DETECTION AND IDENTIFICATION OF MACROMOLECULAR COMPLEXES IN CRYO-ELECTRON TOMOGRAMS USING SUPPORT VECTOR MACHINES

Yuxiang Chen[1, 2], Thomas Hrabe[1], Stefan Pfeffer[1], Olivier Pauly[2, 3], Diana Mateus[2, 3], Nassir Navab[2], Friedrich Förster[1]

[1]Molecular Structural Biology, Max Planck Institute of Biochemistry, Martinsried, Germany
[2]Computer Aided Medical Procedures (CAMP), Technische Universität München, Germany
[3]Institute of Biomathematics and Biometry, Helmholtz Zentrum München, Germany

ABSTRACT

Detection and identification of macromolecular complexes in cryo-electron tomograms is challenging due to the extremely low signal-to-noise ratio (SNR). While the state-of-the-art method is template matching with a single template, we propose a 3-step supervised learning approach: (*i*) pre-detection of candidates, (*ii*) feature calculation, and (*iii*) final decision using a support vector machine (SVM). We use two types of features for SVM: (*i*) correlation coefficients from multiple templates, and (*ii*) rotation invariant features derived from spherical harmonics. Experiments conducted on both simulated and experimental tomograms show that our approach outperforms the state-of-the-art method.

***Index Terms*—** Cryo-electron tomography, template matching, support vector machines, spherical harmonics

1. INTRODUCTION

Cryo-electron tomography (CET) is the highest-resolving imaging technique to visualize biological samples in 3D under near-to-native conditions [1]. In CET, 2D projections of the frozen-hydrated sample are obtained from different tilt angles using a transmission electron microscope (TEM) and the sample's 3D density (tomogram) is reconstructed from those images. The attainable resolution of CET is approximately 5-10nm and is sufficiently high to distinguish individual macromolecular complexes, which allows studying the abundance and interactions of macromolecules [2]. Furthermore, CET also enables studying the structures of macromolecules *in situ*, which requires alignment and averaging of different instances of a macromolecule [3].

Both above-mentioned applications require accurate detection of macromolecules in tomograms. However, their localization is hampered by the low SNR of the tomograms (typically 0.1 – 0.01) and the incomplete sampling in Fourier space caused by the limited tilt range of the specimen ("missing wedge" problem) [1].

The state-of-the-art approach for macromolecule detection in CET is template matching [4]: a tomogram is correlated (most commonly "local correlation" [5]) with a structural template of the molecule under scrutiny in different orientations. The maxima of the correlation function indicate possible locations of the target macromolecule (candidates). The template matching approach is widespread due to three major reasons: (*i*) It is robust to noise compared to many other detection approaches [6]. (*ii*) The handling of the "missing wedge" problem can be integrated into the correlation score [4]. (*iii*) The computation is efficient using Fast Fourier Transforms (FFT) [5].

The performance of template matching is nevertheless limited: (*i*) False positive matches occur due to high-contrast features (*e.g.*, membranes and gold fiducials used for projection alignment). (*ii*) Human interaction is required to set the correlation threshold for putative detections.

To reduce the false positive rate and to avoid subjective thresholding, we present a protocol that uses a supervised learning technique (SVM) for the binary classification of the candidates from template matching. For each candidate, we use two different sets of features: correlation coefficients from multiple templates, and rotation invariant features derived from spherical harmonics expansion of sub-tomograms. The former features utilize the power of matched filtering while the latter can be computed fast.

2. METHOD

For an input 3D tomogram V, the objective of detection and identification of a target macromolecular complex in V is to find a set of sub-tomograms $\{v_1,\cdots,v_n\}$ containing the replicas of the molecule, and their corresponding positions $\{\vec{p}_1,\cdots,\vec{p}_n\}$ (center of mass) and orientations $\{\vec{r}_1,\cdots,\vec{r}_n\}$ (Euler angles).

2.1. Detection and identification workflow

The proposed workflow (Figure 1, lower part) consists of three steps described in the following.

