SOFTWARE TOOL ARTICLE

Generalized EmbedSOM on quadtree-structured self-organizing maps [version 1; peer review: 1 approved, 1 approved with reservations]

Miroslav Kratochvíl1,2, Abhishek Koladiya3, Jiří Vondrášek1

1Institute of Organic Chemistry and Biochemistry of the CAS, Prague, Czech Republic
2Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
3Institute of Hematology and Blood Transfusion, Prague, Czech Republic

Abstract
EmbedSOM is a simple and fast dimensionality reduction algorithm, originally developed for its applications in single-cell cytometry data analysis. We present an updated version of EmbedSOM, viewed as an algorithm for landmark-based embedding enrichment, and demonstrate that it works well even with manifold-learning techniques other than the self-organizing maps. Using this generalization, we introduce an inwards-growing variant of self-organizing maps that is designed to mitigate some earlier identified deficiencies of EmbedSOM output. Finally, we measure the performance of the generalized EmbedSOM, compare several variants of the algorithm that utilize different landmark-generating functions, and showcase the functionality on single-cell cytometry datasets from recent studies.

Keywords
dimensionality reduction, self-organizing maps, single-cell cytometry

This article is included in the RPackage gateway.
Corresponding author: Miroslav Kratochvíl (miroslav.kratochvil@uochb.cas.cz)

Author roles: Kratochvíl M: Conceptualization, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Koladiya A: Investigation, Methodology, Resources, Validation, Visualization, Writing – Review & Editing; Vondrášek J: Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: M.K. and J.V. were supported by ELIXIR CZ LM2015047 (MEYS). A.K. was supported by European Regional Development Fund and the state budget of the Czech Republic (project AllHHP: CZ.02.1.01/0.0/0.0/16_025/0007428, OP RDE, MEYS).

Copyright: © 2019 Kratochvíl M et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Kratochvíl M, Koladiya A and Vondrášek J. Generalized EmbedSOM on quadtree-structured self-organizing maps [version 1; peer review: 1 approved, 1 approved with reservations] F1000Research 2019, 8:2120 (https://doi.org/10.12688/f1000research.21642.1)

First published: 18 Dec 2019, 8:2120 (https://doi.org/10.12688/f1000research.21642.1)
Introduction

EmbedSOM is a dimensionality reduction (DR) algorithm for single-cell cytometry data, designed for high scalability, computational efficiency and performance\(^1\). The design is based off FlowSOM\(^2\), which utilizes unsupervised manifold learning by self-organizing maps (SOMs) to find structure in the high-dimensional data, and process the result into a meaningful and easily interpretable clustering of the dataset. So far, FlowSOM and SOMs in general seem to be the manifold learning and clustering method of choice for all kinds of cytometry based on protein-targeting antibodies, surpassing other clustering methods in precision, speed and scalability\(^3\). EmbedSOM utilizes the same manifold learning method to extract information about the topology of an approximate manifold that describes the high-dimensional cell expression space, and uses it to quickly compute low-dimensional image of the cells that is suitable for visualization.

In this work, we focus on fixing inconsistencies and problems of the first version of EmbedSOM: First, we describe an updated version of EmbedSOM that improves the approximation to achieve perfect smoothness of the projection. The brief description of EmbedSOM provided in the original paper is supplemented here by fully commented pseudocode, in order to aid scrutinization and interpretation of the method. Second, we review EmbedSOM as a generalized function for enriching a projection of selected landmarks to a projection of entire spaces. We demonstrate this by replacing the original SOMs with less-demanding t-SNE on random landmarks. Additionally, we describe GQTSOM, a novel variant of growing self-organizing maps (GSOMs, described e.g. by Rauber et al.\(^4\)) that was designed to alleviate precision and overcrowding problems of the original EmbedSOM. GQTSOMs utilize quad-tree space-partitioning structure to grow inwards, thus allowing the training algorithm to increase the resolution of manifold approximation on demand, and to benefit from the performance gain in early stages of training that is common to all GSOMs.

