Biophysical Journal, Volume 96

Supporting Material

Blind source separation techniques for the decomposition of multiply labeled fluorescence images

Richard A. Neher, Mišo Mitkovski, Frank Kirchhoff, Erwin Neher, Fabian J. Theis, and André Zeug

Supplementary material: Blind source separation techniques for the decomposition of multiply labeled fluorescence images

Richard A. Neher[†], Mišo Mitkovski[‡,§§], Frank Kirchhoff[§,§§], Erwin Neher[¶,§§],

Fabian J. Theis[‖,††], and André Zeug[‡‡,§§,¶¶]

[†] Kavli Institute for Theoretical Physics, University of California, address: Santa Barbara, CA, 93106
[‡] Department of Molecular Biology of Neuronal Signals, Max Planck Institute of Experimental Medicine Göttingen, address: Hermann Rein Strasse 3, 37075 Göttingen, Germany
[§] Glial Physiology and Imaging, Neurogenetics, Max Planck Institute of Experimental Medicine Göttingen, address: Hermann Rein Strasse 3, 37075 Göttingen, Germany
[¶] Department Membranbiophysik, Max-Planck Institute for Biophysical Chemistry Göttingen, address: Am Fassberg 11, 37077 Göttingen, Germany
[‖] Institute for Bioinformatics and Systems Biology, Helmholtz Zentrum München, address: Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany
[††] Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany, address: Bunsenstraße 10, 37073 Göttingen, Germany
[‡‡] Department of Neurophysiology and Cellular Biophysics, University Göttingen, Germany
[§§] Deutsche Forschungsgemeinschaft-Research Center for the Molecular Physiology of the Brain (FZT 103), Göttingen, Germany
[¶¶] present address: Department of Cellular Neurophysiology, Hannover Medical School, Germany, address: Carl-Neuberg-Str. 1, 30625 Hannover, Germany

Correspondence should be addressed to André Zeug (zeug.andre@mh-hannover.de).

Table of contents:

1. Data visualization

In a spectrally resolved image, each pixel corresponds to a vector **y** describing the measurements in the different spectral channels. These vectors are a sum of contributions of the *m* labels in the sample, each of which has a distinct emission spectrum. Since the variation of label concentrations are the main source of signal variation, the vectors **y** lie to a good approximation in a *m*-dimensional plane. This plane is spanned by *m* linearly independent spectra of the dyes and, as it contains the origin, is a subspace of the *n*-dimensional space of possible emission vectors. Variation perpendicular to the *m*-dimensional signal subspace is mainly due to photon shot noise. Since each vector **y** is a combination of nonnegative contributions of the *m* labels, the data points fill only the cone spanned by the spectral vectors of the pure labels.

1.1 Data projection into the plane

While the projection described can be formulated for any number of dyes, we consider the case of *m*=3 dyes in the following. A three-dimensional space cannot be visualized in the plane. However, the absolute signal strength is not relevant for the following consideration. Hence we can normalize all vectors **y** such that $\sum_i |y_i| = 1$ and thereby reduce the relevant subspace to an affine two-dimensional space. To visualize the data in two dimensions, we have to project the data onto the affine 2-*d* subspace and map the latter into the plane.

Assume we are given three spectra $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$ with unit 1-norm that span the signal subspace. For each *n*-dimensional vector **y** with unit 1-norm, we have to determine an orthogonal projection onto the affine 2-*d* space. The projected vector can be written as

$$z = \mathbf{s_3} + \mu_1 \mathbf{t_1} + \mu_2 \mathbf{t_2}, \qquad \text{(S1)}$$

where $\mathbf{t}_1 = \mathbf{s}_1 - \mathbf{s}_3$ and $\mathbf{t}_2 = \mathbf{s}_2 - \mathbf{s}_3$ are vectors within the affine space. From the condition that **y-z** is orthogonal to $\mathbf{t}_1$ and $\mathbf{t}_2$, we can calculate the coordinates within the affine subspace

$$\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} = \mathbf{G}^{-1}\begin{pmatrix}\langle \mathbf{y}-\mathbf{s}_3, \mathbf{t}_1\rangle\\ \langle \mathbf{y}-\mathbf{s}_3, \mathbf{t}_2\rangle\end{pmatrix} \text{ with } \mathbf{G} = \begin{pmatrix}\langle \mathbf{t}_1, \mathbf{t}_1\rangle & \langle \mathbf{t}_1, \mathbf{t}_2\rangle\\ \langle \mathbf{t}_2, \mathbf{t}_1\rangle & \langle \mathbf{t}_2, \mathbf{t}_2\rangle\end{pmatrix}, \tag{S2}$$

