Annals of Combinatorics 12 (2008) 279-290
0218-0006/08/030279-12
DOI 10.1007/s00026-008-0351-4

Annals of Combinatorics

Two Theorems about Similarity Maps

Andreas Dress[1,2], Tatjana Lokot[3], Walter Schubert[4], and Peter Serocka[1,2]

[1]CAS-MPG Partner Institute for Computational Biology, 320 Yue Yang Road, Shanghai 200031, P.R. China

[2]Max-Planck-Institut für Mathematik in den Naturwissenschaften, Inselstrasse 2-26 D-04103 Leipzig, Germany
{andreas, pserocka}@picb.ac.cn

[3]Fakulät für Mathematik, Universität Bielefeld, D-33615 Bielefeld, Germany
tlokot@mathematik.uni-bielefeld.de

[4]Institut of Medical Neurobiology, University of Magdeburg, D-39120 Magdeburg, Germany
Walter.Schubert@med.ovgu.de

Received April 01, 2005

AMS Subject Classification: 54Exx, 54E40, 62P10, 65D18, 68U05, 68U10, 92C50, 92C55

Abstract. One of the problems arising when exploring *toponome* or other *multivariate-image* data is the following: Given a family of n gray-value images of, e.g., a given sample of cell tissue, indexed by a collection of n proteins under investigation (so-called MELK data) — each single image representing the varying local concentration of one of those n proteins at the various sites (pixels) of the given sample, how should one quantify, for any two pixels (or clusters of pixels), the (dis)similarity between the corresponding "vectors" of local protein concentrations in question. Some *(dis)similarity mappings* defined on $\mathbb{R}^n$ allowing for fast *OpenGL* texture mapping turned out to be useful in visual inspection of toponome data. Here, we derive two rather general results regarding similarity and dissimilarity mappings and, as a corollary, the fact that the functions that were used for visual inspection of MELK data are, indeed, metrics. We believe that our results are, however, also of more general interest within the ongoing program of elucidating the structure of metrics from a more abstract point of view.

Keywords: metrics, similarity maps, dissimilarities, MELK, protein localization, protein co-localization, toponome, multivariate images, SGI-type texture mapping, scientific visualization, visual interactive analysis of multivariate images, *Lasagne*

1. Introduction

One of the problems arising when exploring *toponome* data (cf. [7–9]) is the following: Given a family of n gray-value images of a cell-tissue sample, indexed by a collection of n proteins under investigation (so-called MELK data) — each representing the varying

local concentration of one of the n individual proteins at the various sites of the given sample (as represented by the images' pixels), how should one quantify — for any two pixels (or aggregation of pixels) — the (dis)similarity between the corresponding "vectors" of local protein concentrations in question. Some (*dis*)*similarity mappings* defined on $\mathbb{R}^n$ and allowing for fast *OpenGL* texture mapping, turned out to be useful in this context. For example, a particular metric F defined on $[0, 1]^n$ according to Corollary 4.1was used to obtain the eight images shown and discussed below. The program used to obtain these pictures is called *Lasagne*. It was created by PS in collaboration with Sebastian Funke for visual interactive analysis of toponome data. It provides numerous modes for visual inspection, and one of them (the so-called dynamical mode) highlights, almost instantaneously upon mouse click, all pixels whose protein distribution is, as measured by F, sufficiently similar to the distribution at that pixel that was chosen by mouse click (and marked with the cross in pictures below; in these figures, one can also see, in the left lower part of the image, the actual protein distribution at the chosen pixel — the concentration levels of the proteins in question represented by corresponding vertical bars).

Here, we give a rather general definition of *similarity mappings* closely adapted to the definition of metrics, and we show that

- the product of any two non-negative similarity mappings s_1, s_2 defined on the same set X is also a similarity mapping, and
- the composition $f \circ s$ of a similarity map s with a monotonously increasing and convex map f defined on an interval $I \subseteq \mathbb{R}$ that contains all values of s, is also a similarity map.

As a corollary, we show that the functions that were used for visual inspection of MELK data are, indeed, metrics. Moreover, we believe that these two rather general observations may also be of interest within the context of the ongoing program of elucidating the structure of metrics from a more abstract point of view that has been pursued in quite a number of papers in recent years (cf. [1–6]).

