Bioinformatics, 32(5), 2016, 747–754
doi: 10.1093/bioinformatics/btv630
Advance Access Publication Date: 4 November 2015
Original Paper

OXFORD

Bioimage informatics

Fast Optimized Cluster Algorithm for Localizations (FOCAL): a spatial cluster analysis for super-resolved microscopy

A. Mazouchi[1] and J. N. Milstein[1,2,*]

[1]Department of Chemical and Physical Sciences, University of Toronto Mississauga, Mississauga, ON L5L 1C6, Canada and [2]Department of Physics, University of Toronto, Toronto, ON M5S 1A7, Canada

*To whom correspondence should be addressed.

Associate Editor: Robert Murphy

Received on April 7, 2015; revised on September 22, 2015; accepted on October 24, 2015

Abstract

Motivation: Single-molecule localization microscopy (SMLM) microscopy provides images of cellular structure at a resolution an order of magnitude below what can be achieved by conventional diffraction limited techniques. The concomitantly larger data sets generated by SMLM require increasingly efficient image analysis software. Density based clustering algorithms, with the most ubiquitous being DBSCAN, are commonly used to quantitatively assess sub-cellular assemblies. DBSCAN, however, is slow, scaling with the number of localizations like $O(n \log(n))$ at best, and it's performance is highly dependent upon a subjectively selected choice of parameters.
Results: We have developed a grid-based clustering algorithm FOCAL, which explicitly accounts for several dominant artifacts arising in SMLM image reconstructions. FOCAL is fast and efficient, scaling like $O(n)$, and only has one set parameter. We assess DBSCAN and FOCAL on experimental dSTORM data of clusters of eukaryotic RNAP II and PALM data of the bacterial protein H-NS, then provide a detailed comparison via simulation. FOCAL performs comparable and often superior to DBSCAN while yielding a significantly faster analysis. Additionally, FOCAL provides a novel method for filtering out of focus clusters from complex SMLM images.
Availability and implementation: The data and code are available at: http://www.utm.utoronto.ca/milsteinlab/resources/Software/FOCAL/
Contact: josh.milstein@utoronto.ca
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the advent of super-resolved microscopy, images of cellular structure and organization can be acquired with visible light at a spatial resolution well surpassing the diffraction limit. A popular family of techniques, referred to as single-molecule localization microscopy (SMLM), can achieve a lateral resolution down to a few tens of nanometers, with only slightly worse axial resolution. SMLM works by driving a dense population of fluorescent molecules through multiple cycles of photoactivation, photoswitching or blinking and bleaching. Under appropriate conditions, it's possible to temporally isolate the fluorescence of a single-molecule, during each cycle, from adjacent molecules within a diffraction limited volume. The precise location of each isolated molecule is determined, limited primarily by the number of emitted photons, and the process is repeated tens of thousands of times to generate a reconstructed image from the assembled localizations. Examples of SMLM techniques are stochastic optical reconstruction microscopy (STORM) (Rust *et al.*, 2006), direct STORM (dSTORM) (Heilemann *et al.*, 2008), photo-activated localization microscopy (PALM) (Betzig *et al.*, 2006), bleaching/blinking-assisted localization microscopy (BaLM) (Burnette *et al.*, 2011), and so on.

SMLM is increasingly being employed in cells to image and quantitatively analyze various protein complexes forming tens of nm to hundreds of nm assemblies, from membrane receptors to nucleosome bundles (Bar-On *et al.*, 2012; Dudok *et al.*, 2014; Endesfelder *et al.*, 2013; Itano *et al.*, 2014; Pertsinidis *et al.*, 2013; Ricci *et al.*, 2015). However, because of the techniques inherent single-molecule sensitivity, SMLM is prone to a host of artifacts that have little effect on conventional diffraction limited modalities. For membrane proteins or other targets close to the cell surface, illumination via Total Internal Reflection Fluorescence (TIRF) can considerably improve the image quality. Unfortunately, it can be much more difficult to employ SMLM to image intranuclear structures, for example, which may reside deep within the cell. For abundant proteins or extended structures that fill large portions of the cellular volume, alternatives to wide-field illumination, such as HILO (or dirty-TIRF) (Tokunaga *et al.*, 2008) can provide course optical sectioning. Of course, for even more crowded regions, it may be necessary to resort to complicated light-sheet illumination schemes (Cella Zanacchi *et al.*, 2011; Gebhardt *et al.*, 2013; Hu *et al.*, 2013).

Consider the imaging artifacts that arise when using immunofluorescence labeling to target specific molecules within a fixed cell (Haas *et al.*, 2014). Nonspecific interactions of the tagged antibodies with either the growth substrate or cell generate false localizations that add noise to the reconstructed image (Burnette *et al.*, 2011). A static fluorescence background can reduce the precision with which individual target fluorophores can be localized. And blurred, out of focus, blinking flourophores not only increase the local background, but if located close to the focal plane, can be imprecisely localized.

There are a number of approaches to characterizing the spatial organization and nanoscale morphology of protein assemblies imaged with SMLM. On the one hand, there are ensemble methods that provide spatial statistics, such as pair-correlation analysis (Sengupta *et al.*, 2013) or Ripley's L-function (Endesfelder *et al.*, 2014). On the other are methods that directly group localizations into discrete clusters. For SMLM, density based spatial clustering of applications with noise (DBSCAN) is the most commonly employed method of this type (Bar-On *et al.*, 2012; Endesfelder *et al.*, 2013; Pertsinidis *et al.*, 2013). Significant analysis can follow cluster identification, such as determining the cluster size/radius or the number of proteins within each cluster (Cattoni *et al.*, 2013; Deschout *et al.*, 2014; Ghamari *et al.*, 2013).