where $\langle \mathbf{a},\mathbf{b}\rangle$ is the scalar product between two vectors **a** and **b**. The visualization of the data in the plane should preserve the angles and the relative length of points in the affine 2-d subspace. Both of these requirements are met, if the coordinates v_1 and v_2 in the plane are calculated from μ_1 and μ_2 as follows:

$$\begin{pmatrix}v_1\\ v_2\end{pmatrix} = \begin{pmatrix}\|\mathbf{t}_1\| & \|\mathbf{t}_2\|\cos\alpha\\ 0 & \|\mathbf{t}_1\|\sin\alpha\end{pmatrix}\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} \text{ with } \alpha = \arccos\left(\frac{\langle \mathbf{t}_1, \mathbf{t}_2\rangle}{\|\mathbf{t}_1\|\|\mathbf{t}_2\|}\right). \tag{S3}$$

Note that only relative locations and angles have a meaning, the absolute orientation and position of the representation of the data points in the plane are arbitrary; here for simplicity we mapped one edge onto the first unit vector.

The analogous projection for two-dimensional data maps the data onto a line segment. The end-points of the line segment correspond to pure dyes while the points in the interior correspond to mixtures of dyes. For more than three dyes, the data is mapped onto a *(m-1)*-simplex (e.g. a tetrahedron for *m*=4).

1.2 The positive domain

In addition to visualizing the data, we are interested in the range of eligible nonnegative spectra. All eligible spectra have to lie in the same two-dimensional affine subspace and have nonnegative entries in each of their *n* components vector. Hence, we are looking for the subset of the 2-d subspace that lies in the nonnegative sector of the *n*-dimensional signal space. The boundary of this subset can be projected

into the plane just as the data points. To this end, one has to calculate the intersections of the coordinate hyperplanes with the affine 2D subspace. These intersections are lines, from which we have to extract the line segment that is nonnegative. Its image in the plane can be calculated as above using Eqs. A15 and A16. For each coordinate hyperplane, we get one such line segment, and their union is the boundary of the domain of nonnegative spectra.

2. NMF in Practice

2.1 Pre-processing

Background: Before an image can be successfully analyzed with NMF, any constant background should be removed. Preferably, the constant background should be determined by measuring the signal from a region in the field of view, which does not contain cells. Alternatively, one can also estimate the dark signal by taking the minimal pixel value of the intensity in each spectral channel. This dark signal has to be subtracted from the data.

Autofluorescence: In principle, autofluorescence could be included as an additional dye into the algorithm. Estimation of autofluorescence by NMF, however, is particularly difficult as its spectrum is typically broad, its signal is weak, and it is ubiquitously present. In many cases, autofluorescence is weak enough to be neglected entirely. If not, it is advisable to measure the spectrum of autofluoresence in an unlabeled sample and to include such a component during the NMF optimization as a fixed spectrum. Autofluoresence spectra may differ between different excitation wavelengths.

Signal-to-noise ratio: The quality of the decomposition increases with the signal-to-noise ratio of the data. Noisy pixels are detrimental, since the signal in these pixels may be dominated by contributions not captured by the model (e.g. autofluoresence). It is therefore advisable to restrict the analysis to pixels above a threshold. Similarly, saturated pixels have to be removed as their spectral signature is distorted.

Large Data sets: For single images a typical NMF run takes about a minute on a

personal computer. However, if large data sets such as z-stacks or time series are to be analyzed, it is advisable to perform NMF on a representative subset of the data. The obtained estimates for the spectra can either be used to calculate concentrations of the entire set by linear unmixing or as starting conditions for subsequent runs on a larger data set. This accelerates convergence, since existing estimates are refined much more quickly than the initial estimation from scratch. Another possibility is coarse-graining the images. Coarse-graining has the advantage of increasing the signal-to-noise. On the other hand, it averages over spatial structures and thereby decreases signal modulation. Nevertheless, moderate coarse-graining is often advisable. Results from such runs can be used as starting conditions for higher resolution images.

2.2 Initial conditions

The non-negativity constraints often do not suffice to determine the decomposition uniquely, in which case the outcome of an NMF run can depend on the initial conditions. For three dyes, this ambiguity can be illustrated by projecting the density of data points and the domain of positive spectra into a plane, as described above and in the main text. Every set of nonnegative spectra, the convex hull of which encloses the cloud of data points is an eligible solution and might be the outcome of an NMF. To impose initial conditions, we specify a set of spectra and solve for the corresponding concentrations, either by unmixing or by leaving spectra fixed during the initial 10 iterations. Concentrations are initialized with random numbers.

