally sparsed at the level of the 3D representation. Notice however that such undesired perturbation could be in principle corrected for, provided one can assume a reasonable guess for the standard deviation of the statistical distribution of the density distortions. Concerning the case of the IgG, we shall notice that, by construction, each sphere will be forced to visit one of the Fab arms/Fc lobes at the time. Hence, errors in resolving the geometry of a single tomogram will be solely limited to local rearrangement of the spheres in the selected portion of volume. Those in turn will depend on the overall resolution of the reconstructed image. Based on the estimates performed by Sandin *et al.* (2004), a reasonable working figure for the error on the fitted angles is around a few percents.

We remark that in the present analysis (for both real and PDB-based maps), we observed a robust convergence of the docked structure. Other sets of data could in principle display a pronounced sensitivity to the initial guess of the parameters, thus revealing the presence of an intricate density landscape. This being the case, one should run independent tests by initializing the model scaffold in different conformation and draw a statistic of the final docked profiles.

In conclusion, we stress again that the parametrization chosen here to represent the conformational changes of an antibody is directly inspired by the coarse-grained description proposed in reference Bongini *et al.* (2004). As an additional feature with respect to the latter, we have here explicitly allowed for the radial stretching of the binding arms. However, despite such improvement, the model still neglects other important effects, such as the translational offset associated with the hinge region, an intrinsic properties that can be quantified by segmenting the reconstructed immunoglobulin molecule into independent units. This observation motivates the search for other, more flexible, analysis schemes. One promising such strategy is detailed in the following section.

5. A fuzzy framework allowing for structural interpretation

As briefly described in Section 1 and further emphasized in the previous sections, it is not always the case that a correct model of the macromolecule is known *a priori*. Hence, other additional methods are sought, acting in a more blind way, which could be used to determine the structural characteristics of the macromolecule. The knowledge gained from such analysis can serve as an input to construct a refined mechanical model. Aiming at filling the gap between experimental measurements and data analysis, we have here developed a computerized tool to extract structural information from an unknown density profile. In this section, we describe the general method which is intended to identify the units composing a large clustered agglomerate and quantify the parameters specifying their mutual connections. In our case, the agglomerates are typically macromolecules imaged using Cryo-ET. For this specific reason we need to take into account the fact that the input data are represented by low-contrast images (as the TEM images are recorded in a low-dose setting to prevent radiation damages) and that each blob consists of a small amount of voxels (typically 500 voxels for a Fab arm). The method is applied to the case of IgG: the aim is to identify its three domains, the Fc stem and the two Fab arms, and extract measures corresponding to the angles as well as translation between the domains.

In the end of the next section, we will also discuss in a comparative way other published techniques for identifying subunits of macromolecular assemblies.

5.1 Delineating and decomposing a reconstructed macromolecule into its subparts

In Svensson (2007), a decomposition scheme for fuzzy objects was introduced. An adjusted version was applied to Cryo-ET images of the IgG antibody (Svensson *et al.*, 2006), which allowed for automatic delineation and decomposition into Fc stem and Fab arms. In the following we recall the general framework and its application to IgG (in the next subsection).

One commonly used crisp approach for separating clustered bloblike structures, such as cell nuclei in a fluorescence microscopic image of some tissue, or subunits of a macromolecular assembly, is to identify the most internal part of each blob in the cluster, the *seed* of the part, and then apply a region growing process to the identity labelled seeds

(Vincent, 1993). The seeds can be identified, e.g. as being local maxima in the distance transform of the structure. The distance transform is an image in which each point is assigned a value corresponding to its closest distance to a point outside the structure. Hence local maxima are located internally in each part.

Shape analysis is usually, exactly as described in the previous paragraph, applied to binary images, i.e. where the grey-level image depicting the object to be studied has been segmented into object and background. In cases where low-contrast images are used it is difficult to take a crisp decision on how to delineate the object, i.e. a decision on which voxels belong to the object and which to the background, as the edge information is low. This decision influences the subsequent shape analysis. Moreover, when the studied object consists only of a small number of voxels this decision is even more crucial. In addition, as pointed out in Section 2, the effect of the 'missing valley' of data existing in the image acquisition method used here is that the object have a more fuzzy appearance along the beam direction. Therefore it is for many reasons of interest to, if possible, make the analysis resorting to a fuzzy setting. This is a quickly growing approach within the field of computerized image analysis and is what we will make use of here. A brief introduction, including the concepts necessary for this article, can be found in the following paragraphs.

