IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 22, 2019, accepted October 13, 2019, date of publication November 4, 2019,
date of current version November 14, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2950984

Morphology Clustering Software for AFM Images, Based on Particle Isolation and Artificial Neural Networks

SOLEDAD DELGADO[1], MIGUEL MORENO[2], LUIS F. VÁZQUEZ[3], JOSE ÁNGEL MARTÍNGAGO[3], AND CARLOS BRIONES[2,4]

[1]Department of Computer Systems, Universidad Politécnica de Madrid, 28031 Madrid, Spain
[2]Department of Molecular Evolution, Centro de Astrobiología (CSIC-INTA), 28850 Madrid, Spain
[3]Materials Science Factory, Instituto de Ciencia de Materiales de Madrid (CSIC), 28049 Madrid, Spain
[4]Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), 28029 Madrid, Spain

Corresponding authors: Soledad Delgado (sole@etsisi.upm.es) and Carlos Briones (cbriones@cab.inta-csic.es)

This work was supported in part by the Spanish Ministry of Economy and Competitiveness (MINECO) funded by the EU through the FEDER Programme under Grant BIO2016-79618-R and Grant MAT2017-85089-C2-1-R, in part by the Spanish State Research Agency (AEI) through the Unidad de Excelencia María de Maeztu-Centro de Astrobiología (CSIC-INTA) under Project MDM-2017-0737, and in part by the Comunidad de Madrid under Grant S2018/NMT-4349.

ABSTRACT Advanced microscopy techniques currently allow scientists to visualize biomolecules at high resolution. Among them, atomic force microscopy (AFM) shows the advantage of imaging molecules in their native state, without requiring any staining or coating of the sample. Biopolymers, including proteins and structured nucleic acids, are flexible molecules that can fold into alternative conformations for any given monomer sequence, as exemplified by the different three-dimensional structures adopted by RNA in solution. Therefore, the manual analysis of images visualized by AFM and other microscopy techniques becomes very laborious and time-consuming (and may also be inadvertently biased) when large populations of biomolecules are studied. Here we present a novel morphology clustering software, based on particle isolation and artificial neural networks, which allows the automatic image analysis and classification of biomolecules that can show alternative conformations. It has been tested with a set of AFM images of RNA molecules (a 574 nucleotides-long functinal region of the hepatitis C virus genome that contains its internal ribosome entry site element) structured in folding buffers containing 0, 2, 4, 6 or 10 mM Mg^{2+}. The developed software shows a broad applicability in the microscopy-based analysis of biopolymers and other complex biomolecules.

INDEX TERMS Artificial neural networks, atomic force microscopy (AFM), biomolecules, growing cell structures (GCS), hepatitis C virus (HCV), Image analysis, internal ribosome entry site (IRES), ribonucleic acid (RNA), self-organizing maps (SOM).

I. INTRODUCTION

Microscopy techniques have undergone a steep advance in recent years, allowing ultra-high resolution imaging and three-dimensional reconstruction of the imaged features [1]–[4]. A number of modes of operation and set-ups have been applied to organic and inorganic samples under different environments, in which one of the final goals is to discern their morphological features [5]–[7]. In this sense, distinct free software applications for image analysis have became popular, such as ImageJ [8], [9], WsxM [10], and Gwyddion [11]. They are valuable tools for handling microscopy images, and the last two of them have proven especially useful in the field of atomic force microscopy (AFM). AFM is a type of scanning probe microscopy that allows the structural analysis of a wide range of materials (including biomolecules) at nanometre resolution, and provides a 3D surface profile of the imaged sample without requiring its previous staining or coating [12], [13].

Unveiling the morphology of the imaged features is a challenging task when biological specimens (biopolymers such as proteins or structured folded nucleic acids, macromolecular

The associate editor coordinating the review of this manuscript and approving it for publication was Eduardo Rosa-Molinar.

assemblies, viruses or cells) are involved [14], [15]. This is due to the fact that the organic biomaterial shows, in general, much higher flexibility and plasticity than inorganic one, and also because it can interact with the substrate onto which it is adsorbed. Thus, the conformation of the biomolecules can be distorted when the sample is imaged under ambient (or vaccum) conditions that require a drying process. Additionally, depending on the type of interaction that governs the immobilization of the sample onto the substrate, many geometries are possible for a given biomolecule or macromolecular assembly. For instance, many sites at the surface of the biological material are, regardless of their geometrical distribution, avalaible for the unspecific adsorption process. In turn, if specific immobilization takes place through a directional and selective bonding, the geometry of the deposited molecules will depend on the distribution of the active sites on the substrate surface. Additionally, the buffer used to resuspend the biomolecules (in particular, its pH, ionic strength and concentration of divalent cations) as well as the immobilization conditions of the biomolecules (including variables such as temperature and time), can make them adopt different conformations. Therefore, in any apparentely homogeneous biological sample a certain degree of diversity in the imaged molecular conformations is expected. Due to the fact that valuable biological information can be retrieved from the imaged structure, it is currently required to develop computational methods to analyze in detail the microscopy data of biomolecules [16], [17]. Such a scenario is even more critical when imaging biomolecules using AFM. This is so because the dimensions of the imaged structures, typically of a few nanometers, are of the same size or even smaller than the AFM tip diameter and, therefore, the images show a convolution of both. AFM tips could also present irregular shapes that vary from one to another. Moreover, in many cases the substrate is chemically functionalized and, as a result, its own roughness and morphological features could be misintrepreted by the user as the immobilized biomolecule.

This is especially relevant in the case of RNA imaging, as this nucleic acid is highly flexible and can adopt multiple three-dimensional conformations in solution [18]. Moreover, the sequence-structure relationship affects the activity of functional RNA molecules such as ribozymes, aptamers, riboswitches, viroids or functional regions of viral RNA genomes. Indeed, RNA molecules assemble into tertiary structures and form globular conformations in solution that are stabilized by networks of RNA-RNA interactions. In the cells, folded RNAs are then recognized by different ligands (including Mg^{2+} and other cations, low molecular weight organic molecules, other RNA molecules or RNA-binding proteins) in such a way that the patterns of interactions have a direct effect on the cellular metabolism, the regulation of the flow of genetic information, or the viral replication cycles [19], [20]. Therefore, a high-resolution structural analysis of populations of RNA molecules is required in different fields of biochemistry and molecular biology, and AFM shows the advantage over other microscopy techniques of imaging RNA in native conditions [12], [13], [21].

We have previously analyzed two different kinds of functional RNA molecules using AFM. The first of them was the 5' untranslated region (5'UTR) of hepatitis C virus (HCV) genomic RNA, which contains an internal ribosome entry site (IRES) element required for cap-independent initiation of the viral genome translation. We studied the RNA conformations present in a number of prepararions of this 574 nt long molecule at different ionic conditions. Our results showed that HCV IRES element switches between two alternative conformations: from an 'open' and elongated one at 0–2 mM Mg^{2+} concentration to a 'closed' and comma-shaped morphology at 4–6 mM Mg^{2+} [22]. Subsequently, we analyzed viroids by AFM. Viroids are short (typically, 250-430 nt long) infectious and non-protein-coding circular RNAs that replicate independently in plants. Three different species of viroids belonging to the families *Pospiviroidae* and *Avsunviroidae* were imaged at single molecule resolution. Our AFM analysis revealed their compact, rod-like or spoon-shaped three-dimensional conformations at 0 and 4 mM Mg^{2+}, and has evidenced the role played by some elements of RNA tertiary structure in the structural stabilization of viroids [23].

Both studies have involved the manual image analysis of hundreds of RNA molecules in different preparations and, therefore have evidenced the need for new computational methods to automatically isolate, analyze and group the individual morphologies and shapes present in the samples [24]. Within this context, the goal of the present work is to develop a dedicated software that should be able to routinely isolate and analyze the different structures of biomolecules imaged by microscopy techniques, obtain the metrics of each one and cluster them in groups or types based on their morphological features. Though the proposed software shows a broad applicability in different fields of microscopy, based on our previous experience it has been developed using a set of AFM images of RNA molecules, in particular those of the HCV IRES element structured in folding buffers containing 0, 2, 4, 6 or 10 mM Mg^{2+} [22].

The developed method consists of two phases: isolation of the individual images of the molecules in each sample, and morphology-based clustering of the isolated particles. The automation of the molecule isolation process shows a number of advantages with respect to the manual method, as previously highlighted by other Sánchez and Wyman [24]: i) the time needed to identify and obtain the particles is considerably reduced, as manual procedure is very time consuming even for an expert user; ii) the number of retrieved particles is highly increased, as manual procedure is poorly effective with this regard; iii) the feature isolation is neither subject to any operator bias or subjective selection, nor affected by the performance of the human eye (with possible defects in vision), the type of graphics used (palette of colors, resolution and constrast of the image, etc) and the quality of the computer screen; iv) the visualization of the isolated particles is improved by producing individual image files for

each one, where only the pixels of the molecule are included (without any noise around it), and the further generation of a single image file with all the isolated particles provides the user an immediate visualization of the set of feature morphologies present in the processed image; v) the automatic molecule isolation offers a broad compatibility with AFM images of different sizes and resolutions, which might have been obtained using the main microscopes currently in use and saved in a number of formats, provided that the basic parameters of window size and total height interval are known.

Once the isolation of the individual images of the molecules has been conducted, clustering techniques can be applied to automate the grouping of the morphologies detected in the original image. The clustering of 2D or 3D morphologies by a human expert can lead to erroneous results, especially when the imaged particles have not been previously filtered or are very noisy. In turn, the unsupervised clustering techniques can analyze the types of morphologies in an automatic and unbiased way, also facilitating the visualization of the prototypes of molecule morphologies for each cluster. In the field of image analysis, different automatic clustering methods have been used to group the imaged features regarding their shape, geometry and/or dimensions [25], [26]. Among them, Self-Organizing Maps (SOM) are models of artificial neural networks that have been widely applied to clustering tasks and visual exploratory data analysis [27], [28]. The architecture of the SOM network consists of an input layer that contains as many neurons as the number of features or dimension defines the input space, and an output layer composed of neurons arranged in a low-dimensional topology (usually bi-dimensional) ordered lattice that establishes the connections of neighborhood between the output neurons. In addition, each output neuron has an associated vector (known as 'synaptic vector') with the same dimension and nature than the input space [27]. The operation of the SOM model is very simple: when an input vector is presented to the network, it is distributed among all the output neurons, in such a way that only one of them (the best matching unit, *bmu*, which shows the lowest Euclidean distance between its synaptic vector and the input vector) is activated. Thus, the SOM model produces a nonlinear projection of a high-dimensional input space onto the low-dimensional output mesh. The training algorithm defines the process that creates such a mapping, which is a competitive and unsupervised method that adjusts the synaptic vectors of the output neurons: i) a training dataset is presented iteratively to the network; ii) for each training vector, the *bmu* is calculated and its synaptic vector is modified to bring it closer to the processed pattern; iii) similarly, the synaptic vectors of the neurons that fall in a neighborhood area of the *bmu* are also modified; iv) the neighborhood area is a dynamic factor whose size decreases as the training process of the network advances. In this way, the main objective of the SOM training algorithm is that similar input vectors keep mapped by nearby neurons just as, conversely, neighboring neurons in the map represent only similar input data. Indeed, this is an informal definition of topology preservation [29]. Due to the fact that SOM learning algorithm has a non-deterministic nature (i.e., using the same input data set and configuration of the training parameters, it can produce different values in the synaptic vectors of the neurons on different trainings), the topology preservation factor allows measuring the quality of a SOM network after training, as well as the degree of success in the initial configuration of the network architecture (i.e., number of neurons in the output layer and arrangement of the neighborhood connection).

Usually, the SOM network is trained with a number of input vectors higher than that of the output neurons; hence, the synaptic vectors can be considered as a reduced set of prototype vectors of the input space. In addition, based on the nature of the SOM algorithm, the synaptic vectors are ordered by similarity in the grid of neurons of the output layer. In the first SOM model developed, known as Kohonen's SOM [27], the output neurons were organized in a two-dimensional grid that could be projected in the plane. This feature made the SOM model a powerful tool for graphically analyzing multidimensional data properties [30]. Despite the good performance of Kohonen's SOM, this model exhibits several drawbacks. First, the complete architecture of the network must be fully configured before executing the SOM algorithm. Specifically, the two architectural factors to be highlighted are the number of output neurons and the neighborhood connection topology, which remain constant throughout the entire life cycle of the SOM. This requires having a prior knowledge of the features of the input space in order to properly design the architecture of the network, which is not always possible. Another drawback is related to the compact connection topology of the output mesh, which links all the output neurons via neighborhood connections. For probability distributions where the input patterns are located in several separate regions with a probability density greater than zero, the Kohonen's SOM network may obtain poor values of topology preservation, since some output neurons may be positioned in areas of the input space with low or null probability density [31]. These units are known as 'interpolating neurons', suitable for identifying the boundaries between clusters in the U-matrix map [32], but undesirable for using the SOM as a clustering tool. Furthermore, in the SOM model several synaptic vectors identify a single cluster of input data, in contrast to k-means where each cluster of data is characterized by a single centroid. Due to the compactness of the neighborhood connection topology of Kohonen's SOM, once the network training is completed similar synaptic vectors must be combined to identify the natural clusters underlying the input space [33].

