PS Poly: A chain tracing algorithm to determine persistence length and categorize complex polymers by shape

Elizabeth A. Conley[1], Katherine G. Schaefer[1], Harrison C. Davison[1], and Gavin M. King[1,2]

[1] Department of Physics and Astronomy, University of Missouri-Columbia, Columbia, MO 65211 USA

[2] Joint with Biochemistry, University of Missouri-Columbia, Columbia, MO 65211 USA

Abstract

The fundamental molecules of life are polymers. Prominent examples include nucleic acids and proteins, both of which assume a vast array of mechanical properties and three-dimensional shapes. The persistence length represents a numerical value to classify the bending rigidity of individual polymers. The shape of a polymer, dictated by the topology of the polymer backbone - a line trace through the center of the polymer along the contour path – is also a critical metric. Common architectures include linear, ring-like or cyclic, and branched; combinations of these can also exist, as in complex polymer networks. Determination of persistence length and shape are largely informative to polymer function and stability in biological environments. Here we demonstrate PS Poly, a near-fully automated algorithm to obtain polymer persistence length and shape from single molecule images obtained in physiologically relevant fluid conditions via atomic force microscopy. The algorithm, which involves image reduction via skeletonization followed by end point and branch point detection via filtering, is capable of rapidly analyzing thousands of polymers with subpixel precision. Algorithm outputs were verified by analysis of deoxyribose nucleic acid, a very well characterized macromolecule. The utility of method was further demonstrated by application to a recently discovered polypeptide chain named candidalysin. This toxic protein segment polymerizes in solution and represents the first human fungal pathogen yet discovered. PS poly is a robust and general algorithm. It can be used to extract fundamental information about polymer backbone stiffness, shape, and more generally, polymerization mechanisms.

Introduction

Knowledge of cellular function and misfunction (disease) has advanced through developing a detailed understanding of many semi-flexible polymeric molecules. A prime example is the recently discovered peptide toxin candidalysin (CL), which is the virulence factor secreted by the fungus *C. albicans* [1]. CL form loops in solution which then embed themselves into membranes, leading to host cell damage. When establishing the molecular basis of a disease, such as invasive candidiasis which emanates from C. albicans and has a 50% mortality rate[2], characterizing polymer mechanical properties and topological shape can provide significant insights. The bending rigidity, quantified by the persistence length, sheds light on the minimum diameter of the loops formed by CL. This geometric information can be used to predict what size molecules can pass through the pores that CL creates in the host cell membrane. Additionally, when studying the kinetics of polymer loop formation or branching which often involves secondary polymerization interfaces [3,4], it is informative to separate and quantify polymers by shape so that distinct reactions can be isolated. Topological analyses can be used to build kinetic models of looping and branching, to determine under what conditions polymer cyclization occurs, and to explore how conversion of linear to looped polymers can be controlled.

Atomic Force Microscopy (AFM) is a powerful single molecule imaging modality employed in biological nanoscience investigations and has been used to shed light on polymer persistence length, P [5–8]. Statistical treatment of images containing many polymers are typically used. To perform P calculations, it is necessary to know the coordinates along the chain contour, or "backbone", for each polymer included in the analysis. Once these coordinates are obtained, the

so called “worm-like chain” model can be used to connect the mean-square of end-to-end distances to the persistence length [9].

There is no standard method for automatically obtaining the backbone coordinates for a polymer from single molecule imaging methods such as AFM [10–18]. Algorithms have been developed that use manual tracing, which can be time consuming to perform and introduce a degree of human bias and error. A few automated chain tracing algorithms have been developed, but most still require manual identification of points along the chain contour, then the backbone coordinates are refined automatically. Automated algorithms that do not require manual input of any coordinates within the image are rare and may be sensitive to background noise and chains with irregular intensity along the contour [11,13]. AutoSmarTrace has been developed recently, which uses machine learning as a tool for automatic chain detection and tracing, however the program can only determine persistence length from molecules with contour lengths longer than 30 pixels [18]. We are not aware of an algorithm that can process closed loop (cyclic) and branched polymers and sort these differing topological structures by shape; the complex polymerization processes revealed in a recent study of candidalysin motivated us to develop one [19].

Here we present PS Poly, an automated chain tracing algorithm with subpixel precision which calculates persistence length and classifies complex mixtures of polymeric features by shape. The program is open-sourced with code written in both Python as well as in Igor Pro (WaveMetrics, Inc.). In the current version, the user specifies a height threshold to isolate features from background; there is no further manual input required. A workflow of the algorithm is shown (**Fig. 1**). Briefly, the images are skeletonized, a convolutional filter is used to

identify endpoints, and a pathfinding algorithm is used to store the coordinates of all linear particles used for persistence length analysis. To separate features by shape, filters were developed to robustly identify branch points and distinguish branches from cyclic or looped polymers. Persistence length results for CL and DNA were compared to established values. The results were in agreement with expectations; in particular they were within the margin of error of results obtained using the manual EasyWorm method [10].

FIGURE 1. Program Overview. First, the image is processed and reduced to only polymer backbone coordinates, then individual features are separated based upon their shape. Only measurements of linear polymers are considered when calculating persistence length.