We start by recalling the definition of a fuzzy object. Recall that the fuzzy objects in our case are macromolecules and the parts of a fuzzy object are typically the Fab arms and the Fc stem for the IgG antibody. The theory of fuzzy sets applied to digital images originates in the work by Zadeh (1965). A 3D fuzzy digital object $\mathcal{O}$ is a fuzzy subset defined on $\mathbf{Z}^3$, i.e. $\mathcal{O} = \{(p, \mu_{\mathcal{O}}(p)) \mid p \in \mathbf{Z}^3\}$, where $\mu_{\mathcal{O}} : \mathbf{Z}^3 \to [0, 1]$. This means that we have an image where the grey-level in each voxel corresponds to the degree of membership for the object, a high value indicates that it is likely the voxel belongs to the object while a low-valued voxel is more likely to belong to the background. The fuzzy object can be identified using a fuzzy segmentation method. See the recent review by Udupa and Saha (2003). The first step to delineate the subparts of a macromolecule is to identify the fuzzy object corresponding to it. This can be done, e.g. in the way described for the IgG antibody in Svensson *et al.* (2006).

Once the fuzzy object has been identified, we compute the fuzzy distance transform (FDT) (Levi & Montanari, 1970; Saha *et al.*, 2002) of the fuzzy object. The FDT is a replica of the fuzzy object image where each voxel v in the fuzzy object, i.e. for which $\mu_{\mathcal{O}}(v) > 0$, is assigned the fuzzy distance to its closest voxel u outside the fuzzy object, i.e. the shortest length of a path between v and u. The length is weighted with the membership values of the voxels along the path. This means that the FDT will have high (fuzzy distance) values corresponding to the 'centre' of each blob (nearly convex part) of the fuzzy object, analogous to what was described for the crisp case above. By combining grey-levels, taken from the fuzzy object or even from the original grey-level image, and distance information we further emphasize the internal grey-level structure as well as stress the shape of a subunit. For this reason we detect all local maxima on the FDT. There may be more than one local maximum corresponding to each part. However all are centrally located. We apply fuzzy distance based hierarchical clustering (Gedda & Svensson, 2006) to group the local maxima into seeds, where each seed correspond to one part of the fuzzy object. The reason for choosing hierarchical clustering is because it is a deterministic unsupervized clustering technique which allows to find the number of clusters most suitable for the set. The decomposition scheme for fuzzy objects is a general framework which can be used in other applications, e.g. in Svensson (2007), it was successfully used to identify cell nuclei in a fluorescence microscopic image of a tissue slice from carcinoma of the prostate. To finally identify the part corresponding to each seed we apply region growing in terms of seeded watershed segmentation (Vincent & Soille, 1991) to the FDT.

Also the method described by Volkmann (2002) makes use of watershed segmentation. However, there are some significant differences between the two methods. Volkmann (2002) uses the original grey-level distribution, while we further emphasize the structure by resorting to the FDT and thereby also take into account the shape of a subunit. Furthermore, in Volkmann (2002) seeds correspond to local grey-level maxima in the original image. This set is, as pointed out by the authors, noisy and a sophisticated preprocessing step is required in order to reduce the set of candidate seeds. We instead make use of local maxima detected on the FDT and fuzzy distance based hierarchical clustering. By this we can find natural clusters of local maxima, each corresponding to one seed in the subsequent region growing process, in a more elaborated way, which also enable us to account for *a priori* knowledge, if such exists. In Baker *et al.* (2006), seeding together with region growing is also used. The set of seeds is detected based on the original grey-level data. As for Volkmann (2002) this set needs to be reduced. The reduction is done using a sophisticated filtering technique (gradient vector diffusion by the same authors). Once the seeds are identified, region growing is performed by means of fast marching (Sethian, 1999). The region growing part will give a result similar to what could be obtained with watershed segmentation. To summarize, the three methods, our and the above referred to, all use seeding and region growing, though distinct strategies make the implementation different. In particular, we remark that both the methods we have compared with are developed for Cryo-ET data of large macromolecular assemblies, where each subunit is significantly larger than the Fc stem and the Fab arms. Moreover, they are actually developed for single particle ET and rely on symmetry inherent by the imaging method. In our case, we focus on smaller particles and therefore need to secure a method which is even more robust to small shape changes.