2. Basic Properties of Similarity Mappings

Definition 2.1. *A similarity mapping — or, for short, a similarity — defined on a set X is a map*

$$s\colon X \times X \to \mathbb{R}\colon (x, y) \mapsto s(x, y) =: xy^* \tag{2.1}$$

such that

$$xy \leq xx \quad \textit{and} \quad yx + zx \leq zy + xx \tag{2.2}$$

holds for all $x, y, z \in X$.

We collect some simple facts and observations regarding similarities:

Remark 2.2. Note that (2.2) implies that $xy = yx$ must hold for all $x, y \in X$ for any similarity $s\colon X \times X \to \mathbb{R}$ as above (just put $z := x$ in (2.2), then exchange x and y in (2.2) and put $z := y$).

* To improve readability, we write xy rather than $s(x, y)$ for the value of s at a pair x, y of elements from X.

Remark 2.3. A map $d\colon X\times X\to\mathbb{R}$ is a metric defined on X if and only if the map

$$s=-d\colon X\times X\to\mathbb{R}\colon (x,y)\mapsto -d(x,y)$$

is a similarity defined on X for which $xx=0$ holds for all $x\in X$.

Remark 2.4. A map $s\colon X\times X\to\mathbb{R}$ is a similarity defined on X if and only if for one (or as well for all) univariate map(s) $h\colon X\to\mathbb{R}$, the *Farris transform* (cf. [3])

$$s_h\colon X\times X\to\mathbb{R}\colon (x,y)\mapsto xy-h(x)-h(y)$$

is a similarity defined on X. In particular, every similarity s defined on X is of the form

$$s=-d_h\colon X\times X\to\mathbb{R}\colon (x,y)\mapsto h(x)+h(y)-d(x,y), \tag{2.3}$$

for some univariate map $h=h_s\colon X\to\mathbb{R}$ and some metric $d=d_s$ defined on X, uniquely determined by s $\big($as (2.3) implies that $h(x)=\frac{1}{2}xx$ and $d(x,y)=-xy+h(x)+h(y)=-xy+\frac{1}{2}\big(xx+yy\big)$ must hold for all $x,y\in X\big)$.

In other words, there exists a canonical one-to-one correspondence between the set $\mathrm{Met}(X)$ of all metrics defined on X and the set $\mathrm{Sim}(X)/_{\underset{F}{\sim}}$ of equivalence classes contained in the set $\mathrm{Sim}(X)$ consisting of all similarities defined on X relative to the equivalence relation $\underset{F}{\sim}$ (well-)defined on $\mathrm{Sim}(X)$ by "$s\underset{F}{\sim}s'\Leftrightarrow s'=s_h$ holds for some univariate map $h=h_s\colon X\to\mathbb{R}$".

Remark 2.5. A map $s\colon X\times X\to\mathbb{R}\colon (x,y)\mapsto xy$ is a simalarity if and only if the map

$$\rho\cdot s\colon X\times X\to\mathbb{R}\colon (x,y)\mapsto \rho\, xy$$

is a similarity for one (or as well for all) positive real number(s) ρ.

Remark 2.6. Given a similarity s defined on X and a map $p\colon Y\to X$, the map

$$s\circ(p\times p)\colon Y\times Y\to\mathbb{R}\colon (u,v)\mapsto p(u)p(v)$$

is a similarity defined on Y.

3. Basic Results

Theorem 3.1. *If $s_1,s_2\colon X\times X\to\mathbb{R}$ are two similarities defined on X with $s_1(x,y)$, $s_2(x,y)\geq 0$ for all $x,y\in X$, then their product*

$$s:=s_1\cdot s_2\colon X\times X\to\mathbb{R}\colon (x,y)\mapsto xy:=s_1(x,y)s_2(x,y)$$

is also a similarity.

Proof. With $x,y,z,\ldots$, we will always denote elements of X.

It is obvious that, in view of $0\leq s_1(x,y)\leq s_1(x,x)$ and $0\leq s_2(x,y)\leq s_2(x,x)$, the inequality

$$xy=s_1(x,y)s_2(x,y)\leq s_1(x,x)s_2(x,x)=xx$$

holds for all $x, y \in X$. Furthermore, we have

$$s_1(y, z) \geq s_1(x, y) + s_1(x, z) - s_1(x, x)$$

and

$$s_2(y, z) \geq s_2(x, y) + s_2(x, z) - s_2(x, x),$$

as well as

$$0 \leq s_1(x, y), s_1(x, z) \leq s_1(x, x) \quad \text{and} \quad 0 \leq s_2(x, y), s_2(x, z) \leq s_2(x, x),$$

