

METHODOLOGY ARTICLE

Open Access



Detection and visualization of communities in mass spectrometry imaging data

Karsten Wüllems^{1,2,3*} , Jan Kölling^{1,2}, Hanna Bednarz^{3,4}, Karsten Niehaus^{3,4}, Volkmar H. Hans^{5,6} and Tim W. Nattkemper^{2,3}

Abstract

Background: The spatial distribution and colocalization of functionally related metabolites is analysed in order to investigate the spatial (and functional) aspects of molecular networks. We propose to consider community detection for the analysis of m/z -images to group molecules with correlative spatial distribution into communities so they hint at functional networks or pathway activity. To detect communities, we investigate a spectral approach by optimizing the modularity measure. We present an analysis pipeline and an online interactive visualization tool to facilitate explorative analysis of the results. The approach is illustrated with synthetic benchmark data and two real world data sets (barley seed and glioblastoma section).

Results: For the barley sample data set, our approach is able to reproduce the findings of a previous work that identified groups of molecules with distributions that correlate with anatomical structures of the barley seed. The analysis of glioblastoma section data revealed that some molecular compositions are locally focused, indicating the existence of a meaningful separation in at least two areas. This result is in line with the prior histological knowledge. In addition to confirming prior findings, the resulting graph structures revealed new subcommunities of m/z -images (i.e. metabolites) with more detailed distribution patterns. Another result of our work is the development of an interactive webtool called GRINE (Analysis of **GR**aph mapped **I**mage Data **NE**tworks).

Conclusions: The proposed method was successfully applied to identify molecular communities of laterally co-localized molecules. For both application examples, the detected communities showed inherent substructures that could easily be investigated with the proposed visualization tool. This shows the potential of this approach as a complementary addition to pixel clustering methods.

Keywords: MALDI imaging, Networks, Clustering, Community detection, Visualization, Graphs

Introduction

Matrix-assisted laser desorption ionization mass spectrometry imaging (MALDI-MSI) is a rapidly developing technology for investigating the lateral distribution of molecules in biological samples in form of multivariate bioimages [1].

Due to the technological improvements and the increased utilization of MALDI-MSI, the daily amount of generated data is constantly increasing [2]. Since the complete interpretation cannot be automated, semi-automated and assistive computational methods appear promising and are in the focus of our research.

Different methods for grouping MSI data have already been investigated for the analysis of MSI data, such as: k -means [3], hierarchical clustering [4], hierarchical hyperbolic self-organizing maps [5], high dimensional discriminant clustering [6], or probabilistic latent semantic analysis [7]. Many of these studies focus on clustering of all spectra in one data set to achieve a segmentation

*Correspondence: wuellems@cebitec.uni-bielefeld.de

¹International Research Training Group "Computational Methods for the Analysis of the Diversity and Dynamics of Genomes", Bielefeld University, Universitätsstraße 25, 33613 Bielefeld, Germany

²Biodata Mining Group, Faculty of Technology, Bielefeld University, Universitätsstraße 25, 33613 Bielefeld, Germany

Full list of author information is available at the end of the article



map, i.e. the partition of the image into regions with high intrinsic spectra similarity [5, 6]. In other words: most approaches focus on spectral similarity to group pixels.

The approach presented in this paper focuses on the grouping of molecules into molecular communities. We assume that many functionally related molecules may feature a similar lateral distribution in the sample. Thus, our method groups molecules into communities based on the similarity of their m/z -images. Graphs are well known data structures in biology. Therefore, we propose to use community detection for grouping [8, 9], also known as graph clustering. In our approach, one graph represents one MSI data set of N_V m/z -images. The N_V m/z -images are usually selected by a user and/or an automated selection of N_V peaks. A node v_i of the graph corresponds to one m/z -image $I_{(m/z)_i}$, with $i \in 1, \dots, N_V$, where:

$N_V = \#nodes$ and $\#nodes = \#m/z$ -images.

Each edge $e_k = \{v_i, v_j\}$, with:

$k \in 1, \dots, N_E$ and $i, j \in 1, \dots, N_V$, where:

$N_E = \#edges$

has a weight w_{ij} , which represents the similarity of the spatial signal distribution:

$$w_{i,j} = \text{similarity}(I_{(m/z)_i}, I_{(m/z)_j}) \tag{1}$$

between the m/z -images of nodes v_i and v_j . In its initial form the graph is fully connected. Our goal is to identify communities of similar spatial distribution in order to identify groups of functionally related molecules. The

method is illustrated in Fig. 1 for a hypothetical data set of $N_V = 7$ images and an adjacency matrix leading to three communities.

To the best of our knowledge, community detection is a new approach for MALDI-MSI data. It provides an uncommon view on the data as we focus on groups of similar spatial distributions rather than spectra similarity (pixel similarity). Few previous works have already shown the benefit of the analysis of spatial distributions in MSI ([10, 11]). Moreover, our approach provides a graph structure that serves as an additional source of information.

To tackle the problem of finding communities of m/z -images featuring a similar spatial signal distribution, we developed a modular analysis pipeline consisting of five major blocks : 1. data preprocessing, 2. computation of a $N_V \times N_V$ similarity matrix \mathbf{S} , 3. transforming the similarity matrix into an $N_V \times N_V$ adjacency matrix \mathbf{A} , 4. community detection and 5. interactive visualization. Step 5 aims to obtain additional information from the graph that is not available through the community detection result itself.

Methods

Data sets

MALDI-MSI data forms a three dimensional data cube, where the x -axis and the y -axis represent the lateral coordinates (pixels), which can be represented as intensity images also called m/z -images, while the z -axis represents the mass spectra information. In this study three

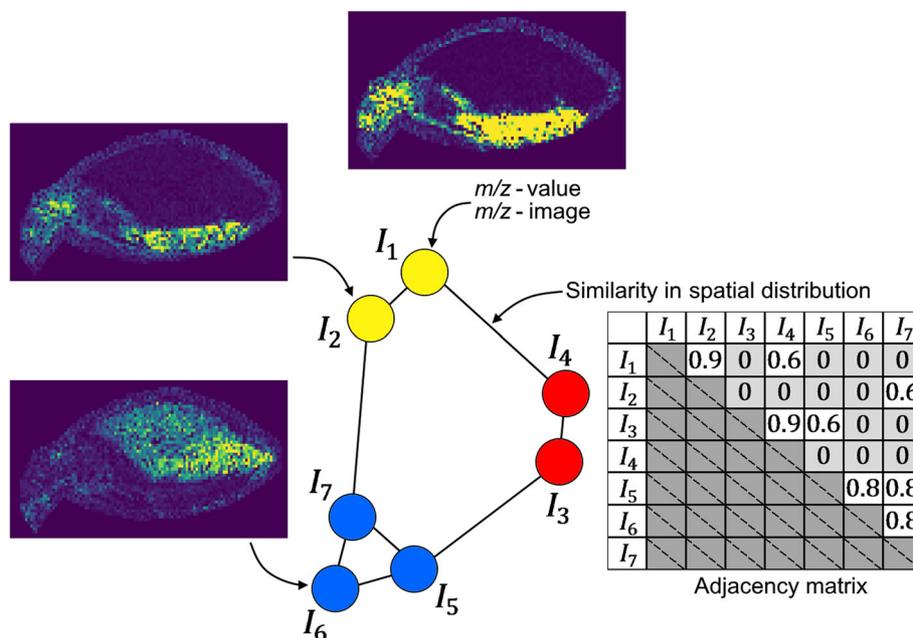


Fig. 1 Structure of the m/z -image similarity graph. Each node represents an m/z -image, each edge represents the similarity between the m/z -images it connects, requiring that this value is above a specific threshold. Each color encodes one community

data sets are used. The first one is a synthetic benchmark data set and consists of nine generated 2D gaussians (D_G) (please find details below), the second data set (D_B) was gathered from a germinating barley seed timeline experiment [12] and the third one (D_T) was recorded from a section of a human glioblastoma tumor [13]. D_B and D_T are in-house produced data sets.