The functionality of the new algorithm is showcased on datasets that were recently used for studying other DR techniques. We show the differences between individual variants of landmark-generating functions, and provide visualizations comparable to those produced by current state-of-art algorithms. Finally, we demonstrate how the dynamic resolution of GQTSOMs aids detection of various small cell populations and rare cell types.

Methods

Landmark-directed embedding

EmbedSOM projection can be viewed as an embedding enrichment method: From a set of landmarks in the high-dimensional space and a set of corresponding landmarks in the low-dimensional space, it produces a smooth function that maps all points from the higher-dimensional space to the low-dimensional space and preserves the relative neighborhoods of the landmarks. EmbedSOM was originally designed to work with simple SOM-originating landmarks, as shown in Figure 1.

We will refer to the high- and low-dimensional landmarks as \(L \in \mathbb{R}^{n \times D}\) and \(l \in \mathbb{R}^{n \times 2}\). EmbedSOM embedding of a single high-dimensional point is achieved by reducing it to a collection of coordinates of its projections into subspaces that are generated by affine combinations of landmark pairs from \(L\), and reconstructing it in low-dimensional space by reversing the process with corresponding landmark pairs from \(l\).

The procedure is detailed as Algorithm 1. First, the algorithm chooses \(k\) landmarks closest to \(X\), which are expected to give sufficient approximation. In lines 2–6 it computes scores for the \(k\) landmarks. The affine projection of \(X\) to a space defined by a pair of landmarks from \(L\) is computed at line 12 as \(d\), its value is used to create a linear equation which has solutions at positions that would project to the same position \(d\) in the affine space generated from corresponding landmarks in \(l\). After adding all parts of the approximation together, the linear system stored in \(M\) is very unlikely to remain singular. The position of embedded point is then obtained by simply solving the linear equation of 2 variables. Alternatively, one can view the algorithm as a minimization of the total squared error in all projected \(d\):

\[
\arg\min_{x \in \mathbb{R}^D} \sum_{i,j} s_{i,j} \left( d_{i,j} - \frac{\langle x - l_i, l_j - l_i \rangle}{\langle l_j - l_i \rangle} \right)^2
\]
Figure 1. Overview of EmbedSOM interaction with landmarks on a toy dataset. Embedding process starts by reducing the input dataset (data flow is visualized as orange arrows) to landmarks (black arrows and dots) in high-dimensional (top) and low-dimensional space (bottom). EmbedSOM quickly places the relatively large amount of individual input points into matching neighborhoods of the low-dimensional landmarks. The landmark-generating methods from left: A simple SOM-based grid, a random selection of input points with 2-D topology reconstructed by t-SNE, and a GQTSOM-based grid. GQTSOM-based landmarks are labeled by their level in the quadtree.

**Algorithm 1:** EmbedSOM projection from $D$-dimensional Euclidean space to 2-D using $n$ landmarks.

1: procedure EmbedSOM($X \in \mathbb{R}^{m \times n}$, $L \in \mathbb{R}^{m \times k}$, $k \in \{4 \ldots n\}$, $m > 0$, $a > 0$)
2: $c \leftarrow$ a sequence of $c_i = ((X - L))$ for $i \in \{1 \ldots n\}$
3: $\alpha \leftarrow$ indexes of $k$ smallest elements of $c$ in order
4: $\mu \leftarrow \sum_{i=1}^{n} \frac{c_{\alpha(i)}}{k}$  \(\triangleright\) estimate the distribution of landmark distances
5: $\sigma \leftarrow \sqrt{\frac{1}{k} \sum_{i=1}^{n} (c_{\alpha(i)} - \mu)^2}$
6: $S \leftarrow$ a sequence of $S_i = \exp \left( \frac{b_i (\mu - c_{\alpha(i)})}{\sigma} \left( 1 - \exp \frac{c_{\alpha(i)} - c_{\alpha(k)}}{m c_{\alpha(k)}} \right) \right)$ for $i \in \{1 \ldots k\}$ \(\triangleright\) compute scores
7: $M \leftarrow \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ \(\triangleright\) accumulator for the linear equation system
Since the squared term is linear in \( x \), the inner function is a quadratic form that can be minimized algebraically by finding zero of its derivation. This procedure gives the formulas used in the algorithm.