The ability of DBSCAN to identify and reject 'noise' is advantageous, which to a large extent accounts for its popularity. However, in addition to the complications of background noise, artifacts appear in SMLM images because the localization density is a function of both the number of tagged proteins in an assembly and the photophysics of the fluorescent labels. Since most fluorophores blink multiple times, and there is some error involved in the localization of each blink, a single fluorescent label will result in a cluster of localizations. Likewise, because the duration of each blink is stochastic, blinks that last for multiple frames will also generate a cluster. Therefore, a single density threshold, as in DBSCAN, may not be sufficient to differentiate between clusters of localizations due to protein assemblies and artifactual pseudo-clusters arising, for instance, from a fluorescently labeled, solitary protein. This is a concern for all SMLM methods, since even with fluorophores referred to as irreversibly photoactivatable, photoblinking is commonly observed (Annibale *et al.*, 2010; Nan *et al.*, 2013). Unfortunately, some clusters may also appear as 'noise' due to limitations in the image reconstruction. For instance, because of the inherently low photo-activation efficiency of the fluorophores, multiple clustered proteins will occasionally produce only a single blink during the acquisition period.

Moreover, in the presence of significant noise and off- target localizations (i.e. localizations not arising from in focus (target) clusters), the performance of DBSCAN critically depends on a pair of set parameters that specify a density threshold (Gebhardt *et al.*, 2013) (i.e. the number of points, minPts, within a search radius ε). A significant drawback to DBSCAN is that the appropriate choice for these parameters is often ambiguous. In practice, clustering results for various sets of DBSCAN parameters are typically generated and the optimum choice is iteratively chosen from an inspection of the results. Beyond the obvious subjectivity of a visual inspection, this procedure may not be practical for localization microscopy when dealing with large data sets of hundreds of thousand to millions of localizations.

To identify neighboring points, DBSCAN calculates all pairwise distances within an image, so the runtime scales with the number of points n as $O(n^2)$. For a large SMLM data set, this analysis can take several hours. Region queries such as grid-based spatial indexing and tree data structures have been implemented to restrict the search for neighboring localizations to obtain an average time complexity of $O(n \log(n))$ (Kłopotek *et al.*, 2006). Despite this improvement, in order to find the algorithm's optimum parameters, DBSCAN should be executed multiple times. In fact, the parameters might vary from image to image due to unavoidable variations in sample preparation such as labeling specificity and heterogeneity of cells. An optimal clustering analysis with DBSCAN, therefore, can be extremely time consuming.

Here we present a rapid clustering algorithm customized for SMLM: Fast Optimized Cluster Algorithm for Localizations (FOCAL). FOCAL is similar in many respects to a gridded version of DBSCAN, and like grid-based versions of DBSCAN, at worst scales like $O(n)$ (Darong and Peng, 2012). But while DBSCAN is an extremely general clustering algorithm, FOCAL incorporates knowledge of artifacts within SMLM images to optimize the analysis. Moreover, we present a well defined method for quickly selecting FOCAL's optimum parameters, which minimizes clustering errors from SMLM image artifacts. In both simulations and experiments on immunofluorescently labeled clusters of RNAPII via dSTORM and fluorescent protein tagged clusters of the histone-like nucleosomal protein (H-NS) by PALM, we compared the performance of FOCAL and DBSCAN in detecting spatial clusters and in dealing with and filtering out image noise and off-target localizations. Our results suggest that FOCAL performs comparable and often superior to DBSCAN while providing a significantly faster analysis of complex SMLM images. In addition, FOCAL provides a novel method for filtering out of focus clusters from complex SMLM images.

2 Methods

Density based clustering techniques find immediate application in analyzing SMLM images due to their ability to identify clusters of arbitrary shapes in the presence of significant off-target signal. DBSCAN is the most popularly employed member of these techniques and is the foundation for many more elaborate clustering methods. We begin by providing a brief overview of DBSCAN, enumerate its limitations, and then move on to a discussion of our rapid clustering method FOCAL.

2.1 Density-based clustering with DBSCAN

In brief, DBSCAN first determines the density of data points in the neighborhood (radius ε) of each point, then exploits a density threshold (minPts) to assign a classification as core, border or

off-target points. Core points have at least minPts neighbors, border points are not core points but contain a core point in their neighborhood, and all other points are classified as off-target. Consequently, neighboring core and border points will be grouped into clusters (Ester *et al.*, 1996).

A quantitative analysis of SMLM data is plagued by DBSCAN's sensitivity to the choice of input parameters, which, in practice, are set either by imprecise heuristics or empirically (Kłopotek *et al.*, 2006). Note, by empirically we mean that the parameters are chosen so the resulting clustering appears more accurate upon visual inspection, an analysis that is both highly subjective and time consuming since it requires an iterative analysis. Nonetheless, an empirical evaluation is the common approach since the heuristically determined parameters tend to perform poorly. For instance, the original DBSCAN article suggests setting minPts $= D + 1$, where D is the dimensionality of the system. In two-dimensional SMLM applications, minPts is frequently set to values ranging from 3-10 (Dudok *et al.*, 2014; Itano *et al.*, 2014; Endesfelder *et al.*, 2013; Pertsinidis *et al.*, 2013), although larger values (e.g. 16 (Bar-On *et al.*, 2012)) are sometimes chosen. Likewise, it is often remarked in the literature that the optimum radius ε can be obtained from a knee, or inflection point, in the *k*-distance plot, which is a plot of the sorted *j*th neighbor distances ($j = 1,2,3\ldots$) of all data points. Unfortunately, not only does ε depend upon the choice of j, which is again arbitrary, but soft-knees often arise in *k*-distance plots that hardly narrow down the optimum value of ε. In fact, as depicted in Figure 1, for SMLM data we obtained of the RNAPII distribution in the nucleus of a mouse cortex cell, the knee spans a wide range of ε (visually, between 20 and 40 nm).