Alternative SOM models have been developed to address these drawbacks, such as Incremental Grid Growing (IGG) [34], Growing Neural Gas (GNG) [35], or Growing Cell Structures (GCS) [36]. These dynamic SOM models share a common feature: they create a basic structure in the

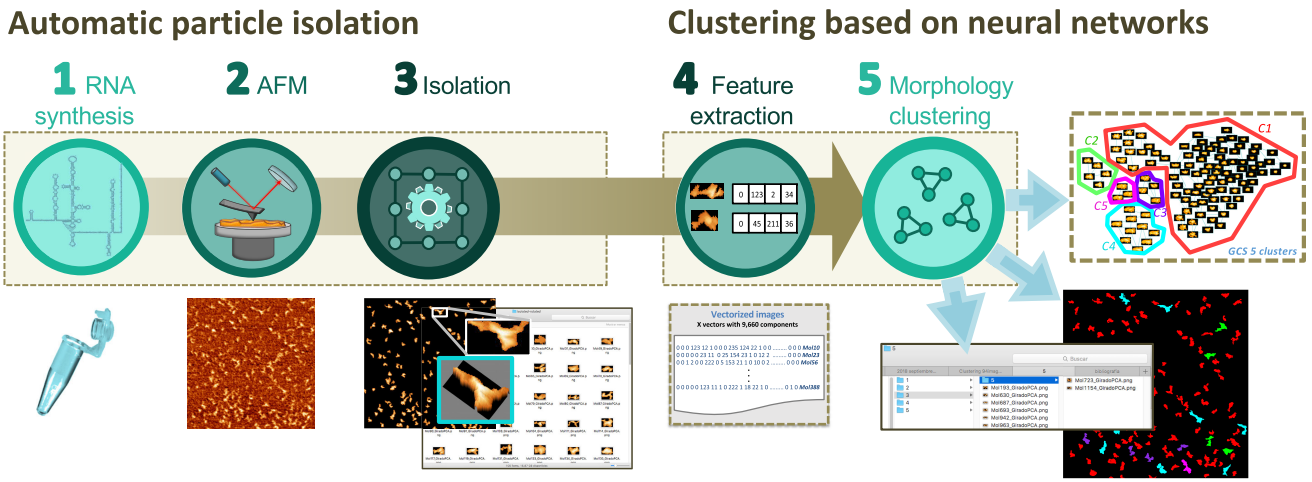

FIGURE 1. **Methodology for the proposed work.**

SOM output layer, and both neurons and neighborhood connections are added or removed during SOM training. In the IGG model, the basic structure is formed by four neurons arranged in a square grid, and new neurons are incrementally added to the regular bi-dimensional grid, thus ensuring that the output layer is drawable. However, this model limits the maximum number of neighboring neurons to four, which can negatively affect the topology preservation for certain probability distributions of the input space. Regarding GNG, its basic structure is formed by two neurons and, during the training process, neurons are added and removed aiming at adapting the dimension of the network topology to the unknown dimension of the input data. As a result, this model is appropriate for clustering and vector quantization tasks. Nevertheless, GNG does not have an output topology with a fixed dimension and thus it cannot be used for visual exploratory data analysis. Finally, in the GCS model, the output layer of the network is formed by basic k-dimensional structures, where k is a factor that can be arbitrarily chosen. Initially, the network consists of a single k-dimensional structure. The insertion and removal of neurons during the training process ensures the topology of basic k-dimensional structures in the output layer of the network. The aim of neuron removal is the elimination of the interpolating ones, which normally produces better topology preserving values in comparison with Kohonen's SOM model. Other advantage of GCS is that the removal of neurons and neighborhood connections can produce several separate subgrids in the output layer, each one identifying a different type of input data. On the other hand, when the network is configured with a $k = 2$ factor, the basic neighborhood structure is a triangle that connects three neurons: in this case, the output layer will be formed by groups of interconnected triangles, which presents a two-dimensional topology that can be used in visual exploratory data analysis [37], [38]. In this way, the output layer of the GCS network can be visualized as a two-dimensional graph, in which some of the characteristics of the training dataset learned by the network are included. Considering that the synaptic vectors of the neurons share dimension and nature with the input data, they can be displayed using the same format than that of the training data: if the dataset consists of images, the synaptic vectors can also be shown as images. For example, in [39] this type of graph is generated from a Kohonen's SOM trained with a dataset of cryoelectron microscopy images of SV40 large T-antigen, whereas in [40] a dataset of electron microscopy images of human muscle phosphofructokinase enzyme is used. All these features support the use of GCS model for the phase of morphology clustering required in this work.

II. METHODOLOGY

The proposed method consists of five steps, the first and second of which are prior to the computer analysis here developed: A) sample preparation; B) AFM analysis; C) particle isolation; D) feature extraction; and E) morphology clustering. An overview of the method is shown in Fig. 1 and the detailed description of each step is given below.

A. SAMPLE PREPARATION

The preparation of the 574 nt-long RNA molecules containing the HCV IRES element (domains I-VI) and their resuspension in the five folding buffers tested (composed of 100 mM HEPES pH 7.4 and 100 mM NaCl, either magnesium-free or containing 2, 4, 6 or 10 mM $MgCl_2$) was previously described [22]. The whole experimental protocol has been conducted using all the precautions required that avoid RNase contamination, thus guaranteeing the highest attainable integrity of RNA. In all cases, RNA adsorption (a 30 μl drop of each preparation, at 0.5 ng/μl concentration) on freshly cleaved muscovite mica Hi-Grade V2 (Monocomp Instrumentation) surfaces was performed, using 3-aminopropyltriethoxysilane (APTES, Sigma-Aldrich). This reagent promotes a tight enough immobilization of RNA molecules on the mica surface through electrostatic

interactions, without damaging or disrupting the solution structure of the biomolecule [41]. However, it is worth mentioning that APTES functionalization induces a certain roughness and morphology on the substrate, which in some cases could make it difficult to analyze the conformation of the adsorbed RNA molecules. After incubation at 25°C for 20 min in a humidity chamber, the excess of RNA was rinsed off using diethyl pyrocarbonate (DEPC)-treated MilliQ water, and the RNA-containing surfaces were air-dried at 25°C for 2 h. Noticeably, dried RNA samples are expected to be protected against further RNase degradation, as the activity of such enzymes require a liquid medium.

B. AFM IMAGING

The AFM data were obtained using two experimental systems, namely a Nanoscope IIIa (Veeco) and an Agilent 5500 PicoPlus (Agilent Technologies) microscopes, with a nominal tip radius of 2 nm. As a general rule, we have kept a resolution of a minimum of 20 pixels per RNA molecule. Thus, different pixel densities were used: larger images (up top 3 x 3 μm) contained 2,048 x 2,048 pixels, whereas smaller ones (500 x 500 nm and 250 x 250 nm) had 512 x 512 pixels. All the images were obtained in the dynamic mode and using low free amplitudes (below 0.6 V), while the setpoint/free amplitude ratios were kept as close as possible to 0.9. These procedures were intended to keep to a minimum the force load exerted on the biomolecule by the AFM tip and, thus, to minimize the eventual distortion of its native conformation. With this aim, silicon cantilevers (Bruker) with constant force in the 1-3 N/m range were employed. The images obtained are a square matrix of heights of the RNA molecules adsorbed onto the functionalized mica. The file header of each image contains its lateral and vertical dimensions in metric scale. Using the Nanoscope or Gwyddion software applications, the images are routinely treated in order to correct an eventual 'general plane' present in the image, corresponding to a tilting of the sample plane with respect to the x-y plane of the scanning tip. Certain discontinuities from one scanning line to the next one, typical of AFM images, are corrected by means of a specific tool implemented in these software packages, known as 'flatten'. In our case, we have used the lowest order of the flatten tool, i.e., 0. Once the obtained image is optimized, a PNG file is exported using a z-hot color scale, which has become a standard method in many AFM software applications since it provides a better visualization of the height profile.

C. PARTICLE ISOLATION

In our model case, two factors must be considered regarding the nature of the images: the underlying morphology of the APTES-functionalized substrate, and the eventual overlap of two or more adsorbed RNA molecules. The success of the particle isolation automatic process will reside in its ability to filter these two sources of noise, which is required to obtain as many image files of individual molecules as potential particles exist in the original image.

Currently, there are several software packages that support the manipulation and transformation of AFM images, such as ImageJ [8] or WSxM v5.0 [10]. They include functions typically used in digital image processing tasks, although the end user must manually select the sequence of steps and algorithms to be applied in each case [22], [23]. The particle isolation stage used in this work addresses the automation of such a task in order to obtain: i) all the individual molecules present in the original image, instead of only a few of them selected by the expert user; ii) an output image of each isolated molecule stored in an individual file, in which only the pixels of the particle are included; iii) the original AFM image without noise, to provide intuitive and fast visual identification of the isolated molecules.

Fig. 2 displays the workflow used for the particle isolation phase. Data input consists of an AFM image (usually, in PNG format) containing a number of HCV IRES molecules (in this example, those resuspended in a folding buffer supplemented with 4 mM Mg^{2+}), with its specific size and resolution. In the first step, the original image in RGB color scale is transformed to gray scale, which is the format used by the thresholding algorithms. Then, 'Thresholding' stage aims at differenciating background pixels from foreground ones, thus producing a black and white image. The algorithms calculate a threshold (a value within the range 0-255) and transform all pixels with an intensity above this value to white pixels, while the rest are set to black ones. ImageJ software package includes several histogram-derived thresholding methods, which have been tested with the images used in this work. Most of them (such as Otsu [42], ISODATA [43], mean threshold [44], Huang and Wang [45], Yen *et al.* [46], Shanbhag [47] or Li and Lee [48]) establish the threshold value based on the assumption that two classes of objects are present in the image: background and foreground. In general, the application of these algorithms to our AFM images produces a threshold value that is either too high (thus, most of the molecule pixels are filtered and lost) or too low (in which case, background pixels due to APTES or substrate irregularities are not filtered and all RNA molecules appear as a single, compact object). ImageJ also includes the multiOtsu thresholding algorithm, a modified version of Otsu's method based on the existence of more than two classes of objects in the original image [49]. Specifically, by setting the number of classes to three, the algorithm looks for two threshold values, which discriminate between background, intermediate and bright pixels. This is the case of the AFM images of HCV IRES molecules under study, as they contain three classes of elements: mica surface (lowest gray levels), APTES layer (intermediate gray levels) and RNA molecules (higher gray levels). We have determined that, by filtering the lowest and intermediate pixels by means of the multiOtsu thresholding algorithm adapted to three classes, the results are comparable to those manually obtained by a human expert. Thus, this has been selected as a thresholding algorithm for the particle isolation stage.

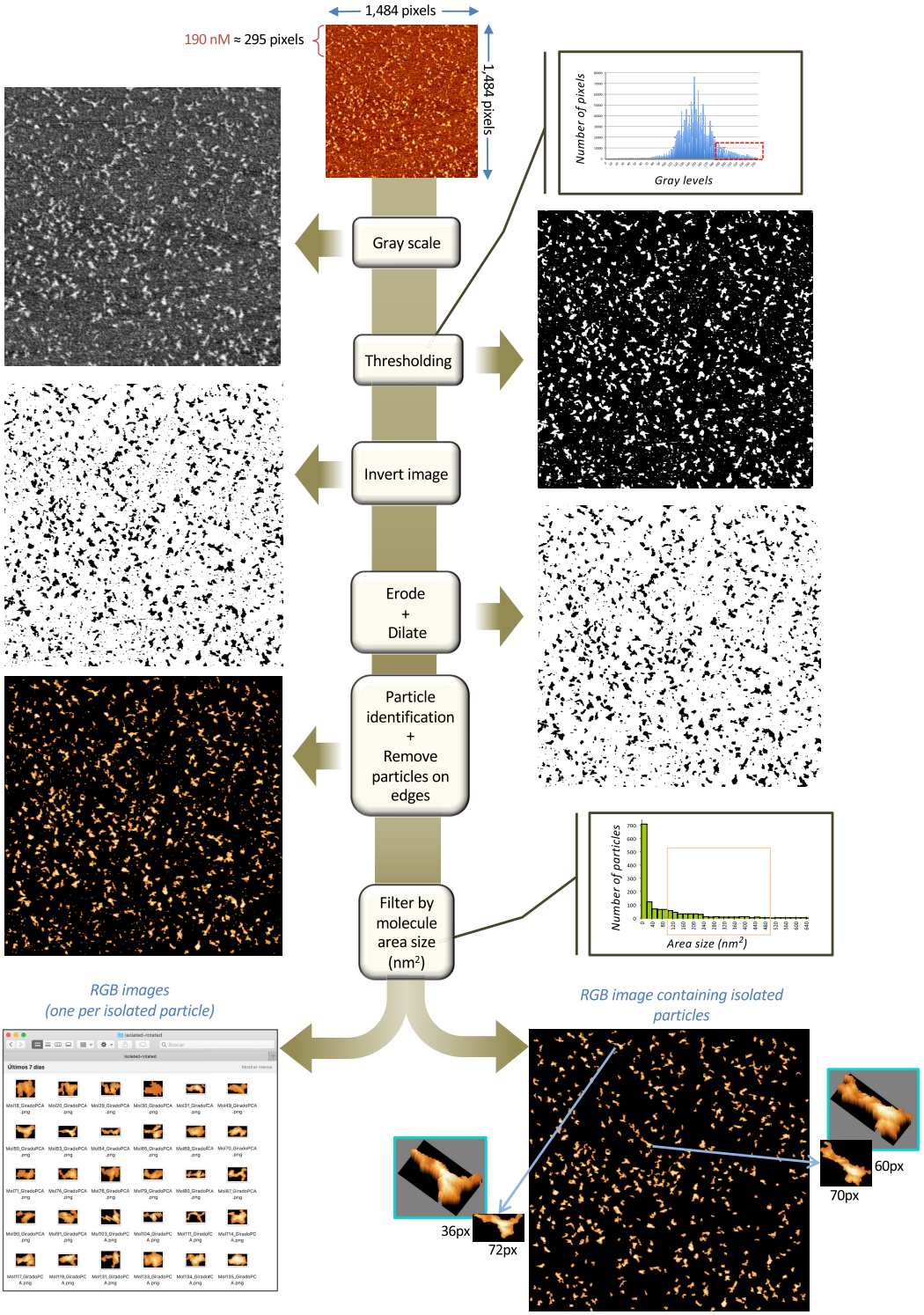

FIGURE 2. **Workflow of particle isolation phase.**

The black and white image resulting from the thresholding stage is subjected to further sequential steps. The 'Invert image' stage transforms white pixels into black, and vice versa: in this way, the resulting image displays more clearly the objects that passed the thresholding filter, considering that the contour of a white object on a black background becomes blurred to the human eye compared to a black object on a white background. Thus, the inversion of the image facilitates the visualization of the smallest particles, as shown in Fig. 2.