The algorithm can be easily expanded to embedding into general \( P \)-dimensional spaces by taking the low-dimensional landmarks \( l \) from \( \mathbb{R}^{n \times P} \), increasing the size of the matrix \( M \) for a linear equation of \( P \) variables, and solving a larger linear system at the end.

Notably, the initial reduction of the input data to one-dimensional projections to affine spaces (\( d \) in the algorithm) prevents various complications from fitting the high-dimensional distances into low-dimensional space, avoiding many problems that arise from dimensionality overhead in other DR algorithms. A similar approach has been taken e.g. by TriMap\(^5\), where the transferred information is reduced to mere binary relations between point distances.

**Embedding parameters** The embedding procedure admits several tunable parameters: \( k \) is the number of nearest landmarks used for the approximation, \( m > 0 \) is an arbitrary parameter that selects the steepness of score decay for distance order approaching \( k \), \( b > 0 \) chooses the steepness of score decay for landmarks far from \( X \), and \( a \) lowers the score of approximations to pairs of relatively far low-dimensional landmarks.

Parameter \( m \) is specifically designed to lower the score of landmarks with distances that approach \( k \)-closest landmark. As a result, small changes in the input point \( X \) can not cause sharp changes in the scores assigned to individual parts of the approximation. Consequently, EmbedSOM function is smooth in \( X \).

Values of parameters \( k, m, m \) and \( a \) must be chosen to avoid singularities and near-singularities when computing the final approximation, which may happen if the set of \( s_j \) contains insufficient number of higher-than-negligible scores. That may be caused mainly by setting too low values of \( k \) or \( m \), or too high value of \( b \). Argument setting of \( k \approx \sqrt{L} \), \( m = 10 \), \( b = e^{-1} \) and \( a = 1 \) worked well in a majority of tested use cases and can be considered a good default.

**Embedding complexity** To compute a \( P \)-dimensional projection of a single point from a \( D \)-dimensional space, EmbedSOM projection conducts the following operations: \(|L|\) measurements of distances in high-dimensional space, sorting the \( k \) smallest elements of the distance vector of size \(|L|\), and conversion of \( k \) distances to scores. On the
landmark pairs, it conducts at most $k^2$ computations of scores $s$, the same number of computations of $d_i$ from 2 dot-products in high-dimensional space, and computation of a partial $P$-by-$(P + 1)$ matrix for solving the linear system in $\mathbb{R}^P$. Finally, the linear system is solved using Cramer’s rule. The total of computation times is thus, in respective order,

$$\mathcal{O}(D \cdot |L|) + \mathcal{O}(\log k \cdot |L|) + k^2 \left( \mathcal{O}(1) + \mathcal{O}(D) + \mathcal{O}(P^2) \right) + \mathcal{O}(P^2).$$

Assuming the default parameter setting and $P \epsilon \{2, 3\}$, this complexity sums to $\mathcal{O}(D \cdot |L|)$. The procedure can be trivially repeated for any number of input points.

**Different distance measures** We have assumed that the metric used in both high-dimensional and low-dimensional spaces is Euclidean. Generally, EmbedSOM behaves well even if the distance measure used for the scoring function is swapped for any function that acts as a metric on vector spaces, including the popular $L^1$ and $L^\infty$ metrics.

Nevertheless, the computation of ‘projections’ using dot-products may then be viewed as a rather questionable reinterpretation of the point coordinates in an inner product space. Fortunately, the minimal-distance projection to a fixed subspace is a linear operator under both $L^1$ and $L^\infty$, which is sufficient for EmbedSOM computation even without requiring the inner product property.

**Generalized landmarks and GQTSOMs**

While the SOMs are a great method to generate landmarks $L$ and $l$ that carry various beneficial properties that simplify human interpretation of the result (notably the regularity of $l$), other methods are admissible as well, as long as they can cover the input space sufficiently by $L$ and generate the corresponding landmarks $l$ in the low-dimensional output space so that the topology is similar to $L$.