Overestimating ε and/or underestimating minPts results in more false clusters and the incorporation of more off-target outliers into the clusters. Not surprisingly, the cluster assignment was found to be significantly inaccurate when analyzing SMLM images in cells and was addressed by a hybrid simulation assisted DBSCAN (SAD) method (Nan *et al.*, 2013). However, SAD only evaluates the statistics of false clusters arising from a random background of localizations, but does not necessarily account for artifacts from out of focus features and photoblinking. Moreover, since the algorithm must be executed multiple times to obtain an optimum parameterization, the computational cost could pose a significant obstacle when facing large data sets containing millions of localizations.

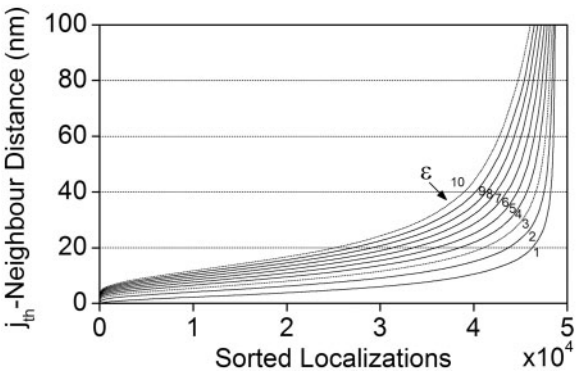

Fig. 1. *K*-distance plot derived from the localization table of the RNAPII distribution in a mouse cortex cell. The optimum ε is commonly estimated from the *j*th-neighbor curves 3–10. Localizations within clusters contribute to the left portion of the curve, while low-density localizations gives rise to the sharp upturn at the far right. The value at the knee between these two regions was originally suggested as an optimized ε for DBSCAN, which is clearly ambiguous in this example

2.2 Fast optimized cluster algorithm for localizations (FOCAL)

The ambiguity inherent in choosing DBSCAN's parameters as well as its high computational overhead is largely due to the algorithms generality. In order to address these limitations, we developed a Fast Optimized Cluster Algorithm for Localizations (FOCAL), specifically designed for single-molecule localization microscopy. FOCAL is a grid-based method with linear time complexity ($O(n)$). This makes it possible to rapidly optimize the algorithm's parameters and analyze data. Moreover, FOCAL only has a single parameter that needs to be optimized: a density threshold (minL). Details of the algorithm are depicted in Figure 2.

SMLM images are built from a localization table, which is a list of the approximate spatial coordinates of fluorescent labels within the sample. We first discretize the localization table by assigning each localization event to a gridded bin (Fig. 2A). A single label produces a spread of localizations, which gives rise to the localization uncertainty (σ), and can be easily estimated for each experiment from a temporal, adjacent-neighbor analysis (Endesfelder *et al.*, 2014). The localization uncertainty σ, therefore, provides a natural scale for the grid spacing. In fact, a bin-size smaller than the localization uncertainty costs more memory without containing additional structural information. For very bright organic dyes such as Alexa-647, a localization uncertainty of ∼10 nm is achievable (Betzig *et al.*, 2006; Rust *et al.*, 2006).

At this stage, contrast between the cluster regions and the background is very low and a simple thresholding cannot efficiently distinguish the two regions. A larger grid would improve the contrast at the expense of losing fine structural information. To increase the contrast several-fold, without changing the grid spacing (i.e. the

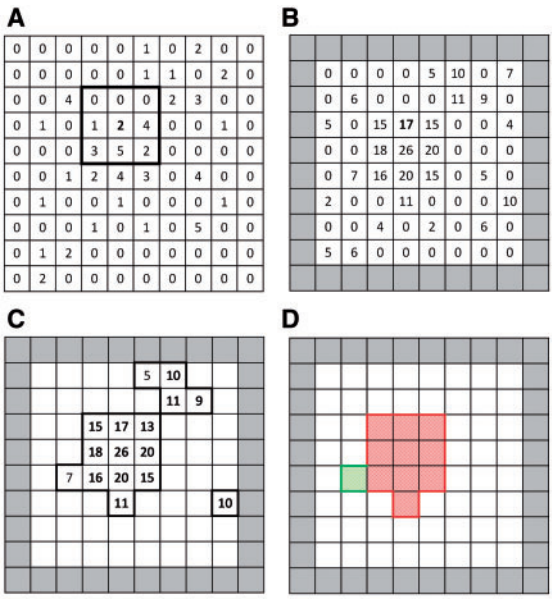

Fig. 2. FOCAL clustering algorithm. The localization table is distributed into bins of size σxσ, where σ is the localization uncertainty (A). A density map is constructed by replacement of each bin value with the sum of the 3×3 array centered at that bin (B). Bins at the edge of the image (grey shaded) are discarded from the analysis. A density threshold (minL), set to 9 in this example, specifies core (bolded) and border (regular) points (C). Only clusters of minimum size 3×3 plus 1 are retained to filter single point sources (D). The identified cluster is shown with red/checkered (green/striped) bins representing core (border) points (Color version of this figure is available at *Bioinformatics* online.)

resolution), a density map is constructed from the binned localizations by substituting each non-zero assignment in the grid with the sum of the 3×3 array centered at that same bin (Fig. 2B). The resulting density map can now be thresholded and those bins of value greater than minL are identified as core points. Of the remaining bins, those that do not satisfy this criterion, but are a 4-connected neighbor to a core point, are considered as border points and included in the cluster candidate map (Fig. 2C). Four-connected core and border points will be regarded as cluster candidates. Individual, blinking or relatively photostable fluorophores would generate many localizations, and would not be removed by the previous threshold. Repeated localizations from such a fluorophore will contribute to bins within a 3×3 array centered at its position. Therefore, to reject these pseudo-clusters, only clusters that have a minimum cluster size of 3×3 plus 1 bins are qualified as a protein cluster (Fig. 2D).