Methods

Polymer Backbone Isolation

Isolating the polymer backbone is the first step of the algorithm, illustrated in **Figure 2**. The program begins by prompting the user to create a threshold corresponding to pixel intensity which is proportional to topographical height, *z*, of the polymer in images taken using an AFM. This creates a binary "mask" image where all values above the threshold are valued as 1 and all values below are valued as 0. Then a copy of this mask is made with a higher pixel density by creating a new image where the value of every pixel on the original image is taken up by 4 pixels on the bigger mask. This allows the result to be obtained with a subpixel level of accuracy which can be valuable for characterizing short polymers. Additional pixels can be added for enhanced resolution at the expense of computational time. A "skeleton" of the mask is created through a 3D surface thinning algorithm which eliminates layers of the image until only single pixel linewidth traces remain (Fig. 2D).

Acquiring polymer coordinates

To obtain separate lists of coordinates for each molecule, we begin by looping through each pixel on a duplicate of the thinned image. Once a one-valued pixel is found, that coordinate is stored. Then a flood-fill algorithm fills-in all one-valued pixels which are continuous with that coordinate. This process continues in a loop until the duplicated image is entirely zero-valued, and the resulting list of coordinates correspond to exactly one "seed" pixel per molecule. Then, depth-first search (DFS) is applied in a square around the seed pixel. This algorithm explores all possible paths stemming from one input coordinate until a path is found to another input coordinate. The implementation of DFS in this program returns 1 if a path is found between the two coordinates, and zero if there is no possible path. The search radius is incremented with each loop iteration, and the loop breaks once all locations continuous with the seed pixel are found. We found that applying DFS by this method, as opposed to checking for continuity with all of the one-valued pixels in the image, reduces computational time significantly.

FIGURE 2. Polymer backbone isolation procedure. Preprocessing steps to perform calculations on an AFM image of candidalysin are shown. Scale bars (yellow) are 100nm. (**A**) AFM image with topographic height, *z,* shown in greyscale. (**B**) Mask created from the AFM data. (**C**) New mask that has been expanded to a higher pixel density. (**D**) Skeleton created from the expanded mask. (**E**) Each molecule is registered as a unique object (pink circles).

<u>Sorting polymers by shape</u>

Polymers are sorted based on shape: linear, branched, or looped. This first requires polymer termination point identification, achieved through convolutional filtering.

Polymer end point determination

The algorithm loops though the image, cropping 9 x 9 pixel areas surrounding each central pixel test point. **Figure 3A** demonstrates the pixel grids that are created for every one-valued pixel in the image. There are 16 possible endpoint configurations because there are 8 possible neighboring pixels, and so for one neighbor, there are 8 possible combinations. For 2 neighbors, there are 16 possible combinations, but only half of them are endpoints because the 2 neighboring pixels must also be adjacent to each other in order to be an endpoint. The pixel grids corresponding to the 16 possible endpoint configurations are shown in **Figure 3B**. Each pixel grid is compared with each of the 16 endpoint grids and if any one of them is an exact match, then that coordinate is considered to be an endpoint.

FIGURE 3. **Convolutional filtering for end point determination**. (**A**) shows 9 x 9 pixel grids that are created from points *i*, *ii*, and *iii* on a skeletonized polymer. (**B)** shows the sixteen different shapes that were used in the convolutional filter that finds endpoint coordinates. Shapes are clustered via number of neighbors.

Branch point identification

Following endpoint detection, branch points are identified through a filter that works by creating an array corresponding to a clockwise 10-pixel-long path which surrounds the test point. A unique path is counted for each one value that is abutted by zero values (**Figure 4A**, blue triangles). If there are three or more unique paths which originate from a single pixel, then that pixel is defined as a branch point. The algorithm is also prevented from overcounting branch points, as not all points with three neighbors are true branch points (**Figure 4B**).

FIGURE 4. Branch point determination. (**A)** Each pixel in the image is centered in a 3 x 3 grid and a string of ones and zeros begins at any point neighboring the test pixel. (**B)** Technique for prevention of branch point overcounting. For test point *A* which has three adjoining positive pixels and is not a branch point, and *B* which has three adjoining pixels is a branch. Each unique path created is marked by a blue triangle. Point *A* has 3 unique paths that stem from it, point *B* has 3 neighboring pixels, but only two paths which originate from the central pixel, one of which is unique. The non-unique path, which appears on separate 3x3 grids, is marked by the orange triangle.

Shape calling and total polymer length determination

If a polymer has exactly two endpoints and no branch points, it is considered to be linear. If there are no endpoints and no branch points, then it is considered to be a loop. Further, polymers that branch and loop are separated from those which branch but do not loop. For polymers that branch and loop, there cannot be more endpoints than there are branch points, and for polymers that branch and do not loop, there will always be more endpoints than branch points. Branch points are differentiated from points where polymers drape over themselves by creating a skeleton with a height threshold of 1.5 times the average value of the heights corresponding to one-valued pixels on the original skeleton. This new skeleton is used to sort polymers as "overlaps" (**Fig. 5**).

FIGURE 5. Process for determining polymer overlap points. (**A**) Original AFM image with height shown in greyscale, the lateral scale bar (yellow) is 20 nm. (**B**) Skeleton created from the data in panel A.

(**C**) New skeleton that incorporates height information. (**D**) Skeleton that is created at a threshold of 1.5 times the average height of the original AFM image.

For overlapped particles, the length is computed by adding the length of the original skeleton with the skeleton made at a threshold of 1.5 times the average height of all pixels on the skeleton with incorporated height information. If over 80% of the particle is above 1.5 times the average height of the image, then it is sorted separately as a noise particle. Such polymers could be aggregates or other artifacts. Polymers with high points that are not overlaps are still sorted by their shape, and they are copied and stored in a separate folder. The total polymerized length is found by applying a pathfinding algorithm which sums all pixels in each feature.