for all $x, y, z \in X$. Hence, if

$$s_i(x, y) + s_i(x, z) - s_i(x, x) \leq 0$$

holds, for given elements $x, y, z \in X$, for some $i \in \{1, 2\}$, and if $j \in \{1, 2\}$ is chosen so that $\{1, 2\} = \{i, j\}$ holds, we get

$$\begin{aligned}
&s_1(x, y)s_2(x, y) + s_1(x, z)s_2(x, z) - s_1(x, x)s_2(x, x) \\
&= s_i(x, y)s_j(x, y) + s_i(x, z)s_j(x, z) - s_i(x, x)s_j(x, x) \\
&\leq s_i(x, y)s_j(x, x) + s_i(x, z)s_j(x, x) - s_i(x, x)s_j(x, x) \\
&= (s_i(x, y) + s_i(x, z) - s_i(x, x))s_j(x, x) \\
&\leq 0.
\end{aligned}$$

Hence,

$$s_1(x, y)s_2(x, y) + s_1(x, z)s_2(x, z) - s_1(x, x)s_2(x, x) \leq s_1(y, z)s_2(y, z)$$

and, therefore, $xy + xz \leq yz + xx$, as claimed. Otherwise, we have

$$0 \leq s_1(x, y) + s_1(x, z) - s_1(x, x) \leq s_1(y, z)$$

and

$$0 \leq s_2(x, y) + s_2(x, z) - s_2(x, x) \leq s_2(y, z),$$

which implies

$$\begin{aligned}
yz &= s_1(y, z)s_2(y, z) \\
&\geq (s_1(x, y) + s_1(x, z) - s_1(x, x))(s_2(x, y) + s_2(x, z) - s_2(x, x)) \\
&= s_1(x, y)s_2(x, y) + s_1(x, z)s_2(x, z) - s_1(x, x)s_2(x, x) \\
&\quad + (s_1(x, y) - s_1(x, x))(s_2(x, z) - s_2(x, x)) \\
&\quad + (s_1(x, z) - s_1(x, x))(s_2(x, y) - s_2(x, x)) \\
&= s_1(x, y)s_2(x, y) + s_1(x, z)s_2(x, z) - s_1(x, x)s_2(x, x)
\end{aligned}$$

$$
\begin{aligned}
&+ (s_1(x,x) - s_1(x,y))(s_2(x,x) - s_2(x,z)) \\
&+ (s_1(x,x) - s_1(x,z))(s_2(x,x) - s_2(x,y)) \\
&\geq s_1(x,y)s_2(x,y) + s_1(x,z)s_2(x,z) - s_1(x,x)s_2(x,x) \\
&= xy + xz - xx,
\end{aligned}
$$

as required in view of the fact that $s_1(x,x) - s_1(x,u), s_2(x,x) - s_2(x,u) \geq 0$ holds for all $u \in X$. ∎

Corollary 3.2. *Given two sets X and Y together with two non-negative similarities*

$$s_X : X \times X \to \mathbb{R} \quad \text{and} \quad s_Y : Y \times Y \to \mathbb{R},$$

the map

$$s := s_X \times s_Y : (X \times Y) \times (X \times Y) \to \mathbb{R} : \big((x,y),(x',y')\big) \mapsto s_X(x,x')s_Y(y,y')$$

is a similarity, too, defined on $X \times Y$.

To continue, recall that a map $f : I \to \mathbb{R}$ defined on an interval $I \subseteq \mathbb{R}$ is called *convex* if

$$f(\alpha x + \beta y) \leq \alpha f(x) + \beta f(y)$$

holds for all $\alpha, \beta \in [0,1]$ with $\alpha + \beta = 1$, and all $x, y \in I$. Our next result states that similarities are transformed into similarities by composing them with a monotonously increasing and convex map.

Theorem 3.3. *Assume that $s : X \times X \to \mathbb{R} : (x,y) \mapsto xy$ is a similarity and that $f : I \to \mathbb{R}$ is a monotonously increasing and convex map defined on an interval $I \subseteq \mathbb{R}$ with $xy \in I$ for all $x, y \in X$. Then, the map*

$$f \circ s : X \times X \to \mathbb{R} : (x,y) \mapsto f(xy)$$

is also a similarity.