The only parameter that needs to be optimized in FOCAL is the density threshold minL. This optimization is achieved by filtering out low density, out of focus clusters. To understand how this works, consider a single, blinking fluorophore that is localized multiple times. The density of localizations will reduce considerably if the fluorphore is defocused. Even a few hundred nm defocusing increases the area of the point spread function by a factor of between 2-10 (DeSantis *et al.*, 2012). The localization uncertainty ($\sigma_{x,y}$) for each blink ranges over a distribution of values (Fig. 3A shows an example distribution for different photon thresholds). While the mean and the spread of the distribution decrease for increasing photon number, because this is a statistical property, a filter on intensity and localization uncertainty is not a perfect discriminator of acceptable localizations (Fig. 3B). When reconstructing an out of focus cluster, many of the localizations originating from the cluster will be rejected either because they emit below the photon threshold or because the width of the localization is too wide. However, statistically, some of the localizations will pass through this filter despite being out of focus, resulting in clusters that are more dilute than those detected in focus. By tuning the density threshold minL, we are able to filter out the lower density clusters, which should result in an overall increase in the localization uncertainty of the acquired image.

With this discussion in mind, our strategy for optimizing the density threshold minL can be depicted as in Figure 4. FOCAL is used to cluster the SMLM data with a low-initial value for minL. All localizations from pixels with densities above or below minL are disregarded. The localization uncertainty is then directly calculated from the filtered data by a temporal adjacent neighbor analysis, as discussed earlier, and the process is repeated at increasing values of minL. Figure 4 shows a monotonic decrease in the localization uncertainty as minL increases. As expected, increasing minL filters out clusters with a lower density of localizations, significantly reducing the overall localization uncertainty of the remaining clustered population. However, setting too high a density threshold will needlessly reject clusters without significant gains in localization uncertainty. A plot of the localization uncertainty versus minL, as in Figure 4, serves as a useful guide for choosing the density threshold minL. While there remains some flexibility, a reasonable and easily automated way to identify an optimum density threshold, minL*, is to select the point at which the localization uncertainty is one standard deviation away from the asymptote of an exponential fit to the curve in Figure 4 (minL* $\sim$ 14 in this example). Note that unlike the inflection point in the *k*-distance plots for DBSCAN, this procedure yields a value for minL that is well constrained and can be rapidly calculated.

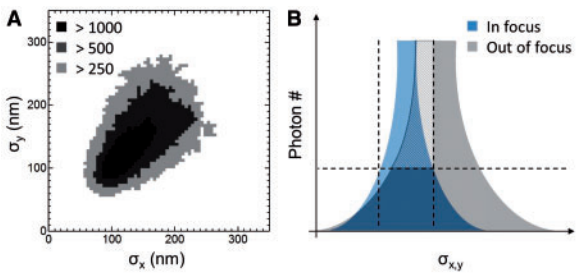

Fig. 3. Sample distributions of localization uncertainties ($\sigma_{x,y}$) for increasing thresholds: 250, 500, and 1000 photons (A). The distribution of $\sigma_{x,y}$ for in focus fluorophores displays a smaller spread, and shifted mean, compared to out of focus localizations. A threshold on photon number and $\sigma_{x,y}$ (dashed lines) still results in an accepted population of out of focus localizations (shaded region) (B) (Color version of this figure is available at *Bioinformatics* online.)

3 Results

3.1 dSTORM microscopy of transcription factories

Transcription factories (TFs) are small (45–100 nm) assemblies of active RNA polymerase II (RNAP II) holoenzymes that form discrete clusters in the eukaryotic cell nucleus (Sutherland and Bickmore, 2009). The number of TFs is estimated from a few hundred to several thousand, dependent on the cell, although these numbers are still a matter of significant debate. Recently, we analyzed the clustering of Alexa-647 immunolabeled RNAP II within the nucleus of several cell types via dSTORM (Davidson *et al.*, in process). These images, acquired deep within the cell, suffer from nonspecific labeling and abundant out of focus fluorescence. A typical SMLM reconstruction is provided in Figure 5A, which is of the RNAP II distribution within a mouse cortex cell (see Supplementary information).

If we heuristically set the parameters in DBSCAN as described in Section 2.1, (i.e. $\varepsilon = 30$ is approximated from the *k*-distance graph of Figure 1 and minPts $= 2 + 1$), the algorithm identifies almost all localizations as members of 1924 (pseudo-) clusters (Fig. 5B). However, a visual inspection shows that FOCAL better rejects artifacts in the image to identify only 377 clusters with an average

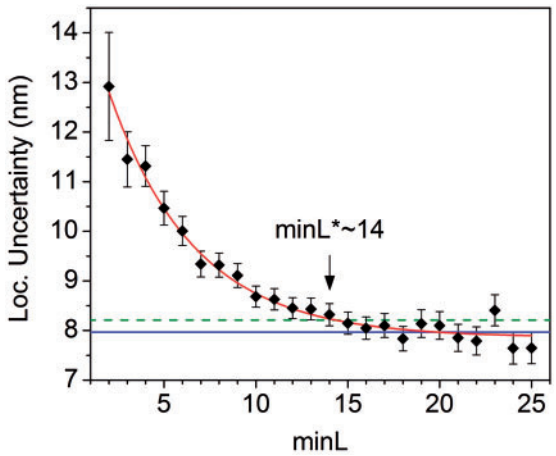

Fig. 4. Average localization uncertainty versus minL. The data points are displayed with error bars +/- std. dev. The solid curve (red) is an exponential fit to the data and the solid line (blue) is a fit to the asymptote. The chosen value, minL*$\sim$14, is taken at the point where the fit is one standard deviation (dashed line) from the asymptote (Color version of this figure is available at *Bioinformatics* online.)

diameter of 47 ± 30 nm (Fig. 5E). For this image, minL was set to 14 from an analysis of the localization uncertainty, which was in the range of ~8–10 nm (we set $\sigma = 10$ nm) (see Fig. 4).