1. **Template matching and peak extraction**. This is the pre-detection step for obtaining the candidates. Given a tomogram V and the structural template T of the target macromolecular complex, we calculate the six-dimensional

local constrained cross-correlation function $LCCC(\vec{p},\vec{r})$ [4]. Orientations are sampled explicitly and translations are efficiently computed using FFT [5]. The calculation is parallelized in our implementation to gain a further speedup. The candidates $C=\{v'_1,\cdots,v'_{n'}\}$ are found by determining the local maxima of the *LCCC* (peak extraction), yielding their corresponding positions and orientations.

2. **Feature calculation**. After obtaining the candidates, their corresponding features are calculated (section 2.2).

3. **Prediction using SVM**. In this step, a SVM is used to discriminate the true and false positives in the candidates. Their class labels will be predicted by the classifiers described in section 2.3. Finally, all the candidates labeled as the positive class: $\{v_1,\cdots,v_n\}\in C, n\le n'$ are the output of our detection and identification approach.

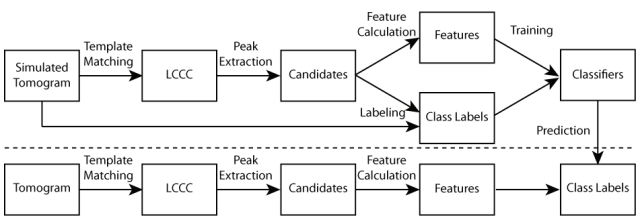

Figure 1. Detection and identification workflow. The panel above the dashed line shows the generation of classifiers.

2.2. Features of sub-tomograms

After the pre-detection step, we characterize the sub-tomograms of candidates using features that are robust to noise and efficient to compute. Here, we propose two sets of features based on correlations from multiple templates and spherical harmonics.

2.2.1. Multi-template based features

Correlation coefficients from multiple templates can reveal more information about the candidates than coefficients from a single template. Given a set of templates $T_1,\cdots,T_n$, the features for the candidate at position $\vec{p}$ are calculated as

$$(\max_{\vec{r}} LCCC_{T_1}(\vec{p},\vec{r}),\cdots,\max_{\vec{r}} LCCC_{T_n}(\vec{p},\vec{r})), \quad (1)$$

where $LCCC_{T_1}$ is the correlation function using T_1 as the template. In practice, the peak value is searched in a small area around $\vec{p}$ in order to account for the possible variations of the peak position due to different templates and the noise. For each sub-tomogram, the computational complexity of these features is $O(TM\cdot N\log N)$, where T is the number of templates, M is the number of sampled orientations, and N is the number of voxels of the sub-tomogram.

2.2.2. Spherical harmonics based features

For the second feature set we use spherical harmonics to construct a rotation invariant descriptor for 3D data [7]. Rotation invariant features are attractive here because they do not require an exhaustive angular search and are hence fast to compute.

The first step of calculating these features for a sub-tomogram v centered at the position of a candidate is to convert v to spherical coordinates:

$$v(x,y,z)=f(r,\theta,\phi). \quad (2)$$

For the spherical coordinates holds $r\in[0,R]$, $\theta\in[0,\pi]$ and $\phi\in[0,2\pi)$. Here, we use spline interpolation for the non-uniformly sampled data points. For each radius, $f(r,\theta,\phi)$ is expanded by spherical harmonics transform and its expansion coefficients are:

$$\hat{f}_l^m(r)=\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi} f(r,\theta,\phi)\overline{Y_l^m(\theta,\phi)}. \quad (3)$$

Herein, Y_l^m is the spherical harmonic of degree l and order m and $\overline{Y_l^m}$ is the complex conjugate of Y_l^m. Since rotating a spherical function will not change the L_2-norm within each frequency l, the vector

$$(\|\hat{f}_0^0(r)\|,\sum_{m=-1}^{1}\|\hat{f}_1^m(r)\|,\cdots,\sum_{m=-L}^{L}\|\hat{f}_L^m(r)\|) \quad (4)$$

is rotation invariant.