The AFM technique produces an image of the surface of the sample using a tip that 'runs' through it making a line-by-line scan. In some cases, this procedure introduces artefacts in the final image in the form of thin lines just 1 pixel thick. Additionally, too small and non-representative imaged features can be due to the sample itself, in our case resulting from incompletely transcribed or even partially degraded RNA molecules, as well as by the formation of APTES aggregates under certain experimental conditions. To remove these sources of noise, which may have not been filtered by the thresholding algorithm, an 'Erode and dilate' step is then performed. The erosion process, in addition to eliminating the smallest particles also removes the thin and isolated lines of pixels, thus separating two or more molecules artifactually joined by such horizontal filaments and smoothing the final contour of the molecules. However, this process produces an undesirable effect on the biochemically relevant objects, because it also modifies them by decreasing their contour. Therefore, a subsequent dilation step is necessary to add pixels to the edges of black objects, thus eliminating the negative effect of erosion on them.

The algorithms used in previous steps work at pixel level. Then, the 'Particle identification' stage will transform pixel unit to particle unit, by grouping the connected pixels that conform a compact block. The implemented algorithm handles the image of masks obtained after the dilatation step as an undirected graph consisting of connected components. Each black pixel is modeled as a vertex of the graph that can have up to eight potential neighboring vertices (black pixels around the current one). The image of masks is scanned from left to right and top to bottom, looking for black pixels. When one of them is found for the first time, a classic breadth-first search is performed from this source vertex, labeling the visited vertices (pixels) in such a way that they will not be further candidates in the image scanning. This procedure visits all the pixels that conform the connected component (i.e., the imaged molecule) to which the source pixel belongs. Then, thanks to the 'Remove particles on edges' step, those masks with one or more pixels in the border of the original image are rejected, since they may represent partially captured molecules. The particle identification step also generates a file that includes the following information for each isolated object: i) id_number (the identifiers are assigned from 0 and are used to name the PNG file of the isolated molecule image); ii) x-y coordinates of the upper left corner of the minimum rectangle that contains the image of the molecule; iii) width and height (in pixels) of such a minimum rectangle; iv) area (number of pixels of the particle contained inside the rectangle); and mass (the sum of all the gray level values of the pixels that conforms the molecule image).

Fig. 2 displays the result of applying this stepwise protocol to the original AFM image, where pixels that do not belong to any particle have been removed. This image still includes small objects (as previously discussed) along with some large ones (presumably due to the eventual overlap of two or more RNA molecules). In order to remove these two types of features, a final step adapted to the characteristics of the HCV IRES molecules has been included. In the previous step, the area of each isolated particle was obtained, expressed as the number of pixels that it contains. However, due to the fact that the method must work with input AFM images captured at different resolutions, pixel unit is not appropriate to compare objects isolated from different AFM images. Hence, based on the original size of each AFM image and the nm/pixel ratio specified in the input data, the particle area expressed in number of pixels is transformed to nm^2. Fig. 2 includes the histogram generated by the 'Filter by molecule area size' step, where a high number of particles in the range 0-40 nm^2, plus some features with an area above 550 nm^2 are evidenced. In fact, an area within the range of 100-500 nm^2 has been experimentally determined for HCV IRES [22], and consequently particles with an area outside these limits are discarded. The selection of such an area range requires a previous knowledge of the imaged molecules, as well as of the technique used. In our case, the maximum theoretical length of the fully unfolded (a situation that is not thermodynamically stable in practice) HCV IRES molecule assayed would be 155 nm, assuming that each of its 574 ribonucleotides is 0.27-nm long (in an A-RNA helical conformation, with a pitch of 3.0 nm and 11 nts per turn). The theoretical diameter of A-RNA is 2.6 nm, a value which in our case would be imaged as about 4.6 nm due to the increase in the lateral dimensions induced by the AFM tip (with a nominal radius of 2 nm) convolution. Thus, in this (very unrealistic) estimation, the area of the fully unfolded RNA molecule imaged by AFM would be around 700 nm^2. Therefore, we can asume that different degrees of compactness in solution would be induced by intramolecular interactions estabilized by the folding buffer, as previously checked by AFM [22]. Thus, an area in the range of 100-500 nm^2 seems a reasonable estimation for our imaged HCV IRES molecule at different Mg^{2+} concentrations. Such an assumption allows discarding the small imaged features that might have been preserved until this stage of the particle isolation process (including fragments of RNA, some protrusions due to either APTES or the salts present in the folding buffer, which also become enlarged by the tip convolution, and even local imaging inestabilities), as well as the statistically scarce, too large structures (e.g., undesired aggregates of RNA, salt crystals or even particles of dust adhered to the sample).

The global process of particle isolation produces two outputs: i) as many individual image files as molecules have been isolated; ii) the original AFM image containing just the RGB information of the isolated molecules, in which all the noise and artifacts have been removed (Fig. 2).

D. FEATURE EXTRACTION

The second objective of this work is the identification and classification of morphologies of the imaged molecules. The clustering method used to accomplish this task is the GCS,

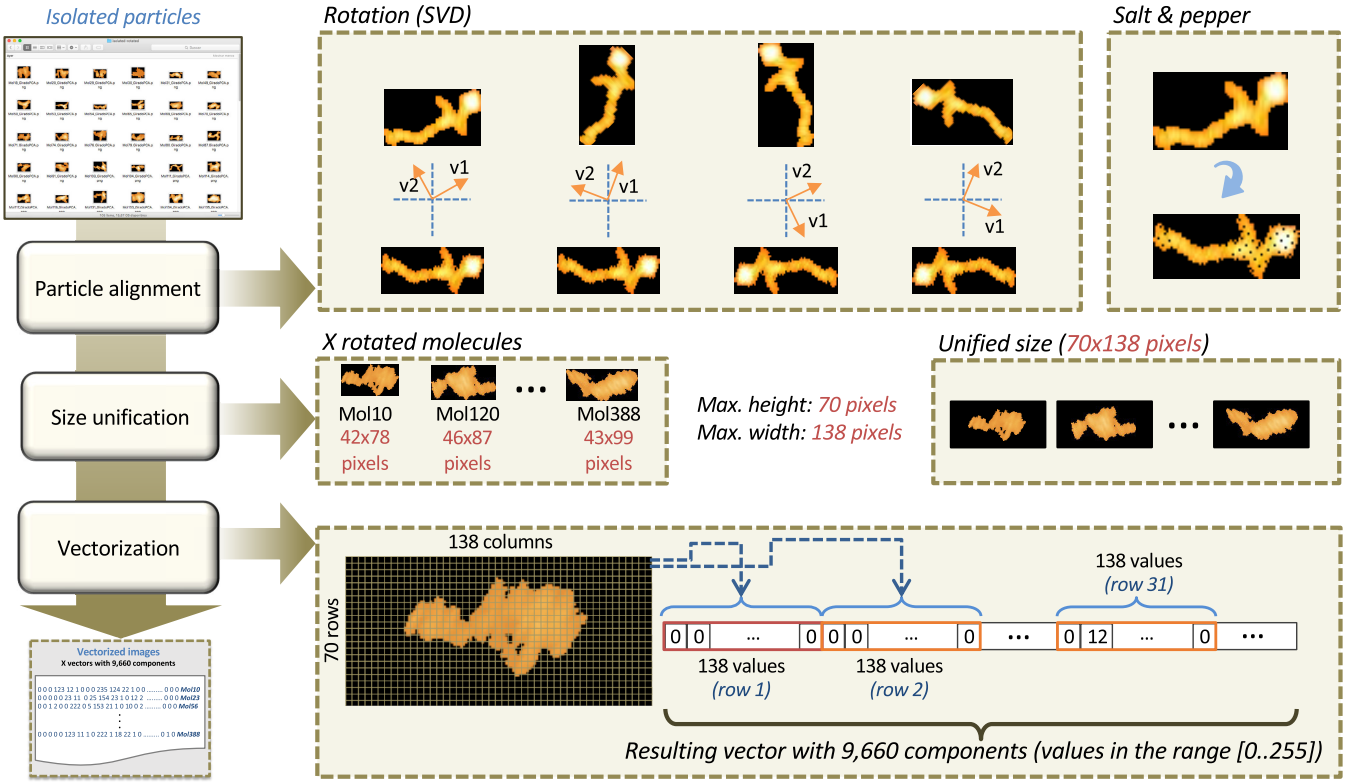

FIGURE 3. Workflow of feature extraction phase.

a dynamic SOM model. In SOM models, input data must be represented in vector format and all vectors must share the same dimension or number of components. In this work, input data consist of the images of the HCV IRES molecules isolated in the previous phase, so its vectorization is a necessary step before being analyzed by means of the GCS model.

A simple way to vectorize an image is to obtain a set of measurements, e.g., number of pixels, mean intensity, maximum value, etc. The resulting vector of this parametrization has as many components as values of the measured image, regardless of the image size. The key advantage of this option is the dimension unification of the vectors, whereas its main drawback is the need to ensure that the selected characteristics are representative and discriminative of the morphology of the imaged molecule.

Other option is to compute the rotational power spectra of the image [50]. This method, based on the Fast Fourier Transform (FFT), produces low-dimensional vectors, because only the first values of the power spectrum are used. Rotational power spectra is adequate to vectorize images that include objects showing some type of rotational symmetry [39], [51]. One of the drawbacks of this type of vectorization is the need to ensure that the origin of polar coordinates lies on the symmetry axis of the object in the image. In addition, two or more objects with the same morphology but different size will produce similar rotational power spectra vectors. It should be noted that Fourier transform has inverse function, that is, it is possible to reconstruct the original image from the complex data of the FFT. However, the rotational power spectra transformation does not have inverse function. Thus, if a SOM method is used to cluster rotational power spectra vectors, it will not be possible to reconstruct the average image represented by a prototype vector. This is not a problem for the clustering algorithm, but the interpretation of the visualization of the rotational power spectra is less intuitive than the visualization of images [39].

In this work, the images of the isolated molecules have been used directly, by applying some previous transformations. Specifically, three treatments have been implemented: particle alignment, size unification of the images, and vectorization (Fig. 3). During the preparation of the AFM samples, the biomolecules are adsorbed onto the APTES-modified mica surface in random orientations. Fig. 3 shows, as an example, an HCV IRES molecule oriented in four of its many possible options. GCS networks trained with images are very sensitive to the orientation of the particles [36]. Thus, to ensure that molecules with similar morphology are grouped in the same cluster of the GCS network, they must have the same orientation. To address this problem, a rotation of the image by means of Singular Value Decomposition (SVD) has been implemented. This technique looks for the two principal eigenvectors of the image, which will conceptually identify the main and secondary pixel distribution of the molecule, respectively. The two eigenvectors are orthogonal

and define the basis of the new coordinate system that will be used to rotate the image. When SVD is computed, the sign of the two eigenvectors is not significant: v1 is one of the two vectors that identifies the main pixel distribution and v2 is one of the two vectors orthogonal to v1. Fig. 3 shows the two eigenvectors obtained for the four orientations of a given HCV IRES molecule, as well as the result of the rotation of the images. Two of the rotated particles differ in a 180ř turn compared to the other two. To ensure an identical orientation, the mass of the left and right half of the rotated molecule is calculated: the image is rotated 180ř if the mass value of the left half is larger than the value of the right half. In the unlikely event that both halves have the same mass, a second validation is computed to ensure that the right half of the image is the one that contains more pixels with the highest RGB value, and the molecule is rotated again 180ř if necessary. When rotation is computed, the (x,y) pixel coordinates are transformed to the new orthogonal basis defined by the eigenvectors. This operation yields real numbers that must be rounded to integer numbers. In this coordinate system transformation, as numbers lower than 0.49 are rounded to 0, several isolated black pixels may appear in the body area of the molecule. Thus, a 'salt-and-pepper' noise arises in the image of the rotated molecule (Fig. 3). To avoid such an effect, a filling process of each of these black pixels is performed, and the average RGB value of its neighboring pixels (up to 4: north, south, east, and west) is computed.

To ensure that all the isolated and rotated images have the same width and height in pixels, the second step in the feature extraction stage is its size unification. This is a simple process in which the maximum width (W) and height (H) of all rotated images is computed. The size of the smaller images is unified by filling their borders with background pixels, and the particle is then centered in this resized image (Fig. 3).