To demonstrate the worst-case scenario, we ran NMF on the sample shown in Figure 1 of the main text with extremely narrow initial spectra. The results are shown in Figure S1 (see legend for a discussion of this run).

According to our experience best results are achieved when starting with spectra that

are broad and slightly red-shifted relative to the peak of the dye. In this case, the algorithm narrows the spectra in such a way that all data points can be described by nonnegative concentrations and avoids the at times substantial ambiguity of too narrow spectra. The red-shift is indicated because most spectra are very asymmetric with long tails at long wavelengths.

A segregation bias as described in the main text forces spectra towards maximal overlap and thereby removes the ambiguity of NMF. A moderate segregation bias applied to the sample of Figure 1 of the main text significantly improves the decomposition. However, since the data points leave some part of the triangle empty, a strong segregation bias result in too broad spectra and too small triangles. Both of these cases are shown in Figure S2.

The two alternative biases discussed in the appendix of the main paper (Eqs. A7 and A9) have qualitatively similar effects, although they differ in their implementation. The segregation bias (A7) operates on the concentrations. This has the advantage that the weight of the bias is easy to tune, since the bias itself scales with the amount of data. A disadvantage is the high degree of non-linearity, which sometimes results in local minima in which the algorithm gets trapped. The bias (A9) targets the overlap of each pair of spectra directly, but is somewhat harder to tune.

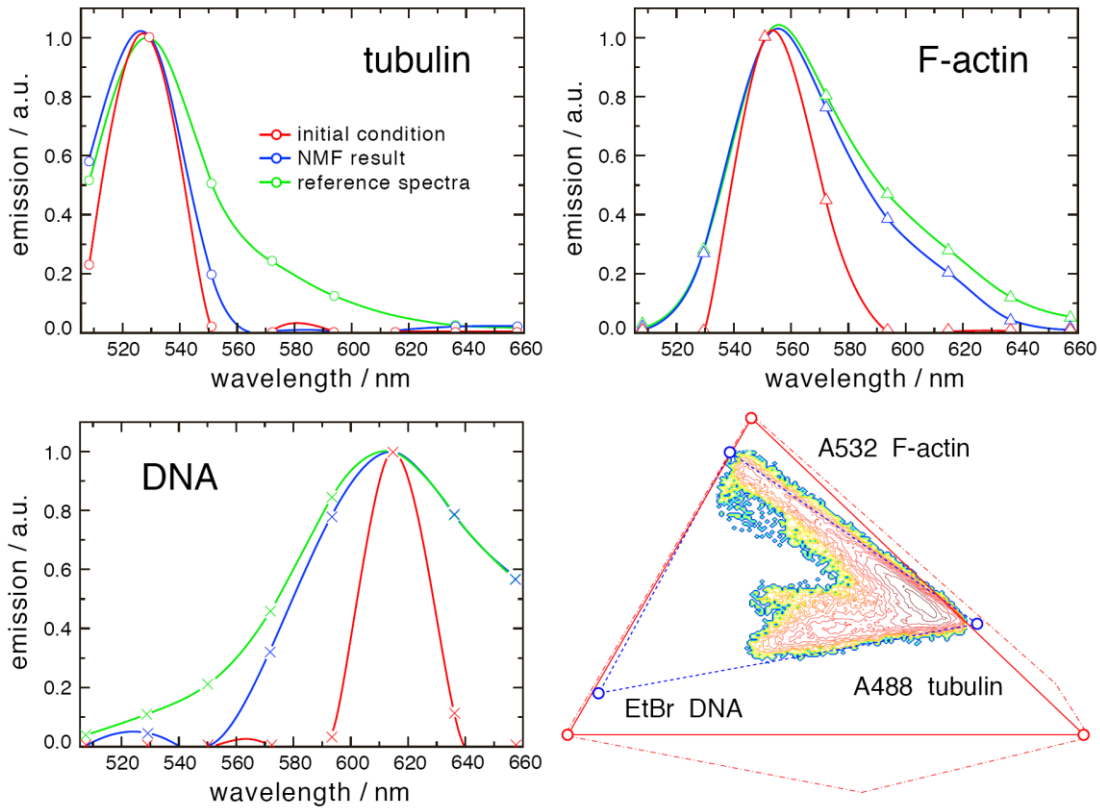

Figure S1: The results of an NMF run with Gaussian initial spectra with a FWHM of 20 nm. The NMF-spectra of A532 **F-actin** and EtBr **DNA** are reasonably close to the true spectra, while the spectrum of A488 **tubulin** is far too narrow. This is a consequence of the vast space between the reference triangle (blue in simplex projection) and the boundary of positive spectra (broken red line). Indeed, the NMF triangle of this run is as large as possible and touches the boundary at each vertex.