Persistence Length Calculation

After finding the contour length for each linear polymer, the mean-square of the end-to-end distances is found from the endpoint coordinates. Then, the polymer is modeled as a worm-like chain with a characteristic persistence length. In particular, we applied the following expression

$$< R^2 > \; = 2sPl(1 - \frac{sP}{l}\left(1 - e^{\frac{-l}{sP}}\right)) \qquad (1)$$

Where R represents end-to-end distance, $s = 2$ is a fitting parameter reflecting the dimensionality of the data [8], P is the persistence length, and l is the contour length [9]. Non-linear least square fitting the data to equation 1 provides an estimate of the persistence length.

Results

Persistence length

DNA represents a convenient benchmark for the algorithm as its persistence length has been well characterized. The persistence length results for DNA were found to be 48 ± 3 nm in an analysis of four AFM images containing 206 linear strands of DNA with data from Hennan *et al* [7]. This is within the margin of error for the widely accepted value for DNA persistence length of 50 nm (**Table 1**). The persistence length for the polymer Candidalysin was determined in an analogous manner and found to be 12.1 ± 0.3 nm using seven images containing 670 linear polymers (**Fig. 6**). Using the Easyworm software developed by Lamour *et al* [10], the results for Candidalysin were found to be 12 +/- 3 nm, which is in good agreement with our algorithm.

Figure 6. PS Poly persistence length output. Plot of mean-square end-to-end distance versus contour length for CL polymers. The persistence length was found from the fit (solid line) to be 12.1 ± 0.3 nm.

Table 1. Results of persistence length analysis comparing PS Poly to other work.

FIGURE 7. The output of PS Poly for shape categorization. (**A)** Input AFM image of Candidalysin. The scale bar is 200 nm and the greyscale spans 12 nm. (**B**) Table showing the output for shape categorization. Two types of artifact are identified, high points (defined as any point on a particle that is above 1.5 times the average height of the polymers in the image) and noise particles (defined as any particle in which 80%

or more of the pixels are above 1.5 times the average height of the polymers in the image). (**C**) Examples of the different particle types identified through PS-Poly. On the left side is the skeleton image of the particle, and on the right is the cropped polymer from the original image. The types shown are linear, looped, branched (without looping), and branched (with looping), moving from top to bottom respectively. The scale bar shown spans 50nm.

After categorizing the polymers based on shape, the coordinates of all endpoints, branch points, and three-dimensional overlaps are stored in the output as well as the total length of each feature, total polymerized length for each image, and total polymerized length for all images. An example AFM image of CL and the resulting PS Poly output are shown (**Fig. 7**).

Conclusions

PS-Poly is an automated persistence length algorithm that is also capable of shape categorization for the study of complex polymers. Persistence length calculations on short polymers are possible due to the interpolation performed on the skeletonized image which allows the program to achieve subpixel precision. The benefit of automating this process is increased accuracy and time saved by the user. We note that it is generally helpful to reduce noise in the images before running the program. Results may vary depending on differences in the user-selected threshold that appears at the beginning of the program. To ensure consistent results, the user can manually input a consistent threshold and evaluate results with the same pixel-density scaling factor. User input is not a fundamental requirement; future implementations could employ automated segmentation techniques such as Otsu's method for thresholding [20] . The runtime for this program for a typical AFM image is about 30 seconds running on Igor Pro 7 (64-bit) on a standard desktop PC (CPU @ 2.60 GHz . We note that computational time is increased as the

pixel-density scaling factor is increased. While PS-Poly has been developed and used with AFM image data, it also has the potential to analyze images taken with other microscopy tools including electron microscopes (EM) [21]. Extension of PS Poly to EM data streams could contribute to nucleic acids studies and many other polymeric systems. It is also possible to combine PS-Poly with other chain tracing algorithms. For example, seed coordinates could be taken from PS-Poly and used as inputs for AutoSmarTrace or other algorithms.

Acknowledgements:

This work was supported by NSF #: 2122027 (to GMK). EAC acknowledges support from the Mizzou Forward Undergraduate Research Training Grant and KGS acknowledges support from the Research Excellence Program at the University of Missouri.

References

1. Moyes, D. L. *et al.* Candidalysin is a fungal peptide toxin critical for mucosal infection. *Nature* **532**, 64–68 (2016).
2. Mayer, F. L., Wilson, D. & Hube, B. Candida albicans pathogenicity mechanisms. *Virulence* **4**, 119–128 (2013).
3. Haque, F. & Grayson, S. The synthesis, properties and potential applications of cyclic polymers. *Nat. Chem.* **12**, 433–444 (2020).
4. Voit, B. I. & Lederer, A. Hyperbranched and Highly Branched Polymer Architectures—Synthetic Strategies and Major Characterization Aspects. *Chem. Rev.* **109**, 5924–5973 (2009).
5. Müller, D. J. *et al.* Atomic Force Microscopy-Based Force Spectroscopy and Multiparametric Imaging of Biomolecular and Cellular Systems. *Chem. Rev.* **121**, 11701–11725 (2021).