The third and last step is the vectorization, a process which consists of linearizing the $W \times H$ pixels of each image: the W pixels of the first row of the image will be the first W components of the vector, the W pixels of the second row will be the second W components of the vector, and so on. The value stored in the vector for each pixel is its gray level (in the range of 0 to 255). In this way, a vector consisting of $W \times H$ components is stored for each image (Fig. 3).

E. MORPHOLOGY CLUSTERING

The morphologies of the HCV IRES molecules are analyzed using the unsupervised artificial neural network model GCS, one of the so-called dynamic SOMs. With the exception of the neighborhood connection topology of the output layer neurons, the architecture of the GCS model is identical to the Kohonen's SOM model (Fig. 4): all the input neurons are connected to all the output neurons, and every neuron in the output layer has a synaptic vector with the same dimension and nature than the input space. The processing dynamics is also identical: the input vector is distributed to all neurons in the output layer and only one (the *bmu*) is activated. However, the GCS training algorithm considerably differs from the

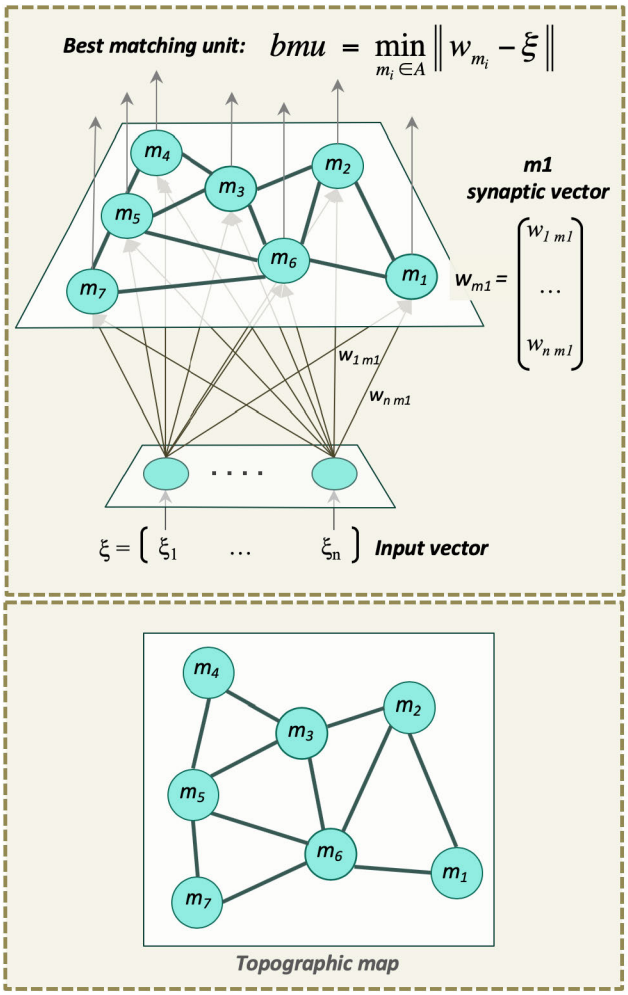

FIGURE 4. Growing cell structures (GCS) architecture (top). GCS topographic map (bottom).

Kohonen's SOM training. At the beginning of the learning process, the output layer of the GCS network consists of a single basic k-dimensional structure (for $k = 2$, three interconnected neurons). A training dataset is presented iteratively to the network: for each training vector, the *bmu* is calculated and its synaptic vector is modified to bring it closer to the processed vector. In addition, the synaptic vectors of the neurons with direct neighborhood connection with the *bmu* are also modified. This eliminates the decreasing neighborhood area factor present in the Kohonen's SOM algorithm. Then, a new neuron is periodically inserted into the output layer of the GCS, and new neighborhood connections are included or removed, ensuring that the output layer is still composed of basic k-dimensional structures [36]. The insertion of neurons can be done using two different criteria: looking for the unknown probability distribution of the input patterns (insertion near the neuron that represents more patterns) or equalizing an accumulated error (insertion near the neuron with the highest accumulated error value) [52]. In addition, neurons from the output layer with a synaptic vector in an area of the input space with low or null probability density

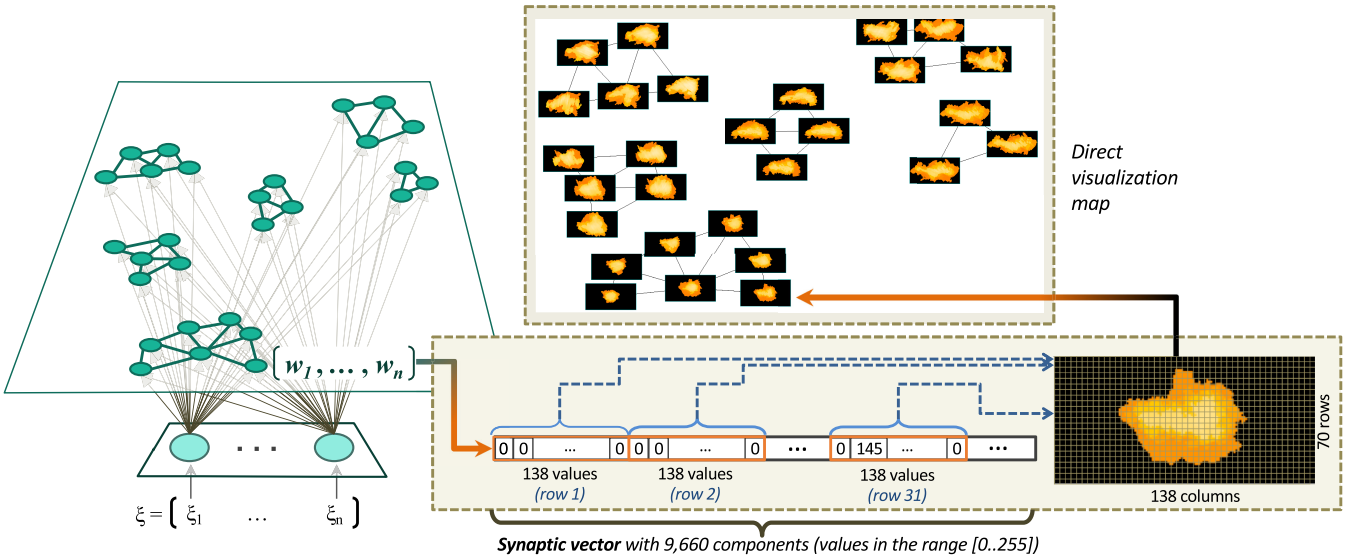

FIGURE 5. GCS direct visualization map.

are periodically removed. This procedure also ensures that the output layer is still formed by basic k-dimensional structures, although it can be divided into several sub-grids. In the experiments carried out in this work, the improved GCS training algorithm proposed in [52] has been used.

When the training patterns are grouped into separated regions of the input space, the removal of neurons causes the output layer to be divided into several sub-grids, each of them formed by neurons with similar synaptic vectors. This feature allows defining a stop-training criterion: the process will end when a pre-defined number of sub-grids in the output layer is achieved. In addition, since the training algorithm itself groups the similar synaptic vectors in the same sub-grid, each one will identify a typology of data, thus determining the different kinds of morphologies present in the input space, in our case, those of the RNA molecules contained in a sample imaged by AFM.

The output layer of GCS networks with architectural factor $k = 2$ consists of groups of interconnected neighborhood triangles in which a neuron is located at each vertex. This topology has a two-dimensional nature and can be projected on a plane, generating what is known as the topographic map, where neighboring neurons have close topographical coordinates and neighborhood connections between neurons do not cross (Fig. 4). The topographic map of the GCS network can be used as the basis to visualize graphics, similar to those generated with the Kohonen's SOM model, such as U-matrix. The algorithm for the construction of the GCS topographic map and the generation of several types of graphics has been described in detail in a previous work [37]. One of the available graphic representations is the direct visualization map, in which the synaptic vectors of the output neurons are showed in some graphic display. Accordingly, a direct visualization map adapted to the characteristics of our input data (images of the HCV IRES molecules) has been developed (Fig. 5). The synaptic vectors of the GCS neurons are prototype vectors of the training dataset, so they have the same dimension and nature: they are thus prototype images of the vectorized HCV IRES molecules. By using the width and height of the training dataset images obtained in the feature extraction stage, the inverse process to the vectorization is performed to generate an image of each prototype vector. The synaptic vector represents a grayscale vectorized image, though in the visualization of this map a conversion of the grayscale to hot-color scale has been applied, as can be seen in Fig. 5. The direct visualization map of Fig. 5 shows six clusters of neurons (sub-grids): in this case, each cluster displays the homogeneity between the images of the synaptic vectors of the neurons that comprise it. If a cluster of neurons contains images with different morphologies, it will be an indicator that such a trained network is not appropriate for discriminating the morphologies of the HCV IRES molecules present in the imaged sample.

To identify the number of different morphologies in the images of the RNA molecules, the scheme depicted in Fig. 6 has been followed. Using the HCV IRES molecules isolated and vectorized, several GCS networks are trained, each formed by c clusters in the output layer. The values used for c ranged from 2 to 10, so nine GCS networks will be available, each of them capable of discriminating as many morphologies as clusters contained in the network. The maximum value $c = 10$ has been experimentally determined by training several GCS networks and analyzing the direct visualization map. It has been found that, when looking for more than 10 types of morphologies, two or more clusters of the trained GCS network identify similar molecule conformations. In any case, although the features of the biomolecules analyzed in this study had led to this upper limit, the developed methodology can be applied to a higher range of clusters when required. To evaluate which of the 9 GCS networks

discriminates better the morphologies of the imaged molecules, one of the most widely used cluster validity indexes has been used: the Davies-Bouldin Index (DBI) [53]. It is defined as a function of the ratio of the within-cluster scatter (measured as the average of the distances of the feature vectors to the centroid of its cluster) and the between-cluster separation (measured as distance between the centroids of each pair of clusters). In this way, the lower the value of DBI, the better the separation between clusters and the homogeneity within each group. DBI is usually calculated on clustered data. However, given that synaptic vectors of a GCS represent a simplified model of the training data, the calculation of DBI directly on the synaptic vectors is proposed in this work, taking into account that they are clustered by the sub-grid to which their neuron belongs. This method of computing DBI is similar to that performed in [54] with Kohonen's SOM networks. Although clustering indexes are a reliable tool for evaluating the quality of a clustering algorithm, none of them has proven to be robust enough for all types of data of different nature and separability of partitions [33]. For this reason, in cases where two or more GCS networks provide similar minimum DBI values, the final decision about which of them differentiates most appropriately the morphologies is taken with the support of the direct visualization maps.

The training algorithm of the GCS networks is non-deterministic, thus two GCS networks trained with the same dataset and training parameters will produce GCS networks slightly different. This is due to the random initialization of the synaptic vectors and the random order of presentation of the training patterns. Therefore, when this type of model is used, it is necessary to train several GCS networks with the same dataset and the same configuration of the training parameters, and then select the best network based on some measure of quality. In the context of the SOM models, quality measures are related to the concept of topology preservation: similar input vectors are mapped by nearby neurons in the map, and neighboring neurons represent similar input data. Some of the metrics proposed in previous works only evaluate the information provided by the SOM network, without considering the structure of the dataset itself [31]. In contrast, Kaski-Lagus function [55] and topographic function [31] deal with this aspect, evaluating both the mapping of the input space in the SOM grid and the precision of the SOM to represent the dataset. The Kaski-Lagus function obtains accurate topology preserving measures, though it is only appropriate to compare SOM networks with the same number of neurons [33]. When the stop-training criterion used is based on obtaining a specific number of clusters of neurons in the output layer, the final number of neurons in the GCS network is unknown, thus it is not feasible to use the Kaski-Lagus measure to compare two GCS networks with the same number of clusters. To address the non-determinism of the GCS training algorithm, in this work 20 GCS networks have been trained for each value of c(Fig. 6). To select the GCS network of c clusters with the best

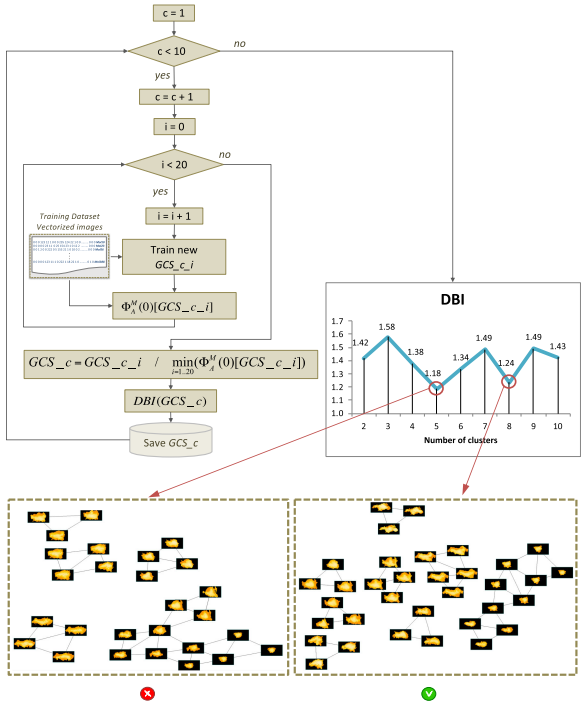

FIGURE 6. Morphology clustering flowchart using GCS.

topology preserving measure, the topographic function [31], $\Phi_A^M(0)$, adapted to the GCS model, has been used [56]. Fig. 6 shows the detailed flowchart of the developed method to analyze the number of clusters of RNA morphologies. Given a dataset composed of the vectorized images of the HVC IRES molecules present in one of the samples analyzed by AFM, the objective is to get 9 GCS networks, each of them with a particular number c of clusters (from 2 to 10) in the output layer. To obtain a single GCS network with c clusters, 20 GCS networks are trained and the one that renders the lowest value of the topographic function is selected. Then, for the 9 GCS networks obtained, the DBI index is calculated. Finally, the direct visualization maps of the networks with the lowest DBI values are generated, and the GCS network that will be used to classify the imaged molecules is chosen.

