Received: 18 June 2018 | Accepted: 15 August 2018

DOI: 10.1002/pamm.201800246

# Feature-based Classification of Protein Networks using Confocal Microscopy Imaging and Machine Learning

**Pouyan Asgharzadeh**[1,2,*], **Bugra Özdemir**[3], **Ralf Reski**[3,4], **Annette I. Birkhold**[1], and **Oliver Röhrle**[1,2]

[1] Institute of Applied Mechanics (CE), University of Stuttgart, Pfaffenwaldring 7, 70569 Stuttgart / Germany
[2] Stuttgart Centre for Simulation Science (SC Simtech), Pfaffenwaldring 5a, 70569 Stuttgart / Germany
[3] Plant Biotechnology, Faculty of Biology, University of Freiburg, Schaenzlestr. 1, 79104 Freiburg / Germany
[4] BIOSS – Centre for Biological Signalling Research, University of Freiburg, Schaenzlestr. 18, 79104 Freiburg / Germany

Fluorescence imaging has become a powerful tool to investigate complex subcellular structures such as cytoskeletal filaments. Advanced microscopes generate 3D imaging data at high resolution, yet tools for quantification of the complex geometrical patterns are largely missing. Here we present a computational framework to classify protein network structures. We developed a machine-learning method that combines state-of-the-art morphological quantification with protein network classification through morphologically distinct structural features enabling live imaging–based screening. We demonstrate applicability in a confocal laser scanning microscopy (CLSM) study differentiating protein networks of the FtsZ (filamentous temperature sensitive Z) family inside plant organelles (*Physcomitrella patens*).



## 1 Introduction

State-of-the-art imaging techniques allow resolving micro-structural details of protein networks. Computational analysis of these images permits resolving and quantifying the components and assembly of these networks [1], and may allow tracking structural changes in the network caused by internal or external stimuli, connecting the structure to functionality of the network or distinguishing between network types. Here, we present an automated method to classify protein networks based on their structural features exploiting a random forest model.

## 2 Materials and Method

### 2.1 Confocal microscopy imaging

The coding sequences of FtsZ1-2 and FtsZ2-1 were fused with Enhanced Green Fluorescent Protein (EGFP). Imaging was conducted directly on live protoplasts between day 4 and 7. A total of $24$ images of protein networks inside chloroplasts ($n = 12$ FtsZ1-2; $n = 12$ FtsZ2-1) were taken with a confocal laser scanning microscopy (Leica TCS SP8 microscope, Leica Microsystems, Wetzlar, Germany; HCX PL APO $100x/1.40$ oil objective, zoom factor of 10.6; pinhole 0.70 AU ($106.1\mu m$). For the excitation a WLL laser was applied at $488$ nm ($4\%$ intensity). The voxel size was $0.021\mu m$ in $x, y$ and $0.240\mu m$ in z dimension. Deconvolution was performed using Huygens Professional version 17.04 (Scientific Volume Imaging, Hilversum, Netherlands).

### 2.2 Computational feature extraction

Images were segmented using an adaptive local threshold, $T = m + k\sqrt{\frac{1}{NP}\sum_{i=1}^{NP}(p_i - m)^2}$, with $NP = 10 * 10 * 10$ numbers of pixels in the local window, $m$ as average value of pixel intensity in the window, $p_i$ as intensity of pixel $i$ and the constant value $k = 10$ [2]. The wrapped hull of the segmented network is determined. Then, the shape matrix representing the covariance of the coordinates of the voxels of the wrapped hull is calculated as: $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{M}(i) \bigotimes \mathbf{M}(i)]$, where, $\mathbf{M}(i) = \mathbf{X}(i) - \mathbf{C}$, $\mathbf{X}(i)$ is the coordinates of voxel $i$ and $\mathbf{C}$ is the center of the mass. Eigenvalues and eigenvectors of $\mathbf{S}$ are determined. To quantitatively describe the gross morphology of the network as a whole, shape descriptors are calculated: enclosed volume of the network, network volume, network volume density, greatest and smallest diameters of the network, stretch of the network, and oblateness of the network. To analyze the details of the structural components a spatial graph consisting of points, nodes and segments is extracted based on: (i) Detection of edges in the segmented image. (ii) Calculating a distance map (iii) Finding the centerline of each filament based on this distance map. (iv) Placing points at the centerlines in any position, where change in either the thickness or the direction of the filament occurs. Elemental descriptors are determined: number of nodes in the network, node thickness, node density, node-to-node distance, node-to-surface distance, node-to-center distance, node-to-surface to node-to-centre distance ratio, compactness, total number of segments, segment length, segment curvature, mean segment thickness, segment inhomogeneity, mean point-to-point distance, mean number of connections per node, number of open nodes, percentage of open nodes, mean angles between segments for connections with 3 and 4 segments.