6. Weaver, D. R. & King, G. M. Atomic Force Microscopy Reveals Complexity Underlying General Secretory System Activity. *Int. J. Mol. Sci.* **24**, 55 (2023).
7. Heenan, P. & Perkins, T. Imaging DNA Equilibrated onto Mica in Liquid Using Biochemically Relevant Deposition Conditions. *ACS Nano* **13**, 4220–4229 (2019).
8. Rivetti, C., Guthold, M. & Bustamante, C. Scanning Force Microscopy of DNA Deposited onto Mica: EquilibrationversusKinetic Trapping Studied by Statistical Polymer Chain Analysis. *J. Mol. Biol.* **264**, 919–932 (1996).
9. Rubinstein, M. & Colby, R. *Polymer Physics*. (Oxford University Press Inc., New York, 2003).
10. Lamour, G., Kirkegaard, J. B., Li, H., Knowles, T. P. & Gsponer, J. Easyworm: an open-source software tool to determine the mechanical properties of worm-like chains. *Source Code Biol. Med.* **9**, 16 (2014).
11. Beton, J. G. *et al.* TopoStats – A program for automated tracing of biomolecules from AFM images. *Methods* **193**, 68–79 (2021).
12. Usov, I. & Mezzenga, R. FiberApp: An Open-Source Software for Tracking and Analyzing Polymers, Filaments, Biomacromolecules, and Fibrous Objects. *Macromolecules* **48**, 1269–1280 (2015).
13. Konrad, S. F. *et al.* High-throughput AFM analysis reveals unwrapping pathways of H3 and CENP-A nucleosomes. *Nanoscale* **13**, 5435–5447 (2021).
14. Wiggins, P. A. *et al.* High flexibility of DNA on short length scales probed by atomic force microscopy. *Nat. Nanotechnol.* **1**, 137–141 (2006).

15. Faas, F. G. A., Rieger, B., Vliet, L. J. van & Cherny, D. I. DNA Deformations near Charged Surfaces: Electron and Atomic Force Microscopy Views. *Biophys. J.* **97**, 1148–1157 (2009).

16. Brangwynne, C. P. *et al.* Bending Dynamics of Fluctuating Biopolymers Probed by Automated High-Resolution Filament Tracking. *Biophys. J.* **93**, 346–359 (2007).

17. Zhang, H., Söderholm, N., Sandblad, L., Wiklund, K. & Andersson, M. DSeg: A Dynamic Image Segmentation Program to Extract Backbone Patterns for Filamentous Bacteria and Hyphae Structures. *Microsc. Microanal.* **25**, 711–719 (2019).

18. Schneider, M., Al-Shaer, A. & Forde, N. R. AutoSmarTrace: Automated chain tracing and flexibility analysis of biological filaments. *Biophys. J.* **120**, 2599–2608 (2021).

19. Russell, C. M. *et al.* The Candida albicans virulence factor candidalysin polymerizes in solution to form membrane pores and damage epithelial cells. *eLife* **11**, e75490 (2022).

20. Houssein, E. H., Mohamed, G. M., Ibrahim, I. A. & Wazery, Y. M. An efficient multilevel image thresholding method based on improved heap-based optimizer. *Sci. Rep.* **13**, 9094 (2023).