Proof. Clearly, our assumptions imply that

$$(f \circ s)(x,y) = f(xy) \leq f(xx) = (f \circ s)(x,x)$$

holds for all $x, y \in X$. Thus, it remains to prove that also

$$
\begin{aligned}
(f \circ s)(x,y) + (f \circ s)(x,z) &= f(xy) + f(xz) \leq f(yz) + f(xx) \\
&= (f \circ s)(y,z) + (f \circ s)(x,x) \qquad (3.1)
\end{aligned}
$$

holds for all $x, y, z \in X$. One easily sees that the inequality (3.1) holds indeed for the case $xy \leq yz$ or $xz \leq yz$ in view of the monotonicity of f and the fact that $xy, xz \leq xx$ holds by assumption.

So, we need to prove (3.1) only in case $yz < xy$ and $yz < xz$. To this end, we put

$$\overline{xy} := xy + \frac{yz + xx - xy - xz}{2}$$

and

$$\overline{xz} := xz + \frac{yz + xx - xy - xz}{2},$$

and note that, in view of (2.2) and our assumptions, one has

$$yz < xy \leq \overline{xy} = \frac{(yz - xz) + (xx + xy)}{2} < \frac{xx + xy}{2} \leq xx,$$

as well as

$$yz < xz \leq \overline{xz} = \frac{(yz - xy) + (xx + xz)}{2} < \frac{xx + xz}{2} \leq xx,$$

while

$$\overline{xy} + \overline{xz} = yz + xx$$

holds by definition of $\overline{xy}$ and $\overline{xz}$. Hence, putting

$$\alpha := \frac{\overline{xy} - yz}{xx - yz} = \frac{xx - \overline{xz}}{xx - yz}$$

and

$$\beta := \frac{\overline{xz} - yz}{xx - yz} = \frac{xx - \overline{xy}}{xx - yz},$$

we have $\alpha, \beta > 0$, $\alpha + \beta = 1$,

$$\alpha xx + \beta yz = \frac{\overline{xy}\, xx - \overline{xy}\, yz}{xx - yz} = \overline{xy},$$

$$\beta xx + \alpha yz = \frac{\overline{xz}\, xx - \overline{xz}\, yz}{xx - yz} = \overline{xz},$$

and, therefore,

$$\begin{aligned}(f \circ s)(x, y) + (f \circ s)(x, z) &= f(xy) + f(xz) \leq f(\overline{xy}) + f(\overline{xz}) \\ &= f(\alpha\, xx + \beta\, yz) + f(\beta\, xx + \alpha\, yz) \\ &\leq \alpha\, f(xx) + \beta\, f(yz) + \beta\, f(xx) + \alpha f(yz) \\ &= (f \circ s)(x, x) + (f \circ s)(y, z),\end{aligned}$$

as claimed. ∎

4. Applications: Metrics from Metrics

The above results can be used to define a large variety of operators that construct metrics from metrics: Starting, e.g., with just any metric d defined on a set X, one can begin by considering the similarity map $s := -d$, then choose some positive real number ρ and some univariate map $h\colon X \to \mathbb{R}$ and form the Farris transform $s' = s'(d, \rho, h) := (-\rho\, d)_h$

of $-\rho\, d$ relative to h, then choose some monotonously increasing and convex map f defined on an interval $I \subseteq \mathbb{R}$ with $s'(x, y) \in I$ for all $x, y \in X$ (or, if I and f are given, one may try to choose ρ and h so that $s'(x, y) \in I$ holds for all $x, y \in X$) to form the similarity map $s'' = s''(d, \rho, h, f) := f \circ s'(d, \rho, h) = f \circ (-\rho\, d)_h$ for which one may then form the associated metric $d' = d'(d, \rho, h, f) := d_{s''}$. Note that this works, in particular, for all d, ρ, and h as above in case f is the exponential function (because this function is monotonously increasing and convex) and yields a similarity $s''(d, \rho, h, f)$ all of whose values are positive (because $\exp(\rho) > 0$ holds for all $\rho \in \mathbb{R}$).

Similarly, if several metrics $d_1, d_2, \ldots, d_N$ defined on X are given, one may choose the parameters $\rho_1, \rho_2, \ldots, \rho_N, h_1, h_2, \ldots, h_N$ so that all the resulting similarity maps

$$s_1 := s'(d_1, \rho_1, h_1), s_2 := s'(d_2, \rho_2, h_2), \ldots, s_N := s'(d_N, \rho_N, h_N)$$

have non-negative values in which case one may also form their product $\prod_{\nu=1}^{N} s_\nu$ which yields yet another similarity s for which the associated metric $d := d_s$ may be formed.

