A probabilistic template model for finding macromolecules in MET volume images

Lennart Svensson and Ida-Maria Sintorn

Centre for Image Analysis, Uppsala University and Swedish University of Agricultural Sciences, Uppsala, Sweden,
lennart,ida.sintorn@cb.slu.se

Abstract. We introduce and investigate probabilistic templates with particular focus on the application of protein identification in electron tomography volumes. We suggest to create templates with a weighted averaging operation of several object instances after alignment of an identified subpart. The subpart to be aligned should, ideally, correspond to a rigid and easily identifiable part of the object. The proposed templates enables common rigid template matching methods to also find different shape variations without increasing time complexity in the actual search procedure, since a static template is still used. We present general ideas on how to perform the object instance alignment and look specifically at how to do it for the antibody macromolecule IgG.

1 Introduction

Standard template matching uses an image patch to find similar regions in other images, by performing template-image region correlation for different positions and orientations of the template. Many articles focus on what kind of similarity measure that should be used, for example Wriggers [1] and Ding et. al. [2]. The latter gives a concise overview of commonly used similarity measures such as sum of absolute differences, cross-correlation coefficient, geometric distance in chamfer matching, mutual information and invariant moments, used in rigid template matching along with their respective strengths and weaknesses. An in-depth and more thorough review of different similarity measures can be found in Brunelli's book on template matching [3].

Here, we focus on the importance of the template used in the matching process. We present a method to incorporate object variability in the template, creating a probabilistic template or p-template, without increasing the computational cost when performing template search. We look particularly on using the method for identifying and analysing proteins and their shape variations in *Molecular Electron Tomography* (MET) volume images. The principle to create a p-template is simple: perform alignment of a specific part of a number of object instances and then create a, possibly weighted, averaged template model. An important aspect is the careful choice of the object subpart to align, in order to create a high contrast template while at the same time cover shape variability. Previously, averaged templates have for example been used to incorporate radial

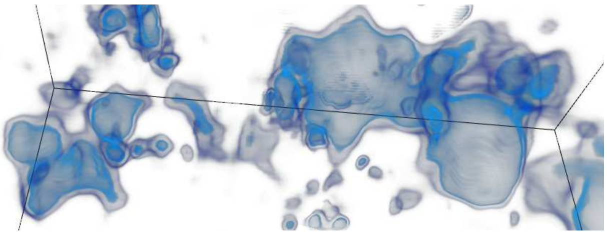

Fig. 1. An example of an IgG in solution in a MET volume image. To the right in the image is a larger protein complex (EGFR). This is connected to the smaller IgG protein to the left.

intensity differences for the detection of approximately circular symmetric virus particles in 2-D [4].

The MET volumes are volume images reconstructed from 2-D electron microscopy images, acquired of a specimen at different tilt angles. MET is the only imaging modality in which the 3-D flexibility and interaction of proteins and protein complexes can be studied in solution and in situ. The images are characterized by a low signal to noise ratio, and limited, if any, ground truth data is available regarding how to interpret the volumes. An example of a MET volume is given in Figure 1.

Rigid registration was the first MET registration method to be used extensively [5]. Templates are usually created from atomic data for a crystallized state of the molecule. To make the registration more flexible, techniques have been introduced that either divide a molecule into rigid sub parts attached to each other with different force constraints, or incorporate some degree of model elasticity [6][7]. More recently, Trabuco introduced a registration method with regularization based on molecular simulation [8]. All the non-rigid registration methods are however more or less computationally heavy and at the current hardware level less suited for volume screening purposes. This work is part of an effort to develop semi-automatic image analysis and visualization methods which are adapted to this kind of data and suitable for interactive volume processing.

3-D reconstruction by averaging is also used in single particle analysis using subtomogram averaging [9], which aims at reconstructing high resolution tomograms for different discretized structural classes of a certain macromolecule. The p-template does not describe any single particular state of a molecule, but rather cover the molecular variation while at the same time keeping the contrast in the template high, which are suitable properties for a template used in initial template matching screening of a MET volume. Another difference to subtomogram averaging is the suggested alignment procedure. Normally, cross correlation is used to align the subtomograms, where we instead use a feature based technique.

In the following sections the p-templates and a suggested sub-part alignment method are explained, followed by a demonstration on synthetic 2-D image data and on a real MET case consisting of images of the antibody molecule Immunoglobulin G (IgG) in solution.

2 p-template construction

Normally, the templates used in rigid template matching in MET volumes are derived from atomic data in the *Protein Data Bank* (PDB). The PDB data give a high resolution atomic map, but only for one snapshot state of the molecule, and does not account for the shape flexibility. An example of such a template of IgG (from PDB entry 1IGT) is shown in Figure 6 to the left.