21. Hagerman, P. FLEXIBILITY OF DNA. *Ann Rev Biophys Biophys Chem* **17**, 265–286 (1988).

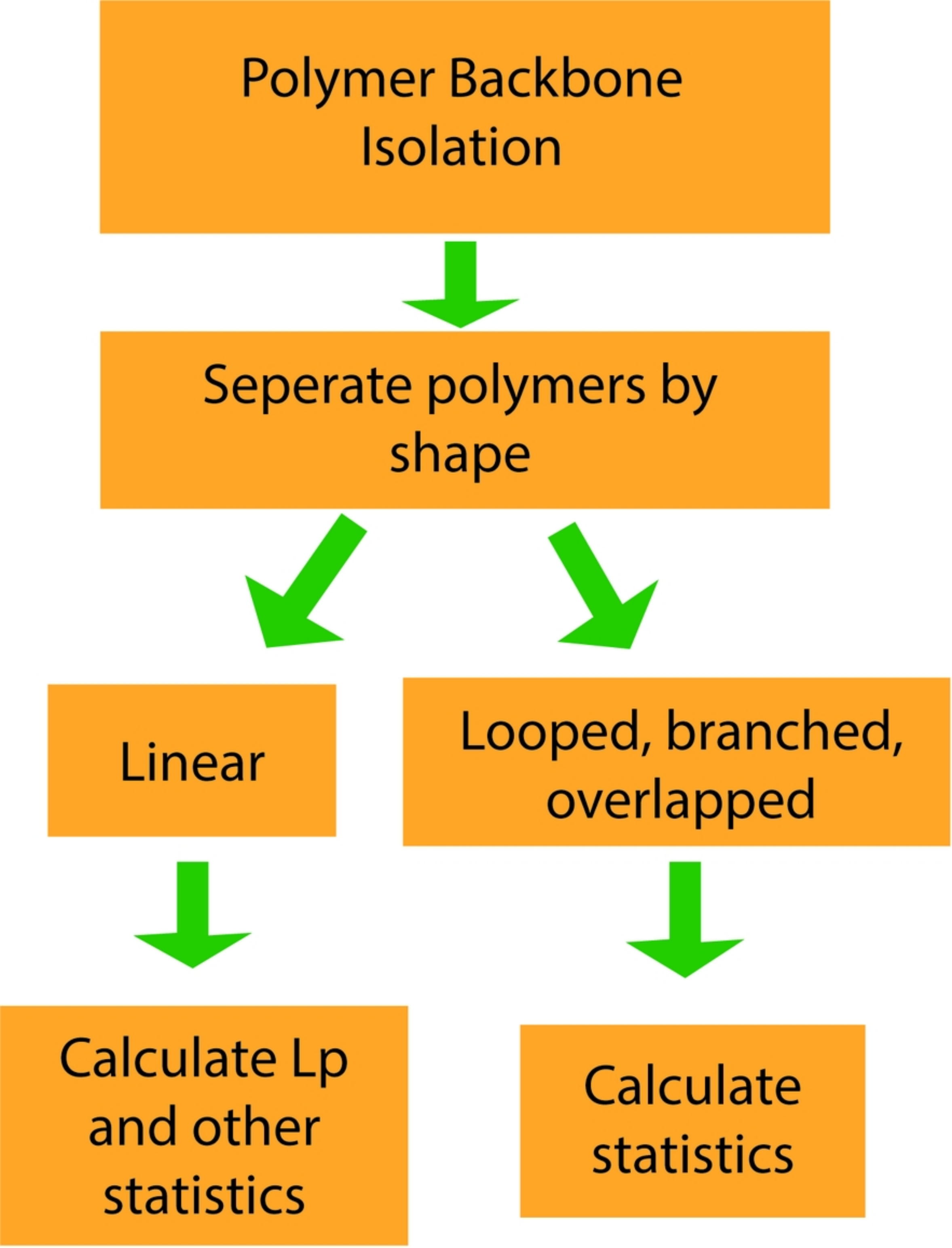

Polymer Backbone Isolation
Seperate polymers by shape
Linear
Looped, branched, overlapped
Calculate Lp and other statistics
Calculate statistics