It could be interesting to check how geodesics of such metrics are related to geodesics of the component metrics.

Finally, we mention the special case that has been used in the software developed for interactive exploration of MELK data.

Corollary 4.1. *Given any family $f_1, \ldots, f_n\colon [0, 1] \to [0, 1]$ of convex, monotonously decreasing functions with $f_i(0) = 1$ and $f_i(1) = 0$ for all $i = 1, \ldots, n$ $\big($e.g., $f_1(x) = \cdots = f_n(x) := (1-x)^\rho$ for some $\rho > 1\big)$, the function*

$$F\colon [0, 1]^n \times [0, 1]^n \to \mathbb{R}\colon ((x_1, \ldots, x_n), (y_1, \ldots, y_n)) \mapsto 1 - \prod_{i=1}^{n} f_i(|x_i - y_i|)$$

is a metric on $[0, 1]^n$.

Proof. Applying our observations and results collected above, we see that

- the maps

  $$s_i\colon [0, 1] \times [0, 1] \to [0, 1] \subseteq \mathbb{R}\colon (x, y) \mapsto f_i(|x-y|) = f_i(-(-|x-y|)) \ (i = 1, \ldots, n)$$

  are similarities because

  - the map $[0, 1] \times [0, 1] \to [-1, 0]\colon (x, y) \mapsto -|x-y|$ is a similarity, and
  - the maps $[-1, 0] \to [0, 1]\colon \rho \mapsto f(-\rho)$ are convex and monotonously increasing for all $i = 1, \ldots, n$,

  which in turn implies that (cf. Corollary 3.2),

- the product

  $$s := s_1 \times \cdots \times s_n\colon [0, 1]^n \times [0, 1]^n \to [0, 1]\colon$$
  $$\big((x_1, \ldots, x_n), (y_1, \ldots, y_n)\big) \mapsto \prod_{i=1,\ldots,n} f_i(|x_i - y_i|)$$

  is a similarity, too, defined on the product space $[0, 1]^n$ for which

- $s\big((x_1,\ldots,x_n),(x_1,\ldots,x_n)\big) = 1$ holds for all $(x_1,\ldots,x_n)$ in $[0,1]^n$ implying that
- the corresponding metric d_s is given by

$$d_s\colon [0,1]^n \times [0,1]^n \to [0,1]\colon$$
$$\big((x_1,\ldots,x_n),(y_1,\ldots,y_n)\big) \mapsto 1 - s\big((x_1,\ldots,x_n),(y_1,\ldots,y_n)\big)$$

and, thus, coincides with F. ∎

5. A Paradigmatic Application in Biology: Synapses in the Rat Perietal Brain Isocortex

Synapses are specialized contacts between nerve cells allowing electrical impulses on the nerve-cell surface to be transduced from cell to cell. On the synaptic level, this is brought about by combinatorial arrangements of proteins (relative abundancies) in the postsynaptic membrane and by specialized molecules (transmitters) that are released on the presynaptic site upon an electrical impulse. Presently 100 different proteins have been identified to be associated with the postsynaptic membrane, however it is completely unknown, according to which local rules these proteins are combined in individual synapses to perform their function(s). To address this problem for the first time, MELK toponome technology (cf. [7–9]) was applied to localize, as a first step, 7 selected synaptic proteins in the rat perietal brain isocortex by analyzing a tissue section of this area. The section was taken tangentially to the isocortex gray matter, approximately crossing the inner pyramidal cell layer.

The MELK toponome technology produced 7 single images, each showing the distribution pattern of each particular protein (data not shown). These images were aligned and then superimposed to be analyzed by *Lasagne*. Figure 3 gives eight examples indicating the relative abundancies of these proteins. One can readily recognize that there are subregions of this cortex area, which are uniquely characterized by similar relative abundancies of the synaptic proteins under investigation, thereby indicating directly different functional states of synapses due to topological restriction defined by the column architecture of the isocortex. Detailed biological conclusions will be discussed elsewhere.

One can employ *Lasagne* using other metrics, too, to obtain alternative representations. However, for the inspection of MELK data, the metrics introduced above appear to produce meaningful data representations more readily than most of the other metrics that we have tested.