Because FOCAL is similar to a grid-based version of DBSCAN, we can use the optimized FOCAL parameters to estimate a comparable set of parameters for DBSCAN. The area of the 3×3 arrays used to construct a density map in FOCAL can roughly be equated to the circular neighborhood in DBSCAN (i.e. $\varepsilon = 3\sigma/\pi^{1/2}$). Similarly, both minL in FOCAL and minPts in DBSCAN set a density threshold. Because DBSCAN excludes each target point under consideration, we set $\text{minPts} = \text{minL}^{-1}$. From these considerations, we adjusted DBSCAN's parameters when analyzing this image to $\varepsilon = 17$ and minPts = 13. DBSCAN now effectively rejects artifacts and considerably reduces the number of detected clusters to 718 (Fig. 5D). This is still ~90% percent more clusters than were detected by FOCAL (these results are summarized in Table 1). Upon closer inspection (see Supplementary Fig. S5), almost all the additional clusters detected by DBSCAN were smaller than the 10 pixel grid minimum imposed in FOCAL, so most likely arise from single labels. Alternatively, these small foci could be tightly condensed clusters, but at this scale a spatial analysis is limited by the resolution of the microscopy.

3.2 PALM imaging of H-NS proteins

To further test the clustering capabilities of FOCAL, we analyzed PALM images of the histone-like nucleosomal (H-NS) protein, fused to the fluorescent protein mEOS3.2, in fixed *E. coli*. H-NS acts as both a global regulator of gene expression and DNA condensation in enterobacteria. Experiments suggest that H-NS can repeatedly dimerize to form extended filaments, which may additionally condense DNA through bridging of adjacent strands (Ali *et al.*, 2012; Dame, 2005).

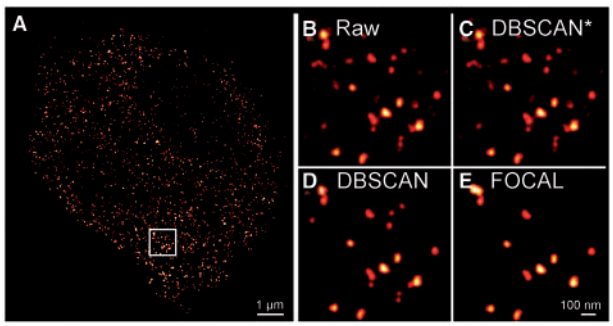

Fig. 5. dSTORM localization image of Alexa-647 immunolabeled RNAP II in a mouse cortex cell (A). A 1 μm square inset is presented: without clustering (B), heuristic DBSCAN* ($\varepsilon = 30$ nm and minPts = 3) (C), optimized DBSCAN ($\varepsilon = 17$ nm and minPts = 13) (D) and FOCAL (minL = 14) (E). Each pixel is 10 nm square and the intensities represent the number of localizations within each pixel (Color version of this figure is available at *Bioinformatics* online.)

Table 1. Tabulated cluster analysis results for RNAPII and H-NS localization imaging: number of clusters and cluster size. Values are presented for heuristic DBSCAN*, optimized DBSCAN and FOCAL

	RNAP II		H-NS	
	# Clusters	Diameter (nm)	# Clusters	Diameter (nm)
DBSCAN*	1924	53 ± 65	194	185 ± 367
DBSCAN	718	38 ± 33	213	39 ± 36
FOCAL	377	47 ± 30	111	47 ± 34

Because bacteria have a spatial depth of approximately ~1 μm, the cells could be imaged by wide-field epifluorescence, and display significantly less background than we encountered imaging RNAPII within the nucleus of a cortex cell. Moreover, since we are imaging fluorescent proteins, the reconstructions do not suffer from nonspecific labeling artifacts that immunolabeling is prone to.

H-NS should be found within the cell as both a monomeric unit and in clusters of varying sizes, all of which should generate localizations above or below the focal plane. Figure 6B shows that a heuristic choice for DBSCAN's parameters ($\varepsilon = 30$, minPts = 2 + 1) accepts virtually every localization as a member of a cluster, so the raw localization image pretty much corresponds to the DBSCAN* filtered image. The parametrically optimized DBSCAN ($\varepsilon = 17$, minPts = 14) significantly filters low density localizations as does FOCAL (minL* = 15, see Supplementary Fig. S3) implying that most of the clusters are either out of focus and/or originate from monomers of H-NS. Again, a comparison between Figure 6C and D shows that both DBSCAN and FOCAL agree on a significant fraction of the detected clusters. FOCAL, however, filters out a number of smaller clusters that DBSCAN accepts since these clusters cannot be distinguished from multiple localizations of a single fluorophore (these results are summarized in Table 1).

3.3 Simulations

We next performed a series of simulations to compare DBSCAN and FOCAL in regards to suppression of off-target localizations (from noise and unclustered, single-fluorophores), cluster identification and the preservation of cluster characteristics. The parameters for these algorithms have been optimized based upon the dSTORM images in the previous section, which we take as representative examples of SMLM data.

We begin with an assessment of false clustering by generating random sets of localizations at increasing densities, within a 10 μm square region, and applying both clustering algorithms (Fig. 7).