D_G consists of nine synthetic m/z -images $I_0^{(gs)} \dots I_8^{(gs)}$ and is a synthetic $9 \times 205 \times 190$ MSI toy data cube. Each image contains a single localized 2D gaussian intensity distribution. The gaussians were initialized with the same size, a slightly different amplitude and were placed in groups of three:

$$K_0^{(gs)} = \{I_0^{(gs)} I_1^{(gs)} I_2^{(gs)}\},$$

$$K_1^{(gs)} = \{I_3^{(gs)} I_4^{(gs)} I_5^{(gs)}\},$$

$$K_2^{(gs)} = \{I_6^{(gs)} I_7^{(gs)} I_8^{(gs)}\}$$

at three different spatial locations $L_0^{(gs)}, L_1^{(gs)}, L_2^{(gs)}$, respectively. The placement is made in such a way that it is ensured that the three groups overlap with each other in all possible combinations. This is followed by a small random distortion of the position, x size and y size, combined with a randomized rotation. A sketch of the gaussians and their variation is shown in Fig. 2.

If we think of a biological analogy for this experiment, each distorted gaussian represents the distribution of a different molecule. Each location L_i^{gs} , with $i = 0, 1, 2$,

represents the area of a spatially bound metabolic process referred to as pseudo-network, meaning that the molecules distributed in this area are likely to take part in this process.

The original data output of D_B and D_T were transformed to the form: $D = N_p \times N_V$, where N_V is the dimension of vector $\mathbf{x} \in \mathbb{R}^{N_V}$, representing the spectrum information and N_p is the dimension of vector $\mathbf{p} \in \mathbb{N}^{(m \times n)}$ with m and n are width and height of the visual field, representing the lateral information. To be more precise, the elements of \mathbf{p} include only the measuring coordinates of the MALDI procedure, i.e. pixel grid cells. Regarding the rendered m/z -image, (x_i, y_j) are pixels matching the area of the measured sample. Furthermore, in our data sets the mass spectra information $\mathbf{x} = (x_0, \dots, x_{N_V-1})$, called m/z -feature vector, does not represent the whole originally measured spectra, since a set of N_V interesting m/z -values were pre-selected by three of the authors (MG, HB, KN) based on their tissue specific and non-homogenous distribution within the tissue section. Applied to D_B and D_T this results in a dimensionality of:

$$D_B = N_p^{(2)} \times N_V^{(2)} = 3422 \times 101 \text{ and}$$

$$D_T = N_p^{(3)} \times N_V^{(3)} = 28684 \times 106.$$

The preprocessing finishes with winsorizing the upper 1% of intensities for each image:

$$x_l = \begin{cases} Q_{99}(x_l), & \text{if } x_l > Q_{99}(x_l), \forall l \in [0, \dots, N_V - 1] \\ x_l, & \text{otherwise} \end{cases}$$

where Q_{99} is the 99th quantile.

Analysis pipeline

To compute the similarity matrix \mathbf{S} we propose to apply the Pearson correlation coefficient:

$$w_{ij} = \frac{\text{cov}(\mathbf{p}_i, \mathbf{p}_j)}{\sigma_{\mathbf{p}_i} \sigma_{\mathbf{p}_j}} \tag{2}$$

where $\text{cov}(\mathbf{p}_i, \mathbf{p}_j)$ is the covariance of the intensity images $\mathbf{p}_i, \mathbf{p}_j$ of the nodes (i.e. metabolites) v_i, v_j and $\sigma_{\mathbf{p}_i}, \sigma_{\mathbf{p}_j}$ are the standard deviations of $\mathbf{p}_i, \mathbf{p}_j$, respectively. The Pearson correlation coefficient is a commonly used similarity measure in the area of MALDI imaging analysis [14–17] and provides a straight forward interpretation. The result is a similarity matrix \mathbf{S} , with $S_{i,j} = w_{ij}$. Please note that also other symmetric similarity measures can be applied here, such as mutual information or cosine similarity. For more information about considered alternatives we would like to refer the interested reader to S17 of the Additional file 1.

Next, we transform the similarity matrix into an adjacency matrix (step 3) $\mathbf{S} \rightarrow \mathbf{A}$, where \mathbf{A} is a much sparser adjacency matrix by thresholding with t_s :

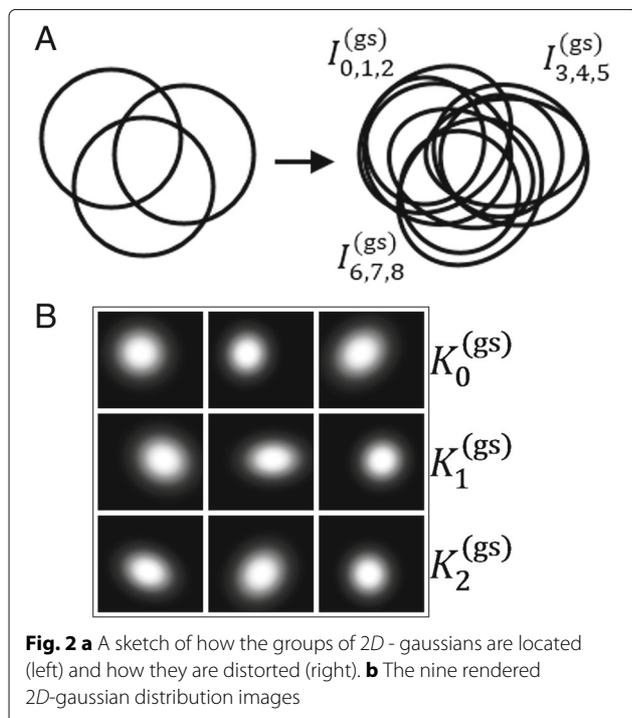


Fig. 2 **a** A sketch of how the groups of 2D - gaussians are located (left) and how they are distorted (right). **b** The nine rendered 2D-gaussian distribution images

$$A_{ij} = \begin{cases} 0, & \text{if } w_{ij} < t_S \\ 1, & \text{otherwise} \end{cases}$$

The objective is to filter out edges with values too low, so that we can assume that these are unlikely to represent a biologically relevant similarity. However, the selection of t_S is a non-trivial task. To avoid time consuming manual tuning we propose a strategy which is inspired by other works on biological network analysis [18–20]. The basic idea is to define an objective function that leads to an adequate threshold after optimization. The objective function is based on quantitative graph properties (QGP). Three QGPs are selected and combined (see [21] for an overview) to determine t_S . The total number of edges N_E , the average clustering coefficient (ζ) [22] and the global efficiency (ξ) [23].

To calculate t_S we define a vector of candidate thresholds:

$$\mathbf{t} = (t_{\min}, \dots, t_{i-1}, t_i, \dots, t_{\max}), \tag{3}$$

where t_{\min} and t_{\max} are the minimum and maximum threshold, respectively and $t_\Delta = t_i - t_{i-1}$ is the step size to reach from t_{\min} to t_{\max} . $[t_{\min}, t_{\max}]$ defines the interval of threshold candidates in which we search for the best possible threshold to reduce the edges in our network. The interval is explored in a discrete manner. This implies that the resolution of the threshold detection is defined by t_Δ , i.e. the distance between two consecutive points t_i to t_{i+1} in $[t_{\min}, t_{\max}]$.