For example, the embedding process can be simplified to a great extent by completely removing SOMs: Instead of constructing $L$ in a complicated way so that it reflects the input space topology, we can take only a small random sample of input points as the landmarks, and use a general DR method to find its topology and arrange landmarks $l$ in a matching way, as shown in Figure 1 on an example with t-SNE. While this is often sufficient, for the purposes of embedding it is more beneficial to find a smaller set of landmarks that provide better description of the various features in the input space than the random sampling.

Many variants of the SOM algorithm have been created to optimize this metric: For example, the Growing SOMs (GSOMs) by Dittenbach et al., start with a simple 2x2 SOM grid, and dynamically add new SOM grid vertices at the SOM perimeter only if it is necessary to keep the total quantization error low. A hierarchical variant of GSOM called GHSSOM introduced by Rauber et al. aims to improve the description of small details in the input data space that were not described sufficiently by GSOMs. Depending on the heuristic, the vertices of GHSSOM grid are converted to small independent versions of GHSSOMs, which map the corresponding local parts of the input space; this continues recursively to create a layered structure of SOMs that describe increasingly fine and subtle details in the data.

**GQTSOMs** Although the GHSSOMs improve the classification of small-scale features in the datasets, the hypertree structure complicates their use as landmarks for planar visualization with EmbedSOM. We propose the Growing QuadTree-structured SOMs (GQTSOMs) to alleviate this problem: The GQTSOMs grow by recursively splitting the nodes to form a hypertree, but unlike GHSSOMs the hypertree shape is restricted to a quadtree, which possesses straightforward interpretation as a 2-dimensional structure.

The nodes in GQTSOMs are identified by their position and depth in the quadtree, represented as an integer triple $(L, x, y)$. The corresponding 2-dimensional coordinates are obtained as $(2x + 1, 2y + 1) \cdot 2^L$. Initial nodes in training occupy positions $(1, 0, 0)$, $(1, 0, 1)$, $(1, 1, 1)$, and $(1, 1, 0)$. Upon growing, a node $(L, x, y)$ is split into 4 nodes identified as $(L + 1, 2x + 2y)$, $(L + 1, 2x + 2y + 1)$, $(L + 1, 2x + 2y + 1)$, and $(L + 1, 2x + 2y + 1)$. Figure 1 shows an example of 3-level GQTSOM in a 2-dimensional space, where the initial 2x2 SOM grew three times to produce 13 landmarks.

GQTSOM training proceeds by batches as in the usual batch SOM training. After each epoch, several nodes with greatest position change in the input data space are split, so that the total number of nodes grows linearly during the whole training. Initial positions for the new nodes are interpolated from the topological SOM neighborhood.
using the same neighborhood function as for training the SOM (e.g. a Gaussian). To avoid overcrowding of the map by small nodes and promote their specialization to fine details, the nodes are penalized by a factor of $L^{-1}$ in the growing heuristic, and by a factor of $4^L$ applied to their neighborhood volume in both input space and SOM space.

**Implementation**

The current version of EmbedSOM is available as R package EmbedSOM from [http://github.com/exaexa/EmbedSOM](http://github.com/exaexa/EmbedSOM), together with the customized versions of SOM and GQTSOM algorithms. The implementations are conducted in C++; the code is independent from the actual package or R language, and can be reused in other environments. The integration into R serves mostly as a bridge to the large number of cytometry-oriented packages in the ecosystem.

Low-level implementation has provided several additional ways to improve the performance of the algorithms when compared to the original implementation: For example, cache-efficient version of the SOM training has improved the performance by up to 15× on large SOMs; SIMD-based acceleration of the vector operations by up to 4×, and parallelization of the batch SOM training and embedding by a factor roughly equivalent to the number of used CPUs.

Overall, the computation time required for typical datasets was reduced by a factor greater than 10× on commonly available hardware, and more than 30× in case of processing complicated datasets using large SOMs.

**Operation**

EmbedSOM is now used primarily from R environment; the package can be downloaded from GitHub using R command `devtools::install_github('exaexa/EmbedSOM')`. The package installation will automatically compile the code that uses the SIMD capabilities if they are enabled on the target platform.