Acknowledgments. These studies were supported by grants from the German BMBF through CELLECT, NGFN-2, NGFN plus, and NBL3, through DFG grants 627/1-8, DFG Schu 627/10-1, the joint support of CAS (China) and MPG/BMBF (Germany) for the CAS-MPG Partner Institute for Computational Biology (PICB/SIBS) in Shanghai, the Science Technology Commission of Shanghai Municipality (Grant 06ZR14048), and the Tschira foundation.

References

1. H.-J. Bandelt, V. Chepoi, and A. Karzanov, A characterization of minimizable metrics in the multifacility location problem, European J. Combin. **21** (6) (2000) 715–725.
2. V. Chepoi and B. Fichet, A note on circular decomposable metrics, Geom. Dedicata **69** (3) (1998) 237–240.
3. A. Dress, K.T. Huber, and V. Moulton, Some uses of the Farris transform in mathematics and phylogenetics — a review, Ann. Combin. **11** (1) (2007) 1–37.
4. A. Dress and T. Lokot, A simple proof of the triangle inequality for the NTV metric, Appl. Math. Lett. **16** (6) (2003) 803–813.
5. A. Dress, T. Lokot, and L.D. Pustyl'nikov, A new scale-invariant geometry on L_1 spaces, Appl. Math. Lett. **17** (7) (2004) 815–820.
6. J.J. Nieto, A. Torres, and M.M. Vázquez-Trasande, A metric space to study differences between polynucleotides, Appl. Math. Lett. **16** (8) (2003) 1289–1294.
7. W. Schubert, Polymyositis, topological proteomics technology and paradigm for cell invasion dynamics, J. Theor. Med. **4** (1) (2002) 75–83.
8. W. Schubert, Topological proteomics, toponomics, MELK-technology, In: Proteomics of Microorganismus: Fundamental Aspects and Application, M. Hecker and S. Müllner, Eds., Adv. Biochem. Engeneer. Biotechnol., Vol. 83, (2003) pp. 189–209.
9. W. Schubert et al., Analyzing proteome topology and function by automated multidimensional fluorescence microscopy, Nat. Biotechnol. **24** (10) (2006) 1270–1278.