Fig. 5. The decomposition scheme illustrated on one slice of a Cryo-ET image: (a) the original image; (b) the fuzzy object for which the support is given in green; (c) the FDT with seeds overlayed and (d) the decomposed fuzzy object.

5.2 The IgG antibody: from decomposition to determination of structural parameters

To find an IgG antibody, in a Cryo-ET image, and decompose it into its Fc stem and Fab arms, we use a properly adjusted version of the scheme outlined above. The fuzzy object is delineated using fuzzy connectedness and the knowledge that the IgG antibody should consist of around 1500 voxels with the used sampling (Creighton, 1993) and thereafter decomposed parts. For the hierarchical clustering used in the decomposition scheme we take advantage of the *a priori* knowledge that three parts are to be identified for the case of the IgG. In Fig. 5, the process is illustrated on a slice from a Cryo-ET image, showing a cross section of the two Fab arms: in (a) the original image is shown; (b) the delineated fuzzy object; (c) the FDT with seeds overlayed and (d) the decomposed fuzzy object. We remark that the method is applied directly to the 3D image. This means that there can be more voxels (local maxima) belonging to the same seed placed in other slices. A detailed description of the method can be found in Svensson, *et al.* (2006).

Once the Fc stem and the two Fab arms for each IgG antibody have been identified, the structural parameters are to be extracted. We illustrate this process by showing how the main orientation of the Fc stem and Fab arms can be estimated in order to determine the interdomain angles for the IgG antibody, as well as the translation between the domains. We make use of not only the shape of the identified IgG antibodies but also the grey-level distribution inside. Each part is represented by its first principal component (PC1), i.e. a vector for which the voxels included is closest to in a least square sense, see e.g. Duda *et al.* (2001). The principal components of a set $\mathbf{X}$ of p random variables $X_1, \ldots, X_p$ are the eigenvectors of the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{X}$. In our case $\mathbf{X}$ is equal to the coordinates of the voxels for the specific part. To give more relevance to voxels having high grey-levels in the fuzzy object, each voxel is weighted proportionally to its grey-level. Taking the first eigenvector for $\mathbf{\Sigma}_{\text{Fab1}}$, $\mathbf{\Sigma}_{\text{Fab2}}$ and $\mathbf{\Sigma}_{\text{Fc}}$, we get a vector representation consisting of PC_{Fab1}, PC_{Fab2} and PC_{Fc} corresponding to the IgG antibody. The interdomain angles are calculated in a straightforward way using PC_{Fab1}, PC_{Fab2} and PC_{Fc}. The validity of the vector representation is given by the eigenvalues corresponding to the eigenvectors. If the variance, i.e. the relative eigenvalue, explained by the first eigenvector is significantly larger than the variance explained by the second eigenvector for $\mathbf{\Sigma}_{\text{Fab1}}$, $\mathbf{\Sigma}_{\text{Fab2}}$ and $\mathbf{\Sigma}_{\text{Fc}}$, it means that PC_{Fab1}, PC_{Fab2} and PC_{Fc} is a suitable representation.

To compute the translation between the two Fab arms, we define the Fab dyad as being a plane perpendicular to the vector connecting centre of mass (COM) for the Fab arms, in the following denoted COM_{Fab1} and COM_{Fab2}, and placed midway between COM_{Fab1} and COM_{Fab2}. The translation is then given as the distance between the points in the plane where PC_{Fab1} and PC_{Fab2} intersect the plane.