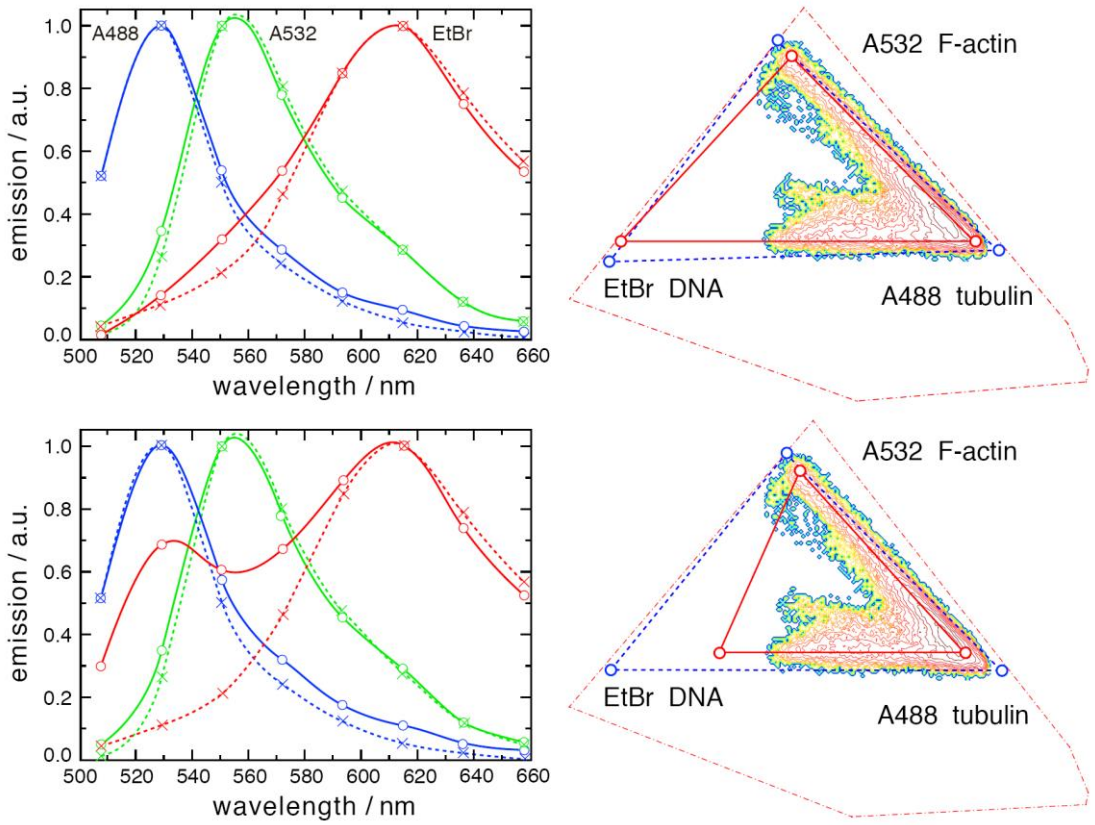

Figure S2: Segregation bias applied to the sample of Figure 1 in the main text. A moderate segregation bias λ=0.005 significantly improves the decomposition and causes the NMF triangle to enclose the data more tightly (top row). A strong segregation bias, however, results a secondary peak in the EtBr spectrum, which now contains admixtures of A488. This is possible, since EtBr - DNA is always colocalized with A488 - tubulin, as apparent in the absence of data points in the lower left of the reference triangle (blue) (bottom row). The segregation bias reduces the triangle to its smallest possible size, which corresponds to mixing A488 into the EtBr spectrum. The resulting secondary peak in the EtBr spectrum is readily removed by the post-processing tool (see below). Initial conditions for these runs are the same as in Figure 1 of the main text, i.e. Gaussian spectra with a FWHM of 75 nm.

2.3 Post-processing of NMF results

Typically, NMF in combination with a segregation bias provides a good initial decomposition. However, secondary peaks as apparent in the EtBr spectrum in Figure S2 can occur when dyes are strongly colocalized. These errors are often obvious and can be straightforwardly corrected using a software tool, which we developed. The use of the tool is illustrated in Figure S3.