The suggested p-template aims at covering protein flexibility normally only covered by more complex deformable models, by superimposing aligned image data from different protein conformations (shape variants) to one averaged molecule. We suggest to perform a rigid alignment between the sample conformations with respect to one stable component (part) of the molecule, instead of finding the alignment with respect to the whole volume. This will produce a template with a strong reference frame and high contrast for the aligned component, that still accounts for all the observed natural flexibility. The method to create a p-template is the following:

Algorithm 1 (P-template generation)

1. Subpart alignment of all instance examples, giving the mapping

$$I_j : \mathbb{R}^3 \rightarrow \mathbb{R} \textit{ to } I'_j : \mathbb{R}^3 \rightarrow \mathbb{R}.$$

2. Creating a template by averaging over the aligned instances

$$T(x,y,z) = \tfrac{1}{N} \sum_j I'_j(x,y,z),$$

where N is the number of images.

3. Gaussian smoothing of the volume template

$$T' = T \star G$$

where G is a Gaussian kernel in 3-D.

The second step in this algorithm can also be performed with a weight factor for each instance. If, e.g., a template derived from PDB is available and included as one of the instance examples, it could be desirable to give it a higher weight compared to other instances derived from MET volume data. The weighting should reflect the confidence of the different instances actually being true molecules of interest as well as possible a priori information about the molecule's flexibility, preference for certain conformations, and the resolution and accuracy of the acquired MET-volumes.

To use the p-template in a correlation search, the procedure is the same as with a standard rigid template search. This can start with generating a well distributed set of rotations, transforming the template after these and then performing a correlation search in Fourier-space, using the fact that a convolution in the real space equals multiplication in the frequency domain. With a GPU implementation a standard MET volume (100x100x100) can be searched within seconds [10]. Since the search procedure is the same as when using a template defined with a static molecule image, the time complexity during the actual search remains the same.

3 Synthetic 2-D test case

The synthetic test case consists of 20 generated 2-D images with three bright regions extending from the center of the image. Three examples from this test series are shown in Figure 2 to the left. The test object is a simplified example of what a cross section of an antibody could look like. The lower arm is intended to be the stable reference component. The angle between the top arms is increased from 80 degrees starting from 10, to in 2-D mimic the molecular flexibility of an IgG molecule. The three examples in Figure 2 are for 10, 50 and 80 degrees between the top arms.

With a synthetic test series describing variations of a shape, the p-template was created using the procedure described in 2. Since the reference arm is at the same position for all instances, they are already perfectly aligned and averaging can be performed directly to generate the p-template. The correlation response between the p-template and each of the instances in the test sequence are shown in Figure 3 with a solid line (the x-axis in the plot gives the top-arm angle). The p-template response is naturally lower than the autocorrelation response of one of the instances, which is 1.0 when using normalized cross correlation.

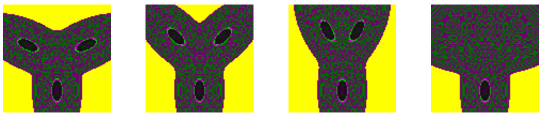

Fig. 2. The first three images on the left show examples from a generated sequence of images that should resemble how a molecule can appear in the MET modality. The top arms gradually moves closer to each other in this image sequence. On the right is the p-template created from this synthetic test set.

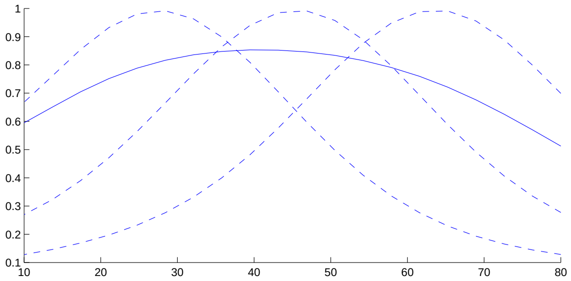

Fig. 3. Correlation responses in the synthetic test case. Each line represent testing one template against the generated images. The x-axis define the angle between a top arm and the vertical axis in degrees. The p-template response curve is shown as a solid line, and the responses from using any of the three instance examples shown in Figure 2 are shown as dashed lines. Using these as templates only give a high response near the same shape.

The correlation response curves show that the response is stable and smooth over the whole range of instances (angles) for the p-template. Using any particular example from the generated set as template instead gives a high response near that shape variant, but naturally a lower response for other shape variations. For example, using the instance with 25° angle as template gives a good response for the instances with angles between 10° and 40° but a low response for instances with high angles above 60° , see Figure 3.

4 IgG test case

The IgG is an antibody protein that consists of three components, the so called Fc-stem and two Fab-arms attached to the stem. All three are of similar size and density. Our sample data set consists of eight manually identified IgG molecules in a MET image of a protein solution [11], see Figure 4.