Generally, the SOM and embedding process can be executed on any real matrix with individual data points in rows, and parameters in columns. This expectation is consistent with many other DR or clustering packages, including FlowSOM, Rtsne and umap. For example, a user may obtain an embedding of the Iris dataset as such:

```r
library(EmbedSOM)
# Load data
d <- iris[,1:4]

# Create a SOM
map <- SOM(d)

# Embed the SOM
e <- EmbedSOM(map=map, data=d)
```

In the code, the landmarks are first created using a SOM and saved in the `map`, which is then passed to the `EmbedSOM` function that produces the final 2-column matrix `e` with embedded coordinates. These can be plotted e.g. using the standard `plot` function.

On data larger than Iris dataset, GQTSOMs may be used to generate the landmarks and a map usable with `EmbedSOM` function in a similar way:

```r
map <- GQTSOM(d, target_codes=500, parallel=T)
```

Here, `target_codes` chooses the desired final number of the landmarks in the fully grown SOM, and parameter `parallel=T` allows the computation to use multiple available CPUs. Functions `SOM` and `EmbedSOM` support parallelization as well, using the same parameter.

Other DR methods may create the landmarks. In case of t-SNE, the following code generates a map object with 500 landmarks suitable for embedding:

```r
library(Rtsne)
landmark_idx <- sample(dim(d)[1], 500)
map <- list(codes=d[landmark_idx,], grid=Rtsne(d[landmark_idx,])$Y)
```
The parameters of the SOM, GQTSOM and EmbedSOM functions are extensively documented in the supplied R manual pages.

Use cases

The output of new EmbedSOM variants is demonstrated on two use-cases: First, using the described variants of landmark-generating functions, we reproduced the visualizations by Becht et al.9 of a dataset that maps specific trafficking and cytokine signatures of human T cells across tissues, created by Wong et al.10. Second, we visualized a human gastrointestinal disorders dataset by van Unen et al.11 using GQTSOMs, showing that EmbedSOM provides a viable alternative to the semi-interactive analysis of rare cell types using the HSNE algorithm12.

The primary purpose of EmbedSOM is to provide quickly available and highly comprehensible data visualization in situations where processing speed is critical. The embedding time of the demonstration datasets was measured on an AMD Ryzen 7 2700U CPU with 16GB of RAM running Debian Linux (Bullseye), R version 3.6.1 compiled with gcc version 9.2.1; the timing is reported in the corresponding figures as t, together with number of cells (n) and landmarks (|L|). As the main result, the measurements show that a high-quality visualization of a data file from a common experiment (around 300 thousand cells) can be obtained in less than 10 seconds using common office hardware.

Alternative landmark-generating methods improve visualization

To visualize the Wong dataset, we have embedded it using EmbedSOM with SOM landmarks (i.e. the original EmbedSOM), EmbedSOM with t-SNE generated landmarks, and EmbedSOM with GQTSOM-generated landmarks. As seen in Figure 2, the original EmbedSOM implementation has managed to separate and visualize both the different cell types and their layout according to source organ. However, the result may seem unsatisfactory due to overcrowding and loss of both detail and global layout, especially when compared to UMAP visualizations of the same dataset [9, Figure 1a,b]. Despite the overcrowding, it is still possible to identify separate clusters of CD69+CD103+ Trms (resident-memory T cells) in all organs except cord blood, and naive (CD69−CD45RA+), central memory (CCR7+CD62L−) and effector memory T cells (CD45RA−CD45RO+CCR7−CD62L+) within both CD4 and CD8 T cell types; this is in agreement with findings of van Unen et al. [11, Figure 3a,b]. Plots of all marker expressions are available as Extended data.