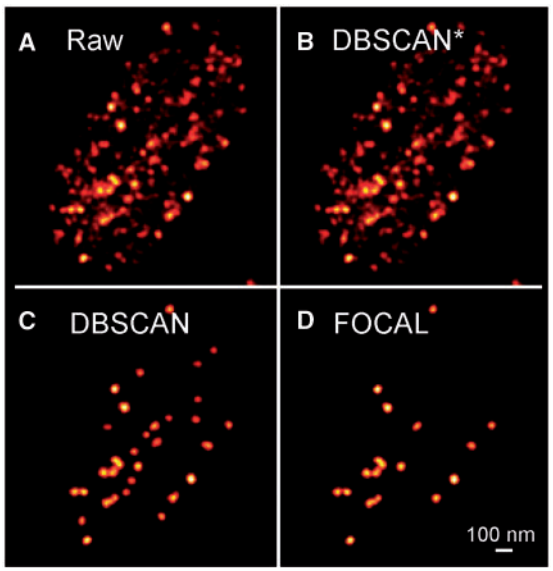

Fig. 6. PALM localization image of H-NS/mEOS3.2 in fixed E. coli presented: without clustering (A), heuristic DBSCAN* ($\varepsilon = 30$ nm and minPts = 3) (B), optimized DBSCAN ($\varepsilon = 17$ nm and minPts = 14) (C) and FOCAL (minL = 15) (D). Each pixel is 10 nm square. The intensities represent the number of localizations within each pixel (Color version of this figure is available at *Bioinformatics* online.)

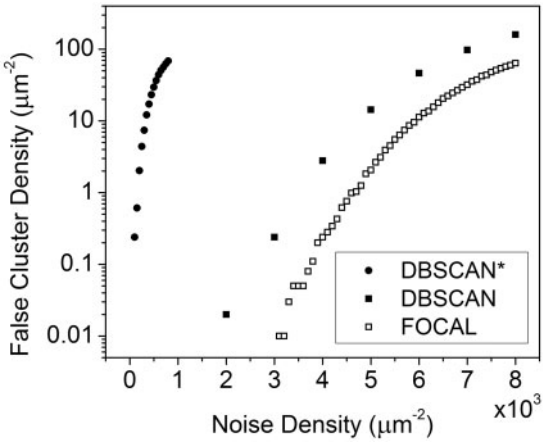

Fig. 7. Density of false clustering as a function of noise density for complete random noise. (Filled Circles) DBSCAN*: optimized via heuristics ($\varepsilon = 30$ nm, minPts = 3). (Filled Squares) DBSCAN: empirically optimized to minimize noise ($\varepsilon = 17$ nm, minPts = 11). (Open Squares) FOCAL: (minL = 12)

Assuming that the off-target localizations we observe in our dSTORM experiments are completely random, we can estimate an empirical density of $\sim 0.2 \times 10^3$ localizations per μm^2. The imaging attained a localization precision of ~10 nm, and the optimum minL fell in the range of 12–18. In the following, since we must pick a parameter set to analyze, we have arbitrarily set $\sigma = 10$ nm and minL = 12, but this choice should not affect our qualitative results. On average, this density gives rise to ~2 false clusters per μm^2 using DBSCAN, with the parameters heuristically set as before ($\varepsilon = 30$ and minPts = 3). Mapping FOCAL's parameters onto DBSCAN, as in the previous section, results in minPts = 11 and $\varepsilon = 17$. These parameters eliminate much of the false clustering, but the gains rapidly disappear as the off-target density is increased. In what follows, we'll employ this optimized set of parameters for DBSCAN.

Now consider the performance of FOCAL on the same simulated random data. Figure 7 shows that FOCAL performs on par with DBSCAN; both are able to reject false clusters up to a density of approximately 4×10^3 localizations per μm^2. At higher densities, however, both algorithms become unreliable. In practice, since the off-target localizations are not truly randomly distributed, local density fluctuations could lead to significantly more false clustering (Nan *et al.*, 2013).

We next introduce a set number of circular clusters among a random background to assess how well the algorithms perform at determining the number and size of the clusters (see Fig. 8). 200 clusters, each composed of 64 localizations scattered over a disk 50 nm in diameter, were distributed in a 10 μm square region (see Supplementary Fig. S7 for a schematic). Again, these numbers were chosen to resemble common experimental conditions. In order to introduce blinking while maintaining a set number of localizations in each cluster, we assumed that each cluster results from 16 uniformly scattered labels and that each label blinks 4 times. The number of blinks was estimated from the ratio of the acquisition time and the fluorophore off-time, the latter of which is calculated from the ratio of the average blink on-time (2.3 frames) and the duty cycle ($\sim 5 \times 10^{-4}$) in our dSTORM measurements (Dempsey *et al.*, 2011).

The position of each localization was found from a normal probability density centered at each label with a standard deviation equal to the localization uncertainty ($\sigma = 10$ nm). At this localization density, both algorithms accurately detect all clusters at low off-target densities. However, as expected from our results in Figure 7,