We calculate N_E , ζ and ξ on each graph of an adjacency matrix $A(t_i)$ and arrange the results in vectors v^{N_E} , v^ζ and v^ξ , respectively. Next, we use $v^{N_E} \mapsto [0, 1]$ as baseline to adjust v^ζ and v^ξ :

$$\begin{aligned} \eta^\zeta &= v^\zeta - v^{N_E} \\ \eta^\xi &= v^\xi - v^{N_E} \end{aligned}$$

We create a mean centered matrix $\mathbf{X} = [\eta^\zeta, \eta^\xi]$ and apply PCA as a weighting method. Therefore we calculate \mathbf{y} , which is the projection of \mathbf{X} on the first PCA component:

$$\begin{aligned} \mathbf{X} &= [\eta^\zeta, \eta^\xi] & \text{and} & & \mathbf{X}^{\text{cov}} &= \text{cov}(\mathbf{X}^c) \\ \mathbf{X}^{\text{cov}} \mathbf{u}_i &= \lambda_i \mathbf{u}_i & \text{and} & & \mathbf{y} &= \mathbf{X} \mathbf{u}_0, \end{aligned}$$

where \mathbf{X}^c is the mean centered version of \mathbf{X} , $\{\mathbf{u}_i\}$ are the eigenvectors of the covariance matrix \mathbf{X}^{cov} of \mathbf{X}^c and λ_i are their respective eigenvalues labeled in decreasing order, $\lambda_0 \geq \lambda_1 \geq \dots$. To determine the final threshold we search for the candidate threshold for which the value of \mathbf{y} is maximized. This leads to maximizing the weighted combination of the baselined average clustering coefficient ζ and the global efficiency ξ . Hence, we can set t_S , with:

$$S = \arg \max_k \{\mathbf{y}_k\}, \quad k = 0, 1, \dots, |\mathbf{y}| \tag{4}$$

Since the primary objective is to achieve dense communities, it is a good choice to optimize a segregation measure like ζ . Nevertheless, we do not want to neglect the information provided from edges between communities and integrate ξ , which scales with integration. We use PCA as a weighting method because by construction ζ shows a higher variance than ξ . This leads to a stronger weighting. The idea to combine segregation and integration is based on the small-world property, which occurs frequently in biological networks [19]. The small-world property describes a graph structure of densely connected subgraphs that are interconnected by a robust amount of edges.

N_E serves as a baseline to avoid the effect that low thresholds produce high values for ζ and ξ , which is induced by the construction of these measures. This way the applied measures scale rather with structural properties than with the amount of edges. Since Pearson correlation (Eq. 2) serves as our similarity measure, we set:

$$t_{\min} = -1, \quad t_{\max} = 1, \quad \Delta = 0.1.$$

For t_{\min} , t_{\max} , and t_Δ one has to balance computation time and resolution.

For considered alternatives we refer the interested reader to the Additional file 1: S17.

Now, \mathbf{A} represents an undirected, unweighted graph \mathbf{G} , which serves as basis for the community detection. In \mathbf{G} each node v_i , with $i = 1, \dots, N_V$, where $N_V = \# \text{nodes}$, corresponds to a single m/z -image and is called m/z -node, while each edge $e_k = \{v_i, v_j\}$ indicates that: $w_{ij} > t_S$, with: $k = 1, \dots, N_E$; $i, j \in \{1, \dots, N_V\}$ and $N_E = \# \text{edges}$.

For community detection we use the leading eigenvector method [8, 9]. This method proceeds in a divisive style and maximizes a measure called modularity [24]. Since this is a divisive method, for initialization each m/z -node v_i is assigned into the same community c , with:

$$c \in 1, \dots, N_C \text{ and } v_i = v_i^{c=1} \forall i,$$

where $N_C = \# \text{communities}$.

Thereafter, the method proceeds with:

- 1 For each existing community c its modularity matrix $\mathbf{M}^{(c)}$ is calculated. Informally speaking, for each pair of vertices (v_i, v_j) the respective modularity matrix entry $M_{ij}^{(c)}$ shows the existing number of edges subtracted by the expected number of edges between these vertices (for more detail see [8, 9]).
- 2 The leading eigenvector $\mathbf{u}^{(c)}$ of $\mathbf{M}^{(c)}$ is calculated, which is the eigenvector corresponding to the largest eigenvalue $\lambda_{\max}^{(c)}$.
- 3 (a) If $\lambda^{(c)} > 0$: All $v_i^{(c)}$ are partitioned into two new communities by:

$$v_i^c = \begin{cases} v_i^{(c)}, & \text{if } \mathbf{u}_i \geq 0 \\ v_i^{(c)}, & \text{otherwise} \end{cases}$$

- (b) else: label $v_i^{(c)}$ as “indivisible” and continue with a divisible community.

The procedure repeats for each community until all are labeled as “indivisible”. $\lambda = 0$ is used as stop criteria as its $\mathbf{u} = (1, \dots, 1)$, which means that the best division is to set all v_i in c and none in c' , i.e. the best division is no division.

It is important to mention that the original work [8, 9] does not explicitly mention how to handle disconnected components. However, for MSI data sets disconnected components can be assumed to be quite common. In order to deal with this problem we propose a slight modification of the algorithm, by changing the initialization. Instead of initializing every m/z -node in one community, we search for connected components and set each connected component in its own community. Using this as initialization we follow the leading eigenvector method as described above.

For alternative community detection methods we would like to refer the interested reader again to S17 of the Additional file 1. To facilitate the description of a community size we will use the terminology of (n)-Community, where n provides information about the size.

Visualization

Molecular communities are characterized by two aspects that need to be explored simultaneously: localization and

network structure. To analyse the computed communities in this regard, we propose an interactive visualization framework that links two visualizations for these two aspects. The tool is referred to as GRINE (Analysis of GRaph mapped Image Data NETWORKs) and can be tested for the data described in this paper using the provided links (availability or supplementary). The interface of the tool is shown in Fig. 3. The functionalities are motivated and described below.

To visualize and explore the **network structure display** the user can choose between two different modes: In **graph mode** the communities’ graph structures are visualized, starting with a community graph G' (see Fig. 3a). Each community forms one node $v_{C_i} = \{v_j\}_i$, where $\{v_j\}_i$ is the set of all m/z -nodes with a community membership of i . Two community nodes are connected by a community edge $e_k^{(C)}$, with:

$$e_k^{(C)} = \{v_{C_i}, v_{C_j}\},$$

if there exists an edge $e_l = \{v_p, v_q\}$, with:

$$v_p \in v_{C_i} \text{ and } v_q \in v_{C_j}.$$

The graph is fully draggable and repositions itself by a force layout. The user has the option to expand a community to show its subgraph and edges $e_k^{(H)} = \{v_{C_i}, v_j\}$ which we refer to as hybrid. Hybrid edges are edges between m/z -nodes and community nodes, meaning that an m/z -node of an expanded community is connected with an