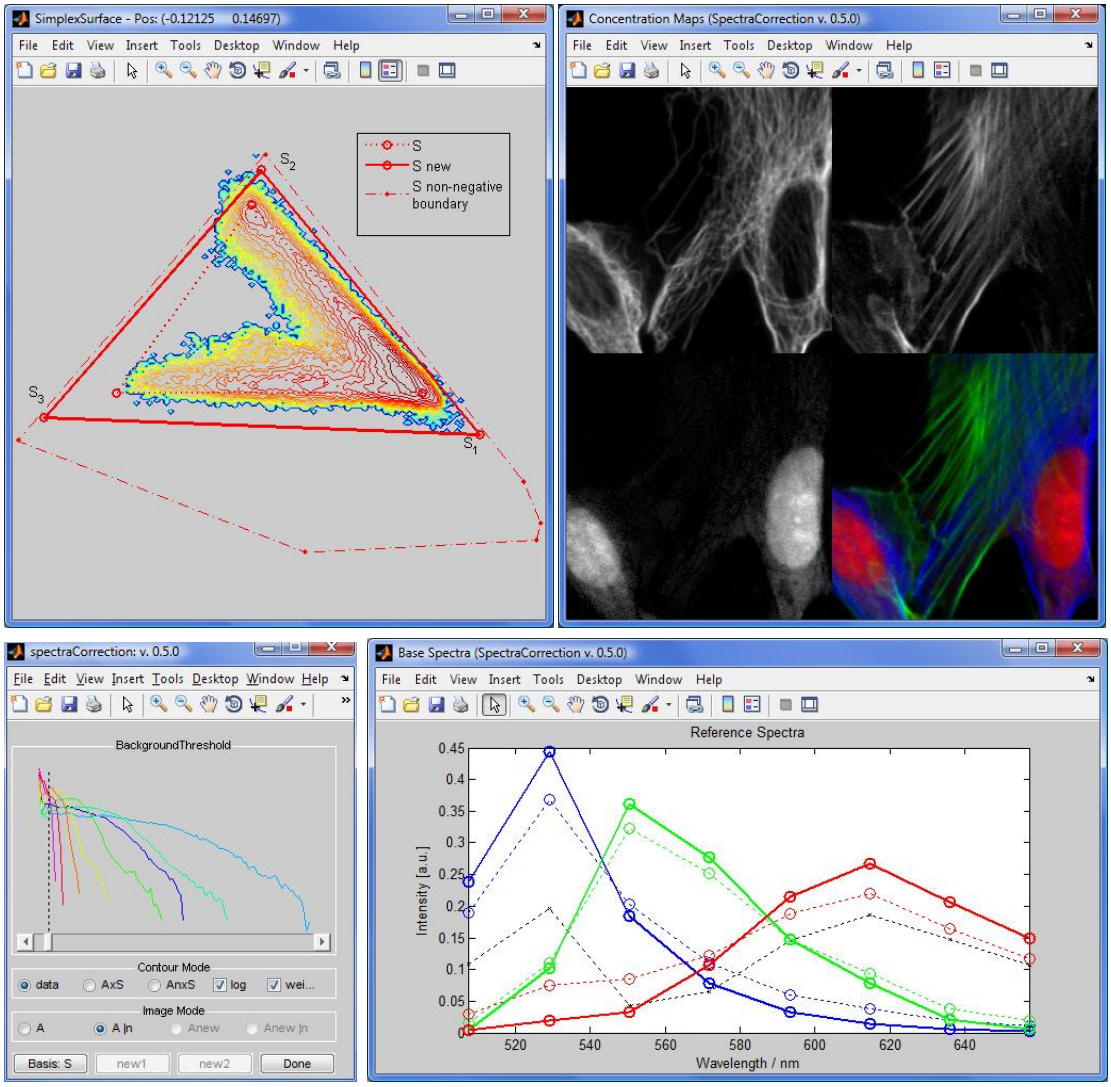

Figure S3: The software tool uses the set of spectra returned by NMF and displays the concentration maps (upper, left), the 2D representation of the spectra and the data (upper right), the intensity distributions in the user interface (lower left) and the spectra simultaneously (lower right). When the mouse cursor is moved within the domain of nonnegative spectra of the 2D simplex projection, the corresponding spectrum is displayed in real time in the spectra plot (dotted black line in spectra plot). When a better spectrum is found, the corresponding vertex of the NMF triangle (solid

red triangle in simplex projection) can be dragged to that new location. The software then recalculates the concentration maps.