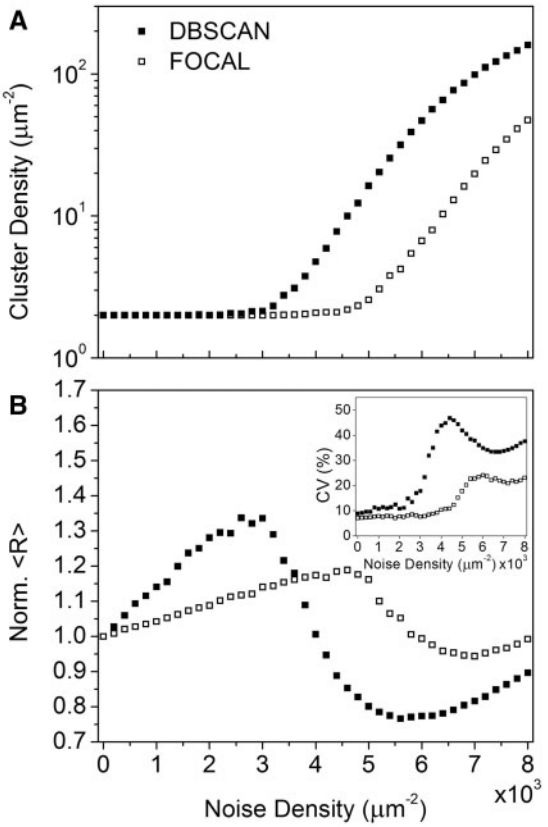

Fig. 8. The detected density of clusters versus noise density (A) and the normalized average cluster radius versus noise density (B). The inset in (B) shows the coefficient of variation of the cluster radii versus noise density

FOCAL and DBSCAN are able to discern the actual clusters from the random background up to densities of $\sim 4 \times 10^3$ localizations per μm^2 (Fig. 8A).

At low to moderate off-target densities, where both algorithms accurately identify the clusters, DBSCAN tends to overestimate the size of the clusters in comparison to FOCAL (Fig. 8B). The size of each cluster is characterized by a radius $R = \sqrt{A/\pi}$, where A is the area of the cluster. For FOCAL, this is simply the area of all grid pixels associated with a given cluster, while for DBSCAN, the area is taken as the convex hull of all localizations within a cluster. In this simulation, at a density of 1×10^3 localizations per μm^2, which is what we observed in our dSTORM experiments, the predicted average cluster size is inflated by 11% in DBSCAN compared to less than 4% in FOCAL. Moreover, the insert of Figure 8B indicates that FOCAL typically detects a more uniform spread of cluster sizes, closer to the simulated population, than DBSCAN. This can be seen in the coefficient of variation $CV = \sigma_R / \langle R \rangle$, where σ_R is the standard deviation in cluster radii and $\langle R \rangle$ is the average detected radius As the off-target density increases further, both algorithms progressively incorporate off-target localizations into the clusters. This initially results in an inflation of the average detected cluster size, but eventually both algorithms start identifying numerous, small pseudo-clusters swinging the average cluster size back down.

Finally, we investigated the performance of both techniques in preserving cluster characteristics in the absence of off-target localizations. Toward this end, we simulated ring shaped clusters of different labeling densities, with inner radii of 60 nm and outer radii varying from 60 to 90 nm (see Supplementary Fig. S8 for a schematic). Labels were distributed uniformly over the rings and, as

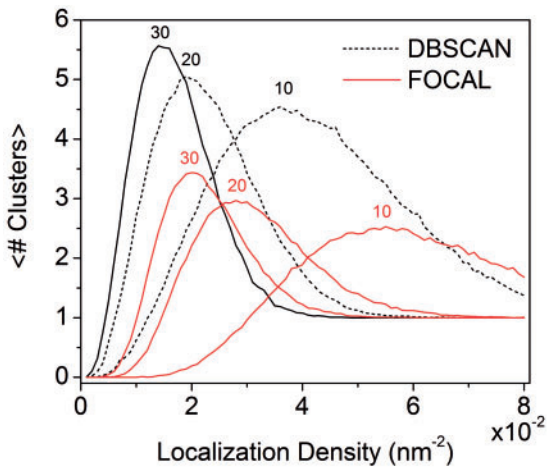

Fig. 9. Performance of FOCAL and DBSCAN in clustering of ring shaped clusters with various thicknesses as indicated (in nm). The figure plots the mean number of clusters detected in each ring shaped cluster as a function of the localization density (Color version of this figure is available at *Bioinformatics* online.)

before, each label gave rise to 4 localizations (blinks) drawn from a normal probability density. At each density, this process was repeated multiple times to obtain the average number of detected clusters. When the localization density (i.e. the number of localizations per unit area) was very low, both techniques failed to detect the clusters while at high enough densities, the full ring clusters could be identified (Fig. 9). However, at intermediate densities, the rings become segmented. Because DBSCAN clusters are not restricted in their size, DBSCAN tended to detect many more component clusters within each ring as compared to FOCAL (Fig. 9). On the other hand, at intermediate densities (i.e. 2–2.5 nm^{-2} for 30 nm thick rings), FOCAL was slightly better at merging the components into a single, larger cluster (Supplementary Fig. S8). Therefore, at intermediate localization densities, FOCAL again seems to preserve the cluster shape better that DBSCAN.

4 Conclusion

Various clustering methods are used to characterize the organization of nanoscale assemblies in localization microscopy images. Unlike statistical approaches, such as Ripley's L-function and pair-correlation analysis that estimate ensemble parameters (Endesfelder *et al.*, 2013), DBSCAN and other derivative density-based clustering algorithms are able to identify and quantify individual clusters. However, DBSCAN is slow, scaling with the number of particles like $O(n \log(n))$ at best; is acutely sensitive to the algorithm's parameters, which are typically selected by visual inspection of the output; and is prone to detecting a large number of false clusters in the presence of a significant background of localizations. To address these problems we developed FOCAL, a rapid $O(n)$, density-based clustering method that is tailored for localization microscopy.

FOCAL is a rapid and efficient tool for density based clustering in super-resolution microscopy data. In contrast to DBSCAN, FOCAL sets a threshold for a minimum cluster size, based upon the image localization uncertainty, to remove pseudo-clusters arising from multiple localizations of single fluorophores. Moreover, FOCAL can be rapidly optimized to minimize false clustering and remove diffuse, out of focus clusters while maintaining cluster integrity. Alternatively, FOCAL could serve as a method for quickly determining the optimum density threshold for a DBSCAN based clustering analysis.