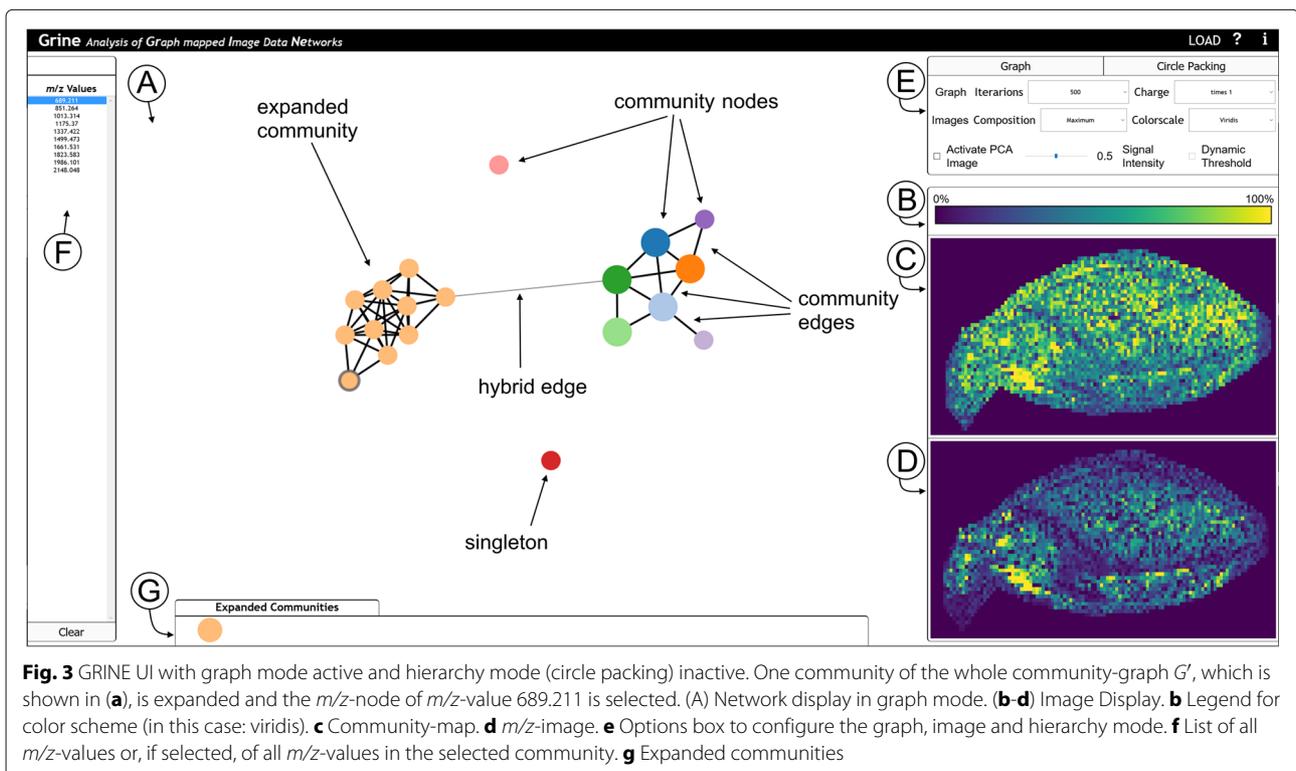


Fig. 3 GRINE UI with graph mode active and hierarchy mode (circle packing) inactive. One community of the whole community-graph G' , which is shown in (a), is expanded and the m/z -node of m/z -value 689.211 is selected. (A) Network display in graph mode. (b-d) Image Display. **b** Legend for color scheme (in this case: viridis). **c** Community-map. **d** m/z -image. **e** Options box to configure the graph, image and hierarchy mode. **f** List of all m/z -values or, if selected, of all m/z -values in the selected community. **g** Expanded communities

m/z -node of a non expanded community. Each node can be selected to activate the image display.

In **hierarchy mode** a circle packing is applied to visualize the networks while hiding the details of the graph structures (i.e. edges). This enables users to focus on community memberships instead (see the Additional file 1: S2 for a screenshot).

To analyse the **localization** of communities and community members, the user selects them either in the graph or in the hierarchy mode, which triggers the visualization of their spatial distribution in the **image display** (see Fig. 3c and d). The upper frame (Fig. 3c) shows the **community map** with a pseudo coloring chosen from a menu (Fig. 3e). The community map is a summary of all images from one selected community $I_{C_i} = D_{p,\{s_l\}i}$, i.e. all m/z -images corresponding to m/z -values s_j that are members of community C_i .

Community maps can be computed and visualized in two modes: In **maximum projection mode** the maximal intensity in the community is displayed for each pixel:

$$\Phi(p'_k) = \max_{s_l}(\Phi(p_{k,\{s_l\}i}),$$

where $\Phi(p'_k)$ is the intensity of pixel p'_k . This mode displays the total area covered by the entire community.

In **averaging mode** the intensity for each pixel is averaged across all images in the community:

$$\Phi(p'_k) = \frac{1}{|\{s_l\}i|} \sum_l \Phi(p_{k,\{s_l\}i}).$$

This emphasizes the quantity of signal coverage.

The lower frame (Fig. 3d) shows the single mass map visualizing one $I_{(m/z)_i}$ image (after selecting this community member in the network display or in the mass list on the far left (Fig. 3f)). The pixel intensities are rescaled for a maximum contrast to enable the visual analysis of weak mass signals.

Furthermore, there is the option to visualize the relation of community localizations with another kind of pseudocolor map, the PCA (principle component analysis) map. This visualization takes the full data set D into account and thus accounts for variances in the entire N_V dimensions. The R, G, B color values in the PCA map are computed with a projection of the full data set onto the three most informative principle components (details given in Additional file 1: S5). This map has been implemented to enable users to integrate global data features. In addition, PCA is a well established and familiar way to analyze high dimensional data so that it can be used as a reference despite its limitations.

Some implementation details can be found in S14 of the Additional file 1.

Finally, we would like to refer the reader to S16 of the Additional file 1 for further information on how the

similarity measure, threshold selection and community detection algorithm influence each other and their impact on the downstream analysis.

Results

Weblinks to all results obtained for data sets: D_G , D_B and D_T can be found under **Availability of data and material**.

Gaussians

For the data set D_G an edge reduction threshold within $t_S \in [0.6382, 0.9397]$ was computed (see Table 1 and Eq. 4). The specific value picked inside of this interval is irrelevant, since the arg max function is maximal over the entire interval. Our community approach detects three communities that corresponds to the groups K_i^{GS} , with $i = 0, 1, 2$, meaning that we can distinguish the gaussians based on their spatial location (see Fig. 4a).

If we discuss this result in relation to our biological analogy, each group K_i^{GS} with distribution at L_i^{GS} consists of molecules that are likely to be representatives of a metabolic process located in this area. Let us remember our initial assumption that functionally related molecules feature a similar lateral distribution within the sample, i.e. metabolic processes are spatially bound. If this assumption holds, the results obtained from D_G indicate that our communities can help to: 1. distinguish metabolic processes based on their spatial location and 2. identify their important molecules.

Figure 4b shows k -means segmentation maps with different k , i.e. clustering of pixel. Even with the correct number of clusters ($k = 4$, i.e. background and three pseudo-networks) the segmentation map cannot distinguish the covered areas at the three different locations.

Compared to k -means clustering or hierarchical clustering, our method does not require to determine the number of groups, which can be considered an advantage.