Figure 1

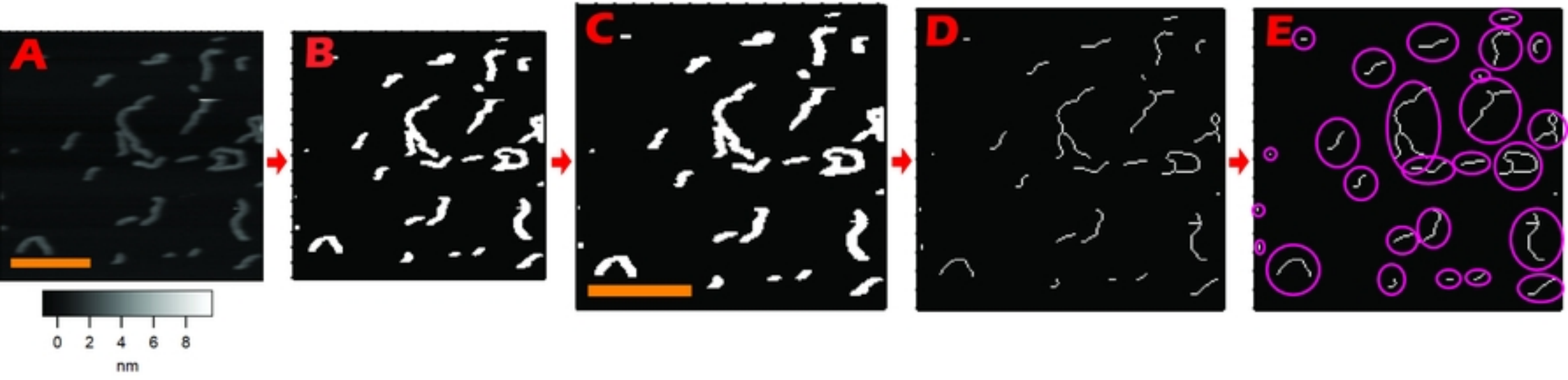


Figure 2

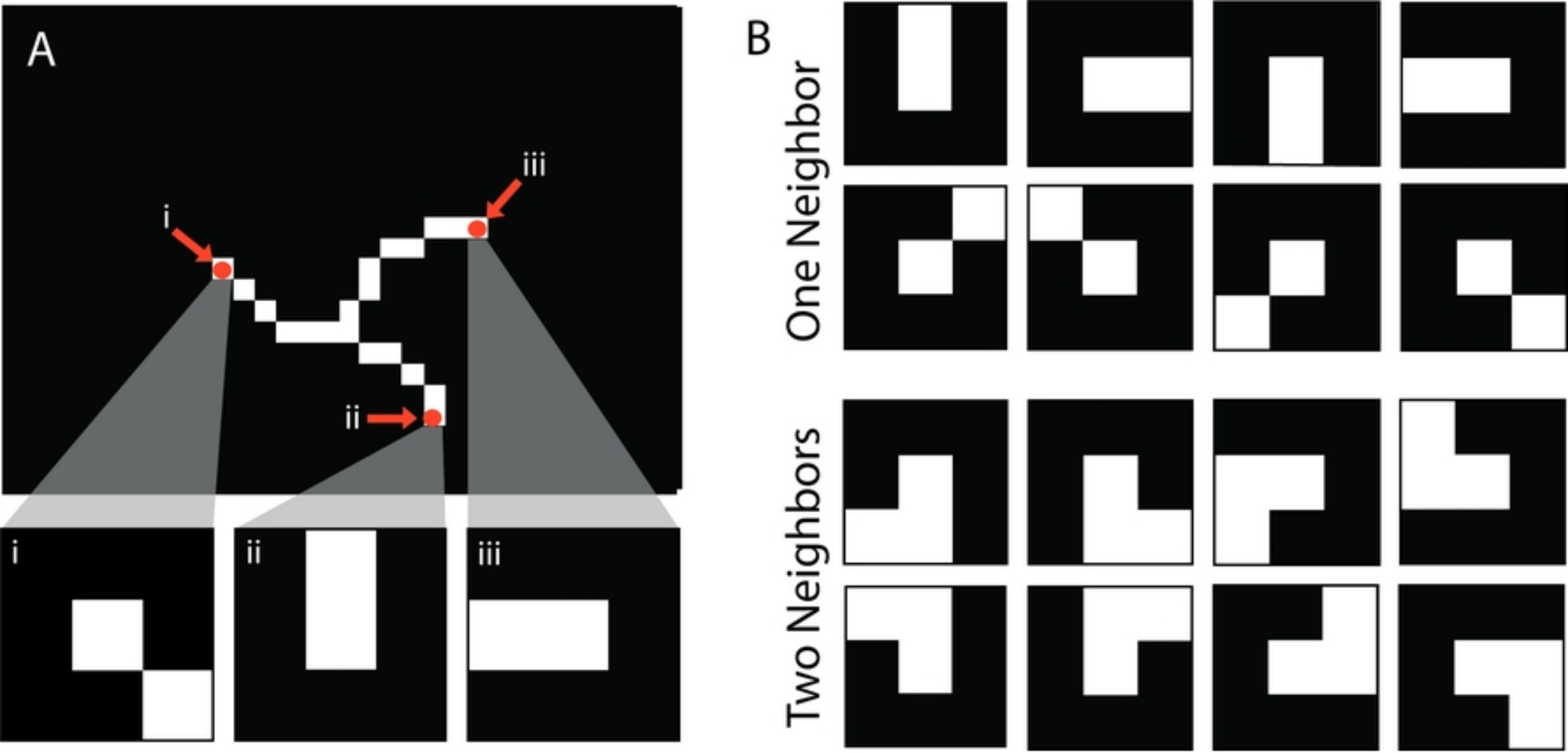


Figure 3

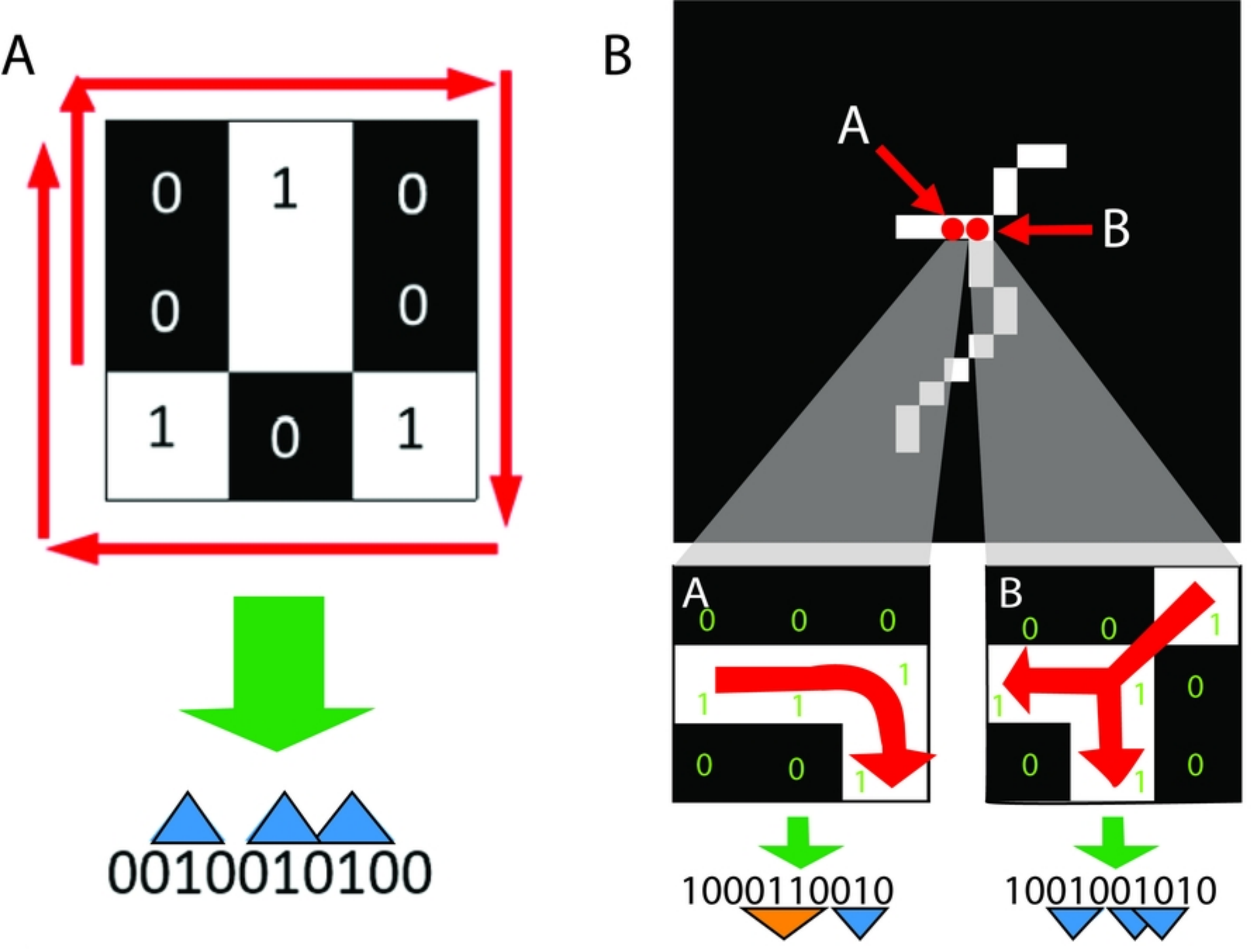