The selected GCS network undergoes an unsupervised labeling process: the neurons belonging to the same sub-grid are assigned the same numerical label (a value in the range 1-c), and the classification of the isolated molecules is carried out using such a labeled GCS network. With that aim, each vectorized image is processed by the labeled GCS, the *bmu* and the second *bmu* (the neuron whose synaptic vector has the second smallest Euclidean distance with the input vector) are determined, and the system checks whether the first and second *bmu* belong to the same cluster. If so, the molecule is classified with the *bmu* label (as the molecule clearly shows a morphology type of the *bmu* sub-grid of neurons). Otherwise, the molecule is classified with a combined label of the first and second *bmu*. In this case, the molecule shows an intermediate morphology between those characterized by

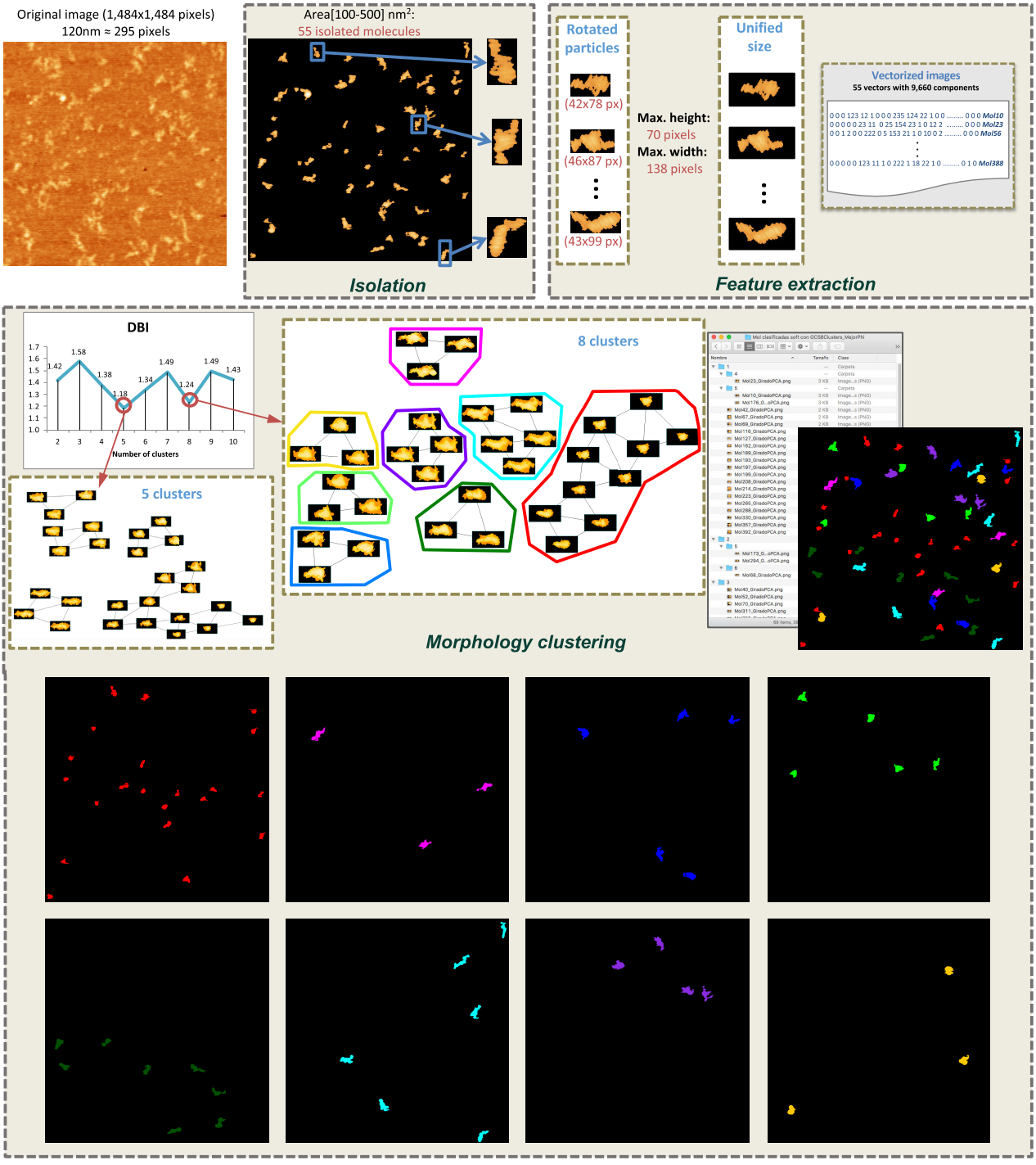

FIGURE 7. Particle isolation and morphology clustering for HCV IRES at 0mM Mg^{2+}.

both clusters, though it will be closer to that represented by the first *bmu*.

Once all the features have been classified, a directory tree is generated to organize the images of the isolated molecules based on their classification. A directory is created for each cluster of the GCS network, and it is named by the label shared by the neurons of such a cluster. The image file of each molecule classified as 'pure' (i.e., that whose first and second *bmu* are in the same cluster) is directly copied into the directory named as the class of the molecule. For each molecule classified as 'soft' (with its first and second *bmu* placed in different clusters), the image of the molecule is copied into a subdirectory created within the directory of the first *bmu*, which is named with the identifier of the second

bmu. In this way, all the isolated molecules are organized by class to facilitate the individual analysis of each morphology. In addition, a PNG file containing the complete AFM image is generated with all the molecules isolated in the sample, each of them identifiyed by the color associated with its class. Finally, it is also possible to obtain as many complete images as types of morphologies have been discriminated, each one visualizing the isolated molecules of a specific class. This last function is very useful when the number of molecules and/or the number of different detected morphologies is large.

III. EXPERIMENTAL EVALUATION, RESULTS, AND DISCUSSION

The method described in the previous section has been implemented in Java. To validate its effectiveness, 5 AFM images of HCV IRES molecules structured in folding buffers containing 0, 2, 4, 6, and 10 mM Mg^{2+}, respectively (which were manually analyzed in our previous report [22]) have been used. All GCS networks trained in the morphology clustering phase have been configured with the same learning parameters (details about them can be found in [52]). The number of input neurons corresponds to the dimension of the training vectors generated in the feature extraction phase, and the stop-training criterion is established by the achievement of a specific number of clusters of neurons in the output layer. A neighborhood connection architecture factor $k = 2$ has been used to generate the direct visualization maps. The GCS training algorithm makes use of two learning rates for the modification of the synaptic vectors, one for the *bmu* (ε_b) and the other for its immediate neighbors (ε_n). The values $\varepsilon_b = 0.06$ and $\varepsilon_n = 0.002$ have been provided (as suggested in [52]). The λ factor determines the periodicity of insertion of a new neuron in the output layer and has been configured to occur at the end of the *epoch*, that is, each time all the vectors of the dataset are processed (thus, λ = size of the dataset). The insertion of neurons is carried out using the criterion of equalizing an accumulated error, so a new neuron is inserted near the one with the highest accumulated error value. Removal of superfluous neurons is performed at the end of the *epoch*. The improvement proposed in [52] about this proceeding has been applied using the recommended threshold value $\mu = 0.001$. For the procedures of insertion and removal of neurons, the GCS training algorithm uses two counters associated with each of the neurons in the output layer. One of them records the number of training patterns for which the neuron has been *bmu*, whereas the other maintains the error accumulated by the neuron (error between the synaptic vector of the neuron and the training patterns for which it is *bmu*). At the end of the *epoch*, an aging factor (α) is applied to both counters in order to 'forget' the oldest accumulated values, in such a way that the insertion and removal of neurons is carried out based on the information of the most recent *epoch*. A value $\alpha = 0.33$ has been used, as recommended in [52].

Fig. 7 summarizes the results of the developed methodology, applied to the particular case of the AFM image of

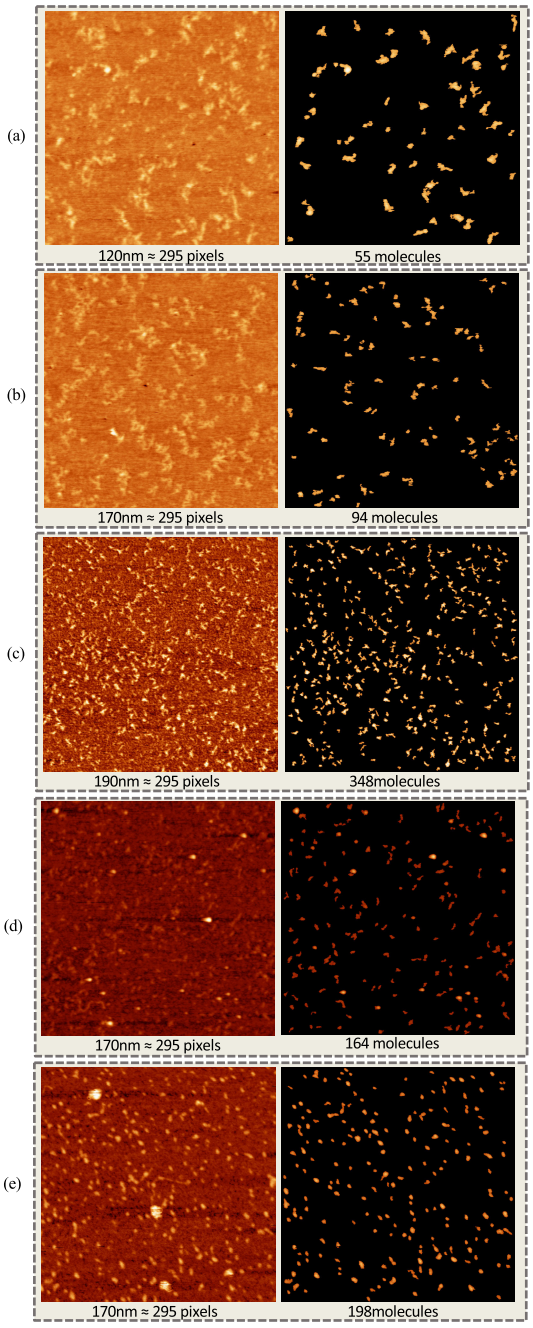

FIGURE 8. AFM images used. Left: Original images (1,484 pixels x 1,484 pixels). Right: Isolation results. The folding buffer used to resuspend the HCV IRES molecules either lacked divalent cations (a) or contained Mg^{2+} at 2 mM (b), 4mM (c), 6 mM (d) or 10 mM (e) concentration.

HCV IRES molecules in folding buffer lacking Mg^{2+}. The original PNG image had a size of 1,484 x 1,484 pixels, with a resolution of 120/295 nm/pixel. As a result of the particle isolation phase, 55 molecules were obtained within the area size range of 100-500 nm^2. Thus, 55 PNG image files were generated (each of them containing an isolated particle), as well as a single PNG image similar to the original AFM image, where all the pixels that do not correspond to any isolated molecule were filtered. Fig. 7 shows the filtered AFM image, and