We have observed that the improved methods of landmark positioning have successfully alleviated both overcrowding and layout problems. In particular, the layout of MAIT (mucosal-associated invariant T) and γδ T cells in the embedding with t-SNE-generated landmarks reflects the expected properties of cell populations, and the individual population clusters are clearly separated by low-density areas with intermediate cell states and noise. Additionally, the smoothness of EmbedSOM projection has displayed even features and similarities that were not captured by t-SNE. That can be seen on the cluster of γδ T cells, where EmbedSOM shows a clear connection of the gut-originating part of the cluster to both the very similar gut-originating CD8+ T cells and the other γδ T cells. In comparison, this connection is preserved by all tested types of SOMs, but ignored by both plain t-SNE and UMAP, which show the population separated to 3 resp. 2 separate clusters [9, Figure 1a]. The embedding based on GQTSOM landmarks has provided similar global layout of the output as the one with t-SNE landmarks, additionally capturing the continuity of γδ T cell cluster and its similarity to MAIT and NK cells, and providing separation of individual clusters differentiated by tissue of origin comparable to that of UMAP. Regardless of that, the main advantage of using GQTSOMs comes with the ability to generate smaller amount of more precise landmarks than plain SOMs. Compared to the SOM used with the original EmbedSOM approach, this resulted in significant computation speed increase (around 50%) and better description of the small and rare cell populations by landmarks (and, therefore, also by subsequent FlowSOM-style clustering). In particular, the small subpopulations of γδ T cells were assigned roughly twice the number of landmarks by GQTSOM than by the standard SOM, which resulted in spatially correct separation of the cell subtypes in the embedding.

GQTSOM landmarks improve display of rare cell types

We showcase the ability of GQTSOM landmark generation method to capture and display various rare cell types using a dataset by van Unen et al.11. The dataset was created as such: A total of 5.2 million single cells were collected from duodenum biopsies, rectum biopsies, perianal fistulas, and PBMC from patients undergoing various gastrointestinal disorders and healthy individuals (as controls). The gastrointestinal disorders included celiac disease (CeD), refractory celiac disease type-II (RCDDII), enteropathy associated T-cell lymphoma type II (EATLII), and Crohn’s disease. Cells were stained using 32 metal conjugated monoclonal antibodies to identify cells within the innate and adaptive immune system. This dataset was later reanalyzed by van Unen et al.12 using
Figure 2. Comparison of EmbedSOM visualizations of the Wong dataset using different landmarks. Top row: cells embedded using 3 different landmark-generating methods, colored by the tissue of sample origin. Middle row: The same embedding colored by major cell types. The colors used for annotation are purposefully reproduced from the article of Becht et al.⁹ to simplify comparison. Bottom row: visualizations of the low-dimensional landmark images, colored by their corresponding marker expressions.

a hierarchical version of t-SNE algorithm called HSNE, showing that the hierarchical dissection of the data was able to identify several rare cell types within the innate lymphoid cell (ILC) compartment.

For the purpose of demonstration, we preprocessed the same dataset by removing debris, doublets and dead cells based on simple thresholds on the DNA, Event length and Viability parameters. The 32 antibody markers of the 4.14 million cleaned cells were then transformed by hyperbolic arcsine and used to train the GQTSOM and produce an embedding. The result in Figure 3 allows easy observation of both the ILC compartment and the CD4⁺ T cell subset, corresponding to the observations produced by second-level HSNE [12, Figures 3 and 5]. Additionally, the embedding shows presence of many clusters from lower levels of the hierarchical dissection:
Figure 3. Display of clusters of rare cell types in GQTSOM-based embedding. Top left: Overview of the cleaned and embedded Unen dataset, colored by expression of main cell lineage markers. The contour based on Gaussian difference is added for easier identification of changes in cell density. Labels mark the rare cell types identified by van Unen et al.\cite{12}, Belkina et al.\cite{13}: (a) CD4\(^+\)CD28\(^-\)CCR7\(^+\), (b) CD4\(^+\)CD28\(^-\)CCR7\(^-\)CD56\(^-\), (c) CD4\(^+\)CD28\(^-\)CCR7\(^-\)CD56\(^+\), and (d) CD7\(^+\)CD3 CD127\(^+\)CD5RA\(^-\) CD56\(^-\) partial. Top right: Expressions of separate markers used for the identification. Bottom: Cells color-coded by sample origin (left) and separated by disease status of the patient (right).