Barley

For data set D_B we computed the threshold $t_S = 0.7085$ (Eq. 4). This results in $N_E = 789$ edges, meaning a reduction of 84.376% (Table 1). Based on the resulting graph, the leading eigenvector method found $N_C = 11$ communities (see Additional file 1: S4). Nine of them are interconnected, while two are singletons, i.e. nodes

Table 1 Summarized graph information

Dataset	t_S	N_E	N_V	N_C	$N_{C_{2+}}$
Gaussian Circles (D_G)	0.6382	9	9	3	3
Barley (D_B)	0.7085	789	101	11	8
Glioblastoma (D_T)	0.5477	2371	106	11	6

Threshold for edge reduction (t_S), number of edges (N_E), number of vertices (N_V), number of communities (N_C) and number of communities of size greater than two ($N_{C_{2+}}$) for D_G , D_B and D_T are shown

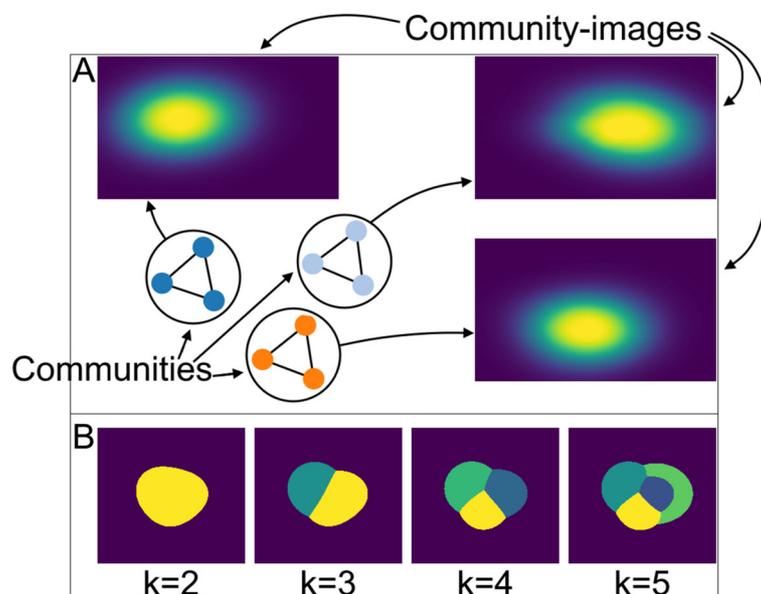


Fig. 4 a Our proposed method was applied to the synthetical D_G data set. The three pseudo-networks were correctly detected as three communities. The communities are displayed as colored graphs (screenshot from the GRINE tool). For each community, the community-map is shown with a viridis color map. **b** k -means segmentation map after clustering of pixel, i.e. m/z -spectra, for $k = 2, \dots, 6$. Each color represents one cluster

without any edge. Eight of the interconnected communities are (n)-Communities, with $n > 1$, the others are (1)-Communities.

Most signal distributions of the community maps (Fig. 5) show a strong correlation to anatomical structures of the barley seed, which is summarized in Fig. 5e.

A view on the graph structure of $C2$ (Fig. 6a) reveals that this community can be divided into more detailed sub-communities (referred to as $C2a - C2c$). $C2b$ shows an increased signal only at the embryo center, while the signal of $C2a$ is less specifically distributed in the entire embryo. $C2c$ is located between both and shows a specific signal

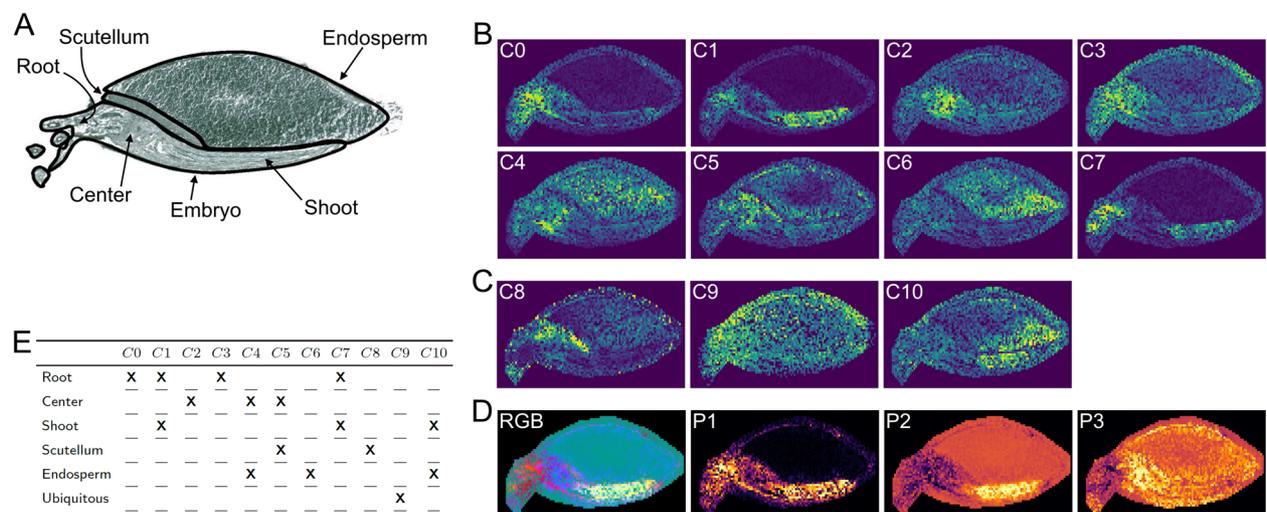
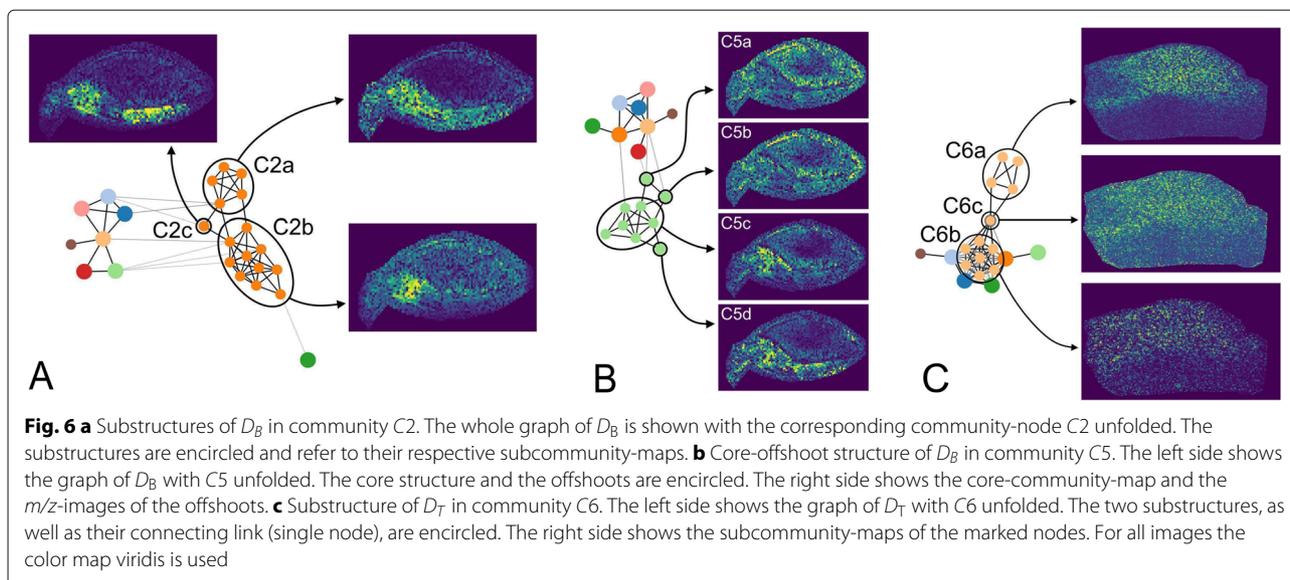


Fig. 5 a Optical image scan with marked and labeled anatomical structures. **b** Average community-maps of all (n)-communities, with $n > 1$ (network in Additional file 1: S4). **c** Images of (1)-Communities (network in Additional file 1: S4). **d** RGB image of the first three PCA projections, where the projections on the eigenvectors of the first, second and third largest eigenvalue is assigned to the red, green and blue channel, respectively and standalone images of these components. PCA images are not scaled like the community-maps and m/z -images. The color map viridis is used for images in (b) and (c) and inferno for images in (d). **e** Correlation between the spatial signal distributions of all found communities and the anatomical structures of the barley seed. **X** indicates that a community shows increased signal in the respective area



distribution at the center and the shoot. A similar observation can be found for C5. The subgraph of C5 (Fig. 6b) shows a structure that can be distinguished into core and offshoots. A core is defined by nodes that are densely interconnected, while offshoots are reaching out from the core and are less interconnected. The core of C5 (C5c) defines the main signal distribution of this community, which extends from the scutellum into the embryo center. The three offshoots C5a, C5b, and C5d deviate from this distribution. A similar core-offshoot differentiation can be observed in C4 (not shown).