The alignment procedure aims at aligning the Fc-stems in the different sample volumes to each other by using three characteristic feature points. We have chosen the binding points for the Fab-arms to the Fc-stem to be feature points as they are relatively stable. The third feature point is the mass center of the Fc-stem.

As a preprocessing step to the p-template generation, everything below a user set intensity level is set to zero. The binding points between the Fab-arms and Fc-stem are then identified from the contour tree for the volume using 6-connectivity, as the two nodes which correspond to the two largest merges of connected components. The contour tree, well described by Carr et.al. [12], can be seen as way of tracing the topological transformations an iso-surface of a volume goes through as the iso-level is changed. It can be constructed by sorting all voxels after intensity, and going through the sorted list while simultaneously building connected components. Each of the two stem-to-arm

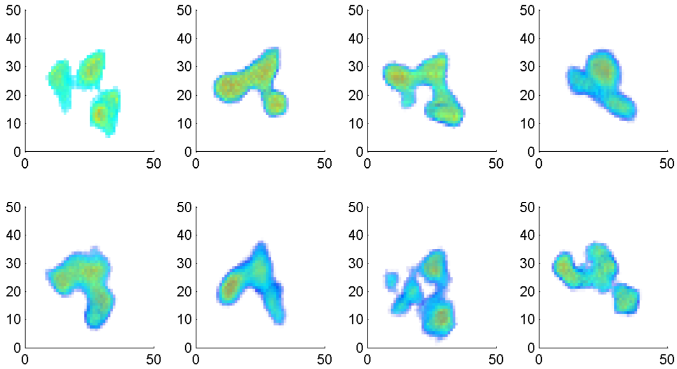

Fig. 4. The used IgG volumes.

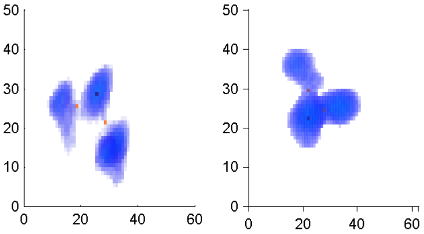

Fig. 5. Extracted feature points on two IgG molecules.

binding points, provides one estimate of the Fc-stem connected component and mass centre. These are averaged to give one estimate of the Fc-stem mass centre. Two examples of the identified IgG feature points are shown in Figure 5.

The three feature points for every sample are aligned using Generalized Procrustes Analysis [13], giving the optimal translation and rotation of each sample. The sample volumes are then transformed using these parameters, and averaged to give the p-template volume.

4.1 Evaluation

For evaluation leave-one-out cross validation was performed. That is, repeatedly a p-template was created using all but one, left out, instance. This instance was instead used to calculate the correlation to the created p-template and also to the reference template. The reference template was derived from the protein's static atom positions deposited in PDB, by creating a zero-set volume and at each atom position adding a 3-D Gaussian [14], with standard deviation 1 nm and weighted by the atomic mass, giving a resolution of 2 nm. The PDB template is seen in Figure 6 to the left.

We tested using both a p-template created in the standard way, with equal weighting for all instances, and also with inclusion of the PDB template as one instance with the weight set to be the sum of the other instances weights. This is called the PDB-adjusted p-template. With the validation procedure one correlation response were acquired per instance and template. Gaussian smoothing was done with kernel size parameter sigma set to 1.5 for these results. This setting

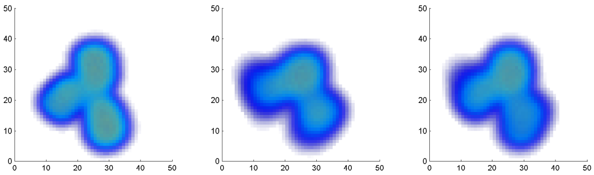

Fig. 6. Left: A template derived from PDB, Middle: A p-template derived from 7 instances, Right: A p-template using both PDB-data and the 7 instances.

was empirically determined by looking at the template responses for varying sigma as shown in Figure 7 to the right. The summarized response results are shown in Figure 7 to the left. The two variants of the p-template give the highest correlation scores on average. A MannWhitney U test, which should be suitable for small sample sizes, was used to test the null hypothesis that the p-template correlation scores and the PDB-template scores are samples from continuous distributions with the same median. It rejects the null hypothesis at the 5% significance level with a p value of 0.028, thus indicating a response difference between the templates.