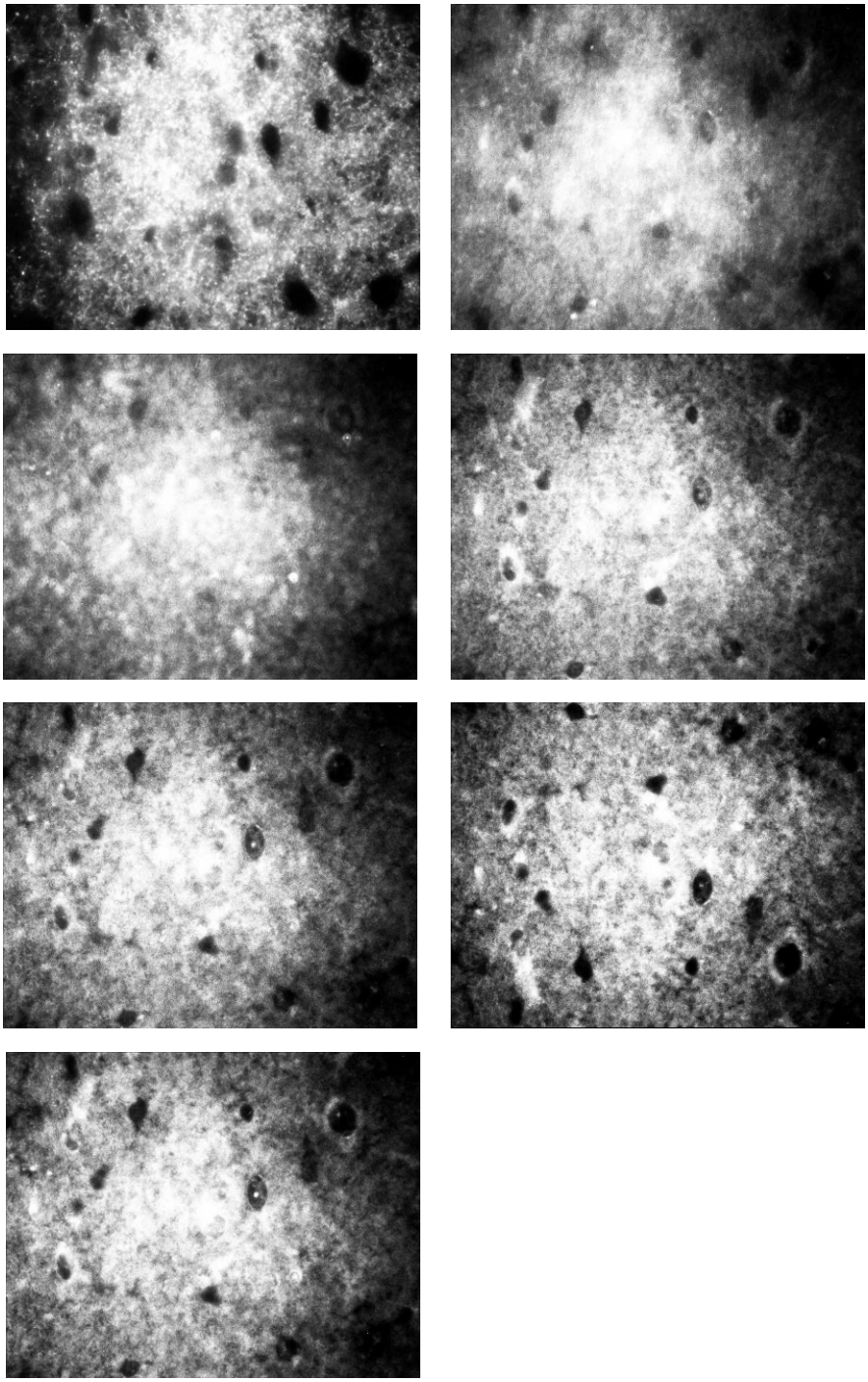

Figure 1: 7 non-processed primary grey-value image data produced by MELK Toponome Technology from one tissue section in one experiment (co-)localizing 7 synaptic proteins.

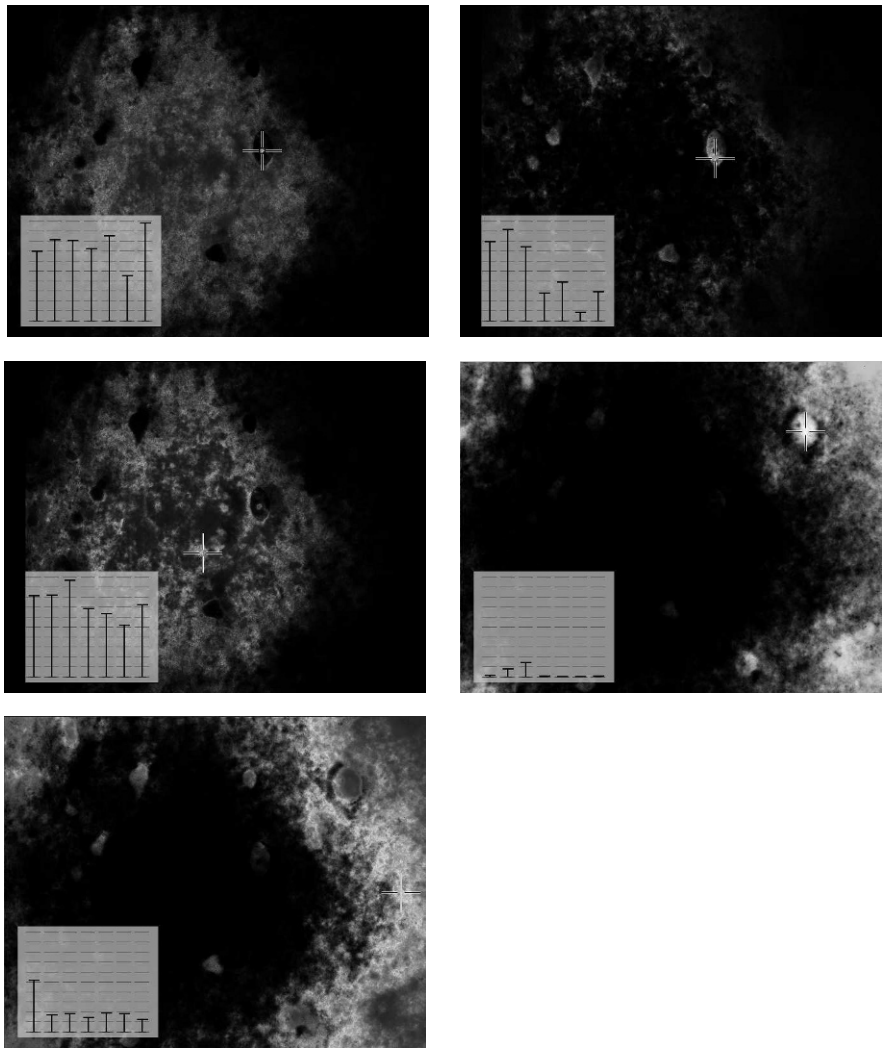

Figure 2: Nerve-cell tissue: 5 images obtained with *Lasagne* from MELK toponome data as shown in Figure 1.