Figure 4

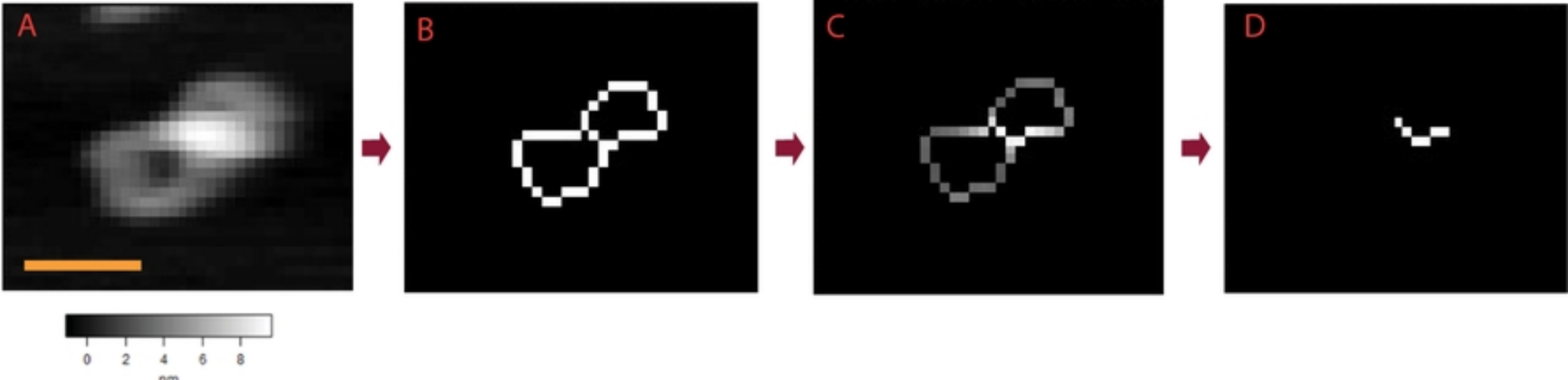


Figure 5

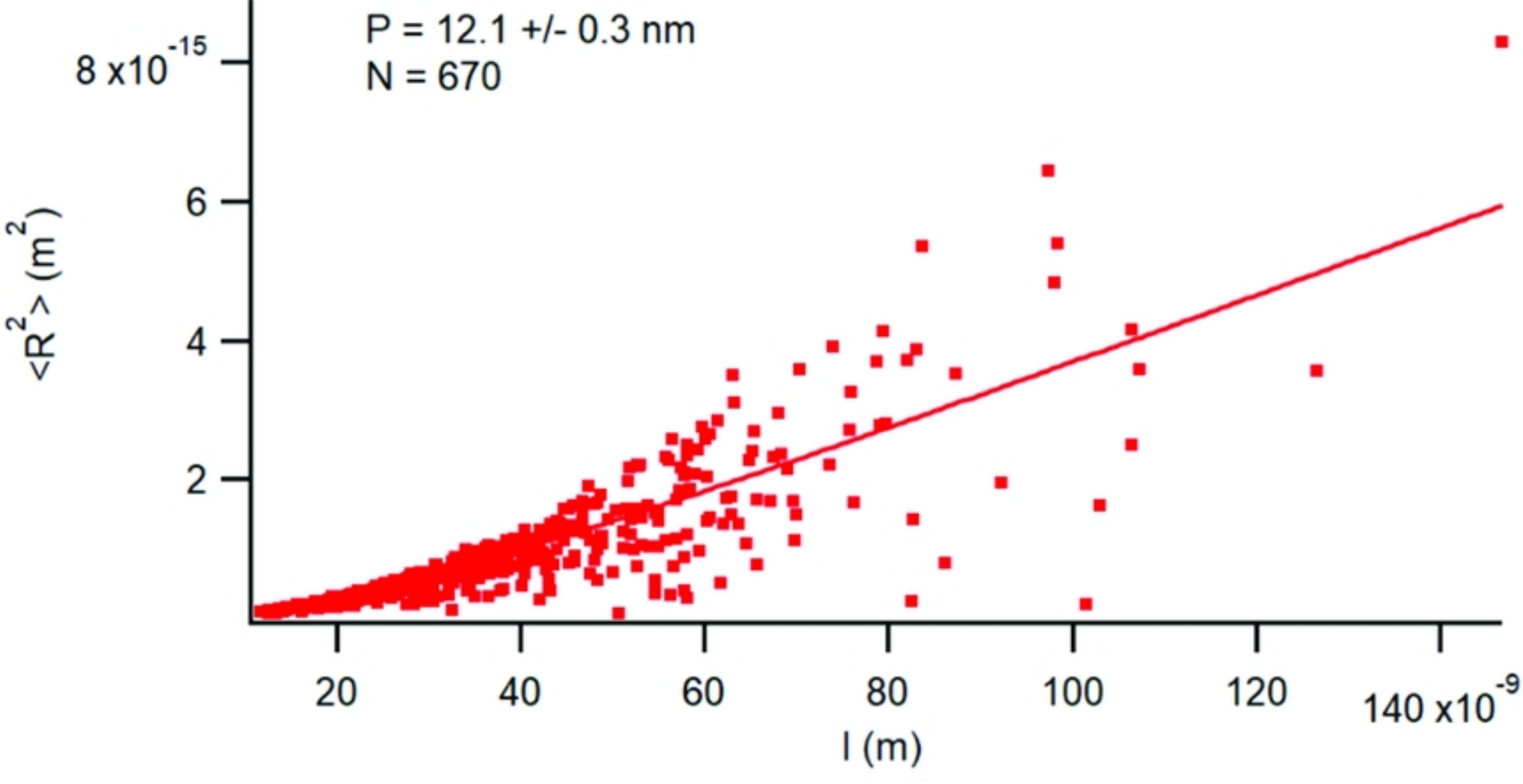
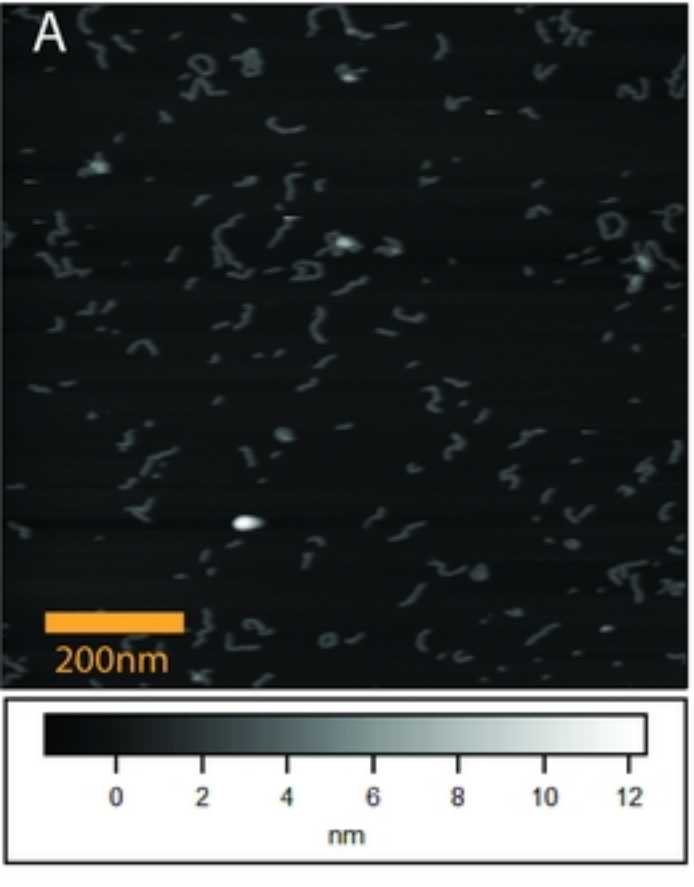