5.3 Results

The method to determine structural parameters described above is developed to make contact with the angles and translation measured in X-ray crystallography experiments. In this section, we verify that this is actually the case by making a comparison with measurements for PDB code 1IGT and 1IGY. Moreover, we show that the proposed method, including decomposition of fuzzy objects and subsequent determination of structural parameters, is stable under changes in resolution. Finally we apply the method to Cryo-ET data.

In Fig. 6, the method is illustrated for 1IGY, constructed at 20 Å. To the left, the result of the decomposition scheme is shown, with the Fc stem in yellow and the two Fab arms in red and blue. PC_{Fab1}, PC_{Fab2} and PC_{Fc}, used to measure the interdomain angles are shown in the middle. The Fab1–Fab2 translation measurement is illustrated by showing the Fab dyad together with PC_{Fab1} and PC_{Fab2} (right). The positions where the plane intersects the PCs are marked out with bullets. For all PCs, the COMs corresponding to their parts are also labelled with bullets.

In Table 2, the interdomain angles and Fab1–Fab2 translation for 1IGT and 1IGY measured using our method are listed. The results are stable under changes in resolution. The only case where the results deviate is observed of 1IGY at 25 Å resolution. As reported in the result part of Section 4, this is due to the smoothing effect present in lower resolution images. In Ollmann Saphire *et al.* (2002), the results from various X-ray crystallography experiments on the IgG antibody are summarized. The relevant measures, i.e. the angles and

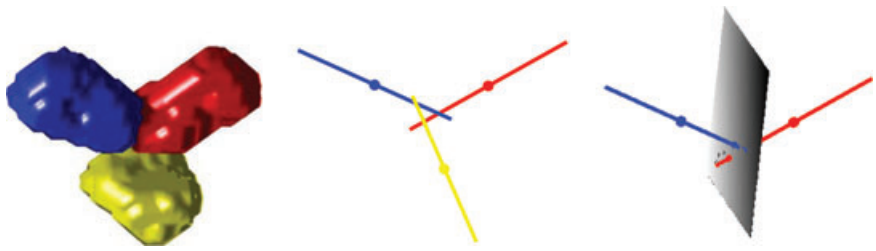

Fig. 6. For PDB code 1IGY constructed at 20 Å resolution: surface rendering of the Fc stem (yellow) and the Fab arms (red and blue); PC_{Fab1}, PC_{Fab2} and PC_{Fc}; and Fab dyad, PC_{Fab1} and PC_{Fab2}.

Table 2. Measurements for PDB data constructed at different resolutions (voxel size 5.24 Å) using the method described in Section 5.2.

Res	Fc–Fab1 Angle	Fc–Fab2 Angle	Fab1–Fab2 Angle	Fab1–Fab2 Translation
		PDB code 1IGT		
10 Å	70°	111°	174°	12 Å
15 Å	74°	110°	172°	12 Å
20 Å	70°	109°	173°	10 Å
25 Å	70°	113°	171°	8 Å
	X-ray crystallography (IgG1 mAb 61.1.3) as reported by Ollmann Saphire *et al.* (2002)			
	66°	113°	172°	9 Å
		PDB code 1IGY		
10 Å	82°	120°	116°	16 Å
15 Å	87°	117°	117°	16 Å
20 Å	80°	114°	118°	14 Å
25 Å	81°	128°	114°	20 Å
	X-ray crystallography (IgG1 mAb 61.1.3) as reported by Ollmann Saphire *et al.* (2002)			
	78°	123°	115°	9 Å

Fab1–Fab2 translation for 1IGT and 1IGY, are listed in Table 2. As can be seen, our measurements, automatically extracted using the method described in this section, correlate well to the homologous parameters determined from X-ray crystallography maps.

The vector representations PC_{Fab1}, PC_{Fab2} and PC_{Fc} used in Table 2 are in all cases such that the variance explained by the first eigenvector is significantly larger than for the second, see Table 3. In the weakest case (1IGT at 15 Å), 47% is explained by the first and 42% by the second eigenvector. For most cases, the difference between the first and the second is at least 20 percentage units.