The identification of m/z -values based on prior examination of barley seed MSI [12] reveals a tendency for communities to mostly contain one class of molecules. C0, C1, C3 and C7 contain only hordatines and hordatine precursors, with one exception in C0, which is a lipid and three exceptions in C3, which are two unknown molecules and one lipid. C2 and C4 contain mostly carbohydrates, with four exceptions (three unknown molecules and one lipid). Further, carbohydrates in C2 are only potassium adducts and in C4 only sodium adducts. C5 and C6 contain mostly lipids, with two exceptions in C5 that are unknown molecules. The (1)-Communities are unknown (C8, C9) and a lipid (C10). This indicates that similar molecules have similar spatial distributions. One reason for this could be that similar molecules are part of the same spatially bound metabolic processes.

The identification also supports the structural features of C2 and C5. C2a is composed of three unknown molecules, one lipid and one carbohydrate, while C2b consists only of carbohydrates. For C5, the two images that fit least to the main signal distribution of the community are both unknown molecules.

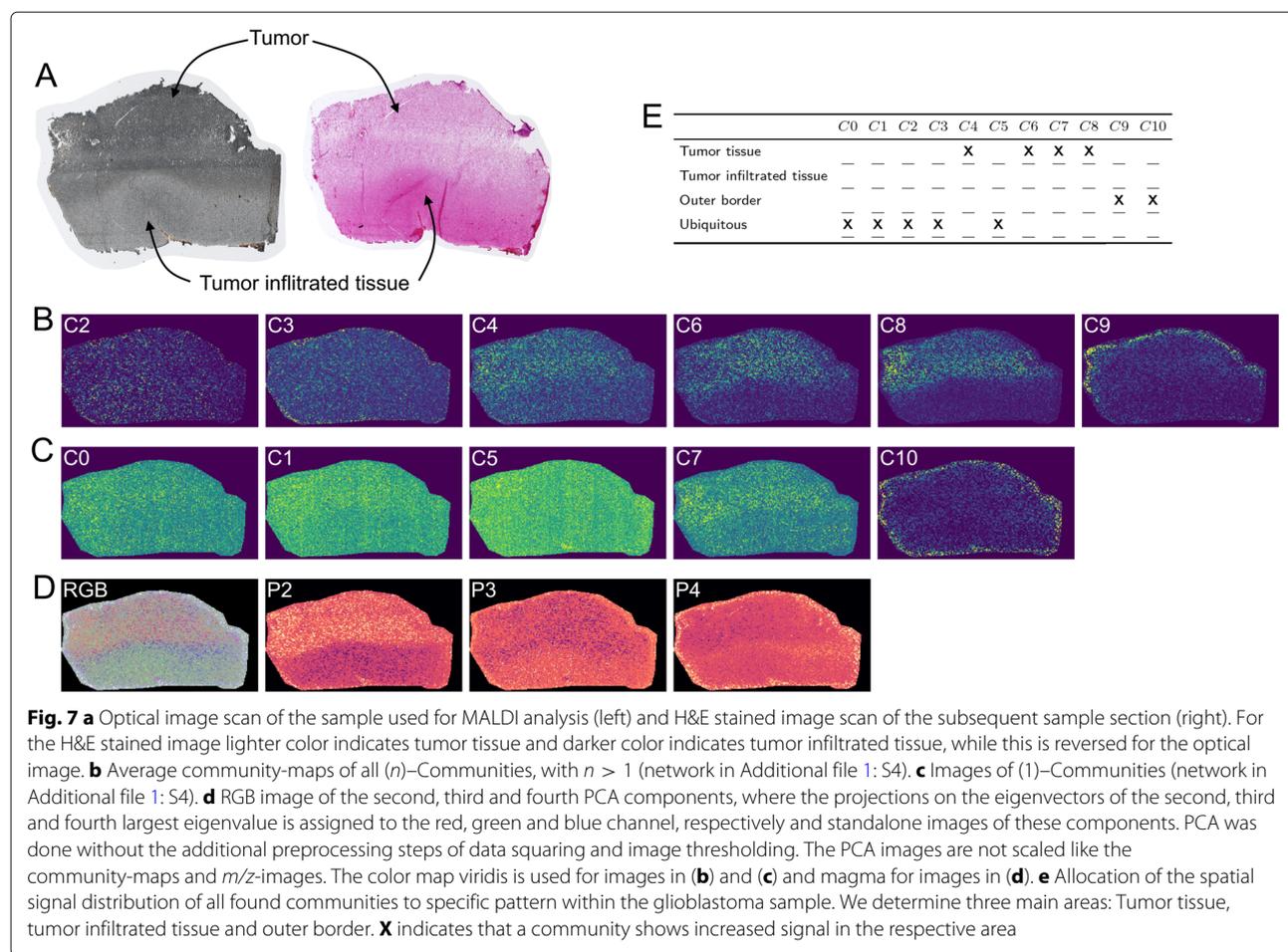
Glioblastoma

For data set D_T we computed the threshold $t_s = 0.5477$ (Eq. 4). The result is $N_E = 2371$ edges, i.e. a reduction of 57.394% (Table 1). Compared to the barley data set the number of edges is clearly higher, although the number of vertices is nearly equal. The reason is a higher general similarity and a lower spread of similarity values, i.e. the algorithm classifies more similarities to be relevant. This indicates a higher degree of complexity for the tissue and its respective network of functionally related molecules. The community detection result shows $N_C = 11$ communities with seven of them interconnected (see Additional file 1: S4). Five are (1)-Communities, the other six are (n)-Communities, with $n > 1$.

The signal distributions (Fig. 7) reveal three main patterns, which are summarized in Fig. 7e.

Similar to the results obtained for barley data, a detailed view on the graph structure reveals more detailed information (Fig. 6c). The subcommunity C6a shows a strong and specific distribution in one half of the sample. C6b is distributed notably less specific, with a slightly biased signal distribution to the same half of the sample as C6a. Both subcommunities are connected by a m/z -image (C6c) that shows a weak similarity to C6a. We assumed that C6c produces a chaining effect during the community detection.

Based on communities C6a and C8 we can conclude that the sample is functionally divided into two halves, which is in line with the PCA result (Fig. 7d) and (more important) the H&E staining information (Fig. 7a), which indicates that the tumor in this sample is side specific. We can presume that at least some molecules of C6a and C8 could be tumor specific.



Results of the publicly available mouse urinary bladder data set from *ms-imaging.org* are shown in Additional file 1: S12. There we provide some basic results without detailed biological interpretation. The results are available for exploration in our webtool. The respective link can be found in Additional file 1: S1.

Discussion

Barley

The analysis of the barley seed data set shows that the community analysis approach delivers reasonable results, i.e. the spatial localizations of the communities reflect biological compartments with distinct functions. This is in accordance with previous findings for this data set [12]. For most communities, we are able to clearly detect correlations with different anatomical structures.