We note the following about this vector: (*i*) The computational complexity is $O(R\cdot L^2\log^2 L)$, where R is the number of radii and L is the maximal degree of the spherical harmonics transform. (*ii*) Strictly speaking, the vector will differ to some extent for sub-tomograms depicting the same macromolecules in different orientations because the unsampled regions in Fourier space ("missing wedge") depend on the orientation of the macromolecules. Nevertheless, as we shall see below the variations are sufficiently small to allow for discrimination of different macromolecules.

2.3. Generation of classifiers

Due to the low SNR of experimental tomograms, it is difficult to obtain a ground truth of the identities of sub-tomograms. Here, we simulate tomograms as the ground truth. To make the simulation as realistic as possible, it is conducted as follows: (*i*) Obtain atomic models of different molecules from the Protein Data Bank (PDB) and convert them to density maps of appropriate pixel size and resolution. (*ii*) Randomly position and rotate duplicates of the density maps in a volume. Moreover, beads of various sizes with the density of gold are added. (*iii*) Project the volume along different directions according to defined angular tilt geometry. (*iv*) Add noise to the projections, convolute them with the contrast transfer function (CTF), and add further noise [8]. The projections are low-pass filtered to the first zero-crossing of the CTF. (*v*) Reconstruct the tomogram from the projections using weighted back-projection.

After the simulation, the steps 1 and 2 described in section 2.1 have to be executed on the simulated tomograms to

obtain the candidates and their features. Subsequently, the class labels of all candidates are determined by their distances to the ground truth locations. If the candidate is closer than a threshold T, it is labeled as positive, otherwise negative. Finally, the labels and corresponding feature vectors constitute the training set of the SVM.

LIBSVM [9] is used as the implementation of SVM. Specifically, we choose the RBF kernel for training and the best parameters C and γ are determined by a grid search. Additionally, five-fold cross-validation is applied to avoid overfitting. To account for the unbalanced training set, different weights are assigned to the classes according to the quantity of the samples in each class. The obtained classifiers are then used to predict the class labels of the incoming candidates from a new tomogram.

3. EXPERIMENTS

We first generated the classifiers as described in section 2.3. Ten tomograms ($512 \times 512 \times 512$ voxels) were simulated, each of which contained 5 different types of abundant objects (30 copies of each): 80S ribosome (PDB ID: 3IZS, 3IZF, 3IZB and 3IZE), 60S ribosome (PDB ID: 3IZB and 3IZE), 20S proteasome (PDB ID: 1PMA), GroEL (PDB ID: 1SS8) and gold beads of different sizes. All of the tomograms were simulated with defocus value 4 μm and pixel size 0.47 nm. The resulting tomograms were finally binned twice (pixel size 1.88 nm) to be consistent with typical processing of experimental tomograms.

In this paper we focus on the identification of 80S ribosomes (positive class). After template matching 1000 peaks (more than three times the amount of 80S ribosomes) were extracted to ensure a high coverage of the positive class. The class labels of the candidates were then determined. In this case, 258 candidates were labeled as the positive class (86% coverage) and the remaining 742 as the negative class.

The two feature sets were computed for all candidates (section 2.2). For the multi-template based features (MT), 3 templates (80S ribosome, 60S ribosome and 20S proteasome) were used, resulting in a 3 dimensional feature vector. For the spherical harmonics based features (SH), the radii for decomposing the sub-tomograms ranged from 1 to 7 voxels, and L was set to 16. Consequently, the dimension of the SH feature space was 119.

The features and the class labels of the candidates formed the training sets for the SVMs. After the training, the obtained classifiers were evaluated both on simulated and experimental tomograms in the following.

3.1. Application to the simulated tomograms

Ten additional tomograms were simulated as the ground truth for the assessment. After the detection and identification protocol was applied with both feature sets, the class labels of the test sets were predicted. As a comparison, we also evaluated the state-of-the-art approach (template matching with a single template followed by thresholding, ST).