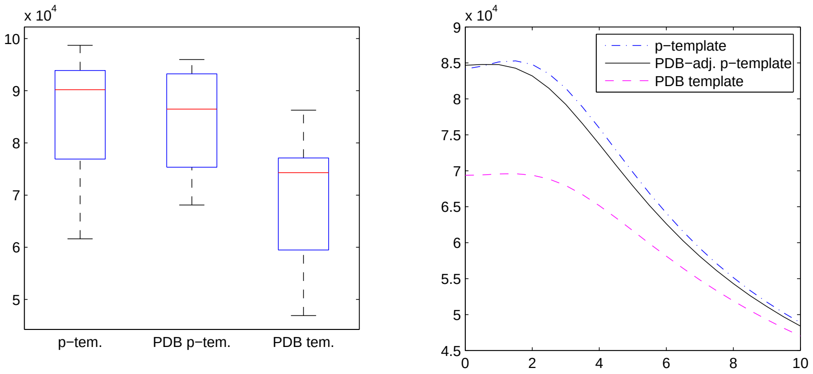

Fig. 7. Left: Correlation responses for the standard p-template, PDB-adjusted p-template and PDB template, respectively. The standard p-template performs the best. Right: Average correlation response when increasing the kernel size (sigma) in the Gaussian filtering. The best correlation response is achieved using sigma = 1.5, but the p-template outperforms the PDB-template when using different smoothing as well.

5 Conclusion and discussion

We show that the p-template is favourable to use when performing correlation in MET images, instead of templates derived from static PDB data. We suggest to use the p-template as a first step when analyzing a MET volume, followed by investigation of high response positions with deformable registration. A current disadvantage with the methodology is that few instance examples are available for different proteins, but as these become more available, the p-templates will be even better and easier to generate.

Concerning the IgG test case, one can note that we only tested on 8 MET volumes altogether. This is of course not optimal but a rather realistic case. MET volumes require a lot of effort to produce and they are also not easy to visually analyze and annotate (which is why improved registration methods are required). It would of course also be beneficial to have a general alignment procedure for any molecule, but that would require a large dataset of molecules for training and verification, and is beyond the scope for this article.

6 Acknowledgements

This work is part of the ProViz project and funded through the *Visualization Program* by the Knowledge Foundation, Vårdal Foundation, the Foundation for Strategic Research, VINNOVA, and Invest in Sweden Agency. Thanks to Anders Brun and Ingela Nyström for feedback on the ideas behind this work.

References

1. Wriggers, W., Chacon, P.: Modeling tricks and fitting techniques for multi-resolution structures. Structure **9** (2001) 779 – 788
2. Ding, L., Goshtasby, A., Satter, M.: Volume image registration by template matching. Image and Vision Computing **19** (2001) 821–832
3. Brunelli, R.: Template Matching Techniques in Computer Vision: Theory and Practice. Wiley Publishing (2009)
4. Sintorn, I.M., Homman-Loudiyi, M., Söderberg-Nauclér, C., Borgefors, G.: A refined circular template matching method for classification of human cytomegalovirus capsids in tem images. Computer Methods and Programs in Biomedicine **76**(2) (2004) 95–102
5. Wriggers, W., Milligan, R.A., McCammon, J.A.: Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. Journal of Structural Biology **125**(2-3) (1999) 185 – 195
6. Wriggers, W., Birmanns, S.: Using situs for flexible and rigid-body fitting of multiresolution single-molecule data. Journal of Structural Biology **133**(2-3) (2001) 193 – 202
7. Tama, F., Miyashita, O., III, C.L.B.: Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. Journal of Molecular Biology **337**(4) (2004) 985 – 999
8. Trabuco, L.G., Villa, E., Mitra, K., Frank, J., Schulten, K.: Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. Structure **16**(5) (2008) 673 – 683
9. Förster, F., Hegerl, R.: Structure determination in situ by averaging of tomograms. In McIntosh, J.R., ed.: Cellular Electron Microscopy. Volume 79 of Methods in Cell Biology. Academic Press (2007) 741 – 767
10. Svensson, L., Nysjö, J., Brun, A., Nyström, I., Sintorn, I.M.: Rigid template registration in met images using cuda. In: VISAPP (1). (2012) 418–422
11. Sandin, S., Öfverstedt, L.G., Wikström, A.C., Wrange, O., Skoglund, U.: Structure and flexibility of individual immunoglobulin G molecules in solution. Structure **12** (2004) 409–415
12. Carr, H., Snoeyink, J., Axen, U.: Computing contour trees in all dimensions. Computational Geometry **24**(2) (2003) 75 – 94
13. Schönemann, P.: A generalized solution of the orthogonal procrustes problem. Psychometrika **31** (1966) 1–10
14. Pittet, J.J., Henn, C., Engel, A., Heymann, J.B.: Visualizing 3D data obtained from microscopy on the internet. Journal of Structural Biology **125** (1999) 123–132