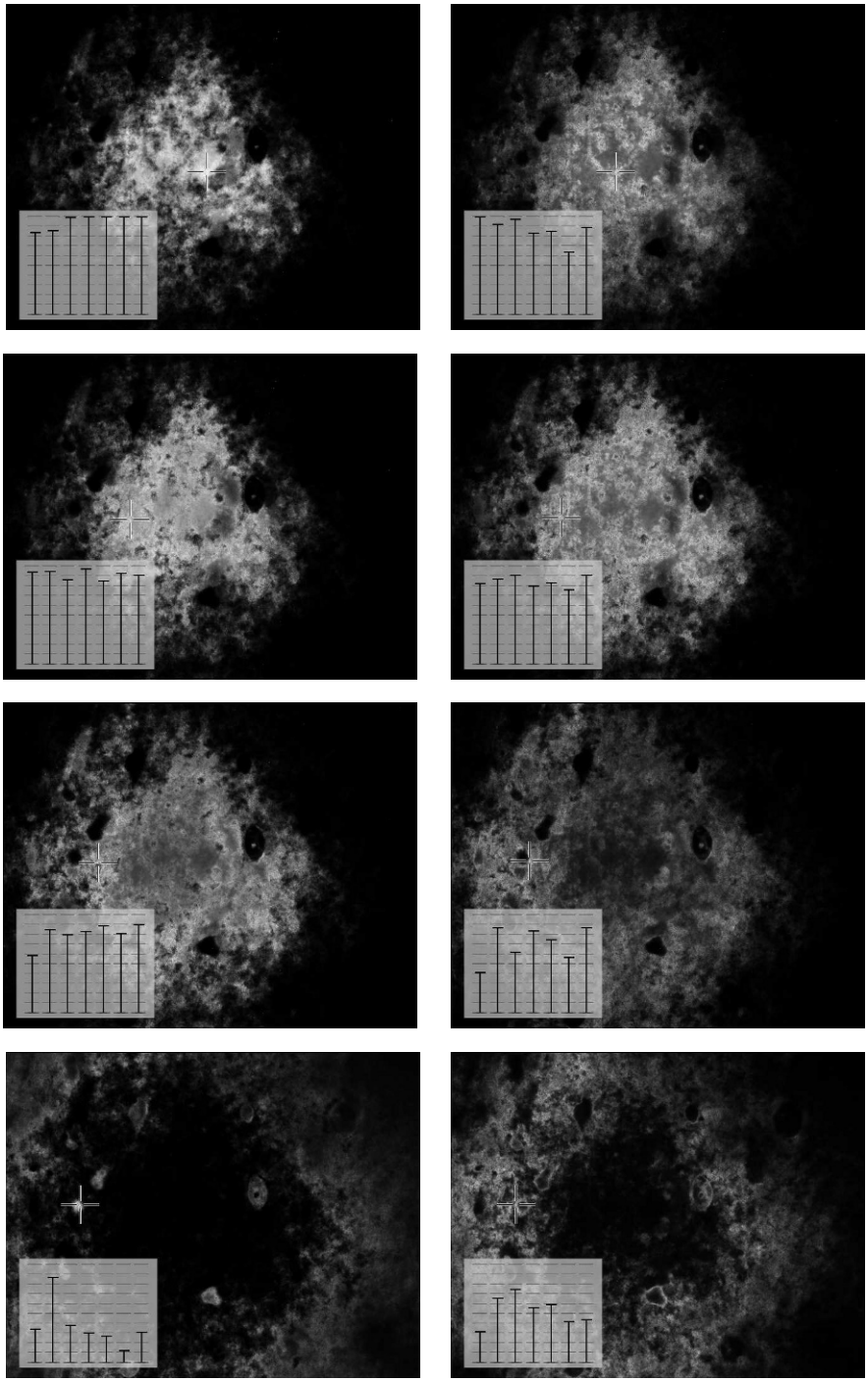

Figure 3: Nerve-cell tissue: 8 further images obtained with *Lasagne* from the same data set.