In Fig. 7, the results for the same IgG antibody as used in Fig. 4 are shown, illustrated in the same way as for 1IGY (Fig. 6): the result of the decomposition scheme with Fc stem in yellow and the two Fab arms in red and blue; PC_{Fab1}, PC_{Fab2} and PC_{Fc}, used to measure interdomain angles; and the Fab dyad together with PC_{Fab1} and PC_{Fab2}, used to measure translation. Also in this case, the vector representation PC_{Fab1}, PC_{Fab2} and PC_{Fc} describes the data well. In the weakest case (Fab1), the variance explained by the first eigenvector is 59% while for the second it is 24%.

Table 3. Variances explained by the eigenvectors.

	PC_{Fc}		PC_{Fab1}		PC_{Fab2}	
Res	1st	2nd	1st	2nd	1st	2nd
			PDB code 1IGT (%)			
10 Å	48	40	66	22	68	22
15 Å	47	42	66	22	69	21
20 Å	47	41	66	21	67	20
25 Å	48	38	64	22	67	20
			PDB code 1IGY (%)			
10 Å	49	39	69	21	67	22
15 Å	49	40	73	18	68	21
20 Å	50	37	71	19	70	20
25 Å	50	34	72	17	69	19

6. DISCUSSION

In this paper, we have outlined two complementary strategies to analyze Cryo-ET reconstructions. The first exploits a Brownian dynamics framework to dock a coarse-grained representation of the biomolecule into the imaged map. Once the reduced model is adjusted into the tomogram, one can proceed with a straightforward extraction of the relevant geometrical information. The second is an automatic decomposition scheme based on a fuzzy representation of the protein, which allows to partly compensate for the 'missing valley' artefact existing in Cryo-ET data. In a fuzzy representation (*fuzzy object*), the grey-level in each point is a measure of the degree of membership the point has to the original object, in this case the protein. By combining distance information and grey-level information, the most internal part of each subpart of the protein can be identified, e.g. the Fab arms and the Fc stem for the case of the IgG antibody. Once this decomposition is done, principal component analysis is used to find the major axis of each part and from that resolve the structural characteristic. The idea of using a Brownian

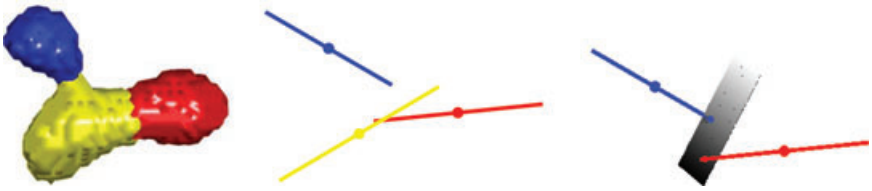

Fig. 7. For one IgG antibody imaged using Cryo-ET: surface rendering of the Fc stem (yellow) and the Fab arms (red and blue); PC_{Fab1}, PC_{Fab2} and PC_{Fc}; and Fab dyad, PC_{Fab1} and PC_{Fab2}. The corresponding angles are 70° (Fc–Fab1), 144° (Fc–Fab2) and 146° (Fab1–Fab2) and the translation 45 Å.

dynamics framework for this type of image data is, according to our knowledge, new. It is promising as it allows for a direct way of expressing the dynamics. The decomposition scheme used in the second approach shares some similarities with e.g. Volkmann (2002) and Baker *et al.* (2006). However, our proposed scheme is more stable for small subunits with respect to existing implementations and does not rely on symmetry. Moreover, it enables one to take advantage of *a priori* knowledge, which can be straightforwardly incorporated into the algorithm, if beneficial.

The fuzzy based method can be seen as a complement to the first when schematic description cannot be designed by visual inspection. In such cases, the decomposition scheme is used in a blind way, meaning that the number of subparts is not given as *a priori* information. Instead the most suitable number is found based on outer shape and grey-level distribution (Svensson, 2007). This decomposition and the interdomain relations described by it can serve as an input to develop, or eventually refine, a simplified mechanical model of the protein. Importantly, a quantitative measure of the degree of internal flexibility opens up the perspective of deriving an effective estimate of the relevant physical interactions that rule the dynamical evolution of the macromolecule under study (Bongini *et al.* 2004).