In the figure, it is possible to identify clusters of CD4\(^+\)CD28\(^-\)CCR7\(^-\)CD56\(^-\) and CD4\(^+\)CD28\(^-\)CCR7\(^+\)CD56\(^-\) rare cell types within the CD4\(^+\) compartment, and of the CD127\(^+\)CD45RA\(^-\)CD56\(^-\) cluster within the ILC (CD7\(^+\) CD3\(^+\)) compartment. These clusters were identified by HSNE at 4\(^{th}\) resp. 3\(^{rd}\) levels of dissection \cite{12, Figures 5b and 3c}. Recently, Belkina et al.\cite{13} showed that the opt-SNE algorithm can additionally identify CD4\(^+\)CD28\(^-\)CCR7\(^+\)CD56\(^-\) rare cell type, which is also clearly separated by the GQTSOM-based embedding, using much less computational resources than optSNE.

The plot of cells separated by disease status in Figure 3 confirms the observation that the rare CD4\(^+\)CD28\(^-\)CD56\(^-\) phenotype is enriched in the samples from patients with Crohn’s disease. Moreover, the plot gives a useful overview for identifying cell types specific for the other diseases, showing two specific and one enriched...
cluster for RCDII, a single specific cluster of CD8⁺CD56⁻CD127⁺c-KIT⁺ cells for EATLII, and one specific and some enriched cell types in patients with CeD.

**Summary**

We have presented an improved and generalized version of EmbedSOM, supported by the new model of quadtree-structured growing self-organizing maps. The functionality of the new algorithm was demonstrated on data and analyses from recent studies, showing that the new combination provides superior embedding speed and good rendering of various cell types, including tissue-specific and rare phenotypes.

**Software availability**

Source code available from: https://github.com/exaexa/EmbedSOM

Archived source code available from: https://doi.org/10.5281/zenodo.3568980

License: GNU GPLv3

**Data availability**

**Underlying data**

The used datasets are freely available from FlowRepository.org under accession IDs:

FR-FCM-ZZTM (Wong dataset; the data was preprocessed exactly as described by Becht et al.⁹)
FR-FCM-ZYRM (Unen dataset)

**Extended data**


Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

**Acknowledgements**

We would like to thank all authors of the original EmbedSOM article for supplying the interesting problems and use-cases that motivated the development of the current version of EmbedSOM.

**References**


Nikolay Oskolkov
Department of Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Lund University, Lund, Sweden

The authors present an improved version of EmbedSOM that was optimized for speed to address extra large single cell data sets. In addition, a novel way of growing SOM based on quad-tree was proposed.

Dimension reduction of large single cell RNAseq and flow cytometry data sets is very challenging due to the crowding problem and algorithm scalability issues. Therefore, developing methods alternative to the current golden standards such as PCA, tSNE and UMAP is of high importance. More specifically, tSNE and UMAP are capable of preserving only local structure while PCA keeps the global structure information. However, no method is currently available that can preserve both local and global structure.

Self-Organizing Maps (SOMs) and the modified EmbedSOM that are discussed in the manuscript represent an interesting and promising algorithm in this respect. However, I would like to raise a few questions and concerns to be addressed by the authors.

First, based on the cost function mentioned on the page 3, the algorithm seems to resemble MDS / PCA type of dimension reduction. Therefore, I would like to see a comparison of EmbedSOM with MDS / PCA. If a connection between the gamma-delta T cells and CD8 T cells was not captured by tSNE and UMAP as it is mentioned on the page 8, probably due to the lack of global structure preservation by tSNE and UMAP, was this connection captured by MDS / PCA?

Second, what would be the benefit of using EmbedSOM compared to PCA / MDS, tSNE and UMAP? Do we discover any new biology using EmbedSOM that is not captured by PCA / tSNE / UMAP? Do we benefit from the computational speed of EmbedSOM compared to PCA / tSNE / UMAP? If so, is it really faster (and how much faster) than PCA? I would like to see a clear formulation of the role of the EmbedSOM among other dimension reduction methods.

Third, I was really impressed by the Figure 1 and how well GQTSOM-based embedding was able to reconstruct the original 3D S-shaped non-linear manifold. To my experience, tSNE / UMAP and especially PCA / MDS would have difficulty reconstructing the 3D S-shaped manifold as 2D embeddings. I have not
found any links to the codes for reproducing this embedding and would be very curious to see whether GQTMSOM / EmbedSOM is really capable of capturing the internal 2D structure of the 3D S-shaped non-linear manifold.