The results are shown in Table 1: both SVM approaches perform vastly superior to the ST approach on simulated data. Interestingly, even though MT has far fewer features than SH, both of them have similar capabilities to distinguish 80S ribosomes. Furthermore, the ROC curves of all classifiers are shown in Figure 2. MT and SH clearly overcome the ST approach. However, the qualitative behaviors of the MT and SH curves differ slightly: the MT approach performs better at low false positive rates whereas the SH approach is superior at high false positive rates.

Features	Accuracy	Precision	Recall
MT	96.2%	94.9%	91.5%
SH	96%	91.8%	94.4%
ST	61.4%	32.4%	33.1%

Table 1. Identification results on simulated tomograms. For a fair comparison, the threshold for ST was set such that the positive class had roughly the same amount as MT and SH.

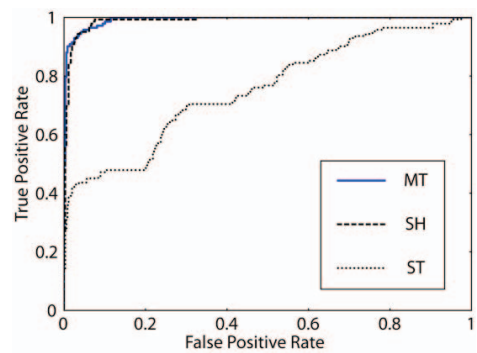

Figure 2. ROC curves of classifications for 80S ribosomes on simulated tomograms. The correlation threshold is varied for plotting the ROC curve of ST.

3.2. Application to an experimental tomogram

The performances of the proposed approaches were further evaluated on an experimental tomogram of endoplasmic reticulum (ER) microsomes derived from canine pancreas (Figure 3a). The tomogram (tilt range: -60° to +60°, 3° increment) was acquired on a FEI Tecnai Polara TEM equipped with a Gatan GIF 2002 energy filter (300 kV acceleration voltage, 4 μm defocus, object pixel size 0.47 nm).

After template matching with the 80S ribosome template (Figure 3b), 500 peaks were extracted and these candidates were subjected to classification. As a result, MT predicted 222 as positives and 278 negatives, while the numbers from SH were 224 and 276, respectively. Due to the lack of the ground truth, we first evaluated the resulting positive classes based on their averages (Figure 3c), as obtained by sub-tomogram alignment [10]. The averages of the MT-

and SH-positives exhibited ribosome-specific features and readily distinguishable ER-membranes, in contrast to the ST average, which was clearly affected by false positives with strong signals, probably gold beads. Thus, the averages suggest improvements of detection accuracy by our approaches.