* Corresponding author: e-mail pouyan.zadeh@mechbau.uni-stuttgart.de, phone +49 711 685 69254, fax +49 711 685 66347

Details can be found in [2, 3]. Statistical analysis revealed that 7 out of the 25 features are significantly different between the two groups.

### 2.3 Classification algorithm

A random forest classification model using the 25 extracted features is built (maximum depth: 10000, minimum split node zone: 10). A total of 18 random images are used for training, 6 random images for validation. Furthermore, the model is trained with only the significant descriptors to evaluate the necessity of training the model with all features. The designed pipeline for feature-based classification of protein networks using confocal microscopy is shown in **Fig**. 1.

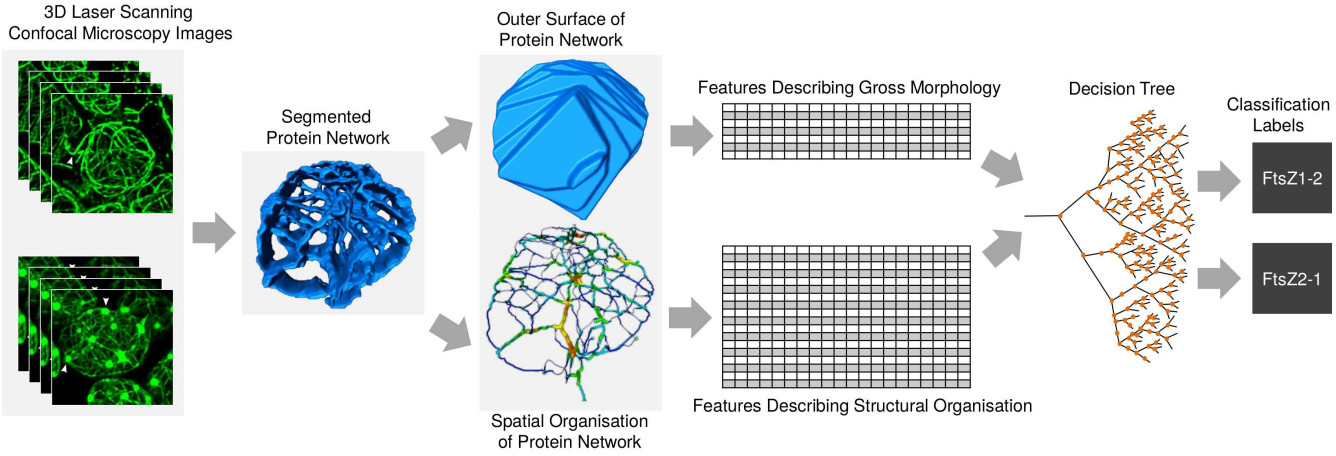


**Fig. 1:** Designed pipeline. 3D CLSM images are segmented, a wrapped hull and spatial graphs are calculated. 25 shape and element descriptors are extracted which are used as input features for training a random forest model to classify FtsZ1-2 and FtsZ2-1 isoforms.

## 3 Results

The 6 randomly selected images for validating the accuracy of the classification method contained 2 FtsZ1-2 and 4 FtsZ2-1 images. The method correctly classified 2 and 3 images of FtsZ1-2 and FtsZ2-1, respectively. If the model was trained only with the significantly different descriptors, then the algorithm classified only one image to be FtsZ1-2 and 2 images to be FtsZ2-1 correctly.

## 4 Discussion

An accuracy comparable to recent classification networks applied to microscopic data [4] was reached. However, in future studies an increased sample size is needed to validate these findings. Extraction of specifically targeted structural features allows reaching high accuracy in classification with few training data while the use of random forest method prevents overfitting. However, our data shows, that training the model with all extracted features produces higher accuracy than training only on significantly different features. Future applications may detect temporal changes in network organization and turnover of filaments, therefore enabling a tracking of structural changes leading to cellular functions, such as shape control for the organelle inevstigated here. Computerized approaches for identification and understanding of complex sub-cellular patterns in microscope images might therefore permit in future applications automated image understanding.

**Acknowledgements** This work was funded by the German Research Foundation (DFG) as a part of Transregio/SFB TRR141, project A09.

## References

[1] P. Asgharzadeh, B. Özdemir, R. Reski, O. Röhrle, and A. I. Birkhold, Acta Biomaterialia **69**, 206–217 (2018).
[2] P. Asgharzadeh, B. Özdemir, S. J. Müller, R. Reski, and O. Röhrle, PAMM **16**(1), 69–70 (2016).
[3] P. Asgharzadeh, B. Özdemir, S. J. Müller, O. Röhrle, and R. Reski (Springer, 2016), pp. 261–275.
[4] O. Z. Kraus, B. T. Grys, J. Ba, Y. Chong, B. J. Frey, C. Boone, and B. J. Andrews, Molecular Systems Biology **13**(4), 924 (2017).

 