Figure 6



B

Particle Type	Number of Features	Average Length (nm)	Percentage of Total Polymerized Length
Linear	164	25.5	62.1
Looped	2	76.7	2.3
Branched (no looping)	18	78.0	20.8
Branched (with looping)	8	125.1	14.8
Overlapped	0	N/A	0
Number of High Points			17
Number of Noise Particles			1

C

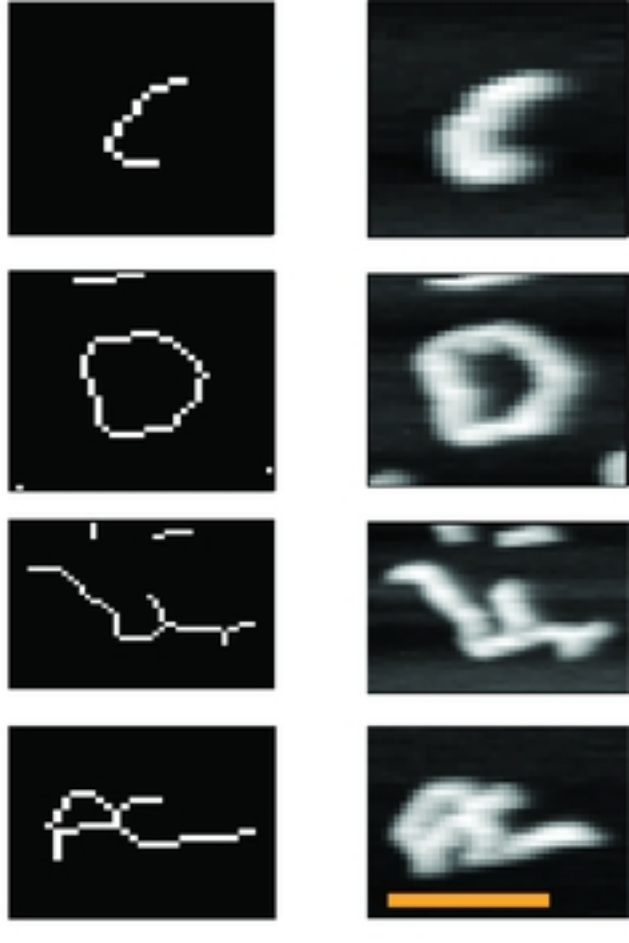


Figure 7

Sample	Persistence Length, Previous work	Persistence Length using PS Poly
DNA	46.8 +/- .6 nm N = 126 (Ref 7)	47.8 +/- 2.6 nm N = 206
Candidalysin	12 +/- 3 nm N = 191 (Ref 10)	12.1 +/- .3 nm N = 752

Table 1