The above two techniques constitute a powerful combination for automatic postprocessing data analysis and naturally complement Cryo-ET for structural determination purposes. The procedures are here validated with reference to the case of the IgG. Both artificial maps, constructed from IgG entries in the PDB (Berman *et al.*, 2000), and real tomogram are analyzed and reported. As concerns the PDB volumes, different resolutions are used (10 Å, . . . , 25 Å) to assess the robustness of the schemes. The conclusion from this analysis is that the methods are robust for changes in resolution, spanning from a resolution close to what can be achieved in X-ray crystallography experiments to that of Cryo-ET.

Automatic analysis of data as presented here is of importance to allow for large-scale reproducible studies. The Brownian dynamics scheme converges after a few iterations, with reference to the IgG case of study: in general, the computational load depends on the amount of details that are to be modelled, the level of coarse-graining being in practise set by the experimental resolution limit. The method is particularly suitable for investigating molecular structure that can be visually segmented in interlaced subunits. The fuzzy based approach constitutes a computationally efficient alternative which enables for a quick extraction of the molecules structural information. The significance of the measured parameters can be quantified through the eigenvalues, as proposed in Section 5.3.

It is worth pointing out that already at 20 Å resolution it is possible to measure the relative orientation of the various domains with results that are shown to match analogous estimates for the corresponding X-ray maps, (Ollmann Saphire *et al.*, 2002), as confirmed by the fuzzy based method, see Table 2. This observation indicates that reliable structural information can be successfully deduced from Cryo-ET data.

Finally, it should be emphasized that the techniques here illustrated, are flexible. Hence, they could with suitable adjustments be successfully employed to quantitatively analyze an ample spectrum of candidate objects, ranging from individual biomolecules to large macromolecular assemblies.

Acknowledgements

This work was supported by grants from the European Union '3D-EM' Network of Excellence, The Agouron Institute, the Karolinska Foundation, the Swedish Foundation for Strategic Research and the Swedish Research Council. Magnus Gedda and Stina Svensson are financially supported by Swedish Research Council (project 621-2005-5540). The Cryo-ET data sets have been provided by Dr Sara Sandin, Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden (currently Division of Structural Studies, MRC Laboratory of Molecular Biology, Cambridge, United Kingdom). We also thank Andrea Guazzini for Fig. 2.

References

Baker, M. L., Yu, Z., Chiu, W. & Bajaj, C. (2006) Automated segmentation of moelcular subunits in electron cryomicroscopy density maps. *J. Struct. Biol.* **156**, 432–441.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000) The protein data bank. *Nucl Acids Res.* **28**(1), 235–242.

Bongini, L., Fanelli, D., Piazza, F., De Los Rios, P., Sandin, S. & Skoglund, U. (2004) Freezing immunoglobulins to see them move. *Proc. Nat. Acad. Sci. (PNAS)*. **101**(17), 6466–6471.

Carragher, B., Potter, C.S. & Sigworth, F.J. (2007) Guest editors of Special Issue: software tools for macromolecular microscopy. *J. Stuct. Biology*. **157**, 1–296.

Creighton, T.E. (1993) *Proteins: Structures and Molecular Properties*. W.H. Freeman and Company, 2nd edition.

Duda, R.O., Hart, P.E. & Stork, D.G. (2001) *Pattern Classification*. John Wiley & Sons, Inc., 2 edn. John Wiley & Sons, New York.

Gedda, M. & Svensson, S. (2006) Fuzzy distance based hierarchical clustering calculated using the A* algorithm. *Combinatorial Image Analysis: 11th International Workshop, IWCIA 2006, Berlin, Germany, June 19-21, 2006. Proceedings*, volume 4040 of *Lecture Notes in Computer Science* (ed. by R. Reulke, U. Eckardt, B. Flach, U. Knauer & K. Polthier), pp. 101–115. Springer-Verlag, Berlin.

Harris, L.J., Larson, S.B., Hasel, K.W. & McPherson, A. (1997) Refined structure of an intact IgG2a monoclonal antibody. *Biochemistry*. **36**, 1581–1597.