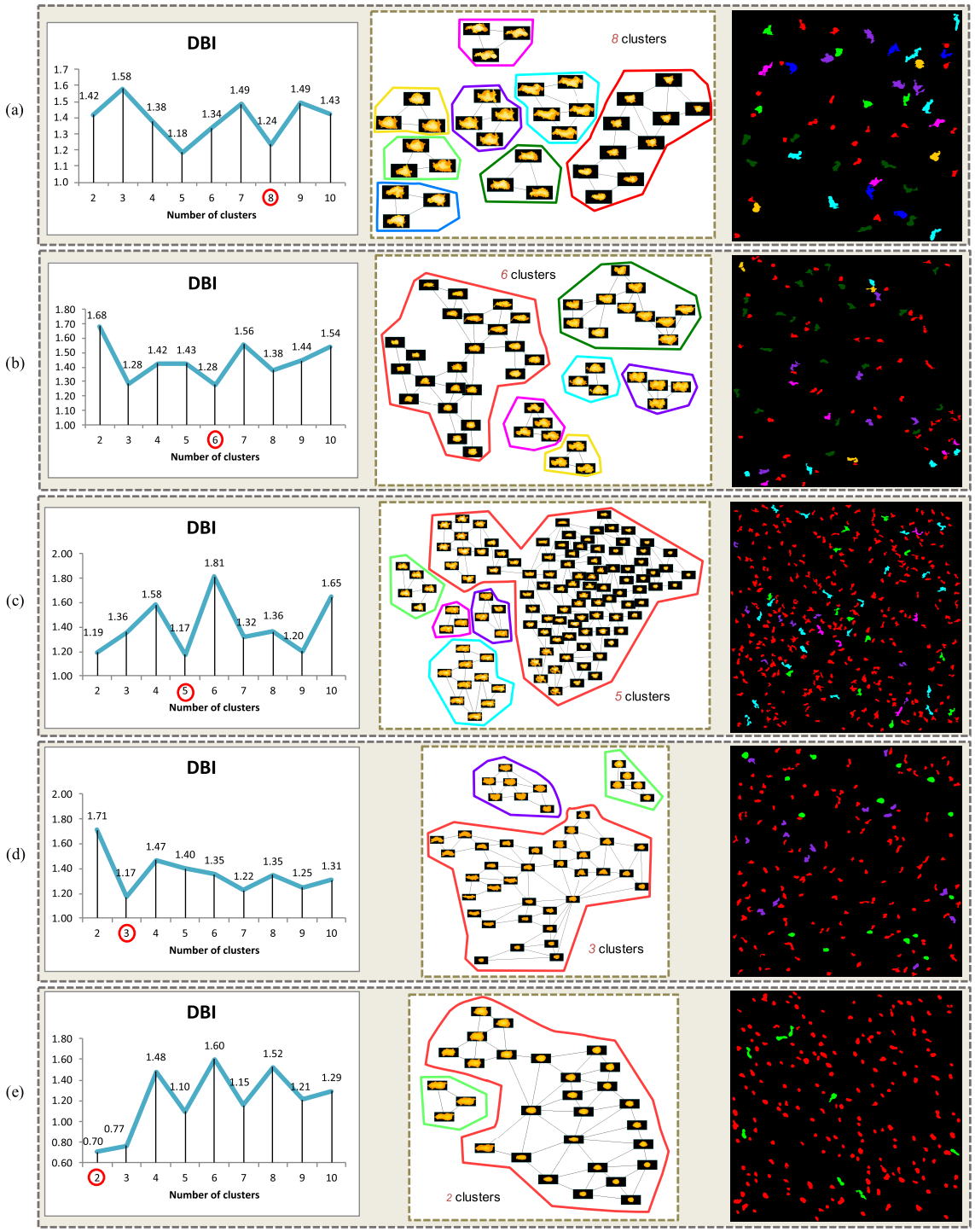

FIGURE 9. Experiments. Left: DBI values obtained for the GCS networks trained for 2 to 10 clusters; Center: Direct visualization map for the selected GCS (neurons belonging to the same cluster has been rounded with a different color for easy identification). Right: Molecules classified by the network (each molecule is colored according to the color code of the cluster in direct visualization map that contains the *bmu* for that particle. The folding buffer used either lacked divalent cations (a) or contained Mg^{2+} at 2 mM (b), 4mM (c), 6 mM (d) or 10 mM (e) concentration.

the enlarged images of three of the isolated particles. In the feature extraction phase, rotation was performed by means of SVD ensuring that the main distribution of pixels of each image is located on the x axis. The result of such a rotation, applied to three of the isolated images, can be also observed in Fig. 7. Subsequently, the width and height of the 55 rotated images were unified, ensuring that the molecule remained centered in the resized image. The unification of the size

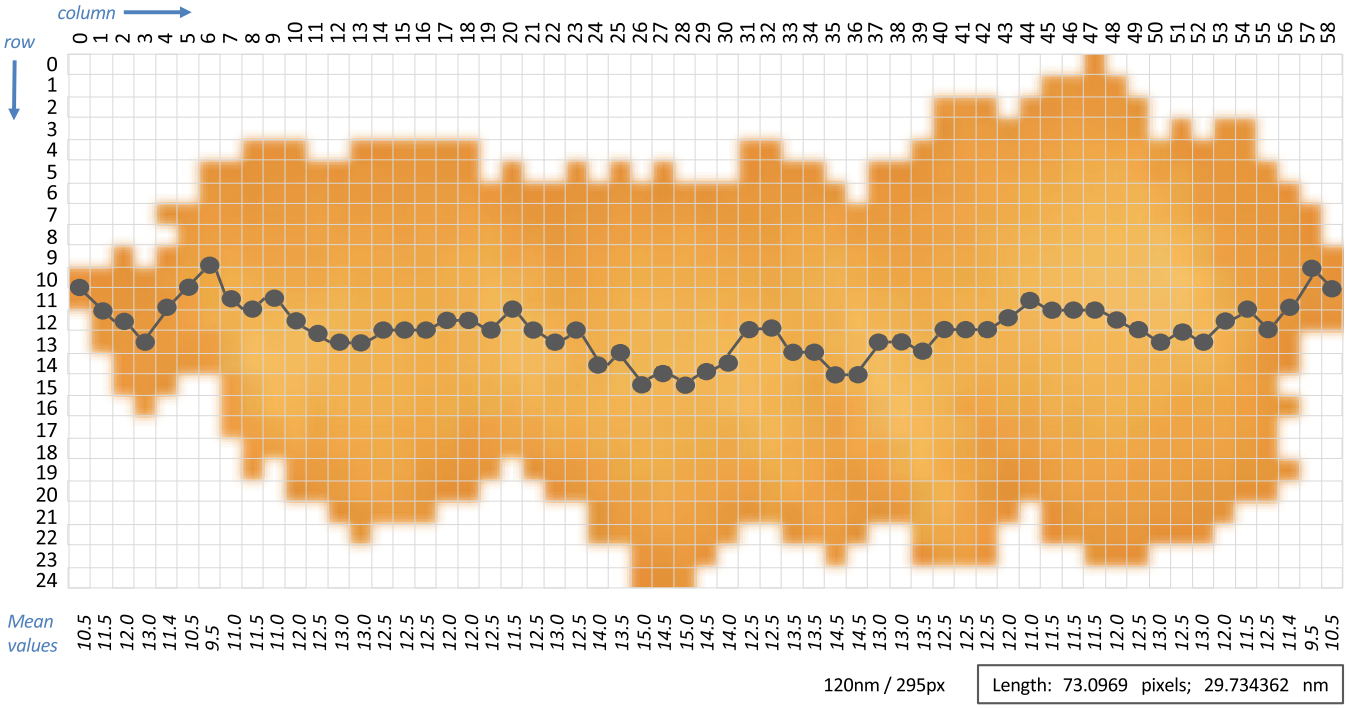

FIGURE 10. Molecule length estimation of HCV IRES at 0mM Mg^{2+}.

produced 55 images with 70 pixels of height (which was that of the highest image) and 138 pixels of width (corresponding to the widest image). Finally, the images were vectorized producing 55 vectors of 9,660 components each, which were used to train the GCS networks in the morphology clustering phase. Following the scheme depicted in Fig. 6, nine GCS networks were obtained, each of them showing a different number of clusters within the range 2-10, and the value of DBI clustering index was calculated. Fig. 7 shows the graph with the DBI values obtained for the nine GCS networks. For the two networks displaying the best DBI values (5 and 8 clusters), the direct visualization maps were generated. The GCS of 5 clusters showed a sub-grid of neurons representing different types of morphologies of the HCV IRES molecule (lower right cluster). However, the GCS of 8 clusters displayed much more compact morphologies in all its subgrids, thus the 55 molecules were accurately classified using such a GCS of 8 clusters. Different colours were used in the direct visualization map depicted in Fig. 7 to identify the morphology class contained at each of the 8 clusters of neurons. As a result, a tree of directories and subdirectories was generated, where the files of the isolated images were organized according to their class. Fig. 7 shows a section of such a directory tree, and some of the image files. Within the directory that groups the files of the molecules classified as 'pure' in 'class 1', two subdirectories ('4' and '5') were created, where the files of the images classified as 'soft' (in this case, with first *bmu* in class 1 and second *bmu* in class 4 or class 5, respectively) have been stored. In addition, a reconstructed image similar to the original AFM one was generated, by filtering all the pixels that do not belong to any particle and coloring each molecule according to its class identified in the direct visualization map (Fig. 7). Finally, 8 images were generated, each of them containing only the isolated molecules of a specific class, to allow a clear visualization of the resulting classification (Fig. 7).

This methodology was also applied to the AFM images of HCV IRES molecules structured in folding buffer supplemented with 2, 4, 6, and 10 mM Mg^{2+}. Figs. 8 and 9 show the details of their processing, together with those obtained for the image corresponding to folding buffer lacking Mg^{2+}, as previously exposed. Fig. 8 depicts the 5 original AFM images (all of them of 1,484 x 1,484 pixels in size) and their resolution (in the range of 120-190 nm per 295 pixels). It also includes the images obtained after the isolation phase, showing the filtered molecules in each one of them. As expected from the fact that the same RNA concentration was used in the preparation of the 5 imaged samples (0.5 ng/μl), the AFM image showing the lowest resolution (corresponding to the sample structured in the presence of 4 mM Mg^{2+}) contained the highest number of isolated particles (348), whereas, conversely, in the AFM image with the highest resolution (sample at 0 mM Mg^{2+}) the smallest number of molecules (55) were retrieved. Table 1 summarizes the results of the feature extraction phase for each image: the width and height values used in the unification of size once the images were rotated (expressed in pixels), the number of vectors produced (as many as isolated molecules), and its final dimensions.

Fig. 9 includes the graphs with the DBI values computed for the 9 GCS networks obtained in each case, after applying

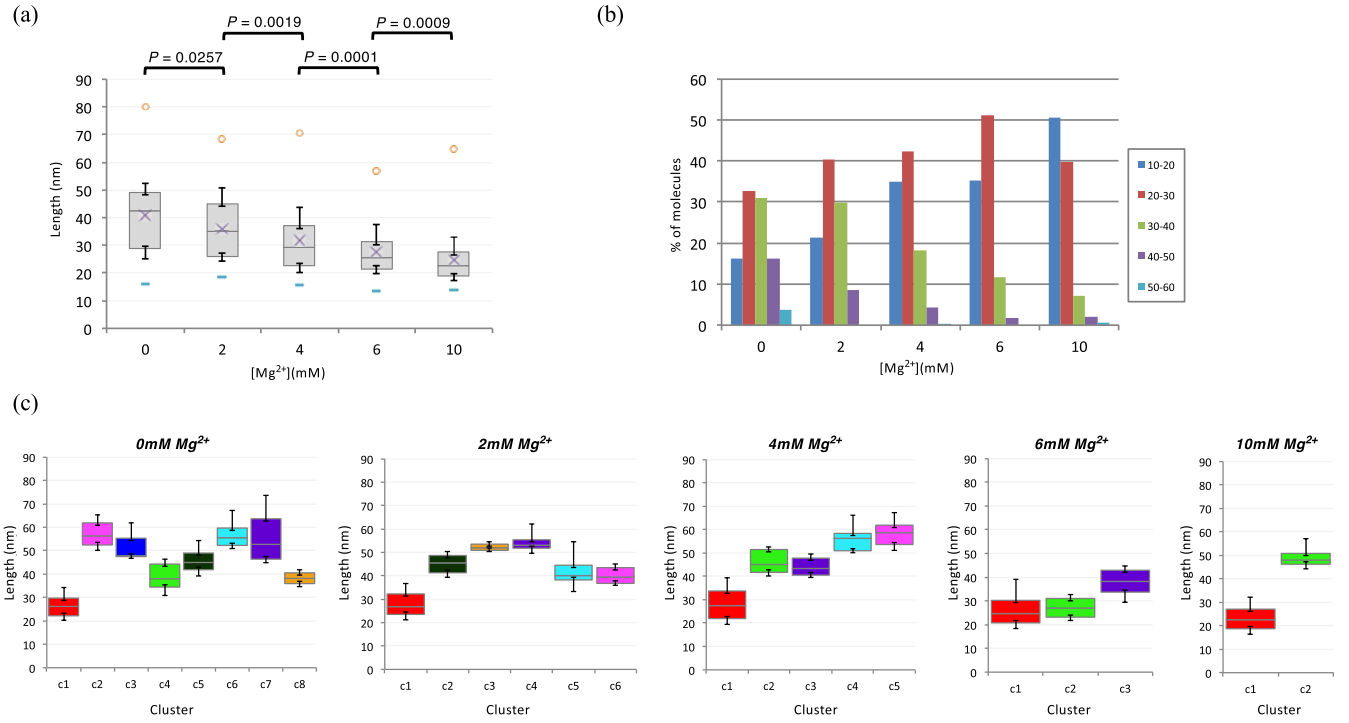

FIGURE 11. Length distribution of HCV IRES molecules (isolated and rotated) at different Mg^{2+} concentrations. (a) Length distribution (computed over all isolated molecules: 55, 94, 348, 164, and 198 for 0, 2, 4, 6, and 10 mM Mg^{2+}, respectively) of the IRES molecule in folding buffer supplemented with 0, 2, 4, 6, and 10 mM Mg^{2+} concentration is depicted in a box plot. Boxes represent 25–75 percentile range, vertical lines span 10–90 percentile range and horizontal bar and cross symbol within the box represent the median and the average value, respectively. Maximum and minimum values are depicted as circles (maximum) and hyphen (minimum). Statistical P-values corresponding to the average length of HCV IRES molecules for 0 versus 2, 2 versus 4, 4 versus 6, and 6 versus 10 Mg^{2+} concentration are shown. (b) Distribution of molecular length versus Mg^{2+} concentration, computed for isolated molecules at each buffer composition. Box: intervals of molecular length (nm). (c) Length distribution box plots of the IRES molecule computed for each group of molecules classified by GCS_8 (0 mM Mg^{2+}), GCS_6 (2 mM Mg^{2+}), GCS_5 (4 mM Mg^{2+}), GCS_3 (6 mM Mg^{2+}), and GCS_2 (10 mM Mg^{2+}), where each box color corresponds to those depicted in Fig. 9. Boxes represent 25–75 percentile range, vertical lines span 10–90 percentile range and horizontal bars within the boxes represent the median.