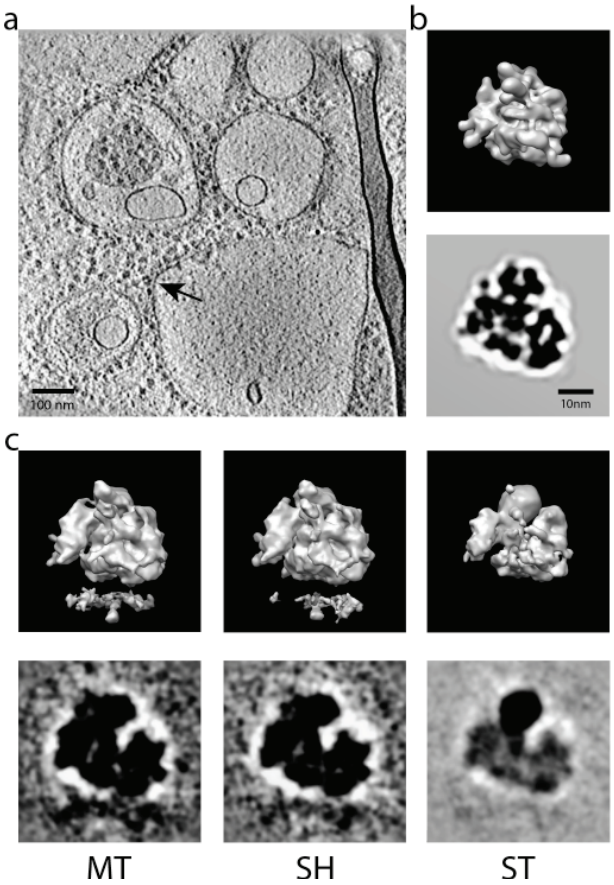

Figure 3. (a) A slice of an experimental tomogram. The arrow points to an ER-associated 80S ribosome. (b) The isosurface of the template of 80S ribosome (upper row) and its corresponding central 2D slice (lower row). (c) The averages of the positive class from various identification approaches. The additional densities in the lower parts of MT and SH maps are the ER-membranes.

Features	Accuracy	Precision	Recall
MT	81%	88.3%	74%
SH	79%	85.7%	72.5%
ST	57%	61.2%	51.7%

Table 2. Identification results on an experimental tomogram based on the ground truth from manual labeling.

For a further assessment, the candidates were manually labeled by experts. According to the subjective decisions, the results were evaluated (Table 2), which also indicate clear improvements over the ST approach, even if the SVM was trained using simulated data. Hence, by improving the quality of simulation further improvements can be expected.

4. CONCLUSION

We presented a protocol to detect and identify macromolecular complexes in cryo-electron tomograms using a 3-step approach with two feature sets: MT and SH. The MT approach accurately incorporates the "missing wedge" problem, but is computationally slower than the SH approach. Moreover, based on the results from simulated tomograms, the MT approach is more powerful when high specificity is targeted whereas the SH approach performs better for high sensitivity. When applied to ribosomes, both approaches yield superior results to the state-of-the-art approach (ST) in simulated and experimental tomograms. It remains to be explored whether combining these two sets of features can provide significant further improvements. While the method was applied to detect ribosomes here, it is expected to provide similar advantages for the detection of other complexes. In the future it will be interesting to determine the minimum mass of complexes that can be detected in tomograms with satisfactory specificity and sensitivity.

5. REFERENCES

[1] V. Lucic, F. Förster, and W. Baumeister, "Structural studies by electron tomography: from cells to molecules," *Annu. Rev. Biochem.*, vol. 74, pp. 833–865, 2005.

[2] S. Nickell, C. Kofler, A. P. Leis, and W. Baumeister, "A visual approach to proteomics," *Nature Rev. Mol. Cell Biol.*, vol. 7, pp. 225–230, 2006.

[3] A. Bartesaghi and S. Subramaniam, "Membrane protein structure determination using cryo-electron tomography and 3D image averaging," *Curr. Opinion Struct. Biol.*, vol. 19, pp. 402–407, 2009.

[4] A. S. Frangakis et al., "Identification of macromolecular complexes in cryoelectron tomograms of phantom cells.," *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 14153-8, 2002.

[5] A. M. Roseman, "Particle finding in electron micrographs using a fast local correlation algorithm.," *Ultramicroscopy*, vol. 94, pp. 225-36, 2003.

[6] Y. Zhu et al., "Automatic particle selection: results of a comparative study," *J. Struct. Biol.*, vol. 145, pp. 3-14, 2004.

[7] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3D shape descriptors," *Symposium on Geometry Processing*, pp. 167–175, 2003.

[8] M. Beck, J. A. Malmström, V. Lange, A. Schmidt, E. W. Deutsch, and R. Aebersold, "Visual proteomics of the human pathogen Leptospira interrogans," *Nature Methods*, vol. 6, pp. 817–823, 2009.

[9] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1-27:27, 2011.

[10] Thomas Hrabe, Yuxiang Chen, Stefan Pfeffer, Luis Kuhn Cuellar, Ann-Victoria Mangold, Friedrich Förster, "PyTom: A python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis," *J. Struct. Biol.*, 2011.