In contrast to other established methods for MSI segmentation, the presented approach offers a very fine identification of the different tissues of a barley seedling based on the mass spectroscopy data. As shown in Fig. 5, the root, the center of the developing seedling, the shoot, the scutellum, and the endosperm could be identified by a

unique combination of communities. This segmentation can be used to analyze the co-localization of specific single mass channels, representing known intermediates of the metabolism.

The fact that certain tissue regions or organs are represented by a number of different communities indicates that these parts of the sample are physiologically more heterogeneous than would be expected if a single m/z -signal were co-localized with that particular tissue or organ. An example for this kind of heterogeneity for the shoot can be seen in the communities C1, C7, and C10. Most interestingly, it shares communities with the root, but not with the scutellum. From a biological point of view, it can be speculated that these differences reflect metabolite compositions that are characteristic for developing tissues, as roots and shoots, versus a tissue, which is metabolically active but not further developing just like the scutellum.

The appearance of substructures in individual communities within the graphs illustrates that our graph approach is able to convey information that would remain hidden if just cluster results were

considered. Interestingly, the three substructures investigated in this study show already three different kinds of motifs: Simple subgroups, core-offshoot structures, and bridging (or chaining) structures. Therefore we believe that substructures are worth further examination.

Glioblastoma

The results of the glioblastoma data set are not as easy to interpret as those of the barley sample, which was to be expected. This is due to its morphological homogeneity, combined with heterogeneity of the cell phenotype. On the other hand the community detection yields at least one clear insight: There are groups of molecules, whose signal distribution correlate with the tumor area that was defined by a pathologist [13]. This provides candidates for subsequent biological experiments.

Regarding their community compositions, the tissue compartments classified as tumor and tumor-infiltrated in data set D_T are much more similar to each other than the different compartments of the barley sample. Five of the eleven communities are categorized as ubiquitous (Fig. 7), reflecting the fact that the tumor tissue is still closely related to the non-tumor tissue. Four communities are tumor-specific (Fig. 7), probably induced by the localization of lactate and other tumor metabolites (see [13]). The last two communities refer to the outer border of the sample (Fig. 7), probably induced by matrix peaks.

We believe that even without any prior knowledge about the sample, like H&E staining, the results offered by this type of analysis provide a good starting point for biologists to set up further experiments.

Visualization

Our visualization tool GRINE is interactive, dynamic and responsive. This makes the usage very intuitive and almost no learning phase is required. The tool shows its main strengths in three areas. First, it combines the information of the graph domain and the image domain. Second, the interaction with the graph facilitates the focus on specific communities and allows to spot structural characteristics. Examples are: Substructures that can indicate more finely resolved communities, cluster ambiguities and potential misclusterings. Third, its possibility to show and hide information, i.e. its interactivity, allows to encode much more information in a clear way than we could achieve with static visualizations [25], e.g. average and maximum images of all communities and correlation with PCA results.

At the current time, the visualization can only deal with distinct communities, whereas the analysis pipeline can also search for overlapping ones.

Comparison to other methodological approaches

A more common approach than the one presented for the analysis of the spatial distribution of imaging data is to employ dimension reduction techniques for segmentation. We compared our method to visualizations of three different dimension reduction techniques: principal component analysis (PCA), non-negative matrix factorization (NMF) and latent dirichlet allocation (LDA) (results are shown and discussed in Additional file 1: S13). We decided for PCA as it is probably the most prominent dimension reduction technique in biology. NMF is also a commonly used technique and does not produce negative intensity values, which can occur in PCA. LDA was chosen because it is a generalization of pLSA (probabilistic latent semantic analysis) that has been previously analysed [7].

The comparison showed that the computed visualizations reveal similar coarse grained structures as our method. It is worth noting that LDA performs better as NMF and NMF performs better than PCA. For D_B and D_T the segmentation maps of LDA reveal the most details and detected structures show the highest contrast. This is followed by the ones obtained with NMF. The PCA maps provide the lowest contrast. All three methods show distributions that correlate with the main structures of the samples. However, compared to our method they fail finding finely detailed structures like the scutellum in D_B .

While the results obtained with PCA, NMF and LDA share similarities with the results obtained by our proposed method, we can report some new favorable features for our approach:

First, the grouping of spatial distributions assigns each image to one group. After analysing the lateral distribution of a community image it is easy and unambiguous to identify which single m/z -images, i.e. molecules, participate in this distribution. This is much harder for PCA, NMF and LDA, where each component image consist of partial combinations of the original m/z -images.

Second, we do not need to determine the number of clusters, i.e. communities, beforehand. Our method chooses this number automatically based on the given optimization criterion (modularity). If needed, a manual decision is still possible. This is different for NMF and LDA. For those methods the number of dimensions, i.e. components, have to be predefined. Finding the most fitting number of dimensions for a given sample is a non trivial task and especially important for NMF and LDA, since the number of dimensions influences the lateral distribution of the resulting components (see Additional file 1: S13).

Third, the community images are based on simple aggregation functions. Therefore, in case of outliers or ambiguities it is easy to re-evaluate the community images without them. The same counts for potential optimizations based on substructures in the clustering space.

Fourth, the network structure can reveal outliers, mis-clusterings, substructures and potential optimizations at first glance and allows an intuitive exploration of the clustering space.

We would also like to add the H²SOM ([5]), as another segmentation method to this comparison discussion. This method is also capable to reveal detailed structures in D_B . A core difference is that in our method a single pixel can be a member of multiple community images. This means we can provide an ambiguous pixel labeling, which is more suitable to represent dynamic biological processes.

Conclusion

In this paper we demonstrated the general applicability of community detection as an unsupervised clustering technique for the analysis of MSI data from different types of samples. We have developed a pipeline to map lateral image data to an image similarity graph. We have also developed a new edge thresholding technique to transform a fully connected graph into a sparse one. Using lateral m/z -images as samples and their pixels as features, we utilized community detection as an example to group molecules with similar lateral distributions. By analysing the network structure with our interactive visualization, we have found finer subclusters within the detected clusters. This offers a possibility for manual refinement.

We stated the initial assumption that functionally related molecules are spatially bound. If this assumption holds, the presented way of clustering lateral imaging data provides a good starting point for targeted biological experiments. For some information on the limitations of this assumption and alternative considerations, we would like to refer to S18 of the Additional file 1.

This paper is designed as proof of concept to demonstrate the general applicability of community detection to MSI data. Therefore, we did not discuss the question of performance. Since we do not want to ignore this question entirely, we refer to S15 of the Additional file 1. There a rough analysis of the complexity is presented.

Finally, a webtool has been implemented to visualize and explain the results and to demonstrate the usefulness and benefits of the approach.

Future research

Further analysis of network patterns of substructures could lead to automated ways to detect finer cluster relationships, like hierarchical structures and reveal and correct misclusterings. Another promising approach, which was not discussed in this paper, is to employ more statistical network properties for the analysis. An example could be the ratio of the in-group degree to the out-group degree to automatically detect very specific or very general lateral patterns. These statistics could also be used to query specific m/z -images for their statistical properties.

Considering the sheer amount of network properties this offers a big area for future research.

Additional file

Additional file 1: Supplementary. (PDF 2459 kb)

Abbreviations

H²SOM: Hierarchical hyperbolic self-organizing map; H&E: Hematoxylin and eosin; GRINE: Analysis of **GR**aph mapped **I**mage data **NE**tworks; LDA: Latent dirichlet allocation; m/z : Mass to Charge; MALDI: Matrix-assisted laser desorption ionization; MSI: Mass spectrometry imaging; NMF: Non-negative matrix factorization; PCA: Principal component analysis; pLSA: Probabilistic latent semantic analysis; QGP: Quantitative graph properties

Acknowledgements

Not applicable.