As SMLM becomes more commonplace, increasingly efficient algorithms will need to be developed to deal with the large amounts of data generated when imaging at such fine resolution. FOCAL can rapidly analyze large SMLM data sets, performing at least on par with, if not better than, DBSCAN in most situations.

Its efficiency in analyzing large image data sets should make FOCAL a suitable replacement to the much slower DBSCAN derived algorithms, and a solid foundation for further clustering analysis.

Acknowledgements

We thank Jennifer Mitchell and Navroop Dhaliwal for providing the labeled mouse cortex cells, and Will Navarre and Hazel Soto-Montoya for cells containing mEOS3.2 fused H-NS. We also thank Yong Wang for providing the data on H-NS and for feedback on the manuscript.

Funding

This work was supported by The Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN:418251-13).

Conflict of Interest: none declared.

References

Ali,S.S. *et al.* (2012) Silencing of foreign DNA in bacteria. *Curr. Opin. Microbiol.*, **15**, 175–181.

Annibale,P. *et al.* (2010) Photoactivatable fluorescent protein mEos2 displays repeated photoactivation after a long-lived dark state in the red photoconverted form. *J. Phys. Chem. Lett.*, **1**, 1506–1510.

Bar-On,D. *et al.* (2012) Super-resolution imaging reveals the internal architecture of nano-sized syntaxin clusters. *J. Biol. Chem.*, **287**, 27158–27167.

Betzig,E. *et al.* (2006) Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, **313**, 1642–1645.

Burnette,D.T. *et al.* (2011) Bleaching/blinking assisted localization microscopy for superresolution imaging using standard fluorescent molecules. *Proc. Natl. Acad. Sci.*, **108**, 21081–21086.

Cattoni,D.I. *et al.* (2013) Super-resolution imaging of bacteria in a microfluidics device. *PLoS One*, **8**, e76268.

Cella Zanacchi,F. *et al.* (2011) Live-cell 3D super-resolution imaging in thick biological samples. *Nat. Methods*, **8**, 1047–1049.

Dame,R.T. (2005) The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Mol. Microbiol.*, **56**, 858–870.

Darong,H. and Peng,W. (2012) Grid-based DBSCAN Algorithm with Referential Parameters. *Phys. Procedia*, **24**, 1166–1170.

Dempsey,G.T. *et al.* (2011) Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. *Nat. Methods*, **8**, 1027–1036.

DeSantis,M.C. *et al.* (2012) Single-image axial localization precision analysis for individual fluorophores. *Opt. Express*, **20**, 3057–3065.

Deschout,H. *et al.* (2014) Progress in quantitative single-molecule localization microscopy. *Histochem. Cell Biol.*, **142**, 5–17.

Dudok,B. *et al.* (2014) Cell-specific STORM super-resolution imaging reveals nanoscale organization of cannabinoid signaling. *Nat. Neurosci.*, **18**, 75–86.

Endesfelder,U. *et al.* (2014) A simple method to estimate the average localization precision of a single-molecule localization microscopy experiment. *Histochem. Cell Biol.*, **141**, 629–638.

Endesfelder,U. *et al.* (2013) Multiscale spatial organization of RNA polymerase in *Escherichia coli*. *Biophys. J.*, **105**, 172–181.

Ester,M. *et al.* (1996) A Density-based algorithm for discovering clusters in large spatial databases with noise. *Second Int. Conf. Knowl. Discov. Data Min.*, **96**, 226–231.

Gebhardt,J.C.M. *et al*. (2013) Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. *Nat. Methods*, **10**, 421–426.

Ghamari,A. *et al*. (2013) In vivo live imaging of RNA polymerase II transcription factories in primary cells. *Genes Dev.*, **27**, 1434.

Haas,B.L. *et al*. (2014) Imaging live cells at the nanometer-scale with single-molecule microscopy: obstacles and achievements in experiment optimization for microbiology. *Molecules*, **19**, 12116–12149.

Heilemann,M. *et al*. (2008) Subdiffraction-resolution fluorescence imaging with conventional fluorescent probes. *Angew. Chem. Int. Ed. Engl.*, **47**, 6172–6176.

Hu,Y.S. *et al*. (2013) Light-sheet Bayesian microscopy enables deep-cell super-resolution imaging of heterochromatin in live human embryonic stem cells. *Opt. Nanoscopy*, **2**, 7.

Itano,M.S. *et al*. (2014) Super-resolution imaging of C-type lectin spatial rearrangement within the dendritic cell plasma membrane at fungal microbe contact sites. *Front. Phys.*, **2**, Article: 46.

Kłopotek,M.A. *et al*. (eds.) (2006) *Intelligent Information Processing and Web Mining*. Berlin/Heidelberg: Springer-Verlag.

Nan,X. *et al*. (2013) Single-molecule superresolution imaging allows quantitative analysis of RAF multimer formation and signaling. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 18519–18524.

Pertsinidis,A. *et al*. (2013) Ultrahigh-resolution imaging reveals formation of neuronal SNARE/Munc18 complexes in situ. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, E2812–E2820.

Ricci,M.A. *et al*. (2015) Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell*, **160**, 1145–1158.

Rust,M.J. *et al*. (2006) Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods*, **3**, 793–795.

Sengupta,P. *et al*. (2013) Quantifying spatial organization in point-localization superresolution images using pair correlation analysis. *Nat. Protoc.*, **8**, 345–354.

Sutherland,H. and Bickmore,W.A. (2009) Transcription factories: gene expression in unions? *Nat. Rev. Genet.*, **10**, 457–466.

Tokunaga,M. *et al*. (2008) Highly inclined thin illumination enables clear single-molecule imaging in cells. *Nat. Methods*, **5**, 159–161.