TABLE 1. Results of feature extraction phase.

	Rotated particles		Unified size	
	[a]*Hmax*	[b]*Wmax*	[c]*NV*	[d]*Dim*
0 mM Mg^{2+}	70	138	55	9,660
2 mM Mg^{2+}	46	86	94	3,956
4 mM Mg^{2+}	48	82	348	3,936
6 mM Mg^{2+}	42	77	164	3,234
10 mM Mg^{2+}	41	89	198	3,649

[a]*Hmax*, maximun height (pixels); [b]*Wmax*, máximum width (pixels); [c]*NV*, number of vectors; [d]*Dim*, vector dimension.

the scheme shown in Fig. 6, where the number of clusters of the selected GCS has been marked with a circle. The direct visualization map is also shown, and the neurons belonging to the same sub-grid have been rounded with different colors for easy identification. Finally, Fig. 9 also includes the complete image of the molecules classified by the corresponding GCS network, in which each molecule has been colored according to the color code of the cluster containing the *bmu* that characterizes each particle. Interestingly, the fact that the number of identified clusters decreases with the Mg^{2+} concentration present in the folding buffer (from 8 clusters at 0 mM Mg^{2+} to only 2 at 10 mM Mg^{2+}) clearly shows that the structural homogeneity and the compactness of HCV IRES molecules are promoted by this divalent cation, in agreement with our previous results [22]. It is also evident that, as observed in our prior manual analysis (which involved the use of several initial AFM images for each Mg^{2+} concentration), elongated and highly branched HCV IRES molecules are mainly present in the buffers with 0 and 2 mM Mg^{2+} concentration, more compact and 'comma-shaped' morphologies abound at 4 and 6 mM Mg^{2+}, whereas compact and 'round-shaped' conformations dominate the sample at the highest Mg^{2+} concentration tested (10 mM).

To make a quantitative study of the effect of Mg^{2+} concentration in the folding process of HCV IRES, the end-to-end length of all isolated, rotated and classified molecules was estimated for each AFM image. The algorithm implemented to estimate the length of a given molecule makes use of the coordinates (row, column) of each pixel of the particle (Fig. 10). It must be taken into account that we have not used the z-coordinate (height) in our current analysis because this value might have been affected by the AFM tip load exerted on the folded biomolecules, as commented in our previous, manual analysis of HCV IRES [22]. Thus, for each column c_i in the image, the average value of the rows containing some pixel of the molecule is calculated, which generates coordinates of the type (r_{i_mean}, c_i), depicted as gray, connected circles in Fig. 10. Then, the sum of the

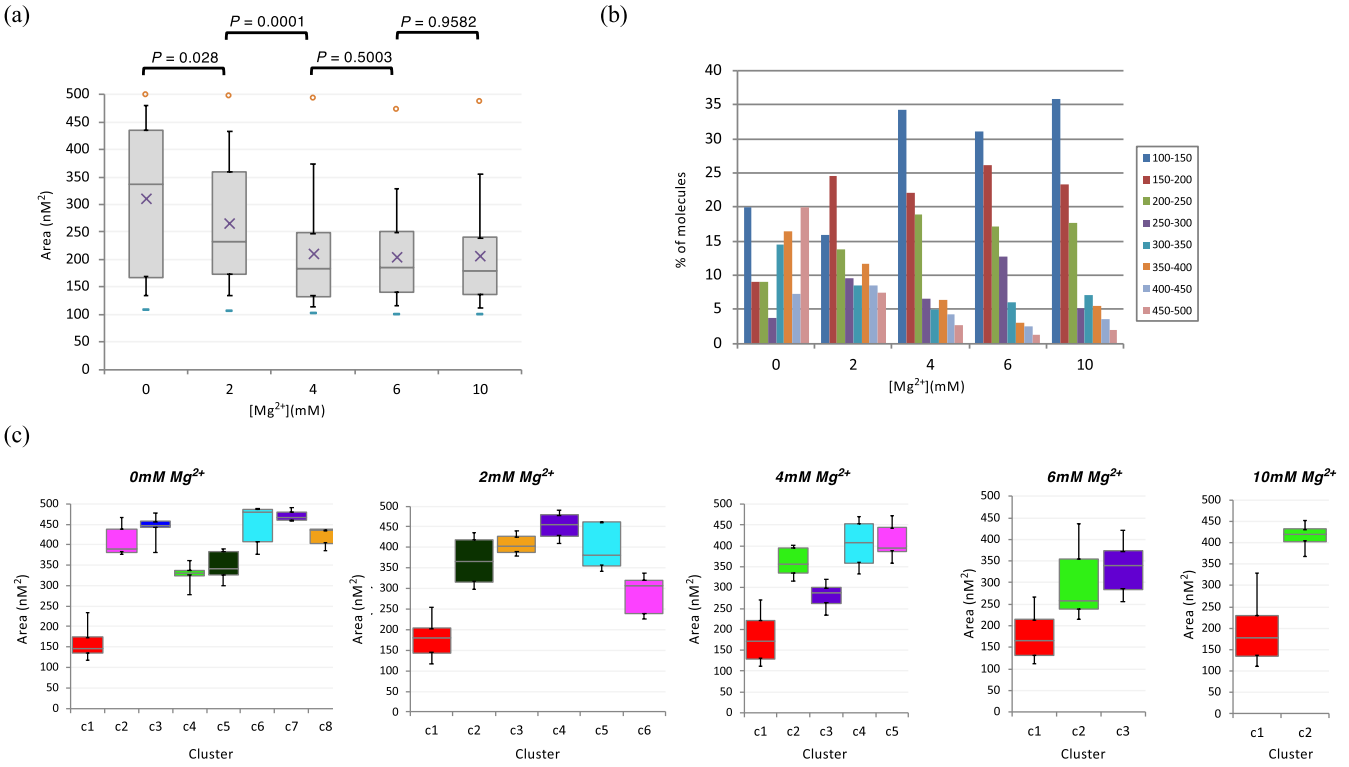

FIGURE 12. Area distribution of HCV IRES molecules (isolated and rotated) at different Mg^{2+} concentrations. (a) Area distribution (computed over all isolated molecules: 55, 94, 348, 164, and 198 for 0, 2, 4, 6, and 10 mM Mg^{2+}, respectively) of the IRES molecule in folding buffer supplemented with 0, 2, 4, 6, and 10 mM Mg^{2+} concentration is depicted in a box plot. Boxes represent 25–75 percentile range, vertical lines span 10–90 percentile range and horizontal bar and cross symbol within the box represent the median and the average value, respectively. Maximum and minimum values are depicted as circles (maximum) and hyphen (minimum). Statistical P-values corresponding to the average area of HCV IRES molecules for 0 versus 2, 2 versus 4, 4 versus 6, and 6 versus 10 Mg^{2+} concentration are shown. (b) Distribution of molecular area versus Mg^{2+} concentration, computed for isolated molecules at each buffer composition. Box: intervals of molecular area (nm). (c) Area distribution box plots of the IRES molecule computed for each group of molecules classified by GCS_8 (0 mM Mg^{2+}), GCS_6 (2 mM Mg^{2+}), GCS_5 (4 mM Mg^{2+}), GCS_3 (6 mM Mg^{2+}), and GCS_2 (10 mM Mg^{2+}), where each box color corresponds to those in Fig. 9. Boxes represent 25–75 percentile range, vertical lines span 10–90 percentile range and horizontal bars within the boxes represent the median.

Euclidean distances between each pair of consecutive coordinates is computed. This length, expressed in pixels, is then transformed to nm using the nm/pixel ratio corresponding to the resolution of each AFM image. However, as previously commented, we are aware that the tip convolution can artifactually increase the length and width of any feature imaged by AFM. Fig. 11 summarizes the main quantitative results regarding the length distribution of the imaged molecules. The box plot (Fig. 11a) evidences a progressive reduction of the HCV IRES length as a result of the increase in the Mg^{2+} concentration used in the folding buffer, which is reflected in the corresponding computed median and average values. Also, the Mg^{2+}-induced homogenization of the RNA conformations (already commented in Fig. 9) becomes evident in the length distribution graphs corresponding to each group of molecules (Figs. 11b and 11c). In turn, Fig. 12 shows quantitative results of the area distribution of the imaged molecules. The decrease in the measured molecular surface and the homogeneization of its distribution as a function of the Mg^{2+} concentration is evident. Indeed, the box plot (Fig. 12a) shows that the distribution of the molecular areas is reflecting more clearly than that of the molecular lengths (Fig. 11a) a Mg^{2+}-induced switch from extended to compact morphologies at 4 mM concentration. Therefore, most of our previous results, derived from the manual analysis of a number of AFM images of HCV IRES molecules [22] are supported by the quantitative data derived from the automatic process of particle isolation and clustering presented in this work.

However, one of the key advantages of the automatic method reported here is that it saves a considerable amount of time compared with the manual procedure previously used. Thus, one researcher took around 170 hours to manually analyse the 5 images included in this paper (150 hours for particle isolation and 20 hours for structure comparison), while the automatic process only took 7 hours of computer time (2 minutes for particle isolation and the remaining time for cluster analysis).

IV. CONCLUSION

We have developed a new morphology clustering software for microscopy images of biomolecules, based on particle isolation and GCS networks. It has been successfully tested using AFM images of functionally relevant RNA molecules

(in particular, an HCV genomic region that contains the viral IRES element) and its results have been compared with the previously obtained ones, derived from our manual analysis of the data. The software application implemented reduces the analysis time by a factor of 24 and requires low intervention from the end user, being the entire process highly automated. In this sense, the inputs that must be established are the resolution of the AFM image used as well as the range of area sizes of the particles to be isolated (in the case of IRES HCV molecules it has been established in 100-500 nm^2, but it must be adapted to the analyzed molecule in other cases). On the other hand, in the clustering phase the user intervention is only necessary to validate the GCS network finally selected by the software (i.e., the one with the lowest DBI index value), which is supported by the application through the very intuitive direct visualization maps (that show the synaptic vectors in image format). Although in the experiments carried out the number of different morphologies present in the AFM images has been selected in the range from 2 to 10, this factor can be customized to evaluate a larger range, simply by indicating the maximum number of clusters to be identified. Such an option will be relevant when the GCS network with the highest number of clusters analyzed obtains the best DBI and its direct visualization map shows disparity of morphologies within one or more subgrids of neurons, thus indicating that the number of clusters is not sufficient and should be expanded. Further improvements of the software developed in this work will include the capabilities to take into account the influence of the AFM tip convolution, as well as the possibility to faithfully compute the height of the imaged molecules. It should be noted that this software could also be applied to images obtained using other AFM modes, such as advanced nanomechanical, PeakForce QNM or electrical ones. However, the usefulness of our method goes above and beyond AFM, as it could be applied to other types of microscopy techniques that provide images with nanoscale resolution of the samples, such as scanning electron microscopy (SEM). Additionally, the automatic process of particle isolation and clustering of biomolecules developed here might constitute the first step towards the reconstruction of a 3D structural model of the biological entities under study.

REFERENCES

[1] S. Pujals, N. Feiner-Gracia, P. Delcanale, I. Voets, and L. Albertazzi, "Super-resolution microscopy as a powerful tool to study complex synthetic materials," *Nature Rev. Chem.*, vol. 3, no. 2, pp. 68–84, Feb. 2019.

[2] L. W. Whitehead, K. McArthur, N. D. Geoghegan, and K. L. Rogers, "The reinvention of twentieth century microscopy for three-dimensional imaging," *Immunol. Cell Biol.*, vol. 95, no. 6, pp. 520–524, Jul. 2017.

[3] M. P. Oxley, A. R. Lupini, and S. J. Pennycook, "Ultra-high resolution electron microscopy," *Rep. Prog. Phys.*, vol. 80, no. 2, Feb. 2017, Art. no. 026101.

[4] J. Dubochet, "On the development of electron cryo-microscopy (nobel lecture)," *Angew. Chem. Int. Ed.*, vol. 57, no. 34, pp. 10842–10846, Aug. 2018.

[5] S. Dai, W. Gao, S. Zhang, G. W. Graham, and X. Pan, "Transmission electron microscopy with atomic resolution under atmospheric pressures," *MRS Commun.*, vol. 7, no. 4, pp. 798–812, Dec. 2017.

[6] T. L. Kirk, "Chapter two—A review of scanning electron microscopy in near field emission mode," *Adv. Imag. Electron Phys.*, vol. 204, pp. 39–109, Jan. 2017.

[7] J. L. Toca-Herrera, "Atomic force microscopy meets biophysics, bioengineering, chemistry, and materials science," *ChemSusChem*, vol. 12, no. 3, pp. 603–611, Dec. 2018.

[8] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "NIH Image to ImageJ: 25 years of image analysis," *Nature*, vol. 9, no. 7, pp. 671–675, Jul. 2012.

[9] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, "Fiji: An open-source platform for biological-image analysis," *Nature Methods*, vol. 9, no. 7, pp. 676–682, Jul. 2012.

[10] I. Horcas, R. Fernández, J. M. Gómez-Rodráguez, J. Colchero, J. Gómez-Herrero, and A. M. Baro, "WSXM: A software for scanning probe microscopy and a tool for nanotechnology," *Rev. Sci. Instrum.*, vol. 78, no. 1, Jan. 2007, Art. no. 013705.