Funding

KW is funded and JK was partially funded by the International DFG Research Training Group GRK 1906. The funding body played no role regarding the design of the study, data collection, analysis or interpretation or the writing of the manuscript.

Availability of data and materials

All webapplication results are accessible at: [Gaussians, Barley, Glioblastoma](#) via Chrome or Firefox. All links are also given explicitly in Additional file 1: S1. Data sets and demo code for network building and community detection: [Data sets, Code](#).

Demo code for the visualization: [GRINE](#)

A small overview video for our webtool is available at:

<https://www.youtube.com/watch?v=l40fe8TUjjs>.

Supplementary information: Supplementary data is available online.

Availability and Requirements

The source code and demos for the community detection workflow the visualization (GRINE) is available on github.

Project name: MSI Community Detection

Project home page:

https://github.com/Kawue/imaging_communitydetection_demo

Operating system(s): Platform independent

Programming language: Python

Other requirements: Python 3.5 or higher

License: GNU GPLv3

Project name: GRINE

Project home page: https://github.com/Kawue/grine_demo

Operating system(s): Platform independent

Programming language: Javascript

Other requirements: Firefox or Chrome

License: GNU GPLv3

Author's contributions

Concept and design of study and software: KW, JK, TWN. Developed and implemented software: KW. Analyzed data: KW. Evaluation and interpretation of results: KW, JK, HB, KN, TWN. Contributed data: JK, HB, KN. Pathological structure analysis of the glioblastoma sample: VHH All authors read and approved the final manuscript.

Ethics approval and consent to participate

The collection and use of human tissue was approved by the ethical review committee of the University Duisburg-Essen. All patients (donors of the human tissue) provided their informed consent in written form. We obtained the data completely anonymised.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹International Research Training Group "Computational Methods for the Analysis of the Diversity and Dynamics of Genomes", Bielefeld University, Universitätsstraße 25, 33613 Bielefeld, Germany. ²Biodata Mining Group, Faculty of Technology, Bielefeld University, Universitätsstraße 25, 33613 Bielefeld, Germany. ³Center for Biotechnology (CeBiTec), Universitätsstraße 25, 33613 Bielefeld, Germany. ⁴Proteome and Metabolome Research, Faculty of Biology, Bielefeld University, Universitätsstraße 25, 33613 Bielefeld, Germany. ⁵Department of Neuropathology, Institute for Clinical Pathology, Dietrich-Bonhoeffer-Klinikum, Salvador-Allende-Straße 30, 17036 Neubrandenburg, Germany. ⁶Department of Neuropathology, Essen University Hospital (AÖR), Hufelandstraße 55, 45147 Essen, Germany.

Received: 28 January 2019 Accepted: 10 May 2019

Published online: 04 June 2019

References

- Herold J, Loyek C, Nattkemper TW. Multivariate image mining. *Wiley Interdiscip Rev Data Min Knowl Disc.* 2011;1(1):2–13.
- Palmer A, Trede D, Alexandrov T. Where imaging mass spectrometry stands: here are the numbers. *Metabolomics.* 2016;12(6):1–3.
- McCombie G, Staab D, Stoeckli M, Knochenmuss R. Spatial and spectral correlations in maldi mass spectrometry images by clustering and multivariate analysis. *Anal Chem.* 2005;77(19):6118–24.
- Deininger S-O, Ebert MP, Fütterer A, Gerhard M, Röcken C. Maldy imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *J Proteome Res.* 2008;7(12):5230–6.
- Kölling J, Langenkämper D, Abouna S, Khan M, Nattkemper TW. White-a web tool for visual data mining colocation patterns in multivariate bioimages. *Bioinformatics.* 2012;28(8):1143–50.
- Alexandrov T, Becker M, Deininger S-O, Ernst G, Wehder L, Grasmair M, von Eggeling F, Thiele H, Maass P. Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. *J Proteome Res.* 2010;9(12):6535–46.
- Hanselmann M, Kirchner M, Renard BY, Amstalden ER, Glunde K, Heeren RM, Hamprecht FA. Concise representation of mass spectrometry images by probabilistic latent semantic analysis. *Anal Chem.* 2008;80(24):9649–58.
- Newman ME. Modularity and community structure in networks. *Proc Natl Acad Sci.* 2006;103(23):8577–82.
- Newman ME. Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E.* 2006;74(3):036104.
- Alexandrov T, Chernyavsky I, Becker M, von Eggeling F, Nikolenko S. Analysis and interpretation of imaging mass spectrometry data by clustering mass-to-charge images according to their spatial similarity. *Anal Chem.* 2013;85(23):11189–95.
- Wijetunge CD, Saeed I, Halgamuge SK, Boughton B, Roessner U. Unsupervised learning for exploring maldi imaging mass spectrometry 'omics' data. In: *Information and Automation for Sustainability (ICIAFS), 2014 7th International Conference On.* IEEE; 2014. p. 1–6.
- Gorzolka K, Kölling J, Nattkemper TW, Niehaus K. Spatio-temporal metabolite profiling of the barley germination process by maldi ms imaging. *PLoS ONE.* 2016;11(3):0150208.
- Giampa M, Lissel M, Patschkowski T, Fuchser J, Hans VH, Gembruch O, Bednarz H, Niehaus K. Maleic anhydride proton sponge as a novel MALDI matrix for the visualization of small molecules (< 250 m/z) in brain tumors by routine MALDI ToF imaging mass spectrometry. *Chem Commun.* 2016;52(63):9801–4.
- Alexandrov T. Maldy imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics.* 2012;13(16):11.
- Widlak P, Mrukwa G, Kalinowska M, Pietrowska M, Chekan M, Wierzgon J, Gawin M, Drazek G, Polanska J. Detection of molecular signatures of oral squamous cell carcinoma and normal epithelium—application of a novel methodology for unsupervised segmentation of imaging mass spectrometry data. *Proteomics.* 2016;16(11–12):1613–21.
- Verbeeck N, Yang J, De Moor B, Caprioli RM, Waelkens E, Van de Plas R. Automated anatomical interpretation of ion distributions in tissue: linking imaging mass spectrometry to curated atlases. *Anal Chem.* 2014;86(18):8974–82.
- McDonnell LA, van Remoortere A, van Zeijl RJ, Deelder AM. Mass spectrometry image correlation: quantifying colocalization. *J Proteome Res.* 2008;7(8):3619–27.
- Zahoránszky-Kóhalmi G, Bologa CG, Oprea TI. Impact of similarity threshold on the topology of molecular similarity networks and clustering outcomes. *J Cheminformatics.* 2016;8(1):16.
- Humphries MD, Gurney K. Network 'small-world-ness': a quantitative method for determining canonical network equivalence. *PLoS ONE.* 2008;3(4):0002051.
- Couto CMV, Comin CH, da Fontoura Costa L. Effects of threshold on the topology of gene co-expression networks. *Mol BioSyst.* 2017;13(10):2024–35.
- Rubinov M, Sporns O. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage.* 2010;52(3):1059–69.
- Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature.* 1998;393(6684):440–2.
- Latora V, Marchiori M. Efficient behavior of small-world networks. *Phys Rev Lett.* 2001;87(19):198701.
- Newman ME, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E.* 2004;69(2):026113.
- Yi JS, ah Kang Y, Stasko J. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Trans Vis Comput Graph.* 2007;13(6):1224–31.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