Levi, G. & Montanari, U. (1970) A grey-weighted skeleton. *Inform. Control*. **17**, 62–91.

Ollmann Saphire, E., Stanfield, R.L., Crispin, M.D.M., Parren, P.W.H.I., Rudd, P.M., Dwek, R.A., Burton, D.R. & Wilson, I.A. (2002) Contrasting IgG structures reveal extreme asymmetry and flexibility. *J. Mol. Biol.* **319**, 9–18.

Pittet, J.-J., Henn, C., Engel, A. & Heymann, J.B. (1999) Visualizing 3D data obtained from microscopy on the internet. *J. Struct. Biol.* **125**, 123–132.

Reif, F. (1965) *Fundamentals of Statistical and Thermal Physics*. McGraw-Hill, New York.

Roux, K.H. (1999) Immunoglobulin structure and function as revealed by electron microscopy. *Int. Arch. Allergy Immunol.* **120**(2), 85–99.

Saha, P.K., Wehrli, F.W. & Gomberg, B.R. (2002) Fuzzy distance transform: theory, algorithms, and applications. *Comp. Vis. Image Understand.* **86**, 171–190.

Sali, A., Glaeser, R., Earnest, T. & Baumeister, W. (2003) From words to literature in structural proteomics. *Nature*. **422**, 216–225.

Sandin, S., Öfverstedt, L.-G., Wikström, A.-C., Wrange, O. & Skoglund, U. (2004) Structure and flexibility of individual Immunoglobulin G molecules in solution. *Structure*. **12**(3), 409–415.

Sethian, J.A. (1999) *Level Set Methods and Fast Marching Methods*. Cambridge University Press, Cambridge.

Skoglund, U., Öfverstedt, L.-G., Burnett, R.M. & Bricogne, G. (1996) Maximum-entropy three-dimensional reconstruction with deconvolution of the contrast transfer function: a test application with adenovirus. *J. Struct. Biol.* **117**, 173–188.

Svensson, S. (2007) A decomposition scheme for 3D fuzzy objects based on fuzzy distance information. *Patt. Recogn. Lett.* **28**(2), 224–232.

Svensson, S., Gedda, M., Fanelli, D., Skoglund, U., Öfverstedt, L.-G. & Sandin, S. (2006) Using a fuzzy framework for delineation and decomposition of Immunoglobulin G in cryo electron tomographic images. *Proceedings The 18th International Conference on Pattern Recognition (ICPR 2006)* (ed. by Y.Y. Tang, S.P. Wang, G. Lorette, D.S. Yeung & Yan, H.), volume 4, pp. 520–523. IEEE Computer Society, California.

Udupa, J.K. & Saha, P.K. (2003) Fuzzy connectedness and image segmentation. *Proc. IEEE*. **91**(10), 1649–1669.

Uhlenbeck, G.E. & Ornstein, L.S. (1930) On the theory of the Brownian motion. *Phys. Rev.* **36**(5), 823–841. Reprinted in "*Noise and Stochastic Processes*" (ed. by N. Wax). New York: Dover, pp. 93–111, 1954.

Vincent, L. (1993) Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. *IEEE Trans. Image Process.* **2**(2), 176–201.

Vincent, L. & Soille, P. (1991) Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans. Patt. Anal. and Mach. Intell.* **13**(6), 583–597.

Volkmann, N. (2002) A novel three-dimensional variant of the watershed transform for segmentation of electron density maps. *J. Struct. Biol.* **138**, 123–129.

Wriggers, W. & Birmanns, S. (2001) Using *Situs* for flexible and rigid-body fitting of multiresolution single-molecule data. *J. Struct. Biol.* **133**, 193–202.

Wriggers, W. & Chacón, P. (2001) Modeling tricks and fitting techniques for multiresolution structures. *Structure*. **9**, 779–788.

Wriggers, W., Milligan, R.A. & McCammon, J.A. (1999) Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.* **125**, 185–195.

Zadeh, L.A. (1965) Fuzzy sets. *Inform. Control*. **8**, 338–353.