[11] D. Necas and P. Klapetek, "Gwyddion: An open-source software for SPM data analysis," *Central Eur. J. Phys.*, vol. 10, no. 1, pp. 181–188, Jan. 2012.

[12] C. Bustamante and D. Keller, "Scanning force microscopy in biology," *Phys. Today*, vol. 48, no. 12, pp. 32–38, Dec. 1995.

[13] H. G. Hansma, K. Kasuya, and E. Oroudjev, "Atomic force microscopy imaging and pulling of nucleic acids," *Current Opinion Struct. Biol.*, vol. 14, no. 3, pp. 380–385, Jun. 2004.

[14] J. Ognjenović, R. Grisshammer, and S. Subramaniam, "Frontiers in cryo electron microscopy of complex macromolecular assemblies," *Annu. Rev. Biomed. Eng.*, vol. 21, no. 1, pp. 395–415, Jun. 2019.

[15] A. T. Wassie, Y. Zhao, and E. S. Boyden, "Expansion microscopy: Principles and uses in biological research," *Nature Methods*, vol. 16, no. 1, pp. 33–41, Jan. 2019.

[16] S. Jonić, "Computational methods for analyzing conformational variability of macromolecular complexes from cryo-electron microscopy images," *Current Opinion Struct. Biol.*, vol. 43, pp. 114–121, Apr. 2017.

[17] Z.-P. Zeng, H. Xie, L. Chen, K. Zhanghao, K. Zhao, X.-S. Yang, and P. Xi, "Computational methods in super-resolution microscopy," *Frontiers Inf. Technol. Electron. Eng.*, vol. 18, no. 9, pp. 1222–1235, Sep. 2017.

[18] C. Hyeon, R. I. Dima, and D. Thirumalai, "Size, shape, and flexibility of RNA structures," *J. Chem. Phys.*, vol. 125, no. 19, Nov. 2006, Art. no. 194905.

[19] M. H. Bailor, X. Sun, and H. M. Al-Hashimi, "Topology links rna secondary structure with global conformation, dynamics, and adaptation," *Science*, vol. 327, no. 5962, pp. 202–206, Jan. 2010.

[20] S. E. Butcher and A. M. Pyle, "The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks," *Accounts Chem. Res.*, vol. 44, no. 12, pp. 1302–1311, Dec. 2011.

[21] P. Schöön, "Atomic force microscopy of RNA: State of the art and recent advancements," *Seminars Cell Develop. Biol.*, vol. 73, pp. 209–219, Jan. 2018.

[22] A. García-Sacristán, M. Moreno, A. Ariza-Mateos, E. López-Camacho, R. M. Jáudenes, L. Vázquez, J. Gómez, J. Á. Martín-Gago, and C. Briones, "A magnesium-induced RNA conformational switch at the internal ribosome entry site of hepatitis C virus genome visualized by atomic force microscopy," *Nucleic Acids Res.*, vol. 43, no. 1, pp. 565–580, Jan. 2015.

[23] M. Moreno, L. Vázquez, A. López-Carrasco, J. A. Martín-Gago, R. Flores, and C. Briones, "Direct visualization of the native structure of viroid RNAs at single-molecule resolution by atomic force microscopy," *RNA Biol.*, vol. 16, no. 3, pp. 295–308, Mar. 2019.

[24] H. Sánchez and C. Wyman, "SFMetrics: An analysis tool for scanning force microscopy images of biomolecules," *BMC Bioinf.*, vol. 16, no. 1, Dec. 2015, Art. no. 27.

[25] R. Xu and D. Wunsch, II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[26] A. Srivastava, S. H. Joshi, W. Mio, and X. Liu, "Statistical shape analysis: Clustering, learning, and testing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 4, pp. 590–602, Apr. 2005.

[27] T. Kohonen, *Self-Organizing Maps*, 3rd ed. Berlin, Germany: Springer-Verlag, 2001.

[28] T. Kohonen, "Essentials of the self-organizing map," *Neural Netw.*, vol. 37, pp. 52–65, Jan. 2013.

[29] T. Martinetz and K. Schulten, "Topology representing networks," *Neural Netw.*, vol. 7, no. 3, pp. 507–522, Jan. 1994.

[30] M. Rubio and V. Giménez, "New methods for self-organising map visual analysis," *Neural Comput. Appl.*, vol. 12, nos. 3–4, pp. 142–152, Dec. 2003.
[31] T. Villmann, R. Der, M. Herrmann, and T. M. Martinetz, "Topology preservation in self-organizing feature maps: Exact definition and measurement," *IEEE Trans. Neural Netw.*, vol. 8, no. 2, pp. 256–266, Mar. 1997.
[32] A. Ultsch, "Kohonen's self organizing feature maps for exploratory data analysis," in *Proc. Int. Neural Netw. Conf. (INNC)*, 1990, pp. 305–308.
[33] S. Delgado, C. Higuera, J. Calle-Espinosa, F. Morán, and F. Montero, "A SOM prototype-based cluster analysis methodology," *Expert Syst. Appl.*, vol. 88, pp. 14–28, Dec. 2017.
[34] J. Blackmore and R. Miikkulainen, "Incremental grid growing: Encoding high-dimensional structure into a two-dimensional feature map," in *Proc. IEEE Int. Conf. Neural Netw.*, Mar./Apr. 1993, pp. 450–455.
[35] B. Fritzke, "A growing neural gas network learns topologies," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 7, 1995, pp. 625–632.
[36] B. Fritzke, "Growing cell structures—A self-organizing network for unsupervised and supervised learning," *Neural Netw.*, vol. 7, no. 9, pp. 1441–1460, Jan. 1994.
[37] S. Delgado, C. Gonzalo, E. Martinez, and A. Arquero, "Visualizing high-dimensional input data with growing self-organizing maps," in *Computational and Ambient Intelligence* (Lecture Notes in Computer Science), vol. 4507, F. Sandoval, A. Prieto, J. Cabestany, and M. Graña, Eds. Berlin, Germany: Springer, 2007, pp. 580–587.
[38] S. Delgado, F. Morán, A. Mora, J. J. Merelo, and C. Briones, "A novel representation of genomic sequences for taxonomic clustering and visualization by means of self-organizing maps," *Bioinformatics*, vol. 31, no. 5, pp. 736–744, 2015.
[39] A. Pascual-Montano, L. E. Donate, M. Valle, M. Bárcena, R. Pascual-Marqui, and J. Carazo, "A novel neural network technique for analysis and classification of em single-particle images," *J. Struct. Biol.*, vol. 133, nos. 2–3, pp. 233–245, Feb. 2001.
[40] I. Arechaga, O. H. Martínez-Costa, C. Ferreras, J. L. Carrascosa, and J. J. Aragón, "Electron microscopy analysis of mammalian phosphofructokinase reveals an unusual 3-dimensional structure with significant implications for enzyme function," *FASEB J.*, vol. 24, no. 12, pp. 4960–4968, 2010.
[41] Y. Lyubchenko, L. Shlyakhtenko, R. Harrington, P. Oden, and S. Lindsay, "Atomic force microscopy of long DNA: Imaging in air and under water," *Proc. Nat. Acad. Sci. USA*, vol. 90, no. 6, pp. 2137–2140, Mar. 1993.
[42] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.
[43] T. W. Ridler's and S. Calvard, "Picture thresholding using an iterative selection method," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, no. 8, pp. 630–632, Aug. 1978.
[44] C. A. Glasbey, "An analysis of histogram-based thresholding algorithms," *CVGIP, Graph. Models Image Process.*, vol. 55, no. 6, pp. 532–537, Nov. 1993.
[45] L. K. Huang and M. J. J. Wang, "Image thresholding by minimizing the measures of fuzziness," *Pattern Recognit.*, vol. 28, no. 1, pp. 41–51, Jan. 1995.
[46] J.-C. Yen, F.-J. Chang, and S. Chang, "A new criterion for automatic multilevel thresholding," *IEEE Trans. Image Process.*, vol. 4, no. 3, pp. 370–378, Mar. 1995.
[47] A. G. Shanbhag, "Utilization of information measure as a means of image thresholding," *CVGIP, Graph. Models Image Process.*, vol. 56, no. 5, pp. 414–419, 1994.
[48] C. H. Li and C. K. Lee, "Minimum cross entropy thresholding," *Pattern Recognit.*, vol. 26, no. 4, pp. 617–625, Apr. 1993.
[49] P.-S. Liao, T.-S. Chen, and P.-C. Chung, "A fast algorithm for multilevel thresholding," *J. Inf. Sci. Eng.*, vol. 17, no. 5, pp. 713–727, 2001.
[50] R. A. Crowther and L. A. Amos, "Harmonic analysis of electron microscope images with rotational symmetry," *J. Mol. Biol.*, vol. 60, no. 1, pp. 123–130, Aug. 1971.
[51] S. H. W. Scheres, M. Valle, R. Nuñez, C. O. S. Sorzano, R. Marabini, G. T. Herman, and J.-M. Carazo, "Maximum-likelihood multi-reference refinement for electron microscopy images," *J. Mol. Biol.*, vol. 348, no. 1, pp. 139–149, Apr. 2005.
[52] S. Delgado, C. Gonzalo, E. Martinez, and A. Arquero, "Improvement of self-organizing maps with growing capability for goodness evaluation of multispectral training patterns," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, vol. 1, Sep. 2004, pp. 564–567.
[53] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
[54] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 586–600, May 2000.
[55] S. Kaski and K. Lagus, "Comparing self-organizing maps," in *Proc. Int. Conf. Artif. Neural Netw.*, 1996, pp. 809–814.
[56] S. Delgado, C. Gonzalo, E. Martinez, and A. Arquero, "A combined measure for quantifying and qualifying the topology preservation of growing self-organizing maps," *Neurocomputing*, vol. 74, no. 16, pp. 2624–2632, 2011.

SOLEDAD DELGADO received the B.S. degree from the Universidad Politécnica de Madrid (UMP), Madrid, Spain, in 1990, the M.S. degree from the Universidad Carlos III, Madrid, in 2000, and the Ph.D. degree from the UMP, in 2010, all in computer science.

From 1991 to 2017, she was an Assistant Professor with the Applied Computing Department and the Organization and Structure of Information Department, UMP. Since 2017, she has been an Associate Professor with the Department of Computer Systems, UMP. Her current research interests include machine learning, artificial neural networks, unsupervised learning, self-organizing maps, pattern recognition, data mining, exploratory data analysis, algorithms and computational complexity, and image processing techniques.

MIGUEL MORENO received the B.S. degree in molecular and cell biology and the Ph.D. degree in biomedicine from the University of Alcalá, Alcala de Henares, Madrid, Spain, in 1998 and 2005, respectively. He was with the University of Alcalá, where he involved in the development of electrochemical biosensors based on DNA-gold capped nanoparticles.

In 2009, he joined the Department of Molecular Evolution, Centro de Astrobiología (CAB, a Joint Research Centre of the Spanish National Research Council (CSIC) and the National Institute of Aerospace Technology (INTA), associated to the NASA Astrobiology Institute), Madrid. He is currently involving in the structural analysis of viral and viroidal RNA using AFM, *in vitro* selection and the evolution of nucleic acids (RNA and DNA aptamers), and in the development of biosensors and nano-biosensors (DNA microarray technology, and PNA- and aptamer-based sensors).

LUIS F. VÁZQUEZ received the Ph.D. degree in physics from ESRF, Grenoble, France, in 1989. Since 1990, he has been a Scientific Staff with the Instituto de Ciencia de Materiales de Madrid-ICMM-CSIC. He has been a CSIC Research Professor, since 2008. His current research interest includes AFM studies. Thus, the part of his research is devoted to the study of growing/etching interfaces, and surface nano-patterning, especially the relationship between the interface morphology and the mechanisms determining it. The AFM characterization of bio interfaces and biomolecules under different imaging conditions, including the study of their mechanical properties. He has coauthored more than 240 articles in international scientific journals and several book chapters.

JOSE ÁNGEL MARTÍNGAGO received the Ph.D. degree in physics, in 1994. He has been a CSIC Research Professor with the ICMM, since 2012, where he leads the ESISNA Research Group. His team addresses interdisciplinary and multitechnique research for understanding the structure and electronic properties of molecules on surfaces and other low-dimensional systems. In 2013, he was awarded with an ERC-Synergy grant for the project Gas and Dust From the Star to the Laboratory: Exploring the Nanocosmos, together with J. Cernicharo's and C. Joblin's groups. He has coauthored more than 160 articles in SCI journals that accumulate more than 4000 citations, and has authored some books in nanotechnology. He is a member of several program review panel boards for European synchrotron radiation facilities, and the President of the Spanish Vacuum Society.

CARLOS BRIONES received the Ph.D. degree in chemistry (majoring in biochemistry and molecular biology) from the Universidad Autónoma de Madrid (UAM), in 1997.

Since 2000, has been with the CSIC Research Group Molecular Evolution, RNA World and Biosensors, which is focused on the origin and early evolution of life, RNA biochemistry (including sequence/structure/function mapping in RNA and AFM-based analysis of RNA structure), *in vitro* evolution of nucleic acids, and biosensor development (aptamer-based and bio nanotechnology-inspired sensors). He is currently a Staff Research Scientist with the Department of Molecular Evolution, CAB (CSIC-INTA). He has authored or coauthored more than 70 articles in SCI journals (which accumulate more than 2000 citations), four books, and more than 30 chapters in specialized books. He is a co-inventor of eight patents and utility models in the fields of biotechnology and biosensor development.