**Is the rationale for developing the new software tool clearly explained?**
Partly

**Is the description of the software tool technically sound?**
Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**
Partly

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**
Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Computational Biology, Bioinformatics, Mathematical Statistics and Machine Learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 11 February 2020

https://doi.org/10.5256/f1000research.23857.r57989

© 2020 Newell E et al. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Evan Newell
Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Hugh MacMilan
Fred Hutchinson Cancer Research Center, Seattle, WA, USA

The authors generalize their EmbedSOM approach to examine two additional ways of selecting the respective sets of landmarks in the high- and low-dimensional spaces, beyond the standard SOM, to address. Example data analyses are appreciated.
That selection is the first stage in the embedSOM approach. The second stage is the actual embedding enrichment process.

The authors explain: "EmbedSOM projection can be viewed as an embedding enrichment method: From a set of landmarks in the high-dimensional space and a set of corresponding landmarks in the low-dimensional space, it produces a smooth function that maps all points from the higher-dimensional space to the low-dimensional space and preserves the relative neighborhoods of the landmarks."

Testing:
- We followed the paper’s guidance on some in-house fcs files and had success with embedSOM and the GQTSOM function.

Some naming confusion:
- By "generalized EmbedSOM" the authors refer to using different ways of generating landmarks, other than the original (self-organizing map) SOM approach.
- It seems preferable to drop the "SOM" rather than refer to these variants as "generalized EmbedSOM" methods. The authors might use the more general notion of landmarks, rather than SOM. As they note, the random-sampling, followed by tSNE, version of “generalized EmbedSOM” doesn’t use SOMs at all.

Re "compacting noise"
- The first reference of the manuscript includes some background on differences between the "generalized EmbedSOM" approach and what the authors call "plain tSNE and UMAP," and attributes these differences to the respective designs of the algorithms.
- In that background paper, the authors explain: “neither UMAP nor tSNE aim to preserve local linearity of the transformation, which allows them to take apart the clusters with noisy data and attach the residual noise to nearest clusters."
- They concluded in that paper: “Compacting the residual or unexplained noise is desirable for providing a clean display of the data for publication. On the contrary, almost-immediate availability of all information about very large datasets, including the (often informative) noise, is more important for producing comprehensive graphics for high-throughput analysis."
- This paper marks an attempt to explore those differences, and the apparent trade-offs, in more detail, so it would benefit from discussing these tradeoffs in the context of the algorithm designs.
- The authors noted in the first reference, "While the observed cluster separation may be desirable if the embedding is expected to approximate the population boundaries, it may be inappropriate if the population environment is relevant for analysis."

GQTSOMs:
- The manuscript introduces a new landmark-generating algorithm that simplifies a hierarchical variant of an adaptive SOM approach, namely, growing quad tree SOMs (GQTSOMs) as a
simplified growing hierarchical SOM (GHSOM), which is in turn a variant of growing SOMs (GSOMs).

- The aim is to identify and incorporate features in the input space more efficiently than random sampling, by using a "layered structure of SOMs". This is a natural thing to do to improve on SOM.

- On page 8 the authors report that GQTSOM leads to using a “smaller amount of more precise landmarks” and thus faster computation, and appears to be a nice contribution.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Cellular immunology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

**Author Response 22 Feb 2020**

**Miroslav Kratochvil**, Institute of Organic Chemistry and Biochemistry of the CAS, Prague, Czech Republic

Thank you for the review and comments. We will wait for the additional reviews and address some of your suggestions in the second version of the manuscript.

Regarding the name of EmbedSOM, we are aware of the issue with textual “specialization” to SOMs which ignores the modifiable parts of the workflow, but since the package is already published over a year and we had not been able to invent a strictly better name so far, we expect that the name will stay. We will gratefully accept suggestions (also from readers) for a reasonably short name that sufficiently characterizes the projection procedure.

**Competing Interests:** No competing interests were